Der *Gemeinsame europäische Referenzrahmen für Sprachen*: Leistung und Grenzen.

Die Bedeutung des Referenzrahmens im Kontext der Beurteilung von Sprachvermögen am Beispiel des semikreativen Schreibens im DESI-Projekt.

Inaugural-Dissertation
zur Erlangung des Doktorgrades
an der
Philologisch-Historischen Fakultät
der Universität Augsburg

vorgelegt von
Claudia Harsch

Augsburg, September 2005

Erstgutachter:	Prof. Dr. Konrad Schröder
Zweitgutachter:	Prof. Dr. Hans Jürgen Heringer
Tag der mündlichen Prüfung:	05. Juli 2006

Danksagung

Viele Menschen haben mich in den letzten drei Jahren während der Entstehung meiner Dissertation auf vielfältige und unterschiedliche Weisen unterstützt und begleitet. Daher möchte ich mich herzlich bedanken

bei meinem Doktorvater Prof. Dr. Konrad Schröder, der die Arbeit betreute, mir mit Rat und Tat zur Seite stand und mir doch genügend individuellen Freiraum ließ;

bei meinem Zweitbetreuer Prof. Dr. Hans-Jürgen Heringer, der sich Zeit für mein Anliegen nahm und mir gute Denkanstöße gab;

bei Herrn Prof. Dr. Dieter Götz dafür, dass er sich bereit erklärte, die Prüfung mit abzunehmen;

bei meinem Kollegen Dr. Rudolf Beck, der Verständnis für mich hatte, mich wiederholt motivieren konnte und für die nötige Bodenhaftung sorgte;

bei meinen Kolleginnen und Freundinnen Dr. Brigitta Mittmann, Corinna Humpfer, Monika Bednarek und Ursula Thum für ihre fachfrauliche und freundschaftliche Unterstützung, ihre konstruktive Kritik und ihre wertvollen Ratschläge, nicht zu vergessen die unzähligen Tassen Tee, die mir über manche Durststrecken verhalfen;

bei Christine Bomball für ihren Beistand in organisatorischen Fragen und bei Dr. Brigitta Mittmann, Gudrun Nelle und Marion Wöhrl für ihre Bereitschaft, Teile der Arbeit Korrektur zu lesen;

bei Stefan Langer, Maciej Golik, Gabriele Kötterle und Ursula Wahl für die tatkräftige Unterstützung bei Programmierarbeiten respektive bei der Beseitigung aller Probleme im Zusammenhang mit Computern oder Internet;

bei allen studentischen Hilfskräften, die ihren Beitrag im Rahmen des DESI-Projekts leisteten, insbesondere bei Eva Stangl, Lena Müller, Olivia Wartha und Sebastian Haffner;

bei den Kollegen und Kolleginnen des DESI-Teams, die mir mit Rat und Hilfe beistanden und immer ein offenes Ohr für mich hatten, vor allem bei Henning Rossa für seine beruhigende Ausstrahlung und bei Dr. Astrid Neumann, Dr. Johannes Hartig und Nina Jude für ihre Geduld in psychometrischen Fragen;

bei allen Kolleginnen und Kollegen, die ich auf Tagungen oder im Rahmen des DESI-Projekts kennen gelernt habe und die willens waren, sich mit mir und meinen Fragen zu beschäftigen;

und nicht zuletzt bei Freundinnen, Freunden und Familie für ihr Verständnis, ihre Freundschaft und die vielen leckeren Mahlzeiten, die mich am Leben erhalten haben:

Thank you all for putting up with me.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Einleitung

Ausgangslage

Die Entwicklung und Sicherung der Bildungsqualität nimmt spätestens seit dem so genannten PISA-Schock[1] einen zentralen Platz in der Bildungspolitik ein. Die internationale Schulleistungsstudie PISA der OECD[2] hat in Deutschland einen Optimierungsbedarf im Bereich des Lernens und Lehrens an den Schulen aufgezeigt. Doch nicht nur die Ausbildung in den durch PISA erfassten Gebieten der Naturwissenschaften und der *reading literacy*[3], des Leseverstehens, sondern auch die schulische Vermittlung der modernen Fremdsprachen bedarf der Verbesserung. Handlungsfähigkeit in mindestens einer modernen Fremdsprache und rezeptive Mehrsprachigkeit sind Schlüsselkonzepte der europäischen Sprachenpolitik[4], um Europas Bürgerinnen und Bürger auf die Bedürfnisse eines zusammenwachsenden europäischen Binnenmarktes mit derzeit 21 offiziellen Amtssprachen und mehr als 60 Minderheitensprachen[5] vorzubereiten. Der Europarat hat auf die Notwendigkeit der Förderung der Mehrsprachigkeit beispielsweise mit der Herausgabe des *Gemeinsamen europäischen Referenzrahmens für Sprachen: lernen, lehren, beurteilen*[6] (im Folgenden mit GER abgekürzt) reagiert, welcher unten vorgestellt wird. Vor dem Hintergrund der Globalisierung sind die Ansprüche der Wirtschaft an Fremdsprachenkenntnisse nicht zu vergessen. Daneben dürfen Migrationsbewegungen im europäischen Kontext nicht vernachlässigt werden. Beispielsweise muss die mit der Migration einhergehende Perspektive des ungesteuerten Zweitspracherwerbs in der schulischen wie außerschulischen Sprachausbildung berücksichtigt werden. Schule ist denn auch, neben Angeboten der Erwachsenenbildung und der freien Wirtschaft, ein Ort der Vorbereitung auf die wachsenden Ansprüche im Bereich der fremdsprachlichen Kompetenzen.

Die Notwendigkeit der Qualitätsentwicklung in der schulischen Fremdsprachenvermittlung ist in Deutschland erkannt worden: Die Kultusministerkonferenz der Länder (im Folgenden mit KMK abgekürzt) hat Zielvorgaben für die naturwissenschaftlichen Fächer und für die erste Fremdsprache entwickelt, die so genannten Bildungsstandards[7], wobei sich die Standards für die erste Fremdsprache an den erwähnten GER anlehnen. Des Weiteren hat die KMK Instrumente eingeführt, die das Erreichen dieser Zielvorgaben überprüfen sollen, wie etwa schulische Vergleichsarbeiten oder das Europäische Sprachenportfolio[8]. Um jedoch die Bildungsqualität im Bereich der fremdsprachlichen Ausbildung entwickeln zu können, ist es notwendig, neben den

---

[1] Vgl. dazu etwa Harsch & Schröder 2005a.
[2] Vgl. Deutsches PISA-Konsortium 2001, für weiterführende Berichte vgl. http://www.pisa.oecd.org/pisa/outcome.htm oder http://www.mpib-berlin.mpg.de/pisa, Zugriff am 2.12.2003.
[3] Das angelsächsische Konzept der *reading literacy* wird im PISA-Projekt wie folgt definiert: “Reading literacy is understanding, using, and reflecting on written texts, in order to achieve one’s goals, to develop one’s knowledge and potential, and to participate in society". (Vgl. http://www.pisa.oecd.org/pisa/read.htm, Zugriff am 2.12.2003).
[4] Vgl. etwa Europäische Union 1995.
[5] Vgl. http://de.wikipedia.org/wiki/Amtssprachen_der_Europ%C3%A4ischen_Union respektive http://de.wikipedia.org/wiki/Minderheitensprache, Zugriff am 19.8.2005.
[6] Vgl. Europarat 2001.
[7] Vgl. Kultusministerkonferenz 2003, oder http://www.kmk.org/schul/Bildungsstandards/1.Fremdsprache_MSA_BS_04-12-2003.pdf, Zugriff am 12.7.2004.
[8] Vgl. http://www.coe.int/portfolio, Zugriff am 3.2.2005.

genannten Zielvorgaben und deren Beurteilungsmöglichkeiten den Stand des fremdsprachlichen Könnens und die schulischen wie außerschulischen Lernbedingungen zu ermitteln. Nur wenn bekannt ist, wo die Lernenden stehen und welche Bedingungen das Lernen fördern, können konkrete Schritte zur Verbesserung des Unterrichts und der Lernkontexte unternommen werden. Diese Evaluation des Ausbildungsstands und seiner Bedingungen kann auf verschiedenen Wegen erfolgen:[9] Denkbar sind interne Beurteilungen durch Lehrende und/oder durch die Lernenden selbst, also Beurteilungen innerhalb des Schulkontextes. Die interne Beurteilung durch die Lehrenden hat zumindest in Deutschland lange Tradition und stellt einen Teil der Selbstbestimmung und Eigenverantwortung der Schulen dar.[10] Der genannte GER hat in diesem Bereich bereits Einzug in die Schulen gehalten,[11] sei es über Curricula, die sich an den GER anlehnen oder sei es über das erwähnte Sprachenportfolio, das im Wesentlichen auf dem GER beruht. Die Perspektive der internen Evaluation sollte jedoch ergänzt werden um die der externen Evaluation, die einen unabhängigeren Standpunkt gegenüber der zu beurteilenden Institution einnehmen kann. Zudem werden dadurch Vergleiche über individuelle schulische Kontexte hinweg ermöglicht, so dass unterschiedliche Bedingungen in ihrer Wirksamkeit analysiert werden können. Zur externen Evaluation lassen sich etwa Schulleistungsstudien einsetzen wie die PISA-Studie oder das DESI-Projekt[12], eine von der KMK im Jahr 2001 in Auftrag gegebene Leistungsstudie, die im fremdsprachlichen Teil auf den erwähnten GER rekurriert. Interne wie externe Evaluation müssen zueinander in Bezug gesetzt werden, um beide Perspektiven möglichst effizient nutzen zu können. Der GER könnte in seiner Funktion als *Referenzrahmen* als ein Hilfsmittel zur Verknüpfung der internen mit der externen Evaluation betrachtet werden, da er in beiden Bereichen eingesetzt werden kann und man im Idealfall die Ergebnisse interner und externer Evaluation auf den GER als *gemeinsames Referenzmittel* beziehen kann. Aufgrund der hier grob umrissenen Bedeutsamkeit (nicht nur) in der schulischen Evaluation soll der GER in der vorliegenden Arbeit näher untersucht werden.

Der Gemeinsame Europäische Referenzrahmen

Der GER ist ein Instrument zur Umsetzung der sprachenpolitischen Ziele des Europarats[13], insbesondere zur Förderung der europäischen Mehrsprachigkeit und der damit verbundenen kulturellen Kompetenzen (vgl. GER 2001: 3). Mit diesem Instrument werden zwei Hauptziele verfolgt: Es will „Praktiker aller Art im Sprachenbereich“ (ebd.: 8) ermutigen, ihr Vorgehen zu reflektieren und es will die Kommunikation, den Erfahrungsaustausch und die Kooperation unter den Praktikern erleichtern (ebd.: 8, 14). Um diese Ziele umsetzen zu können, will der GER „den aktuellen Stand der Fremdsprachenforschung zusammen[fassen]“ (ebd.: 3) und mit den GER-Skalen „ein

[9] Vgl. hierzu etwa Landesinstitut NRW 1999, Ministerium für Schule und Weiterbildung NRW 1997, 1998a, 1998b oder Wiater 2005.
[10] Vgl. etwa Harsch & Schröder 2005a.
[11] Vgl. Landesinstitut für Schule und Weiterbildung NRW 1999.
[12] DESI steht für *Deutsch-Englisch-Schülerleistungen International*, vgl. http://www.dipf.de/desi, Zugriff am 30.3.2005. Diese Studie wird unten vorgestellt.
[13] Diese können beispielsweise in GER-Abschnitt 1.2 nachgelesen werden.

Stufensystem der Sprachbeherrschung“ (ebd.) vorstellen, das die Aspekte und Wissensbestände beschreiben soll, die Sprachlernende „im öffentlichen, beruflichen und privaten Bereich sprachlich handlungsfähig“ machen und „kulturell sensibilisier[en]“ (ebd.). Damit will der GER einerseits „Werkzeuge zur Verfügung“ stellen, die es Praktikern ermöglichen, ihr Vorgehen einzuordnen und auf die Bedürfnisse der Lernenden auszurichten (ebd.: 14); andererseits will er „helfen, die Barrieren zu überwinden, die aus den Unterschieden zwischen den Bildungssystemen in Europa entstehen und die der Kommunikation unter Personen, die mit der Vermittlung moderner Sprachen befasst sind, im Wege stehen“ (ebd.). Die Autoren des GER verstehen den Referenzrahmen als „gemeinsame Basis für die explizite Beschreibung von Zielen, Inhalten und Methoden“ (ebd.) und „… für die Entwicklung von zielsprachlichen Lehrplänen, curricularen Richtlinien, Prüfungen, Lehrwerken usw. in ganz Europa“ (ebd.). Dadurch soll die Transparenz „von Kursen, Lehrplänen und Richtlinien und von Qualifikationsnachweisen“ (ebd.) in Europa erhöht und die „gegenseitige Anerkennung von Qualifikationsnachweisen, die in unterschiedlichen Kontexten erworben wurden“, erleichtert werden (ebd.). So soll die „internationale Zusammenarbeit auf dem Gebiet der modernen Sprachen“ und „auch die Mobilität in Europa“ verbessert werden (ebd.).

Mit seinem Referenzsystem stellt der GER ein Kompetenzmodell bereit, das relevante Teilbereiche kommunikativen Handelns und sprachlichen Könnens kategorisiert und beschreibt. Diese Teilbereiche oder Kategorien, wie sie im GER genannt werden, umfassen Kompetenzen in den Bereichen der Rezeption, Produktion, Interaktion und Mediation. Für diese differenzierten Kategorien werden so genannte Beispielskalen bereitgestellt, in welchen die jeweiligen Könnensbereiche abgestuft auf sechs Niveaus beschrieben werden, wobei sich die Niveaubeschreibungen an Vorarbeiten des Europarats anlehnen (vgl. etwa GER 2001: 33f und 42ff). Neben dem Skalensystem finden sich im GER theoretische Ausführungen zum Selbstverständnis des GER im europäischen Kontext, zur Konzeption und Verwendung des GER-Skalensystems, zum Bereich des Lernens und Lehrens von Fremdsprachen, zur Curriculumentwicklung und zur Thematik des Beurteilens und Bewertens von Sprachvermögen.

Seit seinem Erscheinen wird der GER sowohl auf einer praxisorientierten Ebene als auch in Bezug auf seine theoretische Fundierung diskutiert: Davon zeugen neben Fachpublikationen (vgl. etwa Alderson 2002 oder Bausch/Christ/Königs/Krumm 2003) internationale Tagungen, die den GER thematisieren, wie etwa die Tagung des *British Council* zum Thema “Standards in Language Learning and the Common European Framework“ (Berlin: März 2004), die Tagungen der EALTA[14] (Slowenien: Mai 2004 und Norwegen: Juni 2005) oder die Konferenz der ALTE[15] zum Thema “Language Assessment in a Multilingual Context – Attaining Standards, Sustaining Diversity“ (Berlin: Mai 2005). Die vorliegende Arbeit möchte einen Beitrag leisten zur kritischen Diskussion der Bedeutsamkeit und Reichweite des GER, insbesondere

[14] *European Association of Language Testing and Assessment*, vgl. http://www.ealta.eu.org, Zugriff am 3.2.2005.
[15] *Association of Language Testers in Europe*, vgl. http://www.alte.org, Zugriff am 25.7.2005.

hinsichtlich der Verwendbarkeit des GER bei der Beurteilung des Sprachvermögens durch Schulleistungsstudien, einer Form der externen Evaluation der schulischen Bildungsqualität.

Ziel der vorliegenden Arbeit ist es daher, den GER auf die Schlüsselbegriffe *Sprache, Lernen, Lehren, Beurteilen* und auf seinen *Skalenansatz* hin zu analysieren. Aufbauend auf diesen Analysen wird die Bedeutsamkeit des GER bei der Testerstellung und Testauswertung am Beispiel des semikreativen Schreibens im erwähnten DESI-Projekt konkretisiert und beurteilt. Dadurch sollen positive Impulse des GER herausgearbeitet werden und es soll aufgezeigt werden, was dieses Instrument leisten kann und wo seine Grenzen sind. Neben positiven und kritischen Aspekten des GER sollen Möglichkeiten seiner Weiterentwicklung erörtert werden.

Das DESI-Projekt

Die Schulleistungsstudie DESI soll Aussagen zum Leistungsstand und zu den Fähigkeiten von Schülern und Schülerinnen der 9. Klassen aller Schulformen in Deutschland in den Bereichen Deutsch und Englisch treffen. Dabei werden der aktive und passive Gebrauch des Deutschen und die kommunikative Kompetenz in der Fremdsprache Englisch erfasst. Ein Messwiederholungsdesign zeigt die Veränderungen im Laufe der 9. Klasse auf. Bei der Entwicklung der Testmodule im Bereich der Fremdsprache Englisch wurde der GER nicht als Ausgangspunkt genommen, vielmehr wurden die Tests verankert in curricularen Analysen und theoretischen Modellen. Dennoch wurden Testinhalte, Aufgabenanforderung und Beschreibungen der Kompetenzniveaus mit relevanten Ausführungen im GER abgeglichen, um Aussagen über das Verhältnis der fremdsprachlichen DESI-Module zu den Kategorien und Niveaus des GER treffen zu können.

Diese Studie wird von einem Konsortium, bestehend aus Professoren der Universitäten Augsburg, Berlin (Humboldt), Dortmund, Hamburg, Koblenz-Landau und Oldenburg sowie dem Deutschen Institut für internationale pädagogische Forschung (DIPF, Frankfurt/Main) durchgeführt. Innovativ ist dabei die Zusammenarbeit zwischen Fachdidaktikern und Psychometrikern, die sich ergänzen und unterstützen. Beispielsweise wird das erwähnte Testmodul *Textproduktion Englisch* an der Universität Augsburg von Fachdidaktikern entwickelt und bewertet, während die Skalierung der Bewertungen von den empirischen Bildungsforschern der Universität Berlin durchgeführt wird. Die statistische Seite der Datenauswertung wiederum wird von den Psychometrikern am DIPF vorgenommen, mit fachlicher Unterstützung der Didaktiker bei der Hypothesenbildung und Interpretation der Befunde. Tabelle 1 zeigt Arbeitsteilung und Zuständigkeiten im DESI-Projekt:

Konsorten und Mitarbeiter	Institution	Zuständigkeiten
Prof. Dr. Eckhardt Klieme Dr. Bärbel Beck Dr. Brigitte Steinert Dr. Johannes Hartig, Nina Jude Dr. Hermann Hesse, Dr. Kerstin Göbel	DIPF	Federführung des Projekts Projektkoordination Schulleiterfragebogen Psychometrie, Skalierungen Modul *Interkulturelle Kompetenz*
Prof. Dr. Wolfgang Eichler und Mitarbeiter	Universität Oldenburg, Didaktik des Deutschen	Modul *Sprachbewusstheit Deutsch*
Prof. Dr. Andreas Helmke Dr. Tuyet Vo, Wolfgang Wagner	Universität Koblenz-Landau, Psychologie	Schüler- und Lehrerfragebögen Videographie des Englischunterrichts Skalierungen der Fragebögen
Prof. Dr. Rainer Lehmann Astrid Neumann	Humboldt-Universität zu Berlin, Empirische Bildungsforschung	Modul *Textproduktion Deutsch* Skalierung der Bewertungen der Textproduktionsmodule Deutsch und Englisch
Prof. Dr. Günter Nold Henning Rossa, Kyriaki Chazivassiliadou	Universität Dortmund, Institut für Anglistik	Module *Hörverstehen Englisch, Leseverstehen Englisch, Sprachbewusstheit Englisch, mündliche Kommunikationsfähigkeit Englisch*
Prof. Dr. Hans-Günter Rolff und Mitarbeiter	Universität Dortmund, Institut für Schulentwicklungsforschung	Elternfragebogen Rückmeldestudie
Prof. Dr. Konrad Schröder Claudia Harsch	Universität Augsburg, Didaktik des Englischen	Module *Genereller Sprachstand Englisch* und *Textproduktion Englisch*
Prof. Dr. Günther Thomé	Universität Oldenburg	Modul *Rechtschreibung Deutsch*
Prof. Dr. Heiner Willenberg und Mitarbeiter	Universität Hamburg, Didaktik des Deutschen	Module *Kommunikation und Argumentation Deutsch, Leseverstehen Deutsch*

Tabelle 1: Übersicht über Zuständigkeiten im DESI-Projekt

Die vorliegende Arbeit

Diese Arbeit ist wie folgt aufgebaut: Im Theorieteil (Kapitel 1 mit 3) werden relevante Grundlagen und Konzepte in den genannten Schlüsselbereichen *Sprache*, *Lernen und Lehren*, *Testen von Sprachvermögen* und *Skalen in der Beurteilung des Sprachvermögens* erörtert. Denn um die Bedeutung des GER einschätzen zu können, bietet es sich an, die zu analysierenden Schlüsselbegriffe zunächst jeweils unabhängig von den Aussagen im GER zu erarbeiten. Dann können diese Konzepte im Anschluss als Analyserahmen dienen, innerhalb dessen der GER auf sein Verständnis dieser Begrifflichkeiten untersucht wird. Deshalb werden die GER-Analysen in den ersten drei Kapiteln dieser Arbeit jeweils im Anschluss an die theoretischen Ausführungen zu den genannten Schlüsselbegriffen dargestellt.

Das erste Kapitel befasst sich zunächst mit den fremdsprachendidaktischen Grundbegriffen *Sprache* und *Vermittlung* (*Lernen und Lehren)*; das zweite Kapitel wendet sich der Beurteilung und dem *Testen von Sprachvermögen* zu. Denn ehe man Aussagen zur vielschichtigen Thematik der Sprachbeurteilung treffen kann, muss der Beurteilungsgegenstand in seinen komplexen Facetten beleuchtet werden. Innerhalb des Themenbereichs der Sprachbeurteilung greift das dritte Kapitel den *Skalenansatz in der Beurteilung* heraus, denn diesem kommt zunehmende

Bedeutung bei der Bewertung sprachlicher Leistungen zu. Auch wenn die Arbeit sich in der Regel auf die englische Fachdidaktik bezieht, so treffen die Aussagen doch auf die Fremdsprachendidaktik allgemein zu. Dabei versteht es sich von selbst, dass Modelle und Methoden der Psychometrie, wenn ihnen auch im Bereich des Testens eine wesentliche Rolle zukommt, nicht Gegenstand dieser Arbeit sein können. Sie werden nur insofern erläutert, als es für die Argumentation notwendig erscheint, doch ihre Angemessenheit kann auf didaktischer Basis nicht diskutiert werden.

Im vierten, auf die Praxis ausgerichteten Kapitel dieser Arbeit wird die Entwicklung und Auswertung der semikreativen Aufgabenstellung des Moduls *Textproduktion Englisch* im DESI-Projekt dokumentiert. Dabei wird am Ende eines jeden Unterkapitels Bezug genommen auf den GER und seine Bedeutsamkeit und Verwendbarkeit im jeweiligen Projektabschnitt. Da im Praxisteil einige Analyseergebnisse der vorangegangenen theoretischen Kapitel wiederholt werden, mag es stellenweise zu Redundanzen kommen. Dennoch mögen diese in Kauf genommen werden, um die Kohärenz des vierten Kapitels zu gewährleisten: Somit kann es auch für sich alleine gelesen werden.

Grundlage der Analyse in den ersten drei Kapiteln der vorliegenden Arbeit ist die deutschsprachige Ausgabe des GER aus dem Jahr 2001. Das englische Originaldokument *The Common European Framework of Reference for Languages: learning, teaching, assessment* (Council of Europe 1996a) wurde inzwischen in 18 Sprachen übersetzt[16], so dass in den entsprechenden Ländern diese Übersetzungen im Einsatz sind. Daher soll bezogen auf die Situation in Deutschland auch die deutsche Ausgabe untersucht werden. Zudem sind die Skalen im DESI-Projekt in deutscher Sprache verfasst und rekurrieren auf die deutschsprachigen GER-Skalen. Das englische Original wird jedoch immer dann zu Rate gezogen, wenn Ausführungen im GER nicht nachvollziehbar sind, denn oft lassen sich dadurch Unverständlichkeiten auflösen. Wo immer sich im GER explizite Hinweise auf Hintergrundliteratur oder mit ihm in Zusammenhang stehende Projekte finden, werden diese an der betreffenden Stelle hinzugezogen, so etwa beim Projekt der GER-Skalenkonstruktion.[17] Zusatzdokumente zum GER werden nur analysiert, wenn sie in unmittelbarem Zusammenhang mit dem Thema der Arbeit, der Beurteilung von Sprachvermögen, stehen: Der *User's Guide for Examiners* (Council of Europe ^{2}2002), im Folgenden UGE genannt, wird in Kapitel 2 dieser Arbeit bei der Analyse des Testbegriffs im GER hinzugezogen. Das *Manual for Relating Examinations to the Common European Framework of Reference* (Council of Europe 2003a), kurz *Manual* genannt, wird bei der Analyse der Bedeutung des GER-Skalenansatzes in Kapitel 3 der vorliegenden Arbeit beurteilt. Beide Zusatzdokumente werden bei der Konstruktion des DESI-Schreibtests auch in ihrem praktischen Nutzen bewertet.

[16] Vgl. http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Common_Framework_of_Reference/4linguisticversions.asp#TopOfPage, Zugriff am 22.8.2005.
[17] Beispielsweise wird im dritten Kapitel dieser Arbeit, bei der Erörterung der Kategorien der *proficiency* in der GER-Skalenkonstruktion, die Hintergrundliteratur von North (1996 resp. 2000) und North & Schneider (1998) mit einbezogen, da sie explizit im GER (z. B. im Anhang B) erwähnt wird.

1 Grundlegende Begriffe

Gegenstand des ersten Kapitels ist zum einen die Erörterung der Sprach- und Kompetenzbegriffe, die dieser Arbeit zugrunde liegen, denn beide Konzepte haben grundlegende Auswirkungen auf das Testen von Sprachvermögen: Je nach angesetztem Sprach- und Kompetenzmodell werden die jeweils konstruierten Testformate je andere Aspekte des Beurteilungsgegenstands erfassen. Zum anderen wird im ersten Kapitel ein Lern- und Vermittlungskonzept erarbeitet, das helfen soll, die Prozesse zu erhellen, die der Entwicklung von Sprachvermögen zugrunde liegen. Denn das Testen von Sprachvermögen setzt eine gewisse Sprachlernentwicklung der Probanden voraus und damit eine Auseinandersetzung der Lernenden mit zumindest zwei Sprachen, der Muttersprache und der neuen Sprache. Die der Testentwicklung zugrunde gelegten Vorstellungen von Spracherwerb und Sprachlernen haben Auswirkungen darauf, welches Konstrukt und welches Kompetenzmodell bei einem zu konstruierenden Test angesetzt wird und in welchem Format der Test das Sprachvermögen erfasst; deshalb müssen auch sie vorab erörtert werden.

Bevor also auf testspezifische Überlegungen eingegangen werden kann, muss zuerst ein "workable concept of language", ein Sprachbegriff erarbeitet werden, der die Basis für die Beurteilung von Sprachvermögen bilden muss, will man dieses Vermögen valide beurteilen. Der Sprachbegriff wird unter anderem durch folgende Fragen charakterisiert: Was bedeutet Sprache überhaupt? Welche Vorstellungen der innersprachlichen Organisation gibt es? Wie ist Sprache als internes Wissenssystem strukturiert? Welches Modell von Sprache als Mittel zur Kommunikation und welches Modell der kommunikativen Kompetenz liegt den Tests zugrunde? Gibt es ein Konzept, das den am Sprachenlernen beteiligten Sprachen und Kulturen Rechnung trägt? Wie fließen diese Sprachbegriffe ein in die Vorstellungen von Erwerb und Lernen? Wie lernen Menschen Sprache und wie kann diese erfolgreich vermittelt werden? Welches Vermittlungskonzept wird diesem Bild von Sprache gerecht? Wie können Lernfortschritte valide und verlässlich vermessen werden mit dem Anspruch, Sprache, Lernen und Lernenden gerecht zu werden?

Sieht man Sprache beispielsweise als ein mechanistisches System von teilbaren, isolierten Einheiten, deren Summe wiederum „das Ganze" ergibt, so wird sich dies in der Vermittlung etwa daran zeigen, dass Sprache in portionierbare Einheiten aufgeteilt wird, diese Einheiten punktuell und gewissermaßen isoliert einübt werden und sie dann als erworben gelten, wenn die Lernenden sie wiedergeben können. Der Lernerfolg wird dann möglicherweise kontrolliert in Form von *discrete point tests*[18], die die erwähnten isolierten Teile der Sprache erfassen, ohne je „das Ganze" in den Blick zu nehmen. Betrachtet man Sprache hingegen als „unteilbares Ganzes" und Sprachverwendung als eingebettet in größere Kommunikationszusammenhänge, so wird man die Vermittlung von Sprache integrativ ausrichten und an authentischen Kontexten orientieren, um den natürlichen Auftrittsbedingungen von Sprache und Kommunikation gerecht zu

[18] Vgl. für eine Definition den Eintrag zu *Testformaten* im Glossar oder die Ausführungen in Kapitel 2.2 dieser Arbeit.

werden. Entsprechend dürften zur Lernfortschrittskontrolle integrative und kommunikative Testformate eingesetzt werden, die sich Sprache, Sprachverwendung und Sprachvermögen möglichst ganzheitlich nähern.

Den Sprachtests kommt soziale und politische Bedeutung zu: Oft genug bestimmen Tests über weiterführenden Schulbesuch und Karrieremöglichkeiten. Sprachtests können aber auch ein Mittel der Sprachenpolitik sein, gerade in Europa mit seiner Sprachenvielfalt. Aufgrund dieser Reichweite von Sprachtests ist es unabdingbar, sich über Konzepte der Sprache und Bedingungen des Lehrens und Lernens Klarheit zu verschaffen, ehe man sich dem Testen zuwendet. Auch der oben vorgestellte GER ist solchen Überlegungen verpflichtet, weshalb er in dieser Arbeit am Ende der Unterkapitel 1.2 *Sprachbegriffe* und 1.3 *Lern- und Vermittlungskonzept* hinsichtlich seiner Aussagen zu den dort jeweils erarbeiteten Konzepten analysiert wird.

1.1 Historischer Kontext

Solch grundlegende Konzepte und Begriffe wie die oben genannten sind angesiedelt im jeweils gerade gültigen wissenschaftlichen Paradigma, weshalb sich Sprachbegriffe, Lernkonzepte, Vermittlungsmethoden und Testansätze auch immer einem bestimmten Paradigma zuordnen lassen. Paradigmen sind stets geprägt von politischen, gesellschaftlichen und wissenschaftlichen Vorstellungen der jeweiligen Zeit. Sie entwickeln sich in Form von Pendelschlägen, da die Konzepte und Theorien in der Regel durch Widerlegungen und Antithesen vorangetrieben, erweitert und verbessert werden.[19] Näher auf diese Entwicklungen einzugehen würde den Rahmen der vorliegenden Arbeit sprengen, weshalb hier ein knapper Aufriss erlaubt sei, der es ermöglicht, den notwendigen Rahmen zu skizzieren, der aber naturgegeben ein nur grobes Bild der Paradigmen und der Paradigmenwechsel zeichnen kann, die sich während des letzten Jahrhunderts ausmachen lassen.

Spolsky (1978b) unterscheidet drei Phasen, die sich in der Geschichte des Sprachtestens ausmachen lassen. Zumindest für die beiden jüngsten Phasen möchte ich versuchen, Parallelen auch für die Betrachtung von Sprache und Sprachenlernen zu skizzieren. Spolsky setzt eine “prescientific period“ an, in der man sich der Grammatik-Übersetzungsmethode verpflichtet sah und während derer man sich der Übersetzung und des Aufsatzes als Testformate bediente, wobei es keine statistischen Anforderungen gab, denen Tests genügen mussten. Dieses Paradigma kann man, zumindest bezogen auf Deutschland, bis zur Mitte des 20. Jahrhundert ausmachen.

Ab den 50er Jahren beeinflussten Linguistik und Psychologie mit ihren Ansätzen des Strukturalismus respektive des Behaviorismus auch das Sprachlernen. Diese Phase nennt Spolsky die “psychometric-structuralistic period“. Sprache wurde über kontrastive Analysen in isolierte

[19] Vgl. hierzu u. a. Farhady 1979.

Bestandteile zerlegt, von denen man annahm, dass deren „Summe“ wieder „das Ganze“ ergäbe. Sprache und das Verhalten ihrer Elemente wurden beschrieben mittels eines Systems von Regeln und (oft nicht erklärbaren) Ausnahmen nach dem *Item-and-Process*-Modell[20] bezüglich der Form, Bedeutung und Verteilung der Elemente. Dabei wurden die Elemente auf unterschiedlichen isolierten Ebenen und in klar voneinander trennbaren Kategorien dargestellt (Ebenen wie *sentence – clause – phrase – word – morpheme – phoneme*, vgl. Farhady (1978: 348), und Kategorien wie beispielsweise die Wortarten auf der Wort-Ebene). Spracherwerb wurde mittels der „Kontrastivhypothese“[21] erklärt: Der Erwerb der Zweitsprache L2 werde von der Struktur der Muttersprache L1 bestimmt, wodurch sich bestimmte Phänomene wie „positiver Transfer“ von L1-Strukturen auf L2 oder „Interferenzen“ – nicht korrekter Transfer von L1 auf L2 – erklären ließen. Parallel dazu fasste man Lernen im Sinne des Behaviorismus auf als Einüben bestimmter Gewohnheiten, *habits*, die in so genannten *pattern drills* gefestigt wurden. Dabei wurden Sprachelemente nach den Ansätzen des Strukturalismus isoliert voneinander und ohne Kontext, doch kontrastiv zur Muttersprache dargestellt und vermittelt. Diesem Ansatz war die audio-linguale Methode verpflichtet. Die Annahme, dass der Erwerb isolierter sprachlicher Elemente additiv zur Sprachkompetenz führe, resultierte im Testen eben dieser isolierten Teilbereiche der Sprache mittels so genannter *discrete-point tests*, welche allerdings der Komplexität von Sprache, Spracherwerb bzw. -lernen und Sprachverwendung nicht gerecht werden konnten. Auf *shortcomings* und Kritik dieses Testansatzes wird in Kapitel 2 *Das Testen des Sprachvermögens* näher eingegangen.

Ab den 70er Jahren überwog die Kritik an Strukturalismus und Behaviorismus und neue Erkenntnisse aus Linguistik und Psychologie führten zu neuen Ansätzen in der Sprachbetrachtung, der Sprachvermittlung und beim Sprachtesten. Spolsky nennt diese Zeit die “integrative-sociolinguistic period“. Soziolinguistische Betrachtungen führten zur Kontextualisierung[22] von Sprache, zu einer Beschreibung der sprachlichen Elemente in ihren gegenseitigen Wechselbeziehungen untereinander und zur Situation, dem Kontext, in dem sie auftraten. So wird die Bedeutung eines Wortes beispielsweise nicht mehr als idiosynkratische Eigenart des Lexikons betrachtet, sondern sie ergibt sich immer erst aus dem sprachlichen und außersprachlichen Kontext, in dem das Wort benutzt wird. Sprache wird nicht mehr als System isolierter Elemente verstanden, sondern zunehmend integrativ in ihren Zusammenhängen und Wechselbeziehungen betrachtet.[23] Neueste Ansätze werden unter Kapitel 1.2 *Sprachbegriffe* besprochen.

Kognitive Theorien bezüglich des Erwerbs bzw. Lernens führten zur Entwicklung der „Identitätshypothese“[24], nach der der Erwerb von L1 und L2 nach ähnlichen Prinzipien erfolge, die jedem Sprachenlernen eigen seien. So bestimme nicht die Struktur der L1 den Erwerb, sondern

[20] Dem IP-Modell liegt die Annahme zugrunde, dass das Verhalten sprachlicher Elemente, der *items*, durch Prozesse determiniert wird, die über Regeln und Ausnahmen beschrieben werden können. Vgl. hierzu u. a. Köpcke 1993 und Frey 2001.
[21] Vgl. hierzu u. a. Brown (1994: 48ff) und Bausch et al. (1995: 471ff).
[22] Vgl. hierzu u. a. Halliday & Hasan 1976 oder van Dijk 1977.
[23] Vgl. hierzu u. a. Halliday & Hasan 1989
[24] Vgl. hierzu u.a Bausch et al. (1995: 472f).

Strukturen einer jeden Sprache würden nach ähnlichen Sequenzen, in einer „natürlichen Abfolge“ erworben. Nicht zuletzt aus der Kritik am Behaviorismus entwickelte Chomsky[25] Ende der 50er Jahre seine Hypothesen eines angeborenen Spracherwerbsmechanismus und einer allen Sprachen zugrunde liegenden *universal grammar*.

Psycholinguistisch orientierte Forschung führte zur „*Interlanguage*-Hypothese“[26], die besagt, dass Sprachenlernen zwar gekennzeichnet sein kann durch Transfer von L1 auf L2, aber auch durch von L1 unabhängige Prozesse. Nach dieser Hypothese bilden Lernende ein System von sukzessiven Lernervarietäten, die jeweils in sich kohärent sind und Übergänge darstellen auf dem Weg zur Zielvarietät. Diese Hypothese wurde seit ihrem Aufkommen modifiziert und erweitert und ist auch heute noch eine gültige These in der Theorie des Fremdspracherwerbs.

In den 80er Jahren stellte Krashen[27] fünf Hypothesen zum Spracherwerb auf, u. a. sein „Monitor Modell“, das auch für den Fremdsprachenunterricht bedeutsam ist. Krashen unterschied zwischen dem (unbewussten) Erwerb einer Sprache in natürlichen Kontexten, bei dem das Augenmerk auf den Inhalt der Kommunikation gerichtet sei, und dem (bewussten) Lernen in gesteuerten Situationen, bei dem es um Regelwissen gehe, das über den so genannten Monitor gesteuert und kontrolliert eingesetzt werde. Seine „Input-Hypothese“ besagt, dass bei genügend verständlichem Input Spracherwerb nach ihm eigenen, nicht beeinflussbaren Sequenzen „von selbst“ ablaufen würde. Krashens Modell ist umstritten, da sich Erwerb und Lernen nicht so streng trennen lassen, und da seine postulierten Erwerbssequenzen sprachübergreifend nicht schlüssig nachgewiesen werden konnten.

Bezogen auf den Fremdsprachenunterricht entwickelte sich der „kommunikative Ansatz“, der sich abwandte von einer formalen Betrachtung des Sprachenlernens hin zu einer ganzheitlichen Betrachtung dessen, was es heißt, „eine Sprache zu können“. Bachmanns Konzept der kommunikativen Kompetenz gewann zunehmend an Bedeutung, da es Sprachkompetenz nicht mehr als Summierung der isolierten sprachlichen Einzelfertigkeiten betrachtete, sondern ein weiterreichendes Konzept der an sprachlichen Äußerungen beteiligten Kompetenzen darstellte, das auch der Interdependenz von Sprache, Umwelt, kommunikativer Situation, beteiligten Personen und Persönlichkeiten, und der jeweiligen Funktion der Äußerung Rechnung zu tragen versuchte. Hierauf wird im Detail unter Kapitel 1.2.3 *Sprache als Mittel zur Kommunikation* und Kapitel 1.3.1 *Spracherwerb und internes Wissenssystem* eingegangen, da kommunikative Kompetenz auch heute noch das oberste Richtziel im Fremdsprachenunterricht darstellt und sich die Beurteilung des Sprachkönnens in der Regel auf diese kommunikative Kompetenz bezieht. Parallel dazu gewann auch beim Testen der *integrative approach* immer mehr an Bedeutung. Kritik an den *discrete-point tests*, auf die unter Kapitel 2.2 *Testformate und Auswertungsmöglichkeiten* im Einzelnen eingegangen wird, führte zur Entwicklung von Tests, die die

[25] Vgl. hierzu Chomsky 1959.
[26] Vgl. hierzu u. a. Selinker 1972, Selinker 1992, Bausch et al. (1995: 472f).
[27] Vgl. u. a. Krashen 1982, Krashen 1985.

sprachlichen und außersprachlichen Kontexte, in denen Sprache gebraucht wird, mit einbezogen und die nicht mehr nur formale Korrektheit, sondern auch die kommunikative Wirkung von Sprache betrachteten. Diese „neuen“ Testformate werden ebenfalls in Kapitel 2.2 betrachtet.

1.2 Sprachbegriffe

Sprache kann man unter den verschiedensten Gesichtspunkten beschreiben und definieren, doch diese Arbeit konzentriert sich auf die für Sprachtests relevanten Begrifflichkeiten, da der Klärung und Definition des Sprachbegriffs oder besser der Sprachbegriffe, auf denen Sprachtests basieren, elementare Bedeutung zukommt:

Das Bild der *innersprachlichen Organisation* bestimmt, in welchen „Einheiten“ Sprache vermessen wird: ob diskrete isolierte Elemente geprüft werden oder ob – am anderen Ende des möglichen Spektrums – Sprache als „Ganzes“, als Entität in ihren sprachlichen wie außersprachlichen Kontexten getestet wird. Die Struktur der Sprache als *internes Wissenssystem* wird u. a. bestimmt durch die Struktur der mentalen Repräsentation von Sprache im Gehirn. Diese legt nicht nur fest, wie Sprache überhaupt aufgrund unserer biologische Ausstattung erlernt werden kann, sie sollte im Idealfall auch Auswirkungen haben auf die Art und Weise, wie Sprache vermittelt und vermessen wird. Wieder andere Facetten bieten sich, wenn man Sprache als *Mittel zur Kommunikation* betrachtet: Dabei wird Sprache eingebettet in ihre sozialen Kontexte, um zu sehen, wie Kommunikation und Verständigung ablaufen. In diesem Zusammenhang ist die Struktur der kommunikativen Kompetenzen von Interesse, um darauf aufbauend zu entscheiden, welche Teilkompetenzen man testen will und mit welchen Testformaten man diese erfassen kann. Da Sprachen aber nicht isoliert voneinander und von der Welt existieren, lohnt es sich, auch den Zusammenhang zwischen *Sprachen und Kulturen* näher zu betrachten, der sich sowohl beim Erwerb einer Fremdsprache als auch bei der Vermittlung und beim Testen bemerkbar machen dürfte. *Last but not least* bringt die Betrachtung des Sprachbegriffs aus *Perspektive des Fremdsprachenunterrichts*, und hier im Speziellen der des Englischunterrichts, wiederum neue Aspekte mit sich, die unter Kapitel 1.3 *Lern- und Vermittlungskonzept* erhellt werden sollen, da Sprachtests überwiegend im Fremdsprachenunterricht angesiedelt sind und Schulleistungsstudien auch und insbesondere zur Verbesserung des Fremdsprachenunterrichts dienen können.

1.2.1 Innersprachliche Organisation – Prototypenmodell

Wenn man auch traditionell Sprache als arbiträres Zeichensystem betrachtet hat und morphologische, semantische und grammatische Eigenheiten entweder als idiosynkratische Eigenheiten

des Lexikons angesehen hat oder sie mittels des erwähnten *Item-Process*-Ansatzes zu beschreiben versuchte, so lässt sich in jüngerer Zeit ein Paradigmenwechsel hin zu einer Betrachtung von Sprache als motiviertes und selbstorganisierendes Netzwerk feststellen.

Die innersprachliche Organisation kann mittels eines ganzheitlichen, vernetzten, systemischen Modells der Sprache beschrieben werden: Sprache kann als ein lebendiges, selbstorganisierenden Prinzipien unterliegendes System aufgefasst werden, das aus Netzwerken aufgebaut ist, die wiederum in größere Netzwerke eingebettet sind, wobei das Ganze etwas anderes ist als die Summe der Teile, und wobei die Teile untereinander und mit dem Ganzen auf allen Netzwerkebenen in Interaktion stehen und sich gegenseitig beeinflussen. Es scheint, dass es Organisationsprinzipien in jeder natürlichen Sprache gibt, die die Entwicklungen dieses Netzwerks mit steuern. [28]

Der traditionelle Ansatz, Sprache in klar trennbare Kategorien nach kritischen Merkmalen einzuteilen, ist schon lange kritisiert worden. Beispielsweise hat Wittgenstein die Unzulänglichkeit von Klassifizierungsschemata festgestellt, die scharfe Trennlinien zwischen den Kategorien ansetzen und bei denen besagte kritischen Merkmale darüber entscheiden, ob ein Element in diese oder jene Kategorie fällt.[29] Das so genannte Prototypenmodell[30] der innersprachlichen Organisation hat sich als zutreffendes Modell erwiesen, wie Frey (2001: 24) beschreibt:

> An die Stelle der ‚kritischen Merkmale' und ‚scharfen Trennlinien' tritt bei Wittgenstein ein Kategorisierungskonzept, „das a) durch ein Netzwerk so genannter *Familienähnlichkeiten* und b) durch verschwommene Ränder charakterisiert ist" (Köpcke 1995, 162). Ein solches Modell beschreibt die sprachliche Realität (...) besser als der IP-Ansatz, denn „während bei einem Modell, das mit kritischen Merkmalen operiert, verlangt werden muss, dass jedes Mitglied der Klasse auch jedes der kritischen Merkmale aufweist und dass genau diese Merkmale nicht bei der Kontrastkategorie auftreten, wird bei dem Prototypenmodell davon ausgegangen, dass die Mitglieder einer Klasse nur eine mehr oder weniger große Familienähnlichkeit aufweisen. Zudem können manche der kritischen Merkmale auch bei den Mitgliedern der Kontrastklasse auftreten." (Köpcke 1995, 163f.).

Es werden Klassen um einen Prototypen herum angesetzt, wobei der Prototyp als „der beste Vertreter seiner Klasse" (Köpcke 1995: 163) gekennzeichnet ist durch ein Maximum an Merkmalen und Merkmalsbündeln. Sprachliche Phänomene, auch *items* genannt, ordnen sich um den Prototypen herum an, je nach ihrer Ähnlichkeit: Je ähnlicher ein sprachliches Phänomen dem Prototyp ist, das heißt je mehr Merkmale oder Merkmalsbündel es mit dem Prototyp teilt, desto wahrscheinlicher verhält es sich wie der Prototyp, wenngleich diese Annahme nicht deterministisch zu verstehen ist: Das *item* kann sich auch bei noch so großer Ähnlichkeit ganz anders verhalten als der Prototyp. Zwischen den Kategorien gibt es keine scharfen Trennlinien; vielmehr gilt das Prinzip der *gradience*, also gradueller, „verschwommener" Übergänge: Die Klassenmitglieder sind angeordnet auf einem Kontinuum zwischen den Klassen mit ihren zentralen Prototypen. Je weniger Merkmale ein Phänomen mit dem Prototyp einer Klasse teilt, desto weiter rückt es an die Peripherie dieser Klasse und desto wahrscheinlicher fällt es in eine

[28] Vgl. hierzu Köpckes *guiding principles* in Köpcke 1993.
[29] Vgl. Frey (2001: 23-27) für einen Überblick zu Köpckes Prototypenansatz, entwickelt aus der Kritik an klassischen Kategorisierungstheorien.
[30] Vgl. hierzu u. a. Köpcke 1995 und Frey (2001: 23-27).

andere Klasse. Auch wenn Köpcke dieses Modell aus dem älteren Schemata-Ansatz[31] weiterentwickelt hat, um der Morphologie (insbesondere der deutschen Pluralmorphologie) eine kognitive Basis zu geben, so lässt sich dieses Modell ebenfalls auf andere als morphologische Phänomene, etwa auf grammatische, anwenden, denn auch diese können nur unzureichend mit IP-Regeln beschrieben werden.

Das Lexikon und das grammatische Inventar einer natürlichen Sprache lassen sich über Regeln und (unmotivierte) Ausnahmen, die einem sprachlichen *item* zugeordnet werden, nicht hinreichend beschreiben: Beispielsweise kann das deutsche Pluralsystem auf diese Weise nicht adäquat charakterisiert werden. Eher schon dürften Lexikon und grammatisches Inventar in vielen Bereichen entlang von Prototypen organisiert sein und müssen dementsprechend beschrieben werden, denn bezogen auf das Sprachenlernen fällt es bekanntermaßen schwer, Listen von Regeln und unmotivierten Ausnahmen auswendig zu lernen. Selbst wenn diese dann in einem Test wiedergegeben werden können, so sagt dies noch nichts über die sprachliche Handlungskompetenz im realen Leben aus.

Ein gutes Beispiel für solch eine prototypische Organisation sind die morphologischen Eigenschaften der englischen *ing*-Form, die einen graduellen Übergang von adjektivischen zu substantivischen Eigenschaften aufweisen – diese Eigenschaften sind aber gerade nicht mit traditionellen Klassifizierungen zu fassen. Deshalb ist es für die Vermittlung solcher Phänomene bedeutsam, sie nach einem zutreffenderen Modell der innersprachlichen Organisation darzustellen, damit sie überhaupt von den Lernenden erfasst und als Schemata „abgespeichert“ werden können.

Auch Sprachtests müssen der innersprachlichen Organisation gerecht werden: Wenn man Sprache in „sauber“ voneinander getrennten Bereichen testet, gerade wenn die sprachlichen Bereiche auch derart im Fremdsprachenunterricht dargeboten worden sind, kann dies zwar helfen festzustellen, ob das Vermittelte auch gelernt worden ist, doch erhält man damit keine Aussagen darüber, inwieweit sprachliches Wissen beim Lerner vernetzt ist und inwieweit es in Alltagssituationen anwendbar ist. Beispielsweise könnte man sich vorstellen, dass der Unterschied zwischen *simple* und *continuous form* besprochen wurde, die Regeln des Auftretens der beiden Formen gegeben wurden, einige Beispielsätze und Einsetzübungen behandelt wurden und dann die beiden Formen durch die erwähnten *discrete-point tests* geprüft werden; doch damit erhält man bestenfalls eine momentane Bestandsaufnahme, die im Fremdsprachenunterricht durchaus gerechtfertigt ist, von welcher man jedoch nur bedingt auf das anwendbare Wissen und Können schließen kann. Dazu erhält man eher Zugang, wenn man die betreffenden sprachlichen Elemente in ihren natürlichen Kontexten darbietet und sie von den Lernenden mit Unterstützung der Lehrkraft analysieren lässt im Hinblick auf ihre linguistischen und soziopragmatischen Auftrittsbedingungen, Funktionen und Bedeutungen. Dabei kann durchaus der Vergleich der Versprachlichungsmöglichkeiten der betreffenden Funktionen in anderen

[31] Vgl. dazu Bybee 1976 und 1991.

Sprachen herangezogen werden (hier bietet sich gleichzeitig noch die Gelegenheit, alle anderen für die Lernenden relevanten Punkte zu besprechen, zu wiederholen, zu analysieren). Danach können kontextualisierte Übungen angeboten werden (die sich wiederum nicht nur auf dieses Phänomen beschränken müssen, sondern aufgrund des ganzheitlichen Kontextes auch Gelegenheit zur Übung anderer für die einzelnen Lerner ebenfalls relevanten Phänomene bieten), und bei der Überprüfung wiederum integrative Formate verwendet werden, die es den Lernenden ermöglichen, das holistisch erworbene Wissen in eben solchen Kontexten anzuwenden. Diese Vorgehensweise dürfte auch unserer „biologischen Ausstattung", unserem Gehirn, entsprechen. Im Folgenden soll deshalb kurz darauf eingegangen werden, wie Sprache als Wissenssystem mental repräsentiert ist.

1.2.2 Sprache als internes Wissenssystem – mentale Repräsentation

Wenn die innersprachliche Organisation gemäß dem Prototypenmodell beschrieben werden kann und wenn Sprecher zu einem sprachlichen Phänomen bestimmte „Gestalten" oder Schemata[32] abgespeichert haben, die bestimmte Assoziationen auslösen bezüglich der Funktionen oder des Verhaltens des betreffenden Phänomens[33], so ist es für das Sprachenlernen, Lehren und Beurteilen unabdingbar, dieser „naturgegebenen" Organisation gerecht zu werden. Dazu müsste sowohl die Struktur des Wissenssystems „Sprache" bei kompetenten Sprechern beschrieben werden als auch die kognitiven Strukturen, die diesem Wissenssystem zugrunde liegen.[34]

Die Struktur der internen Wissensbestände, die „kognitive Ausstattung", von der Edmondson (1998) spricht, ist noch nicht hinreichend beschrieben, doch Edmondson gibt einen hilfreichen Überblick über einige „Merkmale des Sprachbesitzes und des Sprachgebrauchs" bei Muttersprachlern, die als „Außenkriterium" dazu beitragen können zu bestimmen, „was es denn heißt, eine Fremdsprache zu können" (Edmondson 1998: 32):

> Es wird in der kognitiven Psychologie zwischen deklarativem und prozeduralem Wissen unterschieden. Diese Unterscheidung zwischen dem „Wissen, dass" und dem „Wissen, wie" ist auf sprachliches Wissen zu übertragen; sie ist natürlich für den Fremdsprachenunterricht von großer Bedeutung. Sprachliches Wissen ist mit Weltwissen und dem episodischen Gedächtnis vernetzt. Sprachliche Formen können mehrmals in unterschiedlichen Repräsentationen gespeichert sein. Das Abrufen sprachlichen Wissens erfolgt normalerweise ganz automatisch, wobei die Automatisierung u. a. eine Funktion der Verwendung ist. Je nach Sachkenntnissen und Gewohnheiten ändert sich die Schnelligkeit der Aufnahme von längeren und syntaktisch komplexeren sprachlichen Elementen; die ganzheitlich verstanden, gespeichert, abgerufen und produziert werden können.

Wenn sprachliches Wissen mit anderen Wissensbeständen verknüpft ist, so stützt dies die oben erwähnte Schemata-These: Sprachliche Formen werden verbunden mit Schemata im Gehirn „gespeichert", so dass die innersprachliche Organisation, wenn sie nach dem Prototypenmodell

[32] Vgl. dazu u. a. Köpcke 1988 und 1993.
[33] Hier sind eben diese Schemata oder Merkmale gemeint, die beim Prototypenmodell das Verhalten des Phänomens in probabilistischer Weise mitbestimmen.
[34] Zum Zusammenhang von Sprachstruktur und Testformaten vgl. Kapitel 2.3.2 dieser Arbeit.

strukturiert ist, über diese Schemata vom Gehirn erfasst werden kann, ohne dass dies den Sprechern bewusst werden müsste.

Wenn sprachliche Formen mehrfach mental repräsentiert sind, so dürfte dies dazu führen, dass das Wissen darum in den unterschiedlichsten Kontexten durch diese Mehrfachrepräsentationen schneller abrufbar wird. Wenn Tests an diese Abrufbarkeit im Sinne von Verwendbarkeit des Wissens in verschiedenen Situationen heranreichen wollen, so müssen die Tests auch solchen Situationen gerecht werden, die das Wissen in natürlichen Situationen aktivieren – ein Testen isolierter Wissensbestände lässt sich nicht auf sprachliche Handlungsfähigkeit im Alltag verallgemeinern. Zur Handlungsfähigkeit gehört auch das automatisierte Verwenden von Sprache, wobei sich Automation und Verwendung vermutlich gegenseitig unterstützen und bedingen – dies muss ebenfalls beim Sprachenlernen und bei der Fremdsprachenvermittlung beachtet werden. Wenn Fremdsprachentests u. a. auch über den Grad der Automation Aufschluss geben sollen, so muss dem in der Testkonzeption Rechnung getragen werden. Wenn schließlich nach Edmondson Sprachverarbeitung ganzheitlich erfolgt, so muss sich diese Art der Verarbeitung, des *language processing*, auch im Fremdsprachenunterricht und in Sprachtests niederschlagen, wie das verschiedentlich ja auch der Fall ist.

Nun lohnt es einen Blick zu werfen auf Erkenntnisse aus der Neurobiologie, um zu sehen wie Sprache im Gehirn organisiert ist und welche der obigen Merkmale in unserer „biologischen Ausstattung“ reflektiert werden. Zur neuronalen Organisation von Sprache bei Erwachsenen fassen Neville & Bavelier (1998) neuere Studien zur Visualisierung der Hirnaktivitäten bei gesunden Menschen zusammen. Traditionell werden drei klar umrissene Areale in der linken Hemisphäre unterschieden, die für das Planen und Ausführen von Sprache, für die Analyse und Identifizierung von Sprache und respektive für die Dekodierung beim Lesen zuständig sein sollen. Die Bedeutung dieser drei Regionen wurde bestätigt, doch es zeigten sich auch neue Aspekte: Die Sprachzentren sind keine klar umrissenen homogenen Regionen, sondern es handelt sich um kleine, voneinander getrennt liegende “focal spots“ (Neville & Bavelier 1998: 254), die auf bestimmte Aspekte von Sprache spezialisiert sind. Das bedeutet, dass Sprache zwar nicht als Gesamtheit im Gehirn abgespeichert ist, doch dass Sprachgebrauch diese *focal spots* gemeinsam aktiviert. Diese sprachbezogenen Aktivierungen wurden nicht nur in den klassischen sprachbezogenen Hirnregionen beobachtet, sondern auch außerhalb dieser Zentren. Es scheint, dass Sprachgebrauch viel mehr aktiviert als rein kognitiv sprachbezogene Gehirnaktivitäten; dies stützt die obige These, dass sprachliches Wissen auch mit anderen Wissensbeständen vernetzt ist.

Des Weiteren zeigte sich, dass die sprachbezogenen Hirnregionen eher ausgerichtet sind auf sprachliche Systeme als auf sprachliche Fertigkeiten: “The functional role of the language-related areas is more accurately characterized in terms of linguistically relevant systems, such as phonology, syntax and semantics, than in terms of activities, such as speaking, repeating, reading and listening.“ (ebd.: 254) Man könnte vorsichtig schlussfolgern, dass Sprache

systemisch abgespeichert ist und dieselben Systeme und Subsysteme jeweils zur Ausführung der unterschiedlichen sprachlich-kommunikativen Aktivitäten herangezogen werden.

Die Rolle der sprachbezogenen Regionen im Gehirn ist keineswegs abschließend erforscht, doch es lassen sich einige generelle Prinzipien ausmachen. Beispielsweise scheint Lexik in Subsystemen organisiert zu sein, die in etwa den Wortklassen wie Verb, Adjektiv und Nomen entsprechen; innerhalb der Wortklassen scheinen die Nomen nach semantischen Relationen organisiert zu sein (ebd.: 254). Das heißt, dass Wortschatzelemente nicht isoliert abgespeichert werden, sondern eher in Feldern, die durch „Familienähnlichkeiten" gekennzeichnet sind. Dies könnte man zur weiteren Bestätigung des Prototypenmodells heranziehen. Ein anderes generelles Prinzip betrifft die syntaktische Analyse: Diese scheint bei der Wort- und Satzverarbeitung auf je einer eigenen Ebene abzulaufen (ebd.: 255). Dies bestätigt die These, dass es neben der traditionellen Satzgrammatik auch eine Ebene der Wortgrammatik[35] gibt. Auch dies macht sich im Fremdsprachenunterricht bemerkbar, beispielsweise bei der Einführung neuen Wortschatzes:

Wenn Wortschatzelemente nicht isoliert abgespeichert werden und es zu syntaktischer Verarbeitung auch auf Ebene der Wortgrammatik kommt, so macht es Sinn, Lexik in Wortfeldern darzubieten, in ihre jeweiligen sprachlichen und außersprachlichen Kontexte eingebettet (hier zeigt sich, welches Wort mit welchen anderen in welchen Konstellationen auftreten kann), um sie eingebettet in ihre „natürlichen" Auftrittsbedingungen darzustellen, abzuspeichern und schließlich anzuwenden. Bezogen auf Sprachtests muss die Folgerung lauten, dass auch diese der mentalen Repräsentation genügen müssen, um Sprachkönnen valide zu erfassen.

1.2.3 Sprache als Mittel zur Kommunikation – Modell der kommunikativen Kompetenz

Sprache existiert nicht um ihrer selbst Willen, sondern stellt innerhalb der Sprechergemeinschaft das Kommunikationsmittel schlechthin dar. Sprache dient u. a. dazu, unsere individuellen wie kollektiven Welten hervorzubringen und zu vermitteln. Hier soll es aber nicht um den individuellen „Besitz" an kommunikativen Verhaltensweisen gehen, sondern um den idealisierten Bestand, die idealisierte kommunikative Kompetenz, die Muttersprachler befähigt, angemessen und effektiv zu kommunizieren. Diese Kompetenz soll in ihren Bestandteilen und, soweit möglich, in ihrer Struktur beschrieben werden, um ausgehend von diesem Zielvorgaben in Anlehnung an *native speakers* Ableitungen für relevante Fragen des Fremdsprachenunterrichts treffen zu können. Fragen wie beispielsweise „Was soll gelehrt werden? Welche Fertigkeiten und sprachlichen Teilgebiete müssen die Lernenden beherrschen?" können erst angegangen werden, wenn die Frage beantwortet ist, was an Kompetenzen bei Muttersprachlern im Idealfall vorhanden ist. Im Gegensatz zu dieser idealisierten Kompetenz stehen die tatsächlichen

[35] Diese These wurde beispielsweise von Prof. Dr. Götz (Universität Augsburg) in seinen Seminaren zu Syntax und Grammatik geäußert. (Proseminar *Syntaktische Analyse* WS 97/98 und Hauptseminar *Von der Grammatik zum Kursmaterial* SoSe 1999).

Kompetenzen der Lernenden, die in Leistungstests erfasst und beschrieben werden können.[36] Dabei fungiert die hier idealisierte kommunikative Kompetenz als Richtzielvorgabe, auf die hin Fremdsprachenunterricht und Leistungstests ausgelegt werden könnten.

Nun kann es bei der Beschreibung dessen, was alles zur Kommunikation beiträgt, nicht nur um die sprachlichen Mittel und Bestände einer Sprache gehen, sondern es müssen auch die Kontexte mit einbezogen werden, in denen Kommunikation auftritt. Denn Sprache funktioniert innerhalb eines breiteren kulturellen Rahmens, in dem sich Sprache und Kultur gegenseitig bedingen. Beide Systeme sind funktional organisiert: Sprache dient der Orientierung in der Umwelt, sie funktioniert als ein „Gefäß" von individuellen und kollektiven Erfahrungen, als Transportmittel, um sich operatives Weltwissen anzueignen. Kulturfunktionen sieht beispielsweise Buttjes (1991) in der „Manipulation der materiellen Welt" und in der Diskussion von Werten, in den „symbolisch-expressiven Dimensionen des Verhaltens", in der „Dramatisierung des täglichen Lebens und den ritualisierten Aspekten sozialer Arrangements" (ebd.: 7f). Dieses Zusammenspiel kann in einem Modell beschrieben werden, das Kultur und Sprache als ein dynamisches, komplexes und vielschichtiges Netzwerk begreift, welches nie in seiner Ganzheit fassbar ist und in dem „das Ganze ... die Teile (bestimmt, und) die Wechselwirkungen der Teile das Ganze (bestimmen)" (Picht 1995: 68).

Sprachliche Äußerungen können und sollten demnach nicht in Isolation betrachtet werden, denn der *context of situation*, der eingebettet ist in den größeren *context of culture*, schafft erst die Rahmenbedingungen, den *context of meaning*, um einer Äußerung Sinn und Bedeutung geben zu können (vgl. u. a. Halliday & Hasan 1989). Die unmittelbare kommunikative Situation lässt also Rückschlüsse auf die Art der Versprachlichung und auf die Bedeutung einer Äußerung zu, genau wie die Äußerung Rückschlüsse auf die Situation und den Kontext zulässt. Diese gegenseitige Vorhersagbarkeit dürfte eine der Grundlagen für das Verstehen sprachlicher Äußerungen sein. Denn das System sprachlicher Formen und das Bedeutungssystem einer jeden Sprache stehen in funktionaler, motivierter Beziehung zueinander, so dass die Wahl der angemessenen sprachlichen Form mitbestimmt wird durch den jeweiligen Kontext, in dem sie eine gewisse Bedeutung annehmen soll.

Des Weiteren spielen bei der Kommunikation (kulturbedingte) Regeln und Verhaltensnormen auch bezüglich non- und paraverbaler Kommunikation eine Rolle, derer sich u. a. Soziolinguistik und Pragmatik angenommen haben. Nicht zu vergessen sind außersprachliche Wissensbestände wie beispielsweise Weltwissen, und Strategien hinsichtlich des angemessenen Wissenseinsatzes, um beispielsweise aus einer Andeutung die richtigen Schlussfolgerungen ziehen zu können.

Wie hängen nun die oben angesprochenen Teilkompetenzen der kommunikativen Kompetenz miteinander zusammen? Gibt es einen generellen Faktor, der Sprachkönnen erklären

[36] Diese werden unter Kapitel 1.3.1.2 der vorliegenden Arbeit beschrieben.

könnte? Wie sieht die Struktur dieser idealisierten Kompetenz aus? Dazu wird ein kurzer Blick auf bisherige Modelle geworfen, um dann auf Bachmanns Modell der kommunikativen Kompetenz zu fokussieren:

In den 60er Jahren trafen Lado (1961) und Caroll (1972) die wichtige Unterscheidung zwischen den *skills*, also den (prozeduralen) Fertigkeiten, und dem (deklarativen) Wissen um Sprachkomponenten; doch sie beschrieben nicht den Zusammenhang und die Wechselwirkungen zwischen „Können" und „Wissen", wie sie auch den Kontext von Sprache nicht in ihr Modell mit aufnahmen. Erst in den 70er Jahren wurde die Bedeutung des sprachlichen, soziolinguistischen und des außersprachlich-situativen Kontextes gewürdigt und beispielsweise von Halliday (1976) und van Dijk (1977) beschrieben. Hymes (1972a) bezog soziolinguistische Faktoren mit in die Sprechsituation ein und betrachtete Performanz als Interaktion zwischen der eigenen Kompetenz (Wissensbestände und Fähigkeit der Wissensanwendung), den Kompetenzen der Interaktionspartner und den Eigenarten der Sprechsituation selbst. Savignon (1983) sah Kommunikation als ein dynamisches Konstrukt, bei dem es um die Bedeutungsaushandlung zwischen den Kommunikationspartnern gehe; Kommunikation sei situationsabhängig und höchst vielfältig und der kommunikative Erfolg hänge u. a. auch von der Einschätzung der Situation und von Vorerfahrungen ab.

Bezogen auf sprachliche Kompetenzen gab es in den 70er Jahren die Diskussion um den so genannten *general language proficiency factor* (vgl. hierzu u. a. Oller 1976) – Spearman hatte Anfang des 20. Jahrhunderts statistisch einen generellen Intelligenzfaktor in einem mechanisch-additiven Modell mittels der von ihm entwickelten Faktorenanalysen[37] errechnet; Oller (ebd.) errechnete den o. g. *general language proficiency factor* mittels Faktorenanalyse, den er mit der *internal grammar* gleichsetzte. Dies kann heute so nicht mehr aufrechterhalten werden, denn zu Sprachkönnen gehört – wie oben erläutert – wesentlich mehr als nur grammatisches Können. Man geht heute i. A. von einem hierarchischen Modell aus, innerhalb dessen sich ein Generalfaktor und zusätzliche Subdimensionen statistisch errechnen lassen. Die generelle Sprachkompetenz i. S. v. kommunikativer Kompetenz lässt sich angemessener als Profil betrachten, wobei es neben der Beschreibung der *skills* und der Wissenskomponenten auch um das notwendige Wissen zum erfolgreichen Sprachgebrauch geht.

Bei der Beschreibung der kommunikativen Kompetenz wird die Schnittstelle zwischen „Sprache als sozialem Gut" (vgl. Edmondson 1998: 29) im Sinn der idealisierten Kompetenz und der je individuellen Ausprägung von Sprache im Sinne eines Kompetenzprofils bei den einzelnen Sprechern einer Sprachgemeinschaft deutlich. Dieses Profil kann in kommunikativen Tests erfasst werden, die in ihrer Entwicklung auf der idealisierten Kompetenz basieren – Näheres dazu unter Kapitel 2.2 *Testformate und Auswertungsmöglichkeiten*. Naturgemäß spielt auch die Persönlichkeit der Kommunikationspartner eine Rolle, doch diese sollte soweit möglich außen

[37] Faktorenanalysen sind ein statistisches Verfahren, um Zusammenhänge zwischen verschiedenen Variablen zu untersuchen; es darf auf das Glossar verwiesen werden.

vor gelassen werden, denn es sollen kommunikative – und nicht personale – Kompetenzen vermittelt und erfasst werden, obwohl beide zugegebenermaßen miteinander in Wechselbeziehungen stehen.

Bleiben wir aber noch bei den Grundlagen und wenden uns den Arbeiten Bachmanns zu, der Basisforschung zur Sprachtestentwicklung geleistet hat.[38] Sein Modell der kommunikativen Kompetenz ist maßgebend für das Testen von Sprache und Kommunikationsfähigkeit. Es basiert auf seinem Konzept der CLA, der *Communicative Language Ability*, die sich zusammensetzt aus einer Wissens- oder Kompetenzkomponente und der Fähigkeit, dieses Wissen umzusetzen in angemessenen, kontextualisierten und kommunikativen Sprachgebrauch:

Abb. 1 (nach Bachmann 1991a: 85): Komponenten der CLA

Nun werden Sprachkompetenzen und strategische Kompetenzen näher betrachtet, da es bei Sprachtests vorwiegend um deren Erfassung geht.

Sprachkompetenzen umfassen bei Bachmann (1991a: 86) sprachliches Wissen, soziokulturelles Wissen und Diskurswissen. Empirische Validierungsversuche dieser Komponenten waren nicht sehr schlüssig, doch Bachmann & Palmer (1987) fanden mittels der Auswertung von Testbatterien Hinweise darauf, dass ihre postulierten Elemente der *communicative proficiency* – namentlich *grammatical competence, pragmatic competence* und *sociolinguistic competence* – voneinander unterscheidbar seien, wobei die Beziehungen zwischen grammatischen und pragmatischen Komponenten enger seien denn zu soziolinguistischen Elementen. Im Folgenden wird Bachmanns schematisches Modell so dargestellt, dass die Interaktionen zwischen den Komponenten etwas deutlicher werden als in seinem Diagramm, denn er schreibt dazu:

> In this case, this diagram represents the hierarchical relationships among the components of language competence, at the expense of making them appear as if they are separate and independent of each other. **However, in language use these components all interact with each other and with features of the language use situation**. (Bachmann 1991a: 86. Herv. d. V.)

[38] Der Begriff der kommunikativen Kompetenz wurde von Hymes in den 70er Jahren geprägt. Doch diese Entwicklung darzustellen, würde hier zu weit führen. Die Beschränkung auf Bachmanns Modell der kommunikativen Kompetenz sei deshalb gestattet, da es für Testerwägungen am meisten bietet.

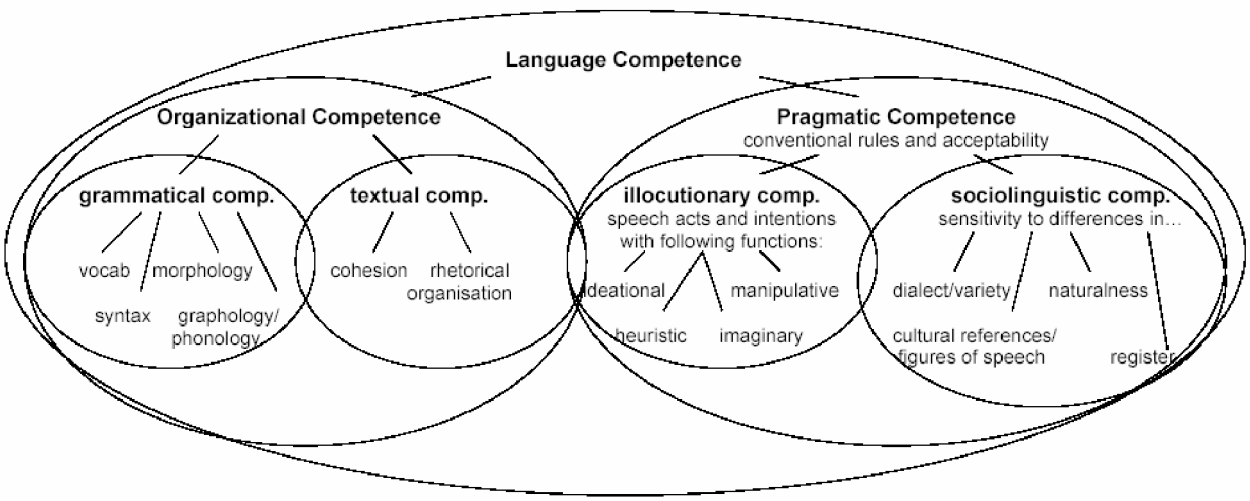


Abb. 2: (Bachmann 1991a: 87): Komponenten der Sprachkompetenz; Ellipsen von Verf. zur Verdeutlichung der Interdependenzen

Strategische Kompetenzen beziehen sich auf den dynamischen Kommunikationsprozess, bei dem unterschiedliche Strategien eingesetzt werden, um angemessen und effektiv zu kommunizieren. Hierunter werden auch die interimsprachlichen Kommunikationsstrategien, die beim Erwerb und Lernen einer Fremdsprache zum Einsatz kommen, verstanden, doch diese Art der Strategien wird erst beim Lernen unter Kapitel 1.3.1 *Spracherwerb und internes Wissenssystem* besprochen; hier soll es wie gesagt um idealisierte Zielvorgaben gehen. Man kann strategische Kompetenzen aus interaktionaler Sicht oder aus psycholinguistischer Sicht definieren. Aus erstgenannter Sicht werden solche Strategien verstanden als Strategien der Bedeutungsaushandlung und als Versuche, das linguistische System effektiv und mit einem Minimum an Aufwand zu nutzen.[39] Canale (1983: 339) sieht sie als "mastery of verbal and nonverbal strategies both (a) to compensate for breakdowns in communication due to insufficient competence or to performance limitations and (b) to enhance the rhetorical effect of utterances."[40] Allerdings lassen diese Betrachtungen die psycholinguistischen Wirkmechanismen außer Acht, die hinter dem Einsatz solcher Strategien stehen, weshalb Bachmann sich auch mit Modellen der Sprachproduktion beschäftigt. Færch & Kasper (1983) haben zu dieser Thematik ein Modell entwickelt, welches die Einschätzung einer Situation, die Planung und die Ausführung einer Sprechhandlung umfasst. [41] Eingeschätzt werden die Informationen, die in einer gegebenen Situation benötigt werden, um das kommunikative Ziel zu erreichen; die sprachlichen Ressourcen, die benötigt werden, diese Informationen effektiv zu transportieren; und der Umfang des Erfolgs, mit dem das Ziel erreicht wurde. In der Planungsphase werden diese Einschätzungen dann in einen Plan umgesetzt und die relevanten Komponenten der sprachlichen Kompetenzen aktiviert, um das kommunikative Ziel zu erreichen. Die Funktion der strategischen Kompetenzen sieht Bachmann (1991a: 102) wie folgt:

[39] Vgl. beispielsweise Tarone 1981.
[40] Zitiert in Bachmann (1991a: 99). Man sieht auch hier wieder die Schnittstelle zwischen idealisierter Kompetenz und individuell ausgeprägten Kompetenzen der einzelnen Sprecher.
[41] Angeführt in Bachmann (1991a: 100ff).

> It is the function of strategic competence to match the new information to be processed with relevant information that is available (including presuppositional and real world knowledge) and map this onto the maximally efficient use of existing language abilities.

Bei der Sprachproduktion interagieren sprachliche Kompetenzen, strategische Kompetenzen und die Sprachgebrauchssituation. Das von Bachmann adaptierte Sprachproduktionsmodell von Færch & Kasper (1983) soll hier vorgestellt werden, um die Prozesse deutlich zu machen, die hinter jeder Sprachproduktion stehen, auch wenn diese in Sprachtests meist nicht erfasst werden können. Dennoch muss man um sie wissen, um valide Tests entwickeln zu können:

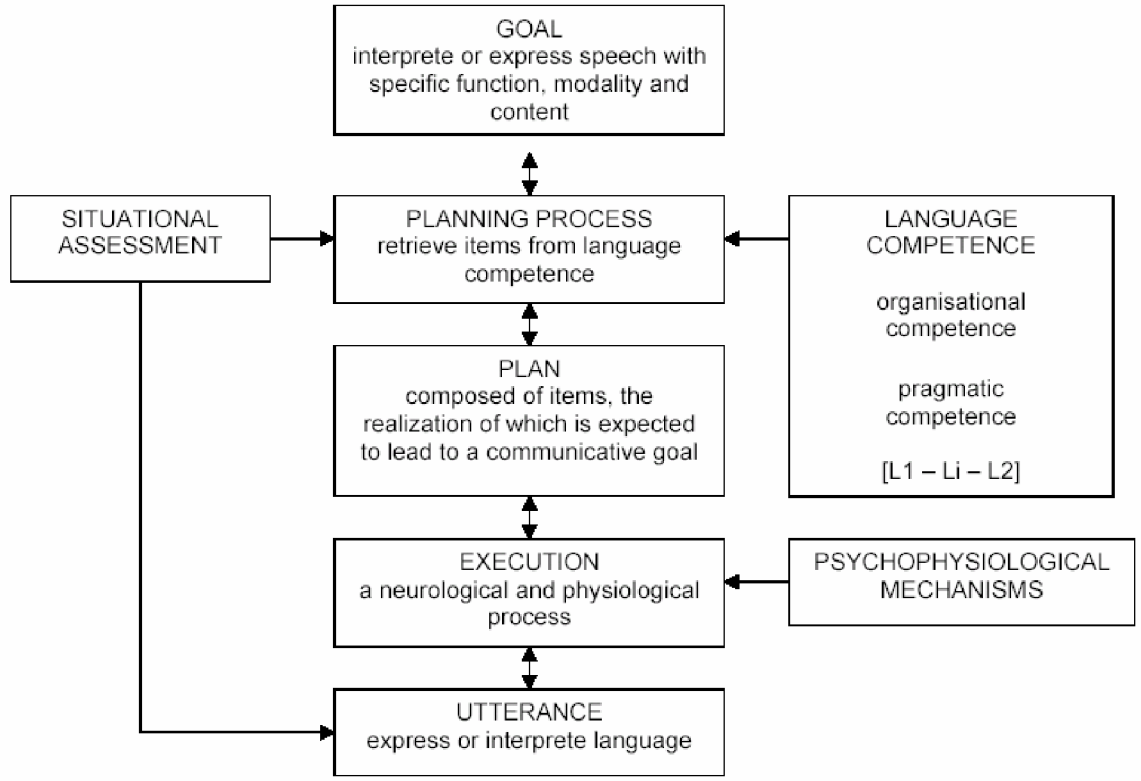


Abb. 3 (Bachmann 1991a: 103): Modell des Sprachgebrauchs

Ausgangspunkt jeder Sprachproduktion ist das kommunikative Ziel, das es mit gegebenen sprachlichen Mitteln zu erreichen gilt. Dazu müssen die Situation und das sprachliche und pragmatische Können eingeschätzt werden, ehe die Äußerung geplant werden kann. Bei der Umsetzung in eine tatsächliche Äußerung spielen neurologische und physiologische Prozesse mit herein: Daneben sind auch psychologische Mechanismen und strategische Kompetenzen bei der Sprachproduktion beteiligt. Erst das Zusammenspiel all dieser Facetten führt zur sprachlichen Äußerung. Da dabei psychologische Mechanismen und strategische Kompetenzen zum Tragen kommen, können letztere auch die Testperformanz beeinflussen und müssen deswegen erkannt und wenn möglich ausgeschaltet oder zumindest kontrolliert werden, doch dazu Näheres unter Kapitel 2.3 bei Überlegungen zur Konstruktvalidität und unter Kapitel 2.6 *Testentwicklung*.

1.2.4 Sprachen und Kulturen – Begriff der Mehrsprachigkeit

Im Folgenden werden die größeren Zusammenhänge von Sprachen und Kulturen diskutiert, um die Wechselwirkungen zwischen ihnen, wie sie in realen Kommunikationssituationen auftreten, auf den Fremdsprachenunterricht übertragen zu können.[42] Zunächst wird der Kulturbegriff betrachtet. Hier soll nicht der Versuch gemacht werden, einen Überblick über vorhandene Definitionen von *Kultur* zu geben oder gar selbst eine solche zu entwickeln, da das Phänomen damit nicht fassbarer wird. Dennoch finden sich beispielsweise bei Thomas (1996) einige interessante Merkmale: Kultur sei ein universelles, für eine Gemeinschaft aber typisches Orientierungssystem, das die Wahrnehmung, das Denken und Handeln seiner Mitglieder beeinflusse, deren Zugehörigkeit definiere und die Handlungsfelder der Individuen strukturiere. Dieses Orientierungssystem stelle die Voraussetzung dar, die Umwelt eigenständig bewältigen zu können. Dabei werden zentrale Merkmale dieses Systems als Kulturstandards bezeichnet, die alle Arten des Denkens, Wahrnehmens, Wertens und Handelns umfassen, die innerhalb einer Gemeinschaft als typisch, normal, selbstverständlich und verbindlich betrachtet werden.

Diese Standards gelten natürlich nur innerhalb einer Gemeinschaft, doch wo sind die Grenzen zwischen zwei Kulturgemeinschaften zu ziehen? Das dieser Arbeit zugrunde liegende Verständnis von Kultur basiert darauf, dass Kultur so vielschichtig ist, dass man – je nach Bezugsrahmen – innerhalb einer Kultur sehr viele „Subkulturen“ finden kann, je nachdem, wie tief man „zoomt“. Man kann innerhalb des Bezugsrahmens „Welt“ zum Beispiel eine westlich geprägte Kultur ausmachen, eine eher östliche, eine afrikanische Kultur und so weiter. „Zoomt“ man auf den Rahmen „westliche Welt“, so wird man darin amerikanische, europäische oder australische Ausprägungen finden. Man kann nun auf Europa mit all seinen Kulturen „zoomen“, innerhalb Europas kann man beispielsweise auf Deutschland „zoomen“ und wird auch innerhalb der „deutschen Kultur“ auf verschiedenen Ausprägungen stoßen. Dieses Modell funktioniert selbst auf Dorfebene, bis hin zum Individuum. Jeder Mensch hat „seine“ persönliche Kultur, genau wie jede Gruppe „ihre“ Kultur besitzt. Für einen Außenstehenden erscheint die jeweilige Gruppe homogen, doch für die Gruppenmitglieder ergeben sich klar abgegrenzte Kulturen innerhalb der Gruppe. (Vgl. u. a. Saunders 1982: 5f). Dabei geht es immer nur um Tendenzen innerhalb der jeweiligen Gruppen, je nachdem, was man vergleichen will oder auf welchen Rahmen man sich bezieht.

Bei Kulturvergleichen gilt das Prinzip des *Kulturrelativismus*, der prinzipiellen Gleichwertigkeit der Kulturen, so wie Sprachen ebenfalls als gleichwertig angesehen werden (vgl. Hansen 2000b: 106-111). Jede Kultur sollte nur „aus sich selbst heraus“ erklärt werden und „nicht nach den Standards der eigenen Kultur“ beurteilt werden (ebd.: 107). Die gemeinsame Grundlage muss man wohl, mangels geeigneterem Maßstab, in den Menschenrechten sehen. Auch wenn diese westlich geprägt sind, so haben sich doch viele Staaten dazu bekannt.

[42] Alle für ein Vermittlungskonzept relevanten Ableitungen werden unter Kapitel 1.3.3 dieser Arbeit dargestellt.

Um bei Vergleichen herausgefundene Unterschiede und Gemeinsamkeiten in positiver Weise nutzen zu können, müsste man es erst schaffen, diese nicht nach eigenen Standards zu bewerten, sondern sie gewissermaßen „wertfrei“ darzustellen und sie als Lernanreiz zu bieten. Dazu bedarf es aber einer kulturellen und persönlichen Identität, die sich durch eben solche Unterschiede nicht „angegriffen“ fühlt. Wie kommt man aber zur eigenen Kultur und Identität? Der Erwerb von eigenkulturellen Meinungen, Erwartungen, Einstellungen, Verhaltensweisen, des Werte- und Normsystems erfolgt in der Gruppe. Man erwirbt also eine spezifische kulturabhängige Orientierung, wobei Weltwissen, Kultur und Sprache miteinander verwoben erworben werden, ohne dass man sich der einzelnen Bestandteile je bewusst werden müsste. Diese Erwerbsprozesse wirken auch bei der Identitätsbildung des Individuums mit. Das einmal erworbene Orientierungs- und Sprachsystem stellt die Ausgangsbasis für jedes weitere Fremdsprachenlernen dar, das immer auch eine „Fremdkulturerfahrung“ bedeutet, welche bis hin zur individuell empfundenen Identitätsbedrohung gehen kann.

In der Begegnung mit Sprechern anderer Sprachen, die somit auch immer Angehörige eines anderen Kulturkreises sind, muss der oben erläuterte Kommunikationsbegriff erweitert werden. Denn die Kommunikationspartner sind in solchen Situationen konfrontiert mit zwei Arten von Orientierungssystemen und damit zwei Arten von Situationsdefinitionen und Situationsdeutungen, wobei das jeweils eigenkulturelle System notgedrungenerweise als Referenzpunkt angesetzt wird. (Vgl. Thomas 1996). Das kann zu Fehleinschätzungen, falscher Antizipation, Unklarheiten im Verhalten und in den Reaktionsweisen führen, ganz abgesehen von den sprachlichen Besonderheiten, die zu ganz eigenen Verständnisproblemen führen können.

Solche Situationen angemessen einzuschätzen und dabei sprachlich wie nonverbal handlungsfähig zu bleiben ist eines der Ziele des Fremdsprachenunterrichts, weshalb unter Kapitel 1.3.3 ein Vermittlungskonzept vorgestellt wird, das neben die kommunikative Kompetenz die interkulturelle Kompetenz stellt. Diese sollte auch Gegenstand von Sprachtests werden, da bloßes Sprachhandeln im „kulturfreien Raum“ in der Realität nicht vorkommt und da den kulturellen Orientierungssystemen in der interkulturellen Begegnung entscheidende Bedeutung zukommt, da sie – wenn auch oft unbewusst – in jeder Kommunikationssituation wirken. Folgendes Modell von Bolten verdeutlicht die Zusammenhänge:

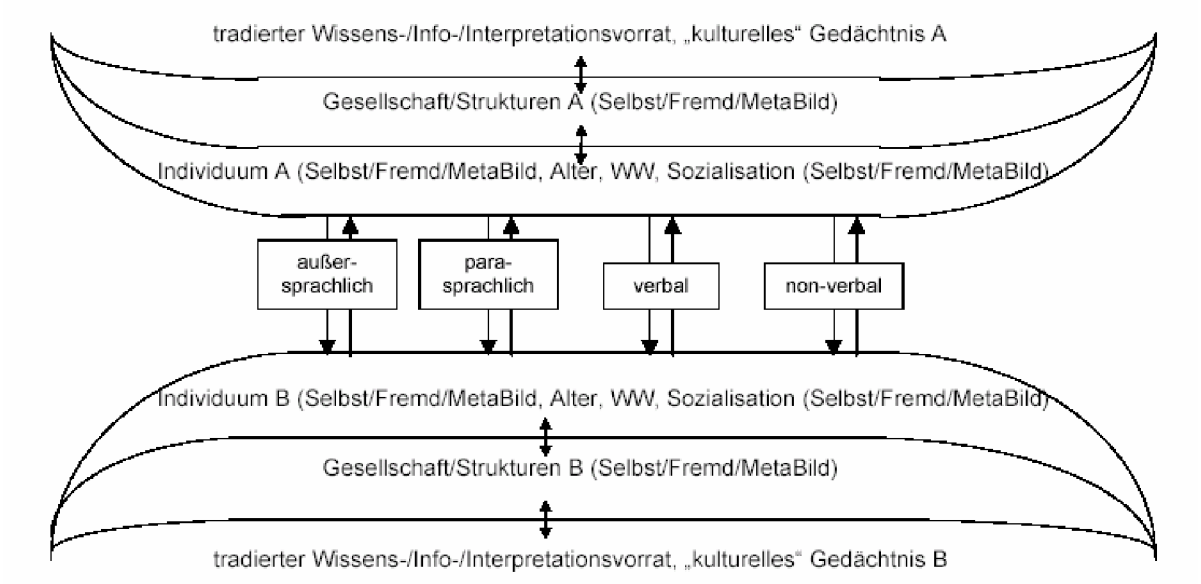


Abb. 4 (Bolten 1994: 29): Kommunikationsmodell

Bolten (1994) sieht den (interkulturellen) Kommunikationsprozess als „Spiel der Lebenswelten", in dem Individuen, Gesellschaften und Kulturen interdependieren. Bezogen auf reale Situationen wird es nur wenigen Lernern gelingen, neben der eigenen in anderen Lebenswelten so „heimisch" zu werden, dass sie als in beiden Welten kompetent gelten könnten. Dennoch ist es möglich, eine gewisse Handlungsfähigkeit, gewisse Schlüsselqualifikationen[43] zu entwickeln, um beispielsweise in einer spezifischen Situation – sei es im Job, in der interkulturellen Kommunikation oder einfach nur im Urlaub – angemessen zu handeln.

Ob nun die betreffenden Sprachsysteme und Kultursysteme mental getrennt oder doch gemeinsam repräsentiert sind, lässt sich nach dem heutigen Stand der Wissenschaft nicht eindeutig sagen. Besagte Studien aus der Gehirnforschung legen aber in Bezug auf Sprache und damit in gewissem Rahmen auch in Bezug auf die kulturellen Systeme, in denen Sprache angesiedelt ist, Folgendes nahe: "The neural representation of different languages is different in bilingual individuals" (Neville & Bavelier 1998: 256f).[44] Es scheint, dass bei jüngeren Lernern die beiden sprachlichen Systeme in überlappenden Regionen abgespeichert werden, während sie bei älteren Lernern in neuronalen Systemen organisiert sind, die sich nur teils oder gar nicht überlappen. Damit konsistent zeigen sich Untersuchungen von Hirnschädigungen: Dabei kann es passieren, dass eine Sprache verloren geht, während die andere erhalten bleibt. Es gibt Hinweise, dass bei automatischem, implizitem *Processing* (wie es in der Erstsprache geschieht) andere Regionen zuständig sind als für bewusstes, kontrolliertes *Processing*, wie es eher in einer gelernten Sprache geschieht. Der Verlust einer der beiden Sprachen bei Schädigung nur

[43] Zu den Schlüsselqualifikationen zählen u. a. Organisationsfähigkeit, Fähigkeit zur Kommunikation und Kooperation, Einsatz von Lern- und Arbeitstechniken, Sicherheit in den Kulturtechniken Lesen, Schreiben und Rechnen, Problemlösungs- und Entscheidungskompetenzen, Selbständigkeit und Verantwortung, Sprachkompetenzen in Fremdsprachen sowie Lernfähigkeit und Lernbereitschaft (vgl. beispielsweise http://212.6.107.180/www02/hk-bremen/baaktuell/2001-09-01/Seite2.pdf, Zugriff am 01.10.2003) – also Qualifikationen, die beitragen zur kompetenten Kommunikation und Handlungsfähigkeit im Alltag. Sie könnten als Richtzielvorgaben für Qualifikationen nicht nur in den Fremdsprachen dienen – sie sind ein fächerunabhängiges Ziel für Bildungsinstitutionen wie Schulen. Welche dieser Vorgaben im Fremdsprachenunterricht umgesetzt werden könnten und welche Bedeutung sie für das Testen von Sprachkenntnissen haben, wird unter Kapitel 1.3.3 resp. Kapitel 2 dieser Arbeit diskutiert.

[44] Hier wurden bilinguale Sprecher untersucht, doch die Hinweise bezüglich des Altersfaktors dürften sich verallgemeinern lassen.

einer spezifischen Region könnte darauf zurückzuführen sein, dass die beiden Sprachen nicht gemeinsam erworben wurden. Das Alter beim Erwerb bzw. Lernen einer zweiten Sprache scheint mit zu entscheiden, wo und wie sie abgespeichert wird. Es gibt aber noch weitere Faktoren, die neben dem Alter die neuronalen Repräsentationen verschiedener Sprachen beeinflussen, jedoch noch nicht hinreichend erforscht sind (ebd.: 256f). Verschiedene Sprachen (und Kulturen) scheinen also nicht zwangsläufig gemeinsam oder getrennt voneinander organisiert zu werden – nicht unbedingt als „ein vermischtes System“[45] und auch nicht unbedingt als zwei getrennte Systeme – dafür gibt es keine Hinweise aus der Hirnforschung, wohl aber Hinweise darauf, dass auch getrennt Gespeichertes vernetzt aktiviert werden kann, und dass die Art der Speicherung altersbedingt sein könnte.

Bezogen auf die Kompetenzen könnte dies bedeuten, dass alle beteiligten Wissensbestände gemeinsam aktiviert werden, relativ unabhängig davon, wie sie im Gehirn repräsentiert sind. Diese Kompetenzen in der interkulturellen Begegnung kann man auch über ein Konzept der Mehrsprachigkeit beschreiben, wie es beispielsweise in der Europäischen Union entwickelt worden ist als Voraussetzung eines gemeinsamen Europas, nach dem Motto „mehrsprachige Bürger in einem vielsprachigen Europa“. Die Mehrsprachigkeit in Europa wird sich notgedrungenerweise auf das Minimum der so genannten rezeptiven Dreisprachigkeit beschränken, denn aufgrund alltäglicher Zwänge dürfte es unmöglich sein, mehr als diese Dreisprachigkeit in größeren Bevölkerungsschichten zu vermitteln:

> Die europäische Union hat in ihrem Weißbuch *Lehren und Lernen. Auf dem Wege zur kognitiven Gesellschaft*, Brüssel, Luxemburg: Amt für amtliche Veröffentlichungen der europäischen Gemeinschaften 1995, besonders S. 72ff, europäische Identität im sprachlichen Bereich definiert als *Trilinguisme*, als Dreisprachigkeit also. Hintergrund der Forderung ist die Einsicht, dass nur so sichergestellt werden kann, dass jeder Unionsbürger zumindest ansatzweise in einer zweiten europäischen Regionalkultur (neben seiner eigenen) zu Hause ist. Da nun jeder Bürger Europas ein gewisses Maß an Fertigkeit in der *lingua franca* Englisch braucht, ergibt sich als Minimalforderung im Sinne des *Trilinguisme* die teilweise Kenntnis (der amtliche europäische Terminus ist “*partial competence*“) einer Nachbarsprache. (Schröder 1999, Anmerkung 1)

Faktisch geht es in erster Linie um die Handlungsfähigkeit in Muttersprache und *lingua franca*[46], und in zweiter Linie um die zumindest rezeptive Kompetenz in wenigstens einer Nachbarsprache. Damit soll Handlungsfähigkeit in einem geeinten Europa erzielt werden.

Wie kann die erwähnte rezeptive Dreisprachigkeit erreicht werden? Der Ort hierfür ist der schulische (und außerschulische) Fremdsprachenunterricht: Dort können mehrsprachige Kompetenzen gefördert werden, die Kenntnisse in der *lingua franca*, in Nachbarsprachen und in der Muttersprache umfassen, ebenso wie Bewusstheit über Sprach- und Kommunikationssystemesysteme und Strategien des Sprachlernens:

[45] wie es beispielsweise der GER auf S.17 oder S.163 darstellt
[46] Zur Diskussion des Für und Wider des Englischen als *lingua franca* in der EU möchte ich mich hier nicht äußern. Aufgrund der realen Tatbestände – beispielsweise ist Englisch häufig die erste Fremdsprache in der Schule – ist es jedoch nicht von der Hand zu weisen, dass Englisch die Stellung einer *lingua franca* bereits de facto besitzt.

> Die Hauptaufgabe der Schule sollte darin bestehen, junge Menschen mit der Vielfalt von Sprachen und mit unterschiedlichen Kommunikationskonventionen dieser Welt vertraut zu machen, ihnen exemplarische Sprachkenntnisse einer kleinen Anzahl ganz unterschiedlicher Sprachen zu vermitteln und somit eine Sprachlernbewusstheit zu trainieren. Solide Grundkenntnisse des Englischen als lingua franca gehören ebenso zu diesem fremdsprachlichen Programm wie Einsichten in die deutsche Sprache. (Edmondson 2003: 69f)

1.2.5 Sprachbegriff(e) im GER

Bevor sich die vorliegende Arbeit den Besonderheiten des Fremdsprachenunterrichts zuwendet und untersucht, wie die oben erläuterten Sprachbegriffe integriert werden können in ein kohärentes Konzept der Fremdsprachenvermittlung, wird analysiert, welcher Sprachbegriff dem GER zugrunde liegt. Denn wenn dort ein Rahmen zum Sprachlernen, -lehren und -beurteilen gesteckt werden soll, so muss dieser auch Aussagen zum zugrunde liegenden Sprachbegriff treffen.

Der GER hält sich bedeckt, wenn es darum geht, eindeutige Positionen zu beziehen, gerade bei Fragestellungen, zu denen es keinen wissenschaftlichen Konsens gibt. Es werden die „Prinzipien einer pluralistischen Demokratie“ bemüht, um den GER „offen, dynamisch und undogmatisch“ zu halten. (GER 2001: 29). Der GER will lediglich einen Rahmen vorgeben, der im Einzelfall von den Beteiligten adäquat gefüllt werden muss. Sprache wird als Gesamtheit betrachtet, die aufgrund ihrer Komplexität notwendigerweise in Einzelkomponenten linguistischer und außersprachlicher Art dargestellt wird. Doch betrachtet der GER diese Einzelkomponenten als interagierend mit der „Entwicklung jedes einzelnen Menschen“, so dass die Bedeutung und Tragweite der Einzelkomponenten jeweils im Einzelfall von den Beteiligten definiert werden muss (GER 2001: 14). Das Klassifizierungsschema für diese Komponenten ist nicht immer transparent, gibt es doch kommunikative Aktivitäten, die eher nach funktionalen Gesichtspunkten kategorisiert sind (GER 2001: Abschnitt 4.4, 62ff), und sprachliche Komponenten, die eher traditionell nach Fertigkeiten und sprachlichen Teilbereichen wie Lexik, Grammatik etc. angeordnet sind. Der GER überlässt es den Benutzern, ob sie eher von den Formen ausgehend zur Bedeutung gelangen, oder umgekehrt nach dem funktional-notionalen Ansatz die Bedeutung als Ausgangspunkt wählen möchten.[47]

H. Christ betrachtet den Sprachbegriff im GER als „Personalisierung der Mehrsprachigkeit“, denn im Mittelpunkt stehen die Sprachhandelnden und deren sprachliche, kommunikative und soziale Kompetenzen, nicht jedoch das theoretisch-abstrakte System von Sprache: „Die Autoren gehen nicht von einem systemischen Sprachenbegriff aus, wenn sie auch linguistische Ordnungskriterien verwenden. Sprache stellt sich ihnen im Gegenteil als personaler und funktionaler Vollzug dar.“ (H. Christ 2003: 58: Verweis auf GER 2001: 14 und 17).

[47] Vgl. GER (2001: 116): Der funktional-notionale Ansatz wurde bei den Publikationen der Lernzielkataloge des Europarats, *Waystage 1990*, *Threshhold Level 1990* und *Vantage Level* gewählt.

Der Sprachbegriff im GER kann zusammenfassend beschrieben werden als handlungsorientiert (GER 2001: 21) und interaktiv, denn Sprache wird kontextualisiert und verwendungsbezogen betrachtet. Sprache wird als variabel in Bezug auf die Kontexte angesehen, in denen sie auftritt. Sie stellt „kein neutrales Instrument des Denkens dar, wie etwa die Mathematik“ (GER 2001: 52).

Eine auffallende Schwäche bei der Konzeptionalisierung von Sprache im GER ist jedoch, dass der GER nicht zwischen den Perspektiven der Sprachverwendung innerhalb einer Sprachgemeinschaft und über Sprachgemeinschaften hinweg unterscheidet. Ebenfalls nicht unterschieden wird zwischen Sprache als individuellem und Sprache als kollektivem, sozialem Gut. Auch zwischen der Spracherwerbs- und Sprachverwendungsperspektive wird nicht unterschieden:

> Sprachverwendung – und dies schließt auch das Lernen einer Sprache mit ein – umfasst die Handlungen von Menschen, die als Individuen und als gesellschaftlich Handelnde eine Vielzahl von ***Kompetenzen*** entwickeln, und zwar ***allgemeine***, besonders aber ***kommunikative Sprachkompetenzen***. Sie greifen in verschiedenen ***Kontexten*** und unter verschiedenen ***Bedingungen und Beschränkungen*** auf diese Kompetenzen zurück, wenn sie ***sprachliche Aktivitäten*** ausführen, an denen (wiederum) ***Sprachprozesse*** beteiligt sind, um ***Texte*** über bestimmte ***Themen*** aus verschiedenen ***Lebensbereichen*** (Domänen) zu produzieren und/oder zu rezipieren. Dabei setzen sie ***Strategien*** ein, die für die Ausübung dieser ***Aufgaben*** am geeignetsten erscheinen. Die Erfahrungen, die Teilnehmer in solchen kommunikativen Aktivitäten machen, können zur Verstärkung oder Veränderung der Kompetenzen führen. (GER 2001: 21)

Ebenso wenig unterscheidet der GER zwischen Sprachverwendung in der realen Welt und Sprachverwendung im Fremdsprachenunterricht[48] – ein Manko bei einem Instrument, das sich ja gerade mit Sprachenlernen und -lehren auseinandersetzt. Diese Vermischung der Benutzerperspektiven zieht sich durch weite Teile des GER, was nicht zur postulierten Transparenz beiträgt.

Im Folgenden werden die Aussagen des GER im Licht der oben in den Kapiteln 1.2.1 mit 1.2.4 entwickelten Kategorien betrachtet, um den Sprachbegriff des GER herauszuarbeiten und aufzuzeigen, wo es Verbesserungsbedarf gibt.

1.2.5.1 Innersprachliche Organisation

Zu dieser Thematik gibt es wenig eindeutige Aussagen im GER. Auf S.116 wird konstatiert, dass Sprache aus Sicht der theoretischen und deskriptiven Linguistik „ein hochkomplexes symbolisches System“ darstelle. Dieses System umfasse das Formeninventar der Sprache so wie das Bedeutungssystem. Zum Verhältnis von Formen und Bedeutung folgt die Feststellung: „Die Kategorien beider Organisationsformen stehen meist in zufälliger Beziehung zueinander.“ Dies stellt m. E. eine nicht unproblematische Tatsachenbehauptung dar, denn es folgen weder Begründung noch nähere Ausführung dazu. Überdies ist diese Behauptung so nicht haltbar, denn wie oben in Kapitel 1.2.3 dieser Arbeit ausgeführt, hängen Form, Bedeutung und Kontext aufs Engste zusammen.

[48] Es gibt beispielsweise keine Thematisierung des *classroom discourse*, vgl. die Darstellungen unten unter Kapitel 1.3.4.4 *Fremdsprache im Unterricht im GER*.

Ein Beschreibungsmodell für die innersprachliche Organisation ist nicht auszumachen. Es ist natürlich legitim, in einem Rahmen, der sich vorwiegend mit Sprachverwendung beschäftigt, auf einen systemorientierten Sprachbegriff zu verzichten, doch können sich daraus Folgeprobleme ergeben. Beispielsweise wird lediglich implizit deutlich, dass der GER traditionellen Sichtweisen verhaftet bleibt: So wird etwa der *Satz* als oberste linguistische Einheit angesetzt (GER 2001: 149), ohne dass damit den Erkenntnissen aus Diskursforschung oder Textlinguistik Rechnung getragen würde. Diese Aussage deutet nicht auf einen systemischen Sprachbegriff hin, denn dort würde die Einheit *Satz* als der Einheit *Text* untergeordnet, jedoch innerhalb des Netzwerks Sprache als miteinander über die Ebenen hinweg interagierend betrachtet werden. Doch textgrammatische Phänomene werden – bis auf die klassische Kohärenz-Kohäsionsperspektive – nicht thematisiert, was in einem handlungsorientierten Ansatz schwer verständlich ist. Die Autoren des GER hätten gut daran getan, zumindest ansatzweise offen zu legen, wie die innersprachliche Organisation betrachtet wird oder zumindest, welche Möglichkeiten hierbei relevant wären, um tatsächlich einen Rahmen zu stecken, innerhalb dessen sich die Benutzer verorten können.

1.2.5.2 Internes Wissenssystem – mentale Repräsentation

Auch hierzu lassen sich nur wenige Aussagen im GER finden. Beispielsweise wird auf S. 17 Folgendes behauptet:

> Diese Sprachen und Kulturen (bezogen auf die Spracherfahrungen eines Menschen, Anm. d. V.) werden aber nicht in strikt voneinander getrennten mentalen Bereichen gespeichert, sondern bilden vielmehr gemeinsam eine kommunikative Kompetenz, zu der alle Sprachkenntnisse und Spracherfahrungen beitragen und in der die Sprachen miteinander in Beziehung stehen und interagieren.

Auf S. 163 dann wird der Gedanke der „gemischten Kompetenz“ wieder aufgegriffen, dem auf derselben Seite allerdings widersprochen wird durch die Aussage, dass „ein Mensch (...) über eine einzige mehrsprachliche und plurikulturelle Kompetenz, die das ganze Spektrum der Sprachen umfasst, die dem Mensch zur Verfügung stehen“ verfügt und „dass die Entwicklung der Fähigkeit, mit anderen Kulturen in Verbindung zu treten, und die Entwicklung der sprachlichen Kommunikationsfähigkeit“ nicht „miteinander verbunden sein müssen“. Nun geht es zwar in letzterer Aussage um die Entwicklung, also um Lernen bzw. Erwerb dieser Kompetenzen, doch ist dabei zu unterstellen, dass so erworbene Komponenten auch mental getrennt repräsentiert werden, wenn sie im Erwerb schon nicht miteinander verbunden sein müssen – und dies widerspricht der Annahme, dass alles Sprach- und Kulturwissen in einer einzigen gemischten Kompetenz aufgehen würde. Wie die Benutzer des GER mit diesen widersprüchlichen Behauptungen umgehen sollen, bleibt ungeklärt.

Im GER-Abschnitt 6.1.3, S.132ff finden sich Aussagen zur Struktur der mehrsprachlichen und plurikulturellen Kompetenzen, die darauf schließen lassen, dass die Autoren des GER diese

Kompetenzen als „ungleichmäßig, sich verändernd“ betrachten, dass allgemeine wie sprachliche Fertigkeiten und Kenntnisse (von solchen im kulturellen Bereich ist gar nicht erst die Rede) in unterschiedlichem Maße zum Einsatz kommen, und dass sich die mehrsprachigen und plurikulturellen Kompetenzen in ihren unterschiedlichen Ausprägungen und Entwicklungsgraden als „partielle Kompetenzen“ begreifen lassen, die „als Teile einer mehrsprachlichen Kompetenz anzusehen [sind] und diese bereichern“ (GER 2001: 134).

Sind diese Aussagen im GER wissenschaftlich fundiert? Wie unter Kapitel 1.2.2 dieser Arbeit ausgeführt, lassen neurobiologische Erkenntnisse darauf schließen, dass verschiedene Komponenten von Sprach- und Kulturerfahrungen in getrennten Bereichen und teils mehrfach repräsentiert werden können, dass sie jedoch gemeinsam aktiviert werden können. House (2003: 98f) erwähnt „empirisch dokumentierte Hypothesen, nach denen eine getrennte Repräsentation und die Möglichkeit der Existenz eines ‚übergeordneten’ konzeptuellen Speichersystems plausibel ist.“[49] Wenn es auch nicht Aufgabe eines Referenzrahmens ist, detaillierte Informationen über die neurobiologischen „Speichermöglichkeiten“ von Sprache zu geben, so würde man doch zumindest einen Verweis auf den derzeitigen Stand der Forschung erwarten, zumal die mentalen Repräsentationsmöglichkeiten eine elementare Grundlage des Sprachen- und Kulturenlernens darstellen.

1.2.5.3 Kommunikative Kompetenz

Zur kommunikativen Kompetenz finden sich im GER schon genauere Aussagen. Wie oben erwähnt gehen in die kommunikative Kompetenz im GER alle Sprach- und Kulturerfahrungen ein, ohne dass explizit eine Unterscheidung hinsichtlich der verschiedenen Sprach- und Kultursysteme getroffen würde (vgl. GER 2001: 17). Dies ist nicht unproblematisch, da nicht gesichert ist, dass tatsächlich alle Erfahrungen in diesem Bereich zusammen abgespeichert werden bzw. auch gemeinsam aktiviert werden. Denn es ist denkbar, dass die gegebenen Transfermöglichkeiten zwischen den verschiedenen Sprach- und Kulturerfahrungen von verschiedenen Sprechern in unterschiedlichem Maß genutzt werden. Wie oben erläutert, kommt es auch altersabhängig zu verschiedenen mentalen Repräsentationen mehrerer Sprachen, je nachdem, wann die zweite Sprache erworben wurde.

Im Sinne eines holistischen Konzepts versteht Barkowski den Kompetenzbegriff des GER „als ein System“ von Teilkompetenzen sprachlicher und kultureller Natur, die sich in einer „sprachlichen Gesamthandlungskompetenz“ integrieren, und „somit qualitativ etwas anderes und mehr als die Summe der aufaddierten Teilkompetenzen“ darstellen (Barkowski 2003: 23). Dem ist, bis auf die gerade erwähnte Problematik der Zusammenhänge der verschiedenen Sprach- und Kultursysteme, nicht zu widersprechen.

[49] House verweist auf Baker 2001.

Das dem GER zugrunde liegende Modell der *language ability* basiert u. a. auf den Modellen von Canale und Swain (1981) und Bachmann (1990 respektive [2]1991a), ohne dass dies im GER transparent dargestellt würde.[50] Zu den Komponenten der kommunikativen Kompetenz findet sich im GER auf S. 22ff ein umfassender Überblick, der sich oft, jedoch nicht immer, mit Bachmanns Modell der kommunikativen Kompetenz deckt. Der GER unterscheidet allgemeine Kompetenzen und kommunikative Sprachkompetenzen, die bei Sprachaktivitäten zum Einsatz kommen, welche in GER-Abschnitt 4 in ihren Charakteristika beschrieben werden:

Erstere allgemeine Kompetenzen umfassen im GER vier Wissensarten: deklaratives Wissen (*savoir*), Fertigkeiten und prozedurales Wissen (*savoir-faire*), persönlichkeitsbezogene Fertigkeiten (*savoir-être*) und Lernfähigkeit (*savoir-apprendre*). Diese Wissensarten werden dann in GER-Abschnitt 5.1 ausführlich thematisiert, doch bedauerlicherweise gibt es für sie im GER keine Deskriptoren, so dass sie nicht über Skalen fassbar werden. Inwieweit solche Deskriptoren überhaupt semantisierbar sind, ist eine andere Fragestellung, der an dieser Stelle nicht nachgegangen werden kann. Zumindest finden die Benutzer des GER genügend Information, um für den jeweiligen Einzelfall zu entscheiden, in welchen Umfang die dort beschriebenen Teilkompetenzen relevant sind.

Die kommunikativen Sprachkompetenzen werden im GER ebenfalls ausführlich dargestellt und umfassen die sprachlichen Kompetenzen sowie soziolinguistische und pragmatische Kompetenzen (vgl. GER 2001: 24f). Letztere Unterscheidung wird nicht begründet und erscheint in dieser Art nicht unbedingt zwingend. Bachmann etwa betrachtet soziolinguistische Kompetenzen als einen Teilbereich der pragmatischen Kompetenzen. „Gesellschaftliche Konventionen“ wie Höflichkeitsnormen werden generell der Pragmatik zugeordnet – der GER führt sie ohne nähere Erläuterung als Teil der soziolinguistischen Kompetenz an.

In GER-Abschnitt 5.2.1 (ebd: 110-118) finden sich Skalen zu allen sprachlichen Teilbereichen in klassischer Aufteilung. Ob es auch möglich gewesen wäre, Skalen nach funktionalen Gesichtspunkten zu entwickeln, sei dahingestellt. Zwar können die am Sprachenlernen Beteiligten an die klassische Aufteilung anknüpfen, dennoch könnte dieser traditionelle Ansatz erweitert werden um eine funktionale Beschreibung der sprachlichen Kompetenzen.

Die soziolinguistischen Kompetenzen werden in GER-Abschnitt 5.2.2 (ebd.: 118-122) unterteilt in „sprachliche Kennzeichnung sozialer Beziehungen“, „Höflichkeitskonventionen“ (diese fallen aber unter Pragmatik), „Redewendungen“, „Registerunterschiede“ und „Varietäten“. Dann

[50] Im *User's Guide for Examiners* (Council of Europe 1996b) findet sich auf S.3 der explizite Hinweis auf das dem GER zugrunde liegende Modell, das angelehnt ist an die oben erwähnten Vorläufer: “The current Framework also contains a model of language ability. Its essence may be presented as a statement about the nature of communicative competence: **communicative competence** (sociolinguistic, linguistic, pragmatic) is a form of **general competence** that leads to **language activity** (interaction, production, reception, mediation) using **tasks**, **texts** and **strategies** in four principal **domains** (public, occupational, educational, personal) in which arise **situations**, consisting of **locations**, containing **organisations** that structure interaction, **persons** with definite roles, **objects** (animate and inanimate) that constitute an environment, **events** that take place in it, and **operations** that are performed (see Chapter 4 of the Framework document).”
In North 2000, North & Schneider 1998 und Schneider & North 2000 finden sich ebenfalls detaillierte Bezüge auf die dem GER-Konzept zugrunde gelegten Modelle. Warum im GER selbst keinerlei Aussagen zur Basis des Kompetenzmodells gemacht werden, ist nicht nachzuvollziehen.

folgt eine Skala „soziolinguistische Angemessenheit“, die, wie die Autoren auf S.121 und S. 212 des GER ausführen, in ihrer Entwicklung problematisch war. Folgerichtig werden auch nur die Deskriptoren angeführt, die sich als skalierbar erwiesen haben. Inwieweit diese Skala verwendbar ist, muss die Praxis zeigen.

Anschließend werden die pragmatischen Kompetenzen beschrieben: Diese umfassen die Diskurskompetenz, die sich auf das Zusammenfügen von Sätzen zu einem Text bezieht, und die Teilkompetenzen „Flexibilität“, „Sprecherwechsel“, „Themenentwicklung“ und „Kohärenz und Kohäsion“ (vgl. GER 2001: 123-130). Dabei werden jedoch verschiedene Bereiche vermischt, denn „Themenentwicklung“ und „Kohärenz und Kohäsion“ würden etwa bei Bachmann nicht unter pragmatische, sondern unter textuelle Kompetenz als ein Unteraspekt der sprachlichen Kompetenzen fallen. Auch wenn es nicht „das eine“ Klassifizierungsschema gibt, so wäre es doch sinnvoller gewesen, sich an ein Modell der kommunikativen Kompetenz zu halten, das weitgehend anerkannt ist, wie beispielsweise das von Bachmann.

In GER-Abschnitt 4.4 (ebd.: 62-92) werden zusätzlich „kommunikative Aktivitäten und Strategien“ beschrieben: Ausgehend von einer funktionalen Klassifizierung werden verschiedenste, auch sprachmittelnde, Aktivitäten (etwa „Vor Publikum sprechen“ oder „Ankündigungen, Durchsagen und Anweisungen verstehen“) und Strategien (beispielsweise „Kooperieren“ oder „Kompensieren“) beschrieben und – soweit möglich – in Skalen dargestellt.

Die Konzeption der kommunikativen Kompetenz im GER wird also im Detail erläutert, wenn auch nicht immer hinreichend begründet. Die Teilkomponenten scheinen manchmal unmotiviert klassifiziert zu sein, was die Benutzung nicht vereinfacht. Das Offenlegen von Beweggründen für die gewählten Klassifizierungen könnte zur Transparenz des Instruments beitragen.

Der dem GER zugrunde liegende Kommunikationsbegriff ist ein idealistisch geprägter, der aus den „Kriterien für ein direktes und effizientes Kommunizieren“ (GER 2001: 123) hervorgeht – es finden sich folgerichtig auch keine Thematisierungen von Kommunikationsabbrüchen oder Missverständnissen, obwohl gerade diese den Kommunikationsprozess bei Beteiligung einer Fremdsprache kennzeichnen. Es werden also idealisierte Kompetenzen beschrieben, ohne dass diesen die Realität gegenübergestellt werden würde.

1.2.5.4 Mehrsprachigkeit

Bei der Konzeption des Mehrsprachigkeitsbegriffs im GER zeigt sich noch deutlicher, dass der GER nicht den Ist-, sondern den Idealzustand eines vielsprachigen und multikulturellen Europas beschreibt: Der Status der einzelnen (europäischen) Sprachen wird nicht betrachtet, ebenso wenig wie der Status des Englischen als *lingua franca*, welcher ein europäisches Faktum darstellt, ob dieser Zustand nun von den Europäern diskutiert wird oder nicht. Völlig ignoriert wird

auch die Stellung der (außereuropäischen) Migrationssprachen innerhalb der einzelnen Mitgliedsstaaten.

Das Konzept der Mehrsprachigkeit in Europa (vgl. oben) ist in den europäischen Gesellschaften meist nicht umgesetzt und wird größtenteils auch nicht als Normalfall betrachtet – die einzelnen Mitgliedsländer verharren zumeist noch in nationalstaatlichem Denken und in der Einsprachigkeit ihrer Bürger und Institutionen als Norm. Diese „monolinguale Grundüberzeugung" (Gogolin 2003: 86 und 1994) zeigt sich auch in der schulischen Realität: Welche Sprachen sind „lehrenswert"? Noch immer werden Nachbarsprachen, Minoritätensprachen und Migrationssprachen benachteiligt, da in der Regel die traditionellen modernen Fremdsprachen Englisch und Französisch sowie die alten Sprachen angeboten werden. Auch Kompetenzen im Umgang mit anderen Sprachen und Kulturen (wie etwa das Überleben in der Zweitkultur) werden meist an den Schulen nicht als solche anerkannt.

Nun ist es nicht Aufgabe eines europäischen Instruments zur Förderung der Mehrsprachigkeit und Plurikulturalität, den Ist-Zustand (politisch) zu diskutieren, doch wenn Ungleichheiten, Schwierigkeiten und Realitäten nicht wahrgenommen werden, wie kann dann ein Referenzrahmen helfen das Ziel der Mehrsprachigkeit zu erreichen? Denn solch ein Instrument könnte ja Wege aufzeigen, wie man vom momentanen Stand zum Idealzustand gelangen könnte. Doch da weder die europäische Union noch die Mitgliedsstaaten integrative Konzepte zur konkreten Umsetzung der europäischen Mehrsprachigkeit entwickelt haben, wie sollte der GER dies leisten können?

Es ist darüber hinaus zu bedenken, dass ein Fördern der Mehrsprachigkeit alleine sicher nicht alle Verständigungsprobleme beheben kann – man müsste sich schon um eine (inter-) nationale Außen-, Friedens- und Kulturpolitik bemühen, um zu einem geeinten und friedlichen Europa zu kommen, in dem aus globaler Perspektive auch Migranten ihren Platz finden können. Doch die Umsetzung bzw. nähere Ausführung einer solchen Politik würde sowohl die Grenzen des (Fremd-)Sprachenunterrichts als auch die der vorliegenden Arbeit bei weitem sprengen, so dass auf eine ausführlichere Darstellung an dieser Stelle verzichtet werden muss.

Wenden wir uns nun dem Begriff der Mehrsprachigkeit im GER zu. Zunächst wird unterschieden zwischen *Vielsprachigkeit* und *Mehrsprachigkeit*: Erstere wird im GER definiert als „Kenntnis einer Anzahl von Sprachen oder (...) Koexistenz verschiedener Sprachen in einer bestimmten Gesellschaft" (GER 2001: 17). Bedenklich ist hierbei die Vermengung zweier völlig unterschiedlicher Perspektiven bei der Definition des Begriffs der Vielsprachigkeit: die Perspektive der Kenntnisse mehrerer Sprachen (also *Mehr*sprachigkeit als individuelles Gut) und die des gesellschaftlichen Zustandes der Koexistenz mehrerer Sprachen (also *Viel*sprachigkeit als soziales Phänomen). Des Weiteren besagt der GER auf S.17, dass die Vielsprachigkeit beispielsweise erreicht werden könne durch vielfältige Sprachlernangebote in einem Bildungssystem oder durch das Erlernen von mehr als einer Sprache. Fraglich dabei ist, ob der Zustand der

*Viel*sprachigkeit in Europa als solcher „erreicht" werden muss, oder ob es sich dabei nicht eher um das Erreichen der *Mehr*sprachigkeit der Bürger handelt. Ehe man sich auf solch impräzise Begrifflichkeiten einlässt, die mehr Fragen aufwerfen denn Klarheit bringen, wäre es sinnvoller, sich auf den in der europäischen Gemeinschaft erarbeiteten Begriff der Vielsprachigkeit, des *Multilingualismus* zu beziehen, der den Zustand des Mit- und Nebeneinader vieler Sprachen bezeichnet, nach dem Motto: „Plurilinguale Bürger in einem multilingualen Europa".

Der Begriff der Mehrsprachigkeit, des *Plurilingualismus*, wird im GER (2001: 17) wie folgt definiert:

> Mehrsprachigkeit (…) betont die Tatsache, dass sich die Spracherfahrung eines Menschen in seinen kulturellen Kontexten erweitert, von der Sprache im Elternhaus über die Sprache der ganzen Gesellschaft bis zu den Sprachen anderer Völker. Diese Sprachen und Kulturen werden aber nicht in strikt voneinander getrennten mentalen Bereichen gespeichert, sondern bilden vielmehr gemeinsam eine kommunikative Kompetenz, zu der alle Sprachkenntnisse und Spracherfahrungen beitragen und in der die Sprachen miteinander in Beziehung stehen und interagieren. In verschiedenen Situationen können Menschen flexibel auf verschiedene Teile dieser Kompetenz zurückgreifen, um eine effektive Kommunikation mit einem bestimmten Gesprächspartner zu erreichen.

Diese Aussage lässt darauf schließen, dass Mehrsprachigkeit im europäischen Sinn als individuelles Gut betrachtet wird. Offen bleibt allerdings die Frage, was man sich unter der „Sprache der ganzen Gesellschaft" vorzustellen hat. Ebenfalls offen bleibt die Frage, wieso Mehrsprachigkeit die Tatsache der Erweiterung der Spracherfahrung in kulturellen Kontexten betonen sollte – doch hierbei hilft ein Blick in das englischsprachige Originaldokument, herauszufinden, welche Tatsache denn ursprünglich betont werden soll (CEF: 4, Herv. d. V.):

> (…) the **plurilingual approach emphasises the fact that** as an individual person's experience of language in its cultural contexts expands, from the language of the home to that of society at large and then to the languages of other peoples (whether learnt at school or college, or by direct experience), **he or she does not keep these languages and cultures in strictly separated mental compartments, but rather builds up a communicative competence to which all knowledge and experience of language contributes and in which languages interrelate and interact**.

Dabei wird deutlich, dass es bei der Übersetzung zu einer Sinnentstellung kam: Im Original wird betont, dass die verschiedenen Sprachen eines mehrsprachigen Individuums *gemeinsam* zur kommunikativen Kompetenz beitragen; in der deutschen Ausgabe jedoch wird die tautologische Tatsache betont, dass sich Spracherfahrung in kulturellen Kontexten erweitert.

Wie wird nun diese „gemeinsame" mehrsprachige Kompetenz im GER verstanden und operationalisiert? In GER-Abschnitt 8 *Sprachenvielfalt und das Curriculum* findet man Belege für das Verständnis der Plurilingualität als „individuelles Gut" (GER 2001: 163):

> Der Begriff ‚mehrsprachige und plurikulturelle Kompetenz' bezeichnet die Fähigkeit, Sprachen zum Zweck der Kommunikation zu benutzen und sich an interkultureller Interaktion zu beteiligen, wobei ein Mensch als gesellschaftlich Handelnder verstanden wird, der über – graduell unterschiedliche – Kompetenzen in mehreren Sprachen und über Erfahrung mit mehreren Kulturen verfügt. Dies wird allerdings nicht als Schichtung oder als ein nebeneinander von getrennten Kompetenzen verstanden, sondern vielmehr als eine komplexe oder gar gemischte Kompetenz, auf die der Benutzer zurückgreifen kann.

Darüber hinaus wird Mehrsprachigkeit „als der Regelfall, Zweisprachigkeit hingegen als Sonderfall der Mehrsprachigkeit angesehen" und die Kommunikationskompetenzen in den verschiedenen

Sprachen als integriert in „eine einzige mehrsprachliche und plurikulturelle Kompetenz“ betrachtet (GER 2001: 163).

Zur Diskussion der mentalen Repräsentationsbestände im GER darf auf Kapitel 1.2.5.2 dieser Arbeit verwiesen werden; zur Diskussion der GER-Auffassung bezüglich der Frage, ob es eine einzige mehrsprachliche kommunikative Kompetenz oder sprach- und kulturabhängige Untersysteme und Kompetenzen gibt, auf Kapitel 1.2.5.3. Zur Frage, über welche Indikatoren diese Kompetenz fassbar werden könnte, findet sich im GER folgende durchaus sinnvolle Aussage (ebd.: 17, Herv. d. V.):

> Zum Beispiel können Gesprächspartner von einer Sprache oder einem Dialekt zu einer oder einem anderen **wechseln** und dadurch alle Möglichkeiten der jeweiligen Sprache oder Varietät ausschöpfen, indem sie sich z. B. in einer Sprache ausdrücken und den Partner in der anderen verstehen. Man kann auch auf die **Kenntnis mehrerer Sprachen zurückgreifen**, um den Sinn eines geschriebenen oder gesprochenen Textes zu verstehen, der in einer eigentlich unbekannten Sprache verfasst wurde; dabei erkennt man zum Beispiel Wörter aus einem Vorrat an **Internationalismen**, die hier nur in neuer Gestalt auftreten. Jemand mit – wenn vielleicht auch nur geringen – Sprachkenntnissen kann diese benutzen, um anderen, die über gar keine verfügen, bei der Kommunikation zu helfen, indem er zwischen den Gesprächspartnern ohne gemeinsame Sprache **sprachmittelnd** aktiv wird.

Wenn man sich nun die Umsetzung dieses Mehrsprachigkeitskonzeptes im GER, beispielsweise in den beschriebenen Teilkompetenzen und Skalen, betrachtet, stellt man schnell fest, dass Mehrsprachigkeit auf reine Zweisprachigkeit im Sinne von Muttersprache und einer weiteren Fremdsprache reduziert wird, was dem europäischen Mehrsprachigkeitskonzept auffällig widerspricht. Sucht man nach stringenter Umsetzung der „interagierenden Gesamtkompetenz“, sei es nun im Sinne des im obigen Zitat thematisierten Wechselns zwischen Sprachen[51] oder des Rückgreifens auf die (rezeptive) Kenntnis mehrerer Sprachen, so treffen die Skalen des GER hierauf gar nicht zu. Selbst die Thematisierung der *Sprachmittlung* in GER-Abschnitt 4.4.4 bleibt einer lediglich zwei Sprachen umfassenden Perspektive verhaftet.

Auch das an die Mehrsprachigkeit angebundene Konzept der *Plurikulturalität* klingt sehr ideell und dürfte noch „weit von der europäischen Realität entfernt“ sein, wie House (2003: 96) bemerkt. Dieses Konzept stellt der GER auf S.18 wie folgt vor:

> Mehrsprachigkeit muss im Kontext der Plurikulturalität gesehen werden. Sprache ist nicht nur ein besonders wichtiger Aspekt einer Kultur, sondern auch ein Mittel des Zugangs zu kulturellen Erscheinungsformen und Produkten. Vieles von dem, was oben gesagt wurde, betrifft in gleicher Weise auch den allgemeineren Bereich der Kultur. Die verschiedenen (nationalen, regionalen oder sozialen) Kulturen, zu denen ein Mensch Zugang gefunden hat, existieren in seiner kulturellen Kompetenz nicht einfach nebeneinander. Sie werden verglichen und kontrastiert, und sie interagieren beim Entstehen einer reicheren, integrierten plurikulturellen Kompetenz; mehrsprachige Kompetenz ist eine ihrer Komponenten, die wiederum mit anderen Komponenten interagiert.

Das GER-Konzept der Plurikulturalität impliziert, dass die kulturelle Vielfalt wahrgenommen wird und sich die Lernenden damit auseinander setzen, um letztlich Zugang zu den anderen Kulturen zu erhalten. Ob sich diese Annahme mit den realen Zuständen in den europäischen Gesellschaften deckt, sei dahingestellt – zumal es nicht Aufgabe eines bildungspolitischen Instruments

[51] House (2003: 95f) beispielsweise interpretiert die besagte ‚Kompetenz-Gesamtheit’ dahingehend, dass sie darin implizit die Fähigkeit zum *code switching* zwischen den verschiedenen Sprachen innerhalb der Gesamtkompetenz vermutet.

sein kann, diese Zustände zu analysieren. Es handelt sich dabei um die Perspektive der Plurikulturalität als „individuelles Gut“ analog dem oben ausgeführten Verständnis des Plurilingualismus, denn das Konzept bezieht sich auf den „Besitz“ von Einzelpersonen und nicht auf den Zustand in einer Gesellschaft. Letzterer könnte durch den Begriff der Multikulturalität charakterisiert werden, doch dieser Begriff wird im GER erst gar nicht erwähnt.

Es mutet in diesem Zusammenhang seltsam an, wenn der GER auf S.12 behauptet, dass GER-Abschnitt 8 „offene Fragen wie: Mehrsprachigkeit und Plurikulturalität“ behandle, wenn sich in diesem besagten Abschnitt lediglich eine einzige Aussage in Bezug auf Plurikulturalität[52] finden lässt – hier muss die postulierte Umfassendheit eingefordert werden. Ebenfalls seltsam mutet es an, wenn sich dafür in GER-Abschnitt 6 *Fremdsprachenlernen und -lehren* ein Unterabschnitt 6.1.3 *Mehrsprachige Kompetenz und Plurikulturelle Kompetenz* findet, in dem dann Merkmale dieser Kompetenzen thematisiert werden (vgl. GER 2001:132ff), die jedoch nicht nur auf diese Kompetenzen zutreffen und somit auch nicht zur Abgrenzung und Charakterisierung der mehrsprachigen und plurikulturellen Kompetenzen herangezogen werden können: Diese Kompetenzen werden im GER als „ungleichmäßig, sich verändernd“ (ebd.: 132) verstanden, ein Merkmal, das auf alle Kompetenzen zutrifft – man denke an das Konzept des *lebenslangen Lernens*; im Gegensatz zur „raschen Stabilisierung“ (ebd.: 133) der muttersprachlichen kommunikativen Kompetenz (es wird allerdings nirgends erläutert oder belegt, wieso diese als sich rasch stabilisierend betrachtet wird) seien diese Kompetenzen gekennzeichnet durch „ein kurzlebiges Profil und eine veränderliche Konfiguration“ (was angesichts der oben beschriebenen Eigenschaften der Lernersprache nicht weiter verwundert; ebd.); die Anwendung dieser Kompetenzen sei gekennzeichnet durch die Nutzung „sowohl [der] allgemeinen als auch [der] sprachlichen Fertigkeiten und Kenntnisse“ (ebd.), als ob dies nicht ein Kennzeichen jeder Kommunikation wäre; einsprachige Kompetenzen würden von diesen Kompetenzen „nicht einfach addiert, sondern [sie ließen] verschiedene Kombinationen und Veränderungen der verschiedensten Art“ zu (ebd.), wobei sich in diesem Zusammenhang die Frage stellt, ob mehrsprachige Kompetenzen rein konzeptionell betrachtet denn überhaupt andere Kompetenzen „addieren“ oder „Kombinationen zulassen“ können. Gemeint ist damit wohl, dass man in der Kommunikation auf alle Teilkompetenzen in unterschiedlicher Art zurückgreifen kann, was sich etwa an der Fähigkeit zum *code switching* zeigen kann; dieses Einsetzen aller Ressourcen ist aber wiederum nicht auf mehrsprachige oder plurikulturelle Kompetenzen beschränkt – es findet sich auch bei „zweisprachiger“ kommunikativer Kompetenz.

Das dem GER zugrunde gelegte Kommunikationskonzept geht von einer idealisierten Sprachverwendung aus, bei der vorausgesetzt wird, dass sich alle Beteiligten um Kooperation bemühen, auf Konsens orientiert sind und sich Wahrheitsprinzipien in der Kommunikation wie

[52] Vgl. GER (2001: 167): Hier wird lediglich konstatiert, dass auch andere als sprachliche Fächer den Zugang zur Plurikulturalität ermöglichen könnten.

etwa den Grice'schen Maximen[53] verpflichtet fühlen, welche explizit in GER-Abschnitt 5.2.3.1 *Diskurskompetenz* erläutert werden (ebd.: 123). Wohl auch deshalb sucht man vergeblich nach Thematisierung von sprachlichen Misserfolgen wie Missverständnissen und anderen *pitfalls of intercultural communication*, die etwa den Kernbereich interkulturellen Kommunikationstrainings ausmachen. Denn der Umgang mit plurikulturellen und mehrsprachigen Realitäten ist geprägt von sprachlichen wie außersprachlichen *critical incidents*, die unbedingt thematisiert werden müssten, sollen denn Kompetenzen in dieser Domäne beschrieben werden. Zwar werden im GER die „kulturspezifisch beeinflussten" persönlichkeitsbezogenen Kompetenzen als „heikle Felder interkultureller Wahrnehmung und Beziehungen" (ebd.: 23) herausgestellt, doch erfolgt keine Umsetzung in den entsprechenden Beschreibungen der Teilkompetenzen: Beispielsweise wird der immer wichtiger werdende Bereich der Reparaturtechniken in der interkulturellen Begegnung nicht thematisiert, ebenso wenig wie Kompetenzen im Umgang mit mehreren Kulturen.[54] Auch der gesamte Bereich der rezeptiven Kompetenzen in mindestens einer Nachbarsprache, wie sie im Weißbuch gefordert ist, wird nicht thematisiert.

Dieses Vorgehen ist bei der Zielsetzung des GER nicht nachvollziehbar, werden hier doch wesentliche Bestandteile mehrsprachiger und plurikultureller Kompetenz außen vor gelassen. Wie solch ein Instrument der Umsetzung europäischer sprachenpolitischer Ziele dienen soll, wenn es diese Ziele und Desiderate nicht thematisiert, bleibt offen.

Es dürfte tautologisch sein, darauf hinzuweisen, dass der GER nur dort eingesetzt werden sollte, wo er auch Gültigkeit hat. Man denke etwa an die Skalen zu Interaktionsstrategien (GER 2001: 87f) – diese sind kulturspezifisch und nicht etwa global zu verwenden. Doch da es sich beim GER um ein europäisches Instrument handelt, das vermutlich überwiegend auch dort eingesetzt wird, dürfte die Verankerung im europäischen Kulturkreis zu keinen größeren Verwendungsproblemen führen.

1.2.5.5 Fazit

Der Sprachbegriff im GER ist vom Ansatz her ein moderner und mehrdimensionaler, der pragmatisch, verwendungs- und handlungsorientiert ausgerichtet ist. Der Mensch als Handelnder steht im Mittelpunkt. Die daraus resultierende Diversität und Komplexität mehrsprachlicher kommunikativer Kompetenzen wird anerkannt in einem Konzept, das die beteiligten Teilkompetenzen und kommunikativen Aktivitäten umfasst. Der GER nimmt zumindest in seiner theoretischen Konzeption die europäische Sprachenpolitik als Ausgangspunkt und illustriert sie über Konzepte wie Mehrsprachigkeit und Erhaltung der sprachlichen und kulturellen Vielfalt. Dadurch kann der GER einen Beitrag leisten zur Veränderung der Wahrnehmung dessen, was den

[53] Vgl. Grice (1975: 41-58).
[54] Zu den linguistisch ausgerichteten Thematisierungen der Kommunikationsstrategien und dem Mitteln zwischen Gesprächsteilnehmern vgl. die Ausführungen dieser Arbeit in Kapitel 1.3.4.1 *Erwerb und Lernen im GER*.

„Normalzustand“ in einer Gesellschaft ausmacht: Europa ist geprägt durch vielsprachige und plurikulturelle Gesellschaften, innerhalb derer allzu oft Parallelgesellschaften existieren, die weit davon entfernt sind, sich gegenseitig wahrzunehmen, geschweige denn miteinander zu kommunizieren. Zunächst müsste der Ist-Zustand der europäischen Gesellschaften wahrgenommen und die Vielfalt als Normalzustand akzeptiert werden. Doch dazu müssen die oben diskutierten Konzepte Europas Bürgerinnen und Bürgern bekannt gemacht werden – dies könnte eine der Aufgaben des GER als sprachenpolitisches Instrument sein, zumindest unter den am Sprachenlehren und -lernen Beteiligten. Doch dazu müssten diese Konzepte zunächst stringent im GER thematisiert und der Ist-Zustand in Europa wahrgenommen werden.

Das mit dem oben skizzierten idealisierten Ansatz des GER einhergehende „Ignorieren“ der europäischen Realität macht sich bemerkbar an den fehlenden Thematisierungen etwa der Stellung und Problematik der Migrantensprachen, der rezeptiven Dreisprachigkeit im Sinne der Forderungen des Weißbuchs oder des Komplexes der interkulturellen Kompetenzen. Hier sollte der GER dem Ist-Zustand gerecht werden, wenn er denn hilfreich sein soll bei der Umsetzung der europäischen Sprachenpolitik.

Die Autoren des GER nehmen allzu oft eine übertrieben neutrale Haltung ein, wenn es um brisante Fragen geht. Wo immer es keinen wissenschaftlichen Konsens gibt, legen sich die Autoren des GER nicht fest. Auch wird das theoretische Konzept nicht stringent in den Beschreibungen der Teilfertigkeiten und in den entsprechenden Skalen umgesetzt. Dazu kommen missverständliche Definitionen und widersprüchliche Aussagen[55], Vermischung von Perspektiven und nicht begründete Klassifizierungsschemata. All dies führt letztlich zu einem impräzisen Sprachbegriff.

Empfehlenswert wäre, bei wissenschaftlich ungeklärten Fragen zumindest die Bandbreite der wissenschaftlichen Theorien anzudeuten oder wo immer möglich klare Stellung zu beziehen – auch wenn dabei der Grundsatz der „pluralistischen Demokratie“ aufgegeben werden müsste. Auch sollten die Begrifflichkeiten systematisiert werden und Klassifizierungen möglichst auf konsensfähigen Modellen basieren.

1.3 Lern- und Vermittlungskonzept

In diesem Kapitel soll Sprache nun im Kontext des fremdsprachliches Lernens und Lehrens näher betrachtet werden. Anschließend wird der Sprachbegriff im Fremdsprachenunterricht erörtert, ehe darauf aufbauend ein Vermittlungskonzept allgemeiner Art für den Fremdsprachenunterricht abgeleitet wird. Auf diesem Hintergrund wird in Kapitel 1.3.4 der vorliegenden Arbeit der

[55] Diese sind teils auf ungenaue oder sinnentstellende Übersetzungen zurückzuführen. Ein Vergleich der Übersetzungen des Instruments wäre ein wichtiger Forschungsbeitrag, kann jedoch nicht Gegenstand dieser Arbeit sein.

GER analysiert im Hinblick auf sein Lern- und Vermittlungskonzept. Es versteht sich von selbst, dass auf alle Fragen des Feststellens von Lernfortschritten und der Beurteilung des Sprachvermögens unter Kapitel 2 *Das Testen des Sprachvermögens* eingegangen wird.

1.3.1 Spracherwerb und internes Wissenssystem

In diesem Zusammenhang interessieren folgende Fragen: Wie findet der Erwerb bzw. das Lernen einer Fremdsprache statt? Welche allgemeinen Prozesse und Lernprinzipien sind dabei wirksam? Wie ist die Struktur des sich dabei herausbildenden internen individuellen Wissenssystems *Lernersprache*? Auch wenn im Folgenden meist von *Sprach*erwerb oder Fremd*sprachen*lernen die Rede ist, so bezieht sich der Terminus *Sprache* doch immer auch auf das *Kultur*system, in das die jeweilige Sprache eingebettet ist. Denn Sprachverwendung – und damit auch Spracherwerb und Sprachlernen – ist wie oben erläutert nicht ohne Kultur denkbar, die mit der neuen Sprache entdeckt und erfahren werden muss.

1.3.1.1 Erwerb und Lernen

Eine Möglichkeit, diese beiden Begriffe voneinander abzugrenzen, findet sich bei Krashen (1982), der unterscheidet zwischen (unbewusstem, ungesteuertem) Erwerb in natürlichen Kontexten und (gesteuertem, bewussten) Lernen in unterrichtlichen Kontexten. Diese scharfe Trennung ist jedoch so nicht haltbar, denn in ihrer „Reinform“ kommen diese beiden Prozesse nicht vor: Beim natürlichen Erstspracherwerb gibt es zwar unbewusste und ungesteuerte Prozesse, doch kommt es etwa über den so genannten *parent talk* zu gesteuerter Rückmeldung. Auf der anderen Seite gibt es auch im gesteuerten und formalisierten Fremdsprachenunterricht ungesteuerte und unbewusste Erwerbsprozesse. Statt der scharfen Trennung der beiden Prozesse wäre es hilfreicher, sich diese als Enden einer Skala mit fließenden Übergängen vorzustellen: Je nach Lerner und je nach Situation dürfte das Verhältnis „Lernen – Erwerb“ individuell geprägt sein.

Betrachtet man neurobiologische Untersuchungen hierzu, finden sich Hinweise darauf, dass bei automatischer, impliziter Sprachverarbeitung (wie sie in der Erstsprache stattfindet) andere Regionen zuständig sind als für bewusste, kontrollierte Sprachverarbeitung, wie sie eher in einer erlernten Fremdsprache stattfindet. Es scheint sowohl sprachspezifische als auch generelle Mechanismen beim Spracherwerb zu geben; unterschiedliche Hirnregionen spielen bei verschiedenen Aspekten des Spracherwerbs je nach Alter unterschiedliche Rollen; noch dazu scheint es Verlagerungen in den Funktionen der verschiedenen Gehirnareale zu geben über den zeitlichen Verlauf des Spracherwerbs hinweg. Es scheint, dass es im Verlauf des Spracherwerbs immer wieder Verlagerungen in der Konfiguration sprachrelevanter neuronaler Systeme gibt; einige dieser dynamischen Veränderungen hängen eher mit der Sprachkapazität zusammen,

andere wiederum mit dem Alter. Diese Veränderungen sind bei unterschiedlichen Aspekten von Sprache je verschieden und noch nicht abschließend erforscht. (Vgl. u. a. Neville & Bavelier 1998: 256).

Deshalb wird im Folgenden *Erwerb* im Sinn eines Überbegriffs für alle undifferenzierten Aneignungsprozesse genutzt, während sich *Lernen* im engeren Sinn auf bewusste Aneignungsprozesse bezieht.

Bei der Betrachtung Erstspracherwerb vs. Fremdsprachlernen fällt auf, dass Kinder mit ihrer Muttersprache auch ihr Weltwissen, ihr Kultursystem, ihr Weltbild und ihre Identität entwickeln. Diese Komponenten werden in ganzheitlichen, sinnstiftenden Kontexten erfahren, erworben und erlernt; Kinder nehmen Sprache in dem Maß auf, wie sie ihre Welt wahrnehmen und erfahren. All diese Komponenten bilden beim Erwerb jeder weiteren Sprache den Hintergrund, die Folie, auf der alle weiteren Lern- und Erwerbsprozesse ablaufen. Beim Fremdsprachenlernen ist das Verhältnis zwischen (Fremd)Sprache und deren Kultur ein anderes, da die Fremdsprache nicht eingebettet in ihre realen Kontexte erfahren werden kann. Zudem besitzen die Lernenden bereits ein (eigenkulturelles) Orientierungssystem, so dass es gerade im Fremdsprachenunterricht eine Herausforderung ist, die Fremdsprache in solch sinnstiftenden und ganzheitlichen Kontexten zu erfahren, in denen auch die Unterschiede und die Tragweite des fremden Kultursystems deutlich werden. Dazu bedarf es besonderer Verfahren, die unter Kapitel 1.3.3 *Vermittlungskonzept* beschrieben werden. An dieser Stelle sei vorwegnehmend auf den kommunikativen Ansatz in der Fremdsprachendidaktik verwiesen, der bewusst authentische Kommunikationssituationen (und sei es nur als Simulation oder Planspiel), handlungsbezogene Erfahrungen und selbstgesteuertes Lernen in den Unterricht integriert, um alle Kanäle im Aneignungsprozess zu nutzen.

Bezogen auf Erwerbserfahrungen im zweit- und mehrsprachlichen Kontext haben sich seit Mitte des letzten Jahrhunderts die Didaktiken beispielsweise des Deutschen als Zweitsprache den Besonderheiten dieses Sprachaneignungsprozesses zugewandt. Nicht umsonst verweist Königs (1995[3]: 431) auf die Notwendigkeit, im Zweitsprachunterricht „besondere, den natürlichen Erwerb mitumfassende Erkenntnisse heranzuziehen“ und schreibt dem „Fremdsprachenunterricht in mehrsprachigen Ländern eine andere Lernqualität“ zu (ebd.: 430). In allen Kontexten, in denen die Lernenden neben den gesteuerten Erfahrungen im Unterricht auch ungesteuerte Erfahrungen in realen, und damit „ungeschützten“ Kommunikationssituationen machen, sollten diese weitmöglichst in den Unterricht integriert werden, um die Vorkenntnisse der Lernenden zu nutzen und um ihnen die Möglichkeit zu bieten, ihre affektiven, handlungsbezogenen oder sprachbezogenen Erfahrungen und möglicherweise damit einhergehende Hypothesen zu reflektieren, in einen größeren Kontext zu stellen und sie gegebenenfalls zu revidieren.

1.3.1.2 Das interne Wissenssystem *Lernersprache*

Die heute gängigste Hypothese des Spracherwerbs ist die Interimsprachenhypothese, auch *Interlanguage*-Hypothese genannt.[56] Danach bilden Lerner sukzessive interimsprachliche Systeme, in sich kohärente, doch variable Zwischenstufen auf dem Weg zum zielsprachlichen System. Diese Interimsprachen oder Lernersprachen sind gekennzeichnet durch Merkmale der Muttersprache wie auch der Zielsprache; dadurch können sich – etwa durch Transfer muttersprachlicher Regeln oder Phänomene auf die Zielsprache – Fehler ergeben, die sich aber im Idealfall auf dem Weg der Annäherung an die Zielsprache verlieren. Diese Fehler stellen den eigentlichen Lernanreiz dar, d. h. sie können überwunden werden, wenn sie vom Lerner erkannt werden und wenn sie im lernersprachlichen System durch das korrekte zielsprachliche Phänomen „ersetzt" werden können. Insofern müssen Fehler im Fremdsprachenunterricht neu bewertet werden. Bei der Interimsprachenentwicklung kommen auch kommunikative Strategien[57] zum Einsatz, die immer dann eingesetzt werden, wenn es zu Kommunikationsbrüchen kommt, die in der Spracherwerbssituation meist durch mangelnde Sprachkenntnisse verursacht werden. Es gibt Strategien, die den Erwerb unterstützen, und solche, die ihn eher behindern; zu den ersteren zählen Strategien wie Paraphrase, Nachfragen, gelungener muttersprachlicher Transfer oder um Hilfe suchen; zu den letzteren zählen etwa Themenwechsel, negativer muttersprachlicher Transfer oder im schlimmsten Fall der Abbruch der Kommunikation. Solche Strategien müssen im Fremdsprachenunterricht thematisiert werden: Erstgenannte, um das Erlernen zu unterstützen; letztgenannte, um Kommunikationsabbrüche zu verhindern.

Es gibt noch weitere, für den Fremdspracherwerb relevante Hypothesen[58], deren Kernaussagen an dieser Stelle nur kurz angerissen werden sollen, um aufzuzeigen, welche Faktoren noch bedacht werden müssen zum erfolgreichen Erwerben und Vermitteln einer Fremdsprache: Es spielen die Form des Inputs (dessen *Bedeutung* eventuell erst *ausgehandelt* werden muss) wie auch der aktive Sprachgebrauch, der Output, eine Rolle; der *Interaktion* zwischen Lehrenden und Lernenden wird Bedeutung im Sprachlernprozess zugesprochen; die *teachability*, die Vermittelbarkeit hängt zusammen mit dem Stand der Interimsprache, auf den hin das Lehrmaterial abgestimmt werden muss.

Die Herausbildung der Interimsprache ist individuell verschieden, da dabei die Muttersprache, alle Vorerfahrungen und die Lernerpersönlichkeit eine Rolle spielen. In neurobiologischen Forschungen (vgl. u. a. Neville & Bavelier 1998: 256) zeigt sich, dass es – auch in Abhängigkeit vom Alter – bei der Abspeicherung von sprachlichen Systemen einen hohen Grad an Variabilität gibt. Die oben unter Kapitel 1.2.2 beschriebenen Charakteristika mentaler Repräsentationen von Sprache gelten auch für die Lernersprache. Interessant ist hierbei jedoch, in welchem Verhältnis

[56] Vgl. Kapitel 1.1 dieser Arbeit. Vgl. auch Selinker 1972 und 1992.
[57] Vgl. Kapitel 1.2.3 dieser Arbeit, Bachmann (1991a: 98ff), Faerch & Kasper 1983.
[58] Zu *Input-Hypothese* vgl. u. a. Krashen 1985; zu *Bedeutungsaushandlungshypothese* vgl. u. a. Long 1983; zu *Output-Hypothese* vgl. u. a. Swain 1985; zu *Interaktionshypothese* vgl. u. a. Henrici 1993; zu *Teachability-Hypothese* vgl. u. a. Pienemann 1989.

die sprachlichen Systeme zueinander stehen: Wird das neue System eingebaut in das bestehende, oder wird ein weiteres (isoliertes oder dependentes) System herausgebildet?[59]

Bei der Verarbeitung neuen Wissens werden aus psychologischer Sicht nach Piaget zwei Prozesse unterschieden: *Assimilation* und *Akkomodation* (vgl. u. a. Edmondson 1998: 35). Wenn neues Wissen in vorhandene Strukturen eingebaut wird, wird es assimiliert; werden hingegen Strukturen dem neuen Wissen angepasst oder neue Wissensstrukturen ausgebildet, so handelt es sich um den Prozess der Akkomodation. Solche neuen Netzwerke werden oft erst ausgebildet, wenn Fehler anzeigen, dass das neue Wissen nicht in vorhandene Strukturen assimiliert werden kann. In Bezug auf die Verfügbarkeit von Wissen unterscheiden Edmondson & House (1993: 229) drei Prozesse:

> Wissen wird „analysiert", d. h. mit vorhandenem Wissen über die Zielsprache koordiniert, „integriert", also mit verschiedenen nicht sprachlichen Schemata und anderen Wissensarten verbunden, und „automatisiert", d. h. blitzschnell und unreflektiert verfügbar gemacht.[60]

Die oben erwähnten Ergebnisse der neurobiologischen Forschung lassen auf eine große Varianz in der Organisation von Sprachsystemen schließen (vgl. Neville & Bavelier 1998: 256):

> Second languages learned late (i.e. after 7 years of age) are organized within neural systems that are partially or completely nonoverlapping with those for the native language. These systems for later-learned languages (...) display a high degree of variability between individuals.

Im Folgenden sollen deshalb aufgrund der hohen Variabilität zwischen Lernenden einige allgemein gültige Lernprinzipien zusammengestellt werden, die die Basis für ein kohärentes Vermittlungskonzept bilden.

1.3.1.3 Lernprozesse und Lernprinzipien

Generell gibt es keine festen Lernwege von der Muttersprache zur Fremdsprache: Das Erlernen einer Fremdsprache ist, wie alle Lernprozesse, gekennzeichnet durch ein hohes Maß an Individualität; die Persönlichkeitsfaktoren (etwa Alter, Vorerfahrungen, Motivation und Einstellung zum Lernen, Ego-Permeabilität; Intro-/Extrovertiertheit usw.) der Lernenden dürften hierbei eine wichtige Rolle spielen. Auch das Interesse an einem bestimmten Sachverhalt und die Bereitschaft, eine Sache zu ergründen, wirken sich positiv auf den Lernprozess aus. In diesem Zusammenhang kommt der Eigenverantwortung in Form von Mitbestimmungsmöglichkeiten der Lernenden in Bezug auf Lernziele und Lerngegenstände eine motivierende Rolle zu. (Vgl. u. a. Broadbent & Oriolo 1991: 308). Motivation und entdeckendes, handelndes, selbstgesteuertes Lernen gelten generell als lernförderlich. Stichworte wie autonomes Lernen oder *learning by doing* mögen an dieser Stelle genügen. Der individuelle Weltwissensbestand führt zu individueller Wahrnehmung des Fremdsprachenlernangebots und auch zu individueller Steuerung der Lernprozesse, die deshalb nicht vereinheitlicht werden können. (Vgl. u. a. Schröder 1999).

[59] Vgl. dazu auch die Ausführungen unter Kapitel 1.2.4 dieser Arbeit.
[60] Zitiert in Edmondson (1998: 35).

Jedes Lernen setzt die Wahrnehmung von Unterschieden zwischen dem momentanen und dem gewünschten Zustand voraus. Dazu muss zunächst der momentane Stand bestimmt werden: Wo stehe ich? Welche Vorerfahrungen bringe ich mit? Welcher Lernertyp bin ich? Welche Strategien setze ich vorwiegend ein? Inwieweit ist meine Identität geprägt durch mein eigenkulturelles Orientierungssystem? Erst dann kann festgestellt werden, wohin man will: Zu welchen Zwecken soll die Fremdsprache gelernt werden? Welche Bereiche sind besonders wichtig, welche treten eher in den Hintergrund? Wie kann oder will man dabei vorgehen? Diese und andere Grundsatzfragen führen dann zu erfolgreichem Lernen, wenn sie von den Lernenden (gegebenenfalls gemeinsam mit den Lehrenden) beantwortet werden. Orientierungshilfe im Lernprozess ist denn auch von entscheidender Bedeutung (vgl. u. a. Buttjes 1995: 146). Man kann nichts aufnehmen, für das man nicht „bereit" ist; um aber diese Bereitschaft (auch im Sinne der Aufnahmefähigkeit, vgl. die Input-Hypothese oben) zu erreichen, werden Orientierungshilfen durch die Lehrenden benötigt. Dabei sollten die Lernenden, ihre Erfahrungen und ihre Identitäten zum Ausgangs- und Zielpunkt eines jeden Lernprozesses werden. (Vgl. u. a. Hansen 2000a: 99-101).

Fremdsprachenlernen ist kein akkumulativer Prozess, bei dem ein „Baustein" nach dem anderen dadurch ins Sprach- (und Kultur-)System aufgenommen wird, dass er kognitiv „auswendig" gelernt wird. Vielmehr handelt es sich um einen zirkulären, nicht-linearen Prozess, bei dem alle Elemente wieder und wieder vorkommen müssen, in immer neuen Verwendungszusammenhängen, die ihrerseits verdeutlichen können, wann und wo welches Element verwendet werden kann (vgl. u. a. Bleyhl 1996). Nach Edmondson (1998: 35) geht „…das ‚Lernen' einzelner Elemente ... kontinuierlich weiter, und dies hauptsächlich dadurch, dass neurologische Netzwerke durch häufige Aktivierungen effizienter werden und gleichzeitig neue Verbindungen schaffen – das vorhandene System wird also ständig umstrukturiert bzw. rekonstruiert, besonders wenn neues Wissen hinzukommt." Welche Aussagekraft haben dann aber Lernzielkontrollen beispielsweise in Form von *discrete-point tests*, die festzustellen versuchen, ob eine bestimmte Form zu einem bestimmten Zeitpunkt korrekt beherrscht wird, wenn sich das System Lernersprache in fortlaufender Um- und Neustrukturierung befinden dürfte? Denn aufgrund des zirkulären Prozesses des Lernens kann es nach Ellis (1994: 343) dazu kommen, dass man zu einem frühen und einem späteren Zeitpunkt dieselben korrekten Formen produziert. Mittels des Abprüfens korrekter Formen kann demnach zwar eine Momentaussage bezüglich des kurzfristigen Erwerbs einer bestimmten Form getroffen werden, aber eine Aussage bezüglich des tatsächlichen Entwicklungsstands der Lernersprache kann mit solch einem Vorgehen nicht getroffen werden. Um an die sprachlichen Bestände zu gelangen, die „erlernt" und in das System Lenersprache eingebaut wurden, müsste schon Sprachproduktion in ganzheitlichen Situationen geprüft werden, in denen das Gelernte in authentischer Weise angewandt werden muss – erst dann zeigt sich, ob es abgerufen und verwendet werden kann.

Dabei ist das Verhältnis zwischen Verstehen und Verwenden gerade beim Spracherwerb nicht gleichwertig: Man wird rezeptiv immer mehr Sprache verarbeiten können als produktiv, sei

es nun im Muttersprachenerwerb oder beim Fremdsprachenlernen. Dieses Ungleichgewicht kann aber über die Darbietung des Lernstoffs in immer neuen Kontexten positiv genutzt werden, indem rezeptive Verarbeitung als Grundlage für Sprachproduktion dient (Näheres dazu unter Kapitel 1.3.3 *Ableitung eines Vermittlungskonzepts*).

Interessant ist aus kognitiver Sicht das Modell der hierarchischen *Skill*-Integration von Schaeffer (1975)[61], nach dem Kinder und Jugendliche aufgrund mangelnder Informationsverarbeitungskapazitäten nach und nach erst die benötigten Fertigkeiten integrativ benutzen können: Zunächst werden die benötigten *skills* je nach kognitiven Möglichkeiten der Lernenden in eine funktionstüchtige (momentane) Fertigkeit integriert, welche erst nach Automation dieser *skills* um weitere benötigte *skills* erweitert werden kann. Beispielsweise nimmt Bereiter (1980) an, dass bei der Entwicklung der fremdsprachlichen Schreibfertigkeit zu Beginn alle Kapazitäten mit Sprachverarbeitungsprozessen ausgelastet sind, so dass sich Lerner in diesem Stadium weniger auf inhaltliche oder strukturelle Aspekte konzentrieren können. Erst wenn etwa Syntaxregeln und ein Grundbestand an Wortschatz automatisiert verwendbar sind, können freiwerdende Kapazitäten auf inhaltliche oder strukturelle Aspekte gelenkt werden. Dies könnte eine mögliche Erklärung für die Tatsache sein, dass Lernende im Anfangsstadium ihre Gedanken eher assoziativ zu Papier bringen, während fortgeschrittenere Lerner mehr Struktur in Geschriebenes bringen. Näheres dazu wird in Kapitel 4.2.2 dieser Arbeit ausgeführt. Dieser Aspekt der Kapazitätsauslastung muss beim Lernen und Lehren beachtet werden, um Lernende nicht zu überfordern und Inhalte sinnvoll zu strukturieren.

Daneben muss das Verhältnis kognitives Lernen – erfahrendes Lernen neu beleuchtet werden: Denn rein kognitives Lernen führt nicht unbedingt zur Verwendbarkeit. Lernen besteht sicherlich auch aus kognitiven Prozessen, doch dürfen affektive und handelnde, erfahrende Prozesse nicht vernachlässigt werden. Man lernt mit den Worten Pestalozzis „mit Kopf, Herz und Hand“, unter Einbezug aller Sinne: Kognitives Herangehen und Analysieren der neuen Phänomene hilft bei der Einordnung in bestehende Wissenssysteme und bei der Feststellung der Bedeutung des neuen Phänomens – immer im Vergleich und Kontrast zu schon Bekanntem; Erfahrungen und Wiederholungen in immer neuen Kontexten helfen bei der (auch assoziativen) Memorierung des Gelernten (man denke an die mehrfachen mentalen Speichermöglichkeiten, die zu schnellerem Abrufen führen, je vernetzter sie miteinander sind); die Anwendung des Gelernten (i. S. des „Selbst-Ausprobieren-Könnens“, *learning by doing*, *automatization by usage*) hilft bei der Automatisierung der Sprachprozesse in der Sprachverwendung und sie hilft, mit den Emotionen umgehen zu lernen, die sich eventuell in realen Kulturbegegnungen ergeben könnten. Durch Fremd- wie Eigenkulturerfahrungen und Reflexion kann erreicht werden, dass neben das eigenkulturelle System ein oder mehrere weitere Kultur- und Wertesysteme treten können, die im Idealfall „gemeinsam“ aktiviert und benutzt werden können.

[61] Angeführt in Bereiter (1980: 83).

In diesem Kontext ist die Forschung des Max-Planck-Instituts[62] zur Rolle der Handlungseffekte im Alltag und beim Lernen interessant: Dort hat man festgestellt, dass die Konzentration auf das Ziel, auf den Effekt einer Handlung, von entscheidender Bedeutung ist: „Man könnte den Lerneffekt beschleunigen, wenn man sich nicht auf das sture Erlernen einzelner Schritte, sondern von Anbeginn an auf den Effekt einer Handlung konzentriert."[63] Dies dürfte auch für sprachliche Handlungen gelten und kann beispielsweise im Projektunterricht genutzt werden: Die Konzentration auf den Effekt, den eine sprachliche Handlung erzielen soll, führt eher zu einem Lernerfolg als ein bloßes Faktenlernen, bei dem die erwähnten Handlungseffekte gar nicht erfahren werden können.

Zum erfolgreichen Lernen gehören demnach unter anderem Reflexion und Handeln, Assimilation und Akkomodation, Systematisierung und Kreativität, Eigenverantwortung und Orientierungshilfe, und die Anerkennung kognitiver, affektiver und handelnder Aspekte des Lernprozesses. Die Frage, ob neues Wissen zusammen mit schon bestehendem oder in eigenen Systemen abgespeichert wird, ist nach den obigen Ausführungen im Detail nicht zu beantworten. Es kommt immer darauf an, ob sich das neue Wissen assimilieren lässt; wenn nicht, so wird es, teils in neuen Strukturen, akkomodiert. Entscheidend für den Lernerfolg ist, wie schnell und in welchen Kontexten das neue Wissen abrufbar ist – und hier scheinen neuronale Netze, die u. a. assoziativ vernetzt sind, eine wichtige Rolle zu spielen: Je ausgeprägter das Netzwerk, je mehr Verbindungen (auch durch Merhfachrepräsentation in verschiedenen Beständen) vorhanden sind, desto schneller kann Wissen aktiviert werden.

1.3.2 Die Fremdsprache im Unterricht

Die oben erarbeiteten Sprach- und Kultur-Begrifflichkeiten werden in diesem Unterkapitel auf die besondere Situation des Fremdsprachenunterrichts übertragen, insbesondere auf den Englischunterricht, da Englisch in der Regel die erste Fremdsprache ist.

Die Fremdsprache stellt zunächst den Lerngegenstand und gleichzeitig das Kommunikationsmittel im Fremdsprachenunterricht dar. Deshalb werden im Folgenden didaktische Ansätze und deren jeweilige Sprachbegriffe gemeinsam erörtert, da sich in diesem Zusammenhang Lernen und Sprache nicht trennen lassen. Denn im Fremdsprachenunterricht soll über die Fremdsprachverwendung ein Wissenssystem herausgebildet werden, das sich in seiner idealen Zielsetzung dem Wissenssystem der Muttersprache annähert. Die Fremdsprache als Kommunikationsmittel funktioniert, wie oben erläutert, auf dem Hintergrund des muttersprachlichen und eigenkulturellen Wissenssystems, weshalb auf fremdsprachliche Kommunikation, die immer zugleich auch interkulturelle Kommunikation beinhaltet, das unter Kapitel 1.2.4 erläuterte Kommunikationsmodell „Im Spiel der Lebenswelten" grundsätzlich zutrifft. Dennoch gelten in der

[62] Vgl. http://www.mpipf-muenchen.mpg.de/MPIPF/forsch_g.htm, Zugriff am 27.02.2003
[63] http://www.3sat.de/nano/cstuecke/42778/index.html, Zugriff am 27.02.2003

unterrichtlichen Kommunikationssituation andere Regeln als in der realen Alltagskommunikation mit Sprechern der Zielsprache: Die Fremdsprache wird zunächst benutzt, um Lernziele zu erreichen oder den Unterricht zu organisieren, und nicht primär zu aus dem realen Leben erwachsenden Kommunikationszwecken. Man denke beispielsweise an die artifizielle Situation der Lehrerfragen, die eine ganz bestimmte Reaktion elizitieren wollen oder an Einsetzübungen, die in dieser Art im Leben nicht vorkommen. Hierbei kommen dem so genannten *classroom discourse* besondere Funktionen zu, wie etwa die der Organisation und Steuerung des Unterrichtsablaufs und des außerunterrichtlichen Geschehens, der Disziplinierungsmaßnahmen oder der fremdsprachlichen Diskursorganisation im Unterricht. Dabei sollte die Fremdsprache möglichst authentisch benutzt werden, etwa in Anlehnung an *native speakers* und deren sprachliches Verhalten im (Fremdsprachen-)Unterricht, um idiomatische, natürliche und angemessene Sprachverwendung seitens der Lehrenden zu erzielen. Denn ansonsten läuft der Fremdsprachenunterricht Gefahr, den Lernenden unangemessenes oder gar falsches Diskursverhalten als Vorbild anzubieten. Auf Konkreta des *classroom discourse* wird unter Kapitel 1.3.3.4 dieser Arbeit näher eingegangen.

Je nach didaktischem Ansatz wird die Funktion der (Fremd-)Sprache im Unterricht aus verschiedenen Perspektiven betrachtet. Deshalb werden im Folgenden einige theoretische Ansätze[64] vorgestellt, die helfen können, den didaktischen Sprachbegriff einzuordnen in das jeweilige Lernkonzept, denn noch gibt es keine umfassende Sprachlerntheorie, die hierfür als Basis dienen könnte.

Im kommunikativen Ansatz wird versucht, den außerunterrichtlichen Begriff der Kommunikation (vgl. Kapitel 1.2.3) auf den Unterricht zu übertragen: Stichworte wie *Authentizität, sinnstiftende Kontexte, pragmatisches Vorgehen* oder *Ausrichtung an Lernerinteressen und -bedürfnissen* verweisen auf das dahinter stehende Lernkonzept: *practice makes perfect*, *learning by doing*, Herausbildung kommunikativer Kompetenz durch authentische Sprachverwendung. Dies stellt zwar eine notwendige Voraussetzung zum erfolgreichen Sprachlernen dar, doch dürfte sich die Realität an deutschen Schulen in der Regel anders darstellen. Zudem müssen, wie oben erläutert, neben Sprachverwendung auch der Kognition und Reflexion Platz eingeräumt werden, um erfolgreiches Lernen zu ermöglichen.

Der kognitive Ansatz versucht, den Begriff der innersprachlichen Organisation (vgl. Kapitel 1.2.1) für didaktische Zwecke zu übernehmen: Die Fremdsprache wird in ihrem Regelsystem bzw. in ihrer Organisation (beispielsweise nach dem oben erwähnten Prototypenmodell) so beschrieben, dass das Regelwerk bzw. das Organisationssystem das Erlernen der betreffenden Sprache erleichtert. Allerdings ist es fraglich, ob man als kompetent in einer Sprache gelten kann, wenn man alle Regeln und Ausnahmen dieser Sprache erlernt hat bzw. wenn man das Organisationssystem internalisiert hat. Denn das so erworbene deklarative Wissen kann nur

[64] Die ersten drei Ansätze sind angelehnt an Edmondson 1998, wobei zu bedenken ist, dass sich die jeweiligen Ansätze nicht gegenseitig ausschließen.

eine Wissensart darstellen, zu der noch prozedurale Wissenskomponenten und Strategienwissen hinzutreten müssen. Eine rein kognitiv ausgerichtete Auffassung des didaktischen Sprachbegriffs wird in ihrer Umsetzung im Unterricht nicht zwangsläufig zu kompetenter Sprachverwendung führen.

Der prozessorientierte Ansatz wiederum betont eher lernfördernde Handlungen und Prozesse – *wie* etwas gemacht wird, und weniger auf Sprache selbst – *was* gemacht wird: “It ain’t what you do, it’s the way that you do it“ (Edmondson 1998: 34). Allerdings stellt dieser Ansatz eher ein allgemeines Lernprinzip (vgl. Kapitel 1.3.1.3) dar, da er die Sprache selbst außen vor lässt und in dieser Ausprägung auch nicht haltbar ist. Denn man wird ohne Fokus auf Sprache als dem eigentlichen Unterrichtsgegenstand im Fremdsprachenunterricht nicht auskommen können.

Der sprachvergleichende Ansatz stellt das Kontrastieren mehrerer Sprachen in den Mittelpunkt, um Gemeinsamkeiten und Unterschiede herauszufinden. Dieser Ansatz ist gerade aus europäischer Perspektive von Bedeutung, da er eine Möglichkeit darstellt, das oben ausgeführte Mehrsprachigkeitsprinzip umzusetzen. Mittels Sprachvergleichen kann beispielsweise gezeigt werden, welche (teils universalen) Funktionen Sprache erfüllt und wie diese Funktionen auf unterschiedlichsten Wegen in verschiedenen Sprachen realisiert werden können. Nicht nur, dass man dadurch Zugang schafft zu tieferem Verständnis dessen, was Sprache ausmacht und wie sie funktioniert; man kann zusätzlich alle im Klassenzimmer vorhandenen Sprachen nutzen, somit das Vorwissen der Lernenden aktivieren und sie dadurch motivieren zum Sprachenlernen. Auch kann der Sprachvergleich helfen, Vorurteile gegen eine Sprache, seien sie nun negativ oder positiv besetzt, abzubauen, indem die Vielfalt der sprachlichen Realisierungsmöglichkeiten aufgezeigt wird und deutlich wird, dass es nicht die eine „überlegene“ Sprache gibt.

Der interkulturelle Ansatz schließlich kann als Erweiterung des sprachvergleichenden Ansatzes betrachtet werden: In ihm fungieren Sprache(n) und Kultur(en) als „Fenster“ zu anderen Sprachen, als „Mittler“ zwischen den Sprachen und Kulturen: Sprach- und kulturübergreifende Betrachtung ermöglicht das Erkennen von Gemeinsamkeiten und Unterschieden. Ein wertfrei gehaltener Vergleich verschiedener Sprachen und Kulturen kann die Funktionsweise dieser Systeme, wie sie in Kapitel 1.2.3 und 1.2.4 erläutert wurde, erhellen und zu mehr Verständnis zwischen den Kulturen führen.

1.3.3 Ableitung eines Vermittlungskonzepts

Um Fremdsprachen erfolgreich vermitteln zu können, müssen obige Sprachbegriffe und Lernprinzipien in ein kohärentes Vermittlungskonzept integriert werden. Dabei sind die Lehrkräfte im fremdsprachlichen Bildungsbetrieb gefordert, ihre Prämissen zu setzen und Position zu beziehen, denn bisher gibt es wie gesagt noch keine umfassende Theorie eines erfolgreichen Sprachlehrkonzepts, auch aufgrund der erwähnten hoch individualisierten Bedingungen und der

Komplexität der beteiligten Faktoren. Im Folgenden wird versucht, Rahmenbedingungen abzuleiten, die von den an der Vermittlung Beteiligten auf den jeweiligen Einzelfall hin ausgefüllt und umgesetzt werden müssen.

Die Bestimmung der Lernziele legt das Vorgehen teilweise mit fest: Vermittlungsmethoden, Inhalte und Darbietung hängen immer auch von den zu erreichenden Zielen ab; in diesem Bereich hat es in den letzten Jahrzehnten in der didaktischen Theoriebildung eine Verschiebung gegeben von der einst geforderten sprachlich orientierten *near nativeness* hin zu einem sprachlich-kulturellen Repertoire an kommunikativen Verhaltensweisen, das zu Handlungsfähigkeit in realen Kommunikationssituationen führen soll.[65] Deshalb kann es nicht mehr alleine um die Vermittlung primär sprachlicher Kenntnisse gehen, wie beispielsweise der Syntax, Idiomatik oder Prosodie, sondern der Fremdsprachenunterricht muss geöffnet und erweitert werden um eben diese kommunikativen und kulturellen Erfahrungen, die auch der GER als für die Mehrsprachigkeit charakteristisch erwähnt. (GER 2001: 17). Das Selbstverständnis des Fremdsprachenunterrichts hat sich denn auch verschoben: Fremdsprachenunterricht wird zunehmend verstanden als ein „Fenster zu anderen Sprachen und Kulturen“ (Schröder 1999: 3f).

Prinzipiell können Lernziele auf drei Ebenen beschrieben werden (vgl. Schröder 1971): Fachunabhängige Ziele beinhalten allgemeine Erziehungsziele, die nicht an einzelne Fächer gebunden sind, wie etwa die Erziehung zur Mündigkeit oder Kritikfähigkeit. Fächerübergreifende Ziele werden von Fachgruppen abgedeckt, wenn sie in keinem eigenen Unterrichtsfach ihre Umsetzung finden; beispielsweise können Lernziele wie interkulturelle Kompetenz oder Mehrsprachigkeit in den sprachlichen Fächern angesiedelt werden. Fachlegitimierende Ziele schließlich sind solche, die einem Unterrichtsfach zugewiesen werden können, wie etwa die kommunikative Kompetenz im Englischen als übergeordnetes Richtziel des Fremdsprachenunterrichts oder der Erwerb bestimmter Grammatikstrukturen als konkretes Feinziel. Lernziele beschreiben je nach Konkretisierungsgrad u. a. den Bedarf der Gesellschaft an Fremdsprachenkenntnissen, die Bedürfnisse der Lernenden, Lernstoff und konkrete Lerninhalte, teils unter Einbezug der Lernprozesse und des Lernverhaltens, und nicht zuletzt die erwünschten Lernergebnisse und Qualifikationen.[66] Welche Ziele im Einzelnen angestrebt werden, hängt so sehr von der jeweiligen Bildungsinstitution und der Lernergruppe ab, dass diese Arbeit keine konkreten Ziele vorstellen kann. Zudem gibt es die Lernzielkataloge des Europarats auf den Niveaus *Waystage*, *Threshold Level* und *Vantage Level*, die für grundsätzliche Zielbeschreibungen zu Rate gezogen werden können.[67]

Im Folgenden sollen, ausgehend einem Überblick über methodische Ansätze, Aussagen zu Auswahl und Darbietung und zu den Besonderheiten des *classroom discourse* getroffen werden,

[65] Vgl. hierzu beispielsweise Schröder 1999 oder Kleppin (2003: 106).
[66] Vgl. hierzu etwa Doyé 1995 oder Krumm 2003[4].
[67] Vgl. van Ek & Trim 1990, 1991, 1997. Auf diese Lernzielkataloge sind auch die Niveaus des Referenzsystems im GER ausgelegt – vgl. dazu Kapitel 3.4 dieser Arbeit.

ehe ein abschließender Blick auf die europäische Perspektive im Fremdsprachenunterricht geworfen wird.

1.3.3.1 Methodische Ansätze

Welche methodischen Ansätze sind für effektiven Fremdsprachenunterricht ratsam? In Bezug auf angemessene Methoden soll wiederum ein kurzer historischer Exkurs[68] helfen, sie in ihre jeweiligen Paradigmen einzuordnen und sie in ihrer heutigen Bedeutung einzuschätzen:

Bis Ende des 18. Jahrhunderts herrschte in der Fremdsprachenausbildung des Adels die so genannte *Konversationsmethode* vor, bei der es vorrangig um die Konversationsfähigkeit im höfischen Leben, um das „Parlieren" ging. Fremdsprachen wurden mittels Konversation gelehrt, häufig durch Muttersprachler – ein Ansatz, der im modernen Fremdsprachenunterricht wieder auflebt. Mit dem Niedergang des höfischen Lebens gegen Ende des 18. Jahrhunderts fand auch diese Methode ihr vorläufiges Ende. Während des nächsten Jahrhunderts zur Zeit des Neuhumanismus trat die *Grammatik-Übersetzungsmethode* in den Vordergrund. Die neuen Fremdsprachen sollten nach dem Modell der alten „Kultursprachen" Latein und Altgriechisch im Geist der Idealbildung unterrichtet werden, wobei die stilistische Analyse fremdsprachlicher Texte und deren Übersetzung im Mittelpunkt stand. Grammatik in Anlehnung an die des Lateinischen war die Basis, Regeln wurden deduktiv vorgegeben und mussten auswendig gelernt werden. Es gab keine Aussprache- oder Konversationsübungen.

Die seit den 1830er Jahren aufkommende Kritik an dieser Methode fasste Viëtor in seiner Schrift „Der Sprachunterricht muss umkehren" im Jahr 1882 zusammen: Die Grammatik-Übersetzungsmethode basiere auf „schöngeistigen", nicht auf authentischen Texten; das Auswendiglernen von Regeln stumpfe die Lernenden ab; es gäbe keine mündlichen Übungen; Sprache sei mehr als Wörter und Regeln auswendig zu lernen, weshalb die Grammatik-Übersetzungs-Methode nicht zu Verständnis für Sprache und deren Regularitäten führen könne und damit auch keine kommunikativen Fertigkeiten erzielen könne – eine sehr modern anmutende Kritik. Viëtor versuchte, sich von den alten Lehrmethoden zu lösen und ein angemessenes Verfahren für die Vermittlung moderner Fremdsprachen zu entwickeln: Seine *direkte Methode* (auch *Reformmethode* genannt), setzte als oberstes Ziel die Befähigung zu freiem Sprechen; er rückte Mündlichkeit und authentische Sprachverwendung in den Mittelpunkt, setzte relevante und sinnvolle Ausspracheübungen und induktive Regelableitung ein, und schlug die Einsprachigkeit im Unterricht vor, um Verständnis für eine Sprache aus dieser selbst heraus zu entwickeln und um diese in Anlehnung an den „natürlichen" Spracherwerb und -Gebrauch zu erwerben und zu verwenden. Seine Vorschläge sind auch heute noch relevant im modernen

[68] Vgl. hierzu u. a. Brown (1994: 16f, 42f, 70f, 95ff, 157ff); Spolsky 1978b; Timm (1998: 319f).

Fremdsprachenunterricht, wenngleich die Forderung nach ausschließlicher Einsprachigkeit unter Fremdsprachendidaktikern umstritten ist.[69]

Die Debatte um die Reformmethode war ab der Jahrhundertwende beigelegt, nicht zuletzt, da die Lehrerausbildung so mangelhaft war, dass die geforderte Einsprachigkeit keinen Sinn machte, und da etwa induktive Regelableitung sich als zu zeitaufwändig herausstellte. Auch die unzureichende Überprüfbarkeit der Lernfortschritte war ein Grund, warum man sich von der Mündlichkeit als oberstem Ziel abwandte. Viëtors Methode fand jedoch Niederschlag in der ab der Jahrhundertwende aufkommenden vermittelnden Methode, die als Kompromiss verstanden werden kann, neue Elemente aufzunehmen, ohne alte Strukturen aufzugeben. Sie war gekennzeichnet durch kognitives Grammatiklernen, wobei Grammatikübungen aber in authentische Situationen gebettet waren, durch Nutzung paraphrasierter Vokabelgleichungen, durch Einbezug von Mündlichkeit und Ausspracheübungen, und nicht zuletzt durch authentische Sprachverwendung, wobei die Basis immer noch Texte bildeten, nun aber authentische Texte.

Die drei hier skizzierten Methoden fallen in die oben in Kapitel 1.1 erwähnte *pre-scientific period*. Ab Mitte des 20. Jahrhunderts beeinflussten das Paradigma des Strukturalismus und Behaviorismus sowie die aufkommenden Sprachlabors die Sprachlehrmethoden und förderten die Entwicklung der *audio-lingualen Methode*: Das oberste Ziel war automatisierte Sprachbenutzung; das Augenmerk lag vorwiegend auf Herausbildung der sprachlichen Kompetenz. Lernen wurde nach dem behavioristischen Lernmodell als imitativ-reaktives Verhalten betrachtet, als *habit formation*, welche über Konditionierung gesteuert werden könne. Auf Basis strukturalistischer Prinzipien wurde Sprache in portionierbare Einheiten eingeteilt und Regeln und Besonderheiten durch Imitation und Wiederholung gedrillt (die Stichworte *slot-filler-technique* und *pattern drill* mögen an dieser Stelle genügen). Die damals an diesem Vorgehen geäußerte Kritik ist auch heute noch gültig: Diese Methode führt nur zu schematischer Kenntnis der Sprache; das Drillen bestimmter Phrasen wirkt demotivierend und lässt keine Einsicht in die Funktionsweise von Sprache oder Kommunikation zu und kann somit auch nicht zu kommunikativer Kompetenz führen. Chomsky (1959)[70] äußerte zudem Kritik am behavioristischen Lernmodell, das auf der Beobachtung von Tieren beruht und keine Rückschlüsse auf komplexen menschlichen Spracherwerb zulässt.

Ab den 60er Jahren des letzten Jahrhunderts, in der *integrative-sociolinguistic period*, fand die so genannte pragmatische Wende statt: Man wandte sich ab von der Konzentration auf rein sprachliche Kompetenz hin zur kommunikativen Kompetenz. Man kann diesem Paradigma keine eigene Methode zuordnen; vielmehr handelt es sich um eine Weiterentwicklung der bisher skizzierten Methoden hin zu einem Methodenmosaik, in dem alle hilfreichen und angemessenen Komponenten früherer Ansätze ihre Verwendung finden. Kennzeichnend für diese Methodik ist

[69] In diesem Zusammenhang vgl. für einen Überblick der Diskussion beispielsweise Ahrendt 1991, oder zu einer Synthese-Position beispielsweise Amor 1999, der je nach Situation und Lernerstand den angemessenen Einsatz von L1 oder L2 vorschlägt.
[70] Eine nähere Ausführung seines Konzepts des *language acquisition device* würde hier zu weit führen.

das Lernziel der kommunikativen Kompetenz, auf das hin alle Komponenten ausgerichtet sind: Je nach Lernergruppe, Zielen, Unterrichtsphase und Lerngegenstand sollten alle Ansätze verwendet werden, die effektiv erscheinen: Eine pluralistische Mischung aus kommunikativem Ansatz, in dem die Fremdsprache möglichst authentisch verwendet wird, kognitivem Herangehen, bei dem Bewusstseinsprozesse durch Reflexion und Analyse ermöglicht werden, affektiv-handelndem Ansatz, der für die notwendigen Erfahrungen sorgt, prozessorientiertem Herangehen, das sich mit lernförderlichen Prozessen und Strategien beschäftigt, sprachvergleichendem Ansatz, der den nötigen pan-sprachlichen Kontext gibt, und nicht zuletzt dem interkulturellem Ansatz, der die kulturelle Komponente mit einschließt, sorgt für die notwendige Vielfalt und Ausgewogenheit aller Elemente im Fremdsprachenunterricht. Dieser Pluralismus gilt auch für den Medieneinsatz, für Sozial- und Arbeitsformen, Themen- und Textauswahl, Übungs- und Testformate, und mögliche Darbietungsformen. Je abwechslungsreicher die eingesetzten methodischen Ansätze, desto eher dürfte man der Vielfalt der Lernenden gerecht werden.

Kennzeichnend für einen Fremdsprachenunterricht, der den oben unter Kapitel 1.3.1 skizzierten Lernbedingungen gerecht werden will und die im Vorangegangenen skizzierten methodischen Ansätze effektiv umsetzen will, sind folgende *core principles*, welche jedoch nicht etwa abschließend zu verstehen sind:

- Zielsetzung: Befähigung zu kommunikativer und akkulturativer Kompetenz, zur Sensibilisierung im Umgang mit fremden Kommunikationssituationen, zur kritischen Toleranz des „nahen Fremden“; Vermittlung von Schlüsselqualifikationen wie Kooperations- und Kommunikationsfähigkeit oder Problemlöse- und Entscheidungskompetenzen;

- Authentizität, Ganzheitlichkeit und Kontextualisierung von Sprache und Sprachhandlungen in ihre sozio-kulturellen Kontexte, um den natürlichen Bedingungen von Sprachverwendung möglichst nahe zu kommen;

- Anerkennung der Notwendigkeit von „Inkubationsphasen“, in denen Sprache erst rezeptiv aufgenommen werden muss: rezeptive Fertigkeiten erhalten gegenüber produktiven mehr Raum;

- Spiralprogression: zyklisches Darbieten des Stoffs statt blockweiser Behandlung;

- Binnendifferenzierung, Methodenvielfalt und Heterogenität, um der Individualität des Lernprozesses, den unterschiedlichen Lernerpersönlichkeiten und Lerntypen gerecht zu werden;

- Lernfördernde Atmosphäre: gemeinsames Lernen in entspannter Umgebung statt eines von Konkurrenz und Fehler-Angst geprägten Lernklimas;

- Lernerautonomie: Selbstbestimmung und Eigenverantwortung: Einbindung der Lernenden in den Lern- und Vermittlungsprozess bis hin zur Befähigung zu lebenslangem Lernen; hierunter fällt auch die Vermittlung von Lern- und Arbeitstechniken sowie Organisationsfähigkeit;

auch das Konzept „Lernen durch Lehren“[71], bei dem Lernende solche Aspekte vermitteln, die sie bereits beherrschen, kann zur Entwicklung von Lernerautonomie beitragen;

- Neue Evaluationsformen: Neuer Umgang mit Fehlern als Lernanreiz; Positivkorrektur; Einbindung aller Beteiligten in evaluative Prozesse, einschließlich der Selbstevaluation.

1.3.3.2 Auswahl und Anordnung

Wenden wir uns nun der Frage zu, *was* im Unterricht gelehrt und gelernt werden soll. Welche Einheiten von Sprache und Kultur müssen abgedeckt werden? Inwieweit können Schlüsselqualifikationen im Fremdsprachenunterricht umgesetzt werden? Über die Auswahl entscheiden letztlich die jeweiligen Ziele, curriculare Vorgaben, die Lernenden, Lehrenden und die Unterrichtssituationen. Im Rahmen dieser Arbeit soll nur ein Minimaldesiderat aufgestellt werden: In jedem Fall müssen Sprache und Kultur Gegenstand des Fremdsprachenunterrichts sein, ebenso wie Lern- und Arbeitstechniken. Gerade die Techniken sind im Hinblick auf lebenslanges Lernen nicht zu unterschätzen, denn nur wenn sich die Lernenden selbst einschätzen können bezüglich ihres Lernstils werden sie die für sie angemessenen Lerntechniken einsetzen können. Dazu müssen grundlegende Arbeitstechniken wie Lesestrategien, der Einsatz von Hilfsmitteln, Sinnerschließungstechniken und mediale Kompetenzen vermittelt werden, um nur einige zu nennen. In Bezug auf die Auswahl von sprachlichen und kulturellen Aspekten muss bedacht werden, dass Sprachen und Kulturen nie in ihrer Gesamtheit Lerngegenstand sein können, da sie niemandem vollständig zugänglich sein können. Vielmehr ist es ratsam, in Absprache mit den Lernenden (soweit dies möglich ist) und in Übereinstimmung mit den Curricula und den Kurs- oder Unterrichtszielen eine angemessene Auswahl zu treffen: „Each individual student should have rights to determine the balance of language and culture he or she chooses to adopt.“ (Broadbent & Oriolo 1991: 308). Diese Forderung wird in der schulischen Situation nur bedingt umsetzbar sein (man denke etwa an Stationenlernen oder Projektarbeit), doch in außerschulischen Sprachkursen und gerade in der Erwachsenenbildung stellt sie ein motivierendes Element dar und kann zur Übernahme von Eigenverantwortung für den Lernprozess anregen. Die Stoffauswahl muss in jedem Fall auf die Lernergruppe hin ausgelegt sein, um die Lernenden dort abzuholen, wo sie stehen und um sie zu motivieren und neugierig zu machen. Beispielsweise dürften sich jugendliche Lerner für Aspekte der Jugendsprache interessieren – wenn sie im Fremdsprachenunterricht erfahren können, wie ihre Altersgenossen sich in der zu erlernenden Sprache ausdrücken, so kann dies motivierend wirken.

Die einmal ausgewählten sprachlichen und kulturellen Elemente sollten sowohl kognitiv als auch affektiv-erfahrend und relflexiv-begleitend abgedeckt werden. Neben Wissensvermittlung muss den Lernenden auch handelndes Erfahren sozusagen „am eigenen Leib“ ermöglicht

[71] Vgl. hierzu beispielsweise Martin & Kelchner 1998.

werden, genau wie ihnen Reflexion über das Erfahrene ermöglicht werden soll. Dann können sich im Idealfall Wissen und Handlungsfähigkeit zusammen mit Erfahrung und Bewusstheit entwickeln. Bei Byram (1991: 20) findet sich folgende Übersicht, die grundlegende Elemente des Sprach- und Kulturlehrprozesses gegenüberstellt:

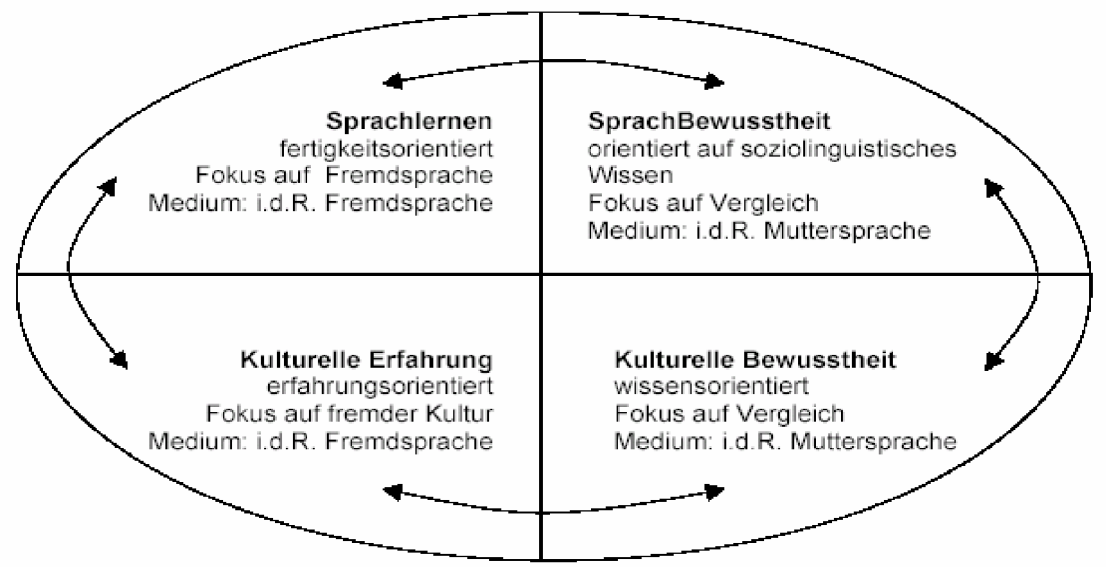


Abb. 5: *The Language and Culture Teaching Process*

Dabei sollten die ausgewählten Bereiche und Gegenstände immer wieder umgewälzt werden in einer spiralförmigen Progression, so dass ein einmal eingeführtes Phänomen nach und nach in immer neuen Kontexten auftaucht und auf diese Weise von den Lernern in seiner ganzen Bedeutung und Verwendbarkeit aufgenommen werden kann. Denn blockweises Abhandeln des Stoffs im Sinne eines Baukastensystems führt nicht automatisch zum dauerhaften Erwerb:

> Sprachphänomene sind eben keine Steine [Anspielung auf die Baukastenmentalität, bei der Stein auf Stein gesetzt wird, Anm. d. V.], sondern „unscharfe Mengen", jene *fuzzy sets* gemäß der Fuzzy-Logik, die aus ihrer Vernetztheit im so komplexen Sprachsystem leben, die die Eingebundenheit in die Situation brauchen und das Mitdenken der Beteiligten erforderlich machen. Sie benötigen den größeren Zusammenhang und die Vernetzung. Sie können nicht einzeln herausgebrochen und isoliert gelernt[72] werden. Deswegen ist der Sprachlernprozess eben nicht-linear. (Bleyhl 1996: 347)

Dieses Vorgehen kann gekoppelt werden mit den schon erwähnten Inkubationsphasen, Gelegenheiten für die Lerner, gesprochene oder geschriebene Sprache rezeptiv aufzunehmen, ehe sie Sprache produzieren müssen. Auch hierbei kann man sich zur Begründung an den Muttersprachenerwerb anlehnen.[73] Denn es braucht Zeit, sprachlichen Input zu verarbeiten, sprachliche Schemata und Prototypen herauszubilden, und die entsprechenden Phänomene zu analysieren, um sie in das bestehende Wissenssystem zu integrieren. Dies hat auch Auswirkungen auf das Verhältnis der traditionellen *four skills*: Den rezeptiven Fertigkeiten sollte mehr Platz eingeräumt werden, gerade im Kontext der europäischen Mehrsprachigkeit und rezeptiven Dreisprachigkeit. Bleyhl (1996: 344ff) beispielsweise hat die „Verstehensmethode", bei der lange rezeptive

[72] Es liegt der Schluss nahe, dass nicht isoliert gelernte features auch nicht isoliert getestet werden sollen, doch dazu mehr unter Kapitel 2 *Das Testen des Sprachvermögens*.
[73] Vgl. auch Bleyhl 1996, der Wandruszka zitiert, welcher das Verhältnis Verwenden – Verstehen in der Muttersprache mit 1:1000 beziffert.

Verstehensphasen der Produktion vorgeschaltet sind, in einem Unterrichtsversuch angewandt und im Vergleich zu traditionellem Unterricht festgestellt, dass erste effektiver zum Ziel führt.

Die erwähnten Inkubationsphasen bieten sich auch deswegen an, da es die Hypothese gibt, dass neue Elemente zuerst ganzheitlich abgespeichert werden als so genannte *chunks*, die erst später einer internen Analyse unterzogen werden und in ihren Einzelheiten erfasst werden, um dann in bestehende Wissenssysteme integriert zu werden oder um neue Systeme herauszubilden.[74] Erst dann ist eine Verwendung jenseits des Abspulens auswendig gelernter Routinen denkbar.

1.3.3.3 Darbietung

Nun, da die grundlegendsten Fragen der Stoffauswahl und Anordnung geklärt sind, wird betrachtet, *wie* der Stoff im Fremdsprachenunterricht dargeboten werden soll, wenn man obige Überlegungen zu Sprache und Lernen mit einbezieht. Es bietet sich aufgrund der oben erwähnten noch nicht umfassend vorliegenden Sprachlehrtheorie an, bei offenen Fragen – wo immer es sinnvoll erscheint – auf die Bedingungen des Muttersprachenerwerb Rekurs zu nehmen. Beispielsweise wird die Muttersprache in ganzheitlichen, sinnstiftenden Kontexten erworben. Analog dazu müsste der Fremdspracherwerbsprozess erleichtert werden, wenn Fremdsprache ganzheitlich eingebettet in ihre natürlichen Kontexte dargeboten und verwendet wird, so dass die Lernenden die natürlichen Bedingungen der Sprachverwendung und die innersprachlichen Organisationsstrukturen (auch unbewusst) aufnehmen können.[75]

Wortschatzelemente zum Beispiel sollten nicht, wie schon erwähnt, als isolierte Wortlisten mit muttersprachlichen Übersetzungen präsentiert werden, sondern immer eingebettet werden in ihre semantischen Felder und in ihre typischen Verwendungskontexte. Ausgehend vom prototypischen Verhalten eines lexikalischen *items* ist es für die Lernenden einfacher, dieses in der Ganzheit seiner Bedeutung(en) und in seinem Auftreten zu erfassen. Die jeweiligen Auftrittsbedingungen eines Wortes oder einer Phrase sollen ebenfalls thematisiert werden: Es gibt oft motivierte Gründe, warum man etwas in einer bestimmten Weise sagt – diese Bedingungen können den Lernenden helfen, die treffenden Ausdrücke zu wählen, beispielsweise im Hinblick auf die Umsetzung von Höflichkeitskonventionen. Die Präsentation neuen Wortschatzes kann und soll auch im Vergleich zur Muttersprache und anderen Sprachen erfolgen, wo immer es Sinn macht: Man kann auf gemeinsame Wurzeln hinweisen, Internationalismen nutzen, Ähnlichkeiten als Lernanreiz bieten, und „nebenbei“ Techniken der Sinnerschließung unbekannter Wörter oder den Umgang mit Lexika einführen.

[74] Vgl. hierzu u. a. Edmondson 1998.
[75] Da Sprache und Gehirn sich koevolutionär entwickelt haben dürften, liegt die Vermutung nahe, dass beide Systeme über natürliche Organisationsprinzipien verfügen, die sich in komplementärer Weise ergänzen und den Spracherwerb positiv unterstützen könnten. Hierzu gibt es keine empirisch fundierten Beweise, doch da jeder Mensch seine Muttersprache quasi „natürlich“ erwirbt, muss es Erwerbsmechanismen geben, die m. E. aber nicht in dem von Chomsky postulierten angeborenen *language acquisition device* angesiedelt sind, sondern wohl eher in der Struktur von Gehirn und Sprache.

Grammatische Phänomene sollten, wie lexikalische, ausgehend von ihrem prototypischen Auftreten und Verhalten eingeführt werden, und zwar auf Wort-, Satz- und Textebene, so dass nach und nach die Netzwerkstruktur der sprachlichen Organisation deutlich wird. Die Funktionen der grammatischen Phänomene müssen ebenfalls thematisiert werden, um die Auftrittsbedingungen zu verdeutlichen. Auch hierbei können motivierte „Regeln" helfen, die Gründe für das Auftreten einer bestimmten Form in einem bestimmten Kontext klarzustellen. Man denke beispielsweise an die englische *continuous form*, die mehrere Funktionen übernehmen kann: Das Erlernen der Form alleine bereitet wenig Schwierigkeiten, doch sagt die korrekte Produktion dieser Form nichts darüber aus, ob auch die Funktionen des Morphems, etwa das Ausdrücken einer gerade andauernden Aktion, einer zukünftigen Aktion oder wiederholter Aktionen, mit erworben worden sind. In diesem Zusammenhang kann der oben schon erwähnte Vergleich der Versprachlichungsmöglichkeiten einer bestimmten Funktion über verschiedene Sprachen hinweg helfen: Ausgehend von den drei grundlegenden Perspektiven, wie Dinge betrachtet werden können (namentlich die durative, iterative und punktuelle Betrachtungsweise[76]), kann man vergleichen, wie welches Konzept in welcher Sprache umgesetzt wird. Im Spanischen und Französischen beispielsweise gibt es Vergangenheitsformen, die für durative Zustände in der Vergangenheit benutzt werden, dem Deutschen und Englischen aber unbekannt sind. Im Englischen hingegen wird die *continuous form* zum Ausdrücken der durativen Perspektive (unabhängig von der zeitlichen Perspektive) benutzt, wohingegen dies im Deutschen über Adverbien wie *gerade* oder Phrasen wie *dabei (gewesen) sein, etwas zu tun* ausgedrückt wird. Durch solch ein Vorgehen erhält das *ing*-Morphem nicht nur den größeren innersprachlichen Kontext, den die Lerner brauchen, um zu erkennen, wann es auftritt – es erhält darüber hinaus einen sprachvergleichenden Kontext, der es den Lernern ermöglicht, universale Funktionen und deren vielfältige Versprachlichungsmöglichkeiten in mehreren Sprachen zu erkennen. Dabei sollten Phasen der grammatischen Deduktion und Induktion abgewechselt werden, um allen Lernenden gerecht zu werden und um sicherzustellen, dass selbst erschlossene Gesetzmäßigkeiten auch korrekt memoriert werden.

Die Einbettung sprachlicher Phänomene in größere sprachliche wie sprachübergreifende Kontexte kann den Lernenden Einblicke in die Funktionsweise von Sprache geben und sie kann ihnen helfen, ihre Spracherfahrungen zu reflektieren. Gleiches trifft zu auf die Präsentation kultureller Phänomene: Wenn Neues auf dem Hintergrund des schon Bekannten aufgenommen wird oder dadurch die Aufnahme erleichtert wird, so sollte man, wann immer es angebracht erscheint, zielsprachliche und zielkulturelle Phänomene mit solchen der Muttersprache und des eigenkulturellen Orientierungssystems kontrastieren. Dadurch kann, wie oben bereits angedeutet, Eigenes im Gegensatz zu Fremdem reflektiert werden und man kann Fenster zu anderen Sprachen und Kulturen auch dadurch öffnen, dass man alle in einer Klasse vorhandenen Muttersprachen und Kulturen in den Fremdsprachenunterricht einbezieht. So können Unterschiede

[76] Vgl. Schröder 2004.

und Gemeinsamkeiten als individuelle Lernanreize angeboten werden. Auch können eventuell existierende Vorurteile durch die Einbeziehung aller Kulturen und Sprachen in einer Klasse aufgefangen werden, indem die Lernenden erfahren, dass ein und dieselben (sprachlichen oder kulturellen) Funktionen auf vielen, grundsätzlich gleichwertigen Wegen umgesetzt und realisiert werden können. Sollte Fremdes dennoch bedrohlich wirken, so ist es wichtig, solche Erfahrungen zu begleiten: Beispielsweise kann man den Lernenden erklären, welche Hintergründe einwirken und man kann ihnen individuelle Wege aufzeigen, die Bedrohung in einen Lernanreiz umzuwandeln.

Die zielsprachliche Kultur wird im Klassenzimmer idealiter über vergleichende Verfahren eingeführt: Wissen über die Zielkultur wird im Kontrast zu anderen Kulturen und zunächst im Zusammenhang mit entsprechenden Versprachlichungsmöglichkeiten vermittelt. Doch das Wissen um andere Kulturen muss, analog zu Sprachwissen und Spracherfahrung, um Momente der Kulturerfahrung ergänzt werden. Nur durch direkte Erfahrung der Beziehungen von Sprache und Kultur, und durch direkte Erfahrung des fremdkulturellen Systems auf dem Hintergrund des eigenen Systems können Lernende auf den Umgang mit fremden Kulturen und auf das Handeln in fremden Kulturen vorbereitet werden. Diese Erfahrungsmomente sollen nicht etwa zur Übernahme fremdkultureller Systeme führen, sondern zur zunächst wertfreien Wahrnehmung und kritischen Toleranz des Fremden, zur Offenheit in der fremdkulturellen Begegnung, die später kulturell wie sprachlich reflektiert werden kann. Dabei können Verfahren wie beispielsweise Projektarbeit, die Nutzung von *Critical Incidents*[77] oder Planspielen und Simulationen in Form von Krisenexperimenten helfen:

Es bieten sich Projekte an, bei denen Kontakte mit *native speakers* herbeigeführt werden: Bahnhöfe, Flughäfen, Jugendherbergen, Universitäten und viele weitere Orte sind geeignet, authentische Kommunikationssituationen zu schaffen, die thematisch auf die Bedürfnisse und Interessen der Lernenden abgestimmt werden können. Motivierend wirken sich auch E-mail Projekte mit *native speakers* aus: Beispielsweise berichten Jost & Multhaupt (1996) von einem Telekommunikationsprojekt mit Kanada, das selbst eine problematische Hauptschulklasse zum authentischen Gebrauch der englischen Sprache anregen konnte. Gaile (1999) wiederum berichtet von einem europäischen Bildungsprojekt im Rahmen von multilateralen Schulpartnerschaften, wobei die beteiligten Schulen zeitgleich ausgewählte Themen von europäischer Relevanz bearbeiteten, um auf diese Weise unterschiedliche Sichtweisen zu vergleichen, Perspektiven zu wechseln und aufgrund der vorwiegend auf Englisch geführten Kommunikation und der Nutzung moderner Medien gleichzeitig die Kommunikationsfähigkeit im europäischen Kontext zu fördern.

Die Verwendung so genannter *Critical Incidents* und deren Aufarbeitung stellt eine weitere Möglichkeit dar, die Lernenden der Fremderfahrung im geschützten Kontext des Klassenzimmers

[77] Zur Verwendung von *Critical Incidents* vgl. beispielsweise Fetscher & Hinnenkamp 1994, Gudykunst et al. 1996, Thomas & Wagner 1999.

auszusetzen. *Critical Incidents* beschreiben interkulturelle Situationen, die potentiell krisengeladen sind und aufgrund unterschiedlicher Wissensbestände der an der Situation Beteiligten zu Fehlinterpretationen oder Missverständnissen führen (können). Solche Situationsbeschreibungen können zur Bewusstmachung wie auch zur Erfahrung potentiell kritischer Umstände genutzt werden, um zu lernen, mit Verhaltensweisen oder Situationen umzugehen, die sich nicht auf eigenkulturellem Hintergrund verstehen lassen, sondern die aus fremdkultureller Perspektive interpretiert und bewertet werden müssen. Es bietet sich eine Reihe von Einsatzmöglichkeiten an, von der Vorgabe einer solchen Situation zur kognitiven Analyse über die eigenständige Erarbeitung von Deutungsmöglichkeiten oder die Erstellung einer kritischen Situation (eventuell aus schon gemachter Erfahrung heraus, die auf diese Weise reflektiert werden kann), bis hin zu freien Ansätzen: Denkbar sind hierbei Rollenspiele mit Angehörigen verschiedener Kulturkreise, die neben der kognitiven auch die affektiv-handelnde und erfahrende Perspektive mit einbeziehen, so dass die Lernenden in relativ geschützter Umgebung Erfahrungen im Umgang mit Krisensituationen sammeln können. So genannte Krisenexperimente beziehen ebenfalls die habituelle und affektive Ebene mit ein; im Gegensatz zu Rollenspielen wissen die Beteiligten aber nicht um ihre Rolle, sondern werden unvorbereitet in eine potentielle Krisensituation gebracht, in der sie auch sprachlich handeln müssen und im Nachhinein, etwa auf Basis einer Videoaufzeichnung, ihr verbales wie nonverbales Verhalten analysieren und reflektieren können.

Generell sollte es genügend Raum geben für selbstentdeckendes und selbstgesteuertes Lernen, um der Individualität der Lernenden gerecht zu werden und eventuell vorhandene Motivationen weitgehend zu nutzen. Hilfreich ist auch ein entspanntes, angstfreies Lernklima, in dem alle als Team mit- und voneinander lernen. Selbstverständlich müssen die Lehrenden sich ihrer Verantwortung als Vermittelnde bewusst sein und Orientierung und Hilfestellung anbieten, wann immer es geboten scheint oder die Lernenden danach verlangen, doch sollte immer auch Raum bleiben für selbständige Analysen und Ableitungen.

1.3.3.4 *Classroom Discourse*

Wenden wir uns nun der Verwendung der Fremdsprache im Unterricht zu steuernden Funktionen zu, dem *classroom discourse*.[78] Wie oben erwähnt sollte sich diese Art des Sprachgebrauchs möglichst an den Gebrauch durch *native speakers* anlehnen, wie er beispielsweise in Cattliff & Thorne (1988) dargestellt ist, um die Lerner erfahren zu lassen, welche Regularitäten in der fremdsprachlichen Kommunikation gelten und wie welche (unterrichtlichen) Funktionen versprachlicht werden können. Welche Funktionen erfüllt nun diese Art des Diskurses? Zuerst einmal wird damit das unterrichtliche Geschehen organisiert: Grüßen und Verabschieden fallen

[78] Die im Folgenden verwendeten englischen Fachtermini aus dem Bereich der Diskursanalyse werden nicht übersetzt, da sie eine ganz bestimmte Bedeutung in diesem Forschungsfeld tragen, die durch deutsche Übersetzungen nicht adäquat wiedergegeben werden kann. Zur Definition dieser Termini darf auf die hier zitierte Literatur verwiesen werden.

hierunter, genau wie die eigentliche Organisation des Stundenablaufs. Dazu gehören Situationen wie Überblicke geben über den Ablauf der Stunde, Planungen des längerfristigen Unterrichtsverlaufs, oder auch Disziplinierungsmaßnahmen. Des Weiteren dient *classroom discourse* der Einführung, Strukturierung, Erklärung oder auch Wiederholung des jeweiligen Unterrichtsgegenstands selbst. Innerhalb dieser Funktion werden Anweisungen gegeben, Aufgaben gestellt, *Feedback* und Korrekturen gegeben, Fragen gestellt und nicht zuletzt wird über den Diskurs motiviert. Unterrichtsgespräche und Diskussionen wiederum verlangen möglichst authentische Diskursorganisation, sei es hinsichtlich der Rederechte (allgemeiner *turn-taking* genannt), des Frageverhaltens oder der Höflichkeitsnormen in der Konversation.

Um lernförderliche Merkmale des unterrichtlichen Diskurses zu identifizieren, kann man sich der Diskursanalyse authentischer Unterrichtssprache in der Zielkultur bedienen. Im Folgenden sind solche lernfördernden Merkmale zusammengestellt, basierend auf nachstehenden Überlegungen: Nach Ellis (1990) ist Spracherwerb u. a. die Folge gemeinsamer Anstrengungen zwischen Lehrenden und Lernenden und beinhaltet ein dynamisches Zusammenspiel von äußeren und inneren Faktoren; dabei beeinflussen sich Lehrer und Schüler gegenseitig im In- und Output und konstruieren gemeinsam einen Diskurs, an dem sich die Art und Weise, wie der Lernprozess abläuft, manifestiert und aufzeigen lässt. Dieser Annahme der Zusammenhänge zwischen Diskurs und Lernprozessen liegen mehrere Spracherwerbshypothesen zugrunde:[79]

- Die „Input-Hypothese“ nach Krashen (1977) besagt, dass Lerner gemäß einer natürlichen Erwerbsordnung Fortschritte machen, wenn sie über genügend fremdsprachlichen Input verfügen, der zudem verständlich sein muss und über dem Niveau ihrer Interimsprache liegen muss; Veränderung des Input verändere die Art des Spracherwerbs.
- Die „Interaktions-Hypothese“ nach Long u. a. baut auf der Input-Hypothese auf, indem sie verständlichem Input große Bedeutung zuschreibt, Spracherwerb aber an möglichst umfassenden interaktionalen Aktivitäten zwischen Lehrenden und Lernenden festmacht; verständlichen Input erreiche man demnach über Vereinfachung desselbigen, möglichst viel sprachlichen wie außersprachlichen Kontext und Modifikationen der interaktionalen Strukturen.
- Die „Output-Hypothese“ nach Swain (1985) stellt sozusagen eine Erweiterung der Interaktions-Hypothese dar, da sie besagt, dass ohne aktiven Gebrauch der Fremdsprache Spracherwerb nicht möglich sei, denn ein in der unterrichtlichen Situation „erzwungener“ Output mobilisiere die sprachlichen Ressourcen, er diene der Überprüfung von Hypothesen, und er zwinge zur Beachtung formaler Sprachmerkmale, wohingegen Input die Aufmerksamkeit eher auf Bedeutung lenke.

Ausgehend von der Annahme, dass Interaktion zwischen Lernenden und Lehrenden zu lernförderndem Diskurs führt, können Diskursmerkmale wie die Aushandlung von Lerninhalten, gegenseitiges Verständnis, Gleichberechtigung der Partner, Initiierung neuer Gedanken, transparentes *Feedback*-Verhalten u. a. als lernfördernd festgestellt werden (vgl. beispielsweise Henrici 1995). Im Folgenden sind die Ergebnisse einiger Diskursanalyse-Forscher kurz zusammengestellt:

- Sinclair & Coulthard (1975) stellen fest, dass „qualitativ hochwertiger“ Diskurs im Sinne der gerade angeführten Merkmale lernfördernd ist.

[79] Nach Ellis 1990, insbesondere Kapitel 5.

- Swain (1985) schreibt, dass eine „erzwungene“ Schüleräußerung (“pushed output“) in Form einer Selbstkorrektur das Reformulieren der ursprünglichen Äußerung erfordert und somit die Weiterentwicklung der Lernersprache fördert.
- Nach Chaudron (1988) wird Lerneffektivität u. a. erreicht durch die Anpassung der Lehrersprache an das Niveau der Lernenden; durch Wiederholungen und Redundanz in der Lehrersprache; durch Transparenz und Konsistenz im *Feedback*-Verhalten; und durch interaktive Fragen, die das Aushandeln von Verständnis seitens Lehrenden und Lernenden ermöglichen.
- Ellis (1990, 1994) sieht lernfördernde Merkmale in Bedeutungsaushandlungen zwischen Lehrenden und Lernenden; im Einsatz möglichst vieler verschiedener Sprechakte der am Vermittlungsprozess Beteiligten; im verständlichen Input durch modifizierte Lehrersprache; in Gelegenheiten für Sprachpraxis seitens der Lernenden; in interaktionalen Rückfragen wie den so genannten *comprehension or confirmation checks* und *clarification requests*. Er stellt darüber hinaus fest, dass lehrerkontrollierter pädagogischer Diskurs zu formalen Fertigkeiten führt und lernerzentrierter natürlicher Diskurs zu mündlichen Fertigkeiten.
- Henrici (1995) betrachtet folgenden Sachverhalt als Indiz für kurzzeitigen Erwerb fremdsprachlicher Phänomene: Die sprachliche Bearbeitung eines Problems von beiden Partnern, wobei die Lösung seitens der Lehrenden ratifiziert wird, deutet dann auf kurzfristigen Erwerb, wenn die ratifizierte sprachliche Lösung [in verändertem sprachlichen Kontext] seitens der Lernenden wieder verwendet wird. Diese Wiederverwendung deutet Henrici als Dokumentation, dass ein sprachliches Phänomen/Problem verstanden wurde und daher seitens der Lernenden kurzfristig (re-)produziert werden kann.

Betrachten wir nun die Lehrersprache näher, die im Fremdsprachenunterricht von entscheidender Bedeutung ist: Wie können *teacher talk*, Frage- und *Feedback*-Verhalten lernfördernd gestaltet werden? Lehrersprache sollte möglichst authentisch gehalten werden, sei es im Frageverhalten, hinsichtlich des *turn-taking* oder der kommunikativen Abläufe. Eine zu künstliche Sprache oder ein zu einfach gehaltener so genannter *caretaker talk* spiegeln die Realität nicht wider und können deshalb nicht auf diese vorbereiten. Allerdings muss die Lehrersprache auf das Niveau der Lernenden hin angepasst werden, wenn sie auch leicht über diesem liegen darf, um die Interimsprache herauszufordern. Die Fragen im Unterrichtsgespräch müssen ebenfalls adäquat für das Niveau der Lerner formuliert sein und die Lehrkraft sollte diese gegebenenfalls modifizieren können. Offene Fragen lassen mehr Aushandlungsspielraum, doch hängt die Art der Fragestellung von der jeweiligen Unterrichtsphase ab. Dem *Feedback*-Verhalten kommt ganz entscheidende Bedeutung im Lernprozess zu: Die Lernenden können ihre Hypothesen über die Zielsprache im Lernprozess nur überprüfen, wenn sie *Feedback* erhalten. Es ist zwar nicht abschließend erforscht, wer wann welche Fehler auf welche Art kommentieren sollte, doch lassen sich einige allgemeine Prinzipien (nach Chaudron 1988: 132-152) feststellen:

- Selbstkorrektur geht vor *peer correction,* und diese vor Lehrerkorrektur;
- Hinweise zur Selbstkorrektur sind hilfreicher als die Korrektur selbst;
- Selektives Vorgehen ist ratsam, da nicht alle Fehler von gleicher Tragweite sind: Kommunikationsbelastende Fehler oder Fehler kultureller Art haben meist weitreichendere Folgen als etwa reine Grammatikfehler, die den Sinn einer Äußerung nicht entstellen. Systematische Fehler sollten kommentiert werden.
- *Feedback*-Verhalten sollte von den jeweiligen Unterrichtsphasen abhängig gemacht werden (formbezogene vs. kommunikative Phasen) und vielfältig sein (Nutzung von Hinweisen,

Kommentaren, Reparaturen, Korrekturen, Nachfragen, Denkanstößen, Wiederholungen, etc.); selbstverständlich darf *Feedback* nie verletzend oder gar demütigend sein;
- Loben wird als motivierend empfunden, wenn es nicht übertrieben eingesetzt wird.

Wichtig ist, den Lernenden auch im Rahmen des *Feedback* Reparaturhilfen anzubieten und ihre Sensibilität für (interkulturelle) Missverständnisse zu schulen, um auftretende Probleme in echten Kommunikationssituationen rechtzeitig wahrzunehmen und entschärfen zu können. Dabei bietet es sich an, den Bereich der sprachlichen Reparaturtechniken einzuführen, damit die Lernenden adäquat nachfragen oder ihren Standpunkt erklären können, sollte es zu einer kritischen Situation gekommen sein. Auch sprachliche Möglichkeiten des Entschuldigens sollten thematisiert werden. Dies kann beispielsweise im oben gesteckten Rahmen bei der Verwendung von *Critical Incidents* geschehen, so dass Sprach- und Kulturarbeit verzahnt werden.

Zur Frage, ob der Unterricht ganz einsprachig gehalten werden soll, darf auf den oben unter Kapitel 1.3.3.1 skizzierten Disput verwiesen werden. Man kann vorsichtig feststellen, dass es verschiedene Phasen gibt, in denen es unterschiedlich sinnvoll ist, in der Einsprachigkeit zu verharren: Alle reflexiven Phasen oder Phasen der Unterrichtsorganisation könnten, gerade zu Beginn der fremdsprachlichen Ausbildung, in der Muttersprache gehalten werden; doch wenn erkennbar ist, dass die Lernenden dem Geschehen auch in der Fremdsprache folgen können, so sollte sie wann immer möglich auch verwendet werden. Hier muss die Lehrkraft jeweils mit Feingefühl entscheiden, was für ihre Lernenden im gegebenen Moment am günstigsten ist.

1.3.3.5 Die europäische Dimension

Abschließend sei noch ein Blick auf die europäische Dimension im Fremdsprachenunterricht gestattet. Wenn die regionale und kulturelle Vielfalt Europas bewahrt und erhalten werden soll, so ist ein möglicher Weg der, die Vielfalt der Sprachen zu erhalten: „Sprachtod ist ein ökologisches Problem, denn mit der Sprache stirbt auch die Kultur, die in der Sprache ihren Ausdruck fand.“ (Schröder 1999: 3).

Der Fremdsprachenunterricht bietet sich als einer der Orte an, die die Bürger Europas zu einer „sprachenteiligen“ mehrsprachigen Gesellschaft führen könnten. Im Fremdsprachenunterricht könnten neben den traditionellen Sprachunterricht so genannte rezeptive Sprachkurse oder Sprachklubs treten, beispielsweise in den europäischen Nachbarsprachen, um ein grenzüberschreitendes Verstehen anzubahnen. Auch „Schnupperkurse“, in denen „neue“ Fremdsprachen ausprobiert werden könnten, beispielsweise im Rahmen von Projektwochen, bieten sich an. Bilingualer Unterricht, früher Fremdsprachenunterricht schon an der Grundschule, spät beginnende Fremdsprachenangebote mit dem Fokus auf rezeptive Fertigkeiten (ab der Oberstufe) oder so genannte Kompetenzkurse, in denen Dokumente im Original gelesen werden (beispielsweise die französische Revolution mit französischen Originaltexten erarbeiten) stehen als

weitere Möglichkeiten zur Verfügung. Mittels solcher Angebote könnte man die Forderung der EU nach rezeptiver Mehrsprachigkeit ausgehend von einem Dreisprachigkeitskonzept umsetzen. Ergänzend sei auch die Einführung von europäischen oder internationalen Sprachzertifikaten erwähnt als ist eine Möglichkeit, sprachliche und interkulturelle Kompetenzen international zu zertifizieren und damit vergleichbar zu machen.

Darüber hinaus sollte Englisch als *lingua franca* in „relativ kulturneutraler" Form (Schröder 1999: 3) ebenfalls seine Beachtung finden, denn es ist unmöglich zu einem geeinten Europa zu kommen, ohne auf eine gemeinsame Sprache wenigstens in den Alltagsbereichen rückgreifen zu können. Allerdings müssen die Grenzen dieses Modells für Europa erkannt werden: Eine dominante Sprache in Europa wird sich auch aufgrund der historischen Entwicklung nicht durchsetzen, da dies die Überlegenheit einer Sprache und damit Kultur über die anderen europäischen Sprachen und Kulturen implizieren würde und somit gegen die europäischen Maximen der Freiheit, Toleranz und Demokratie spräche. Auch wäre die Vorteilsstellung dieser Sprache nicht zu rechtfertigen.[80] Nichts desto trotz kann man de facto die Situation des Englischen als (weltweite) Verkehrssprache immer dann, wenn es keine gemeinsame andere Sprache gibt, nicht ignorieren, wenngleich man die Probleme im Bereich der minimalsprachlichen Kommunikation nicht übersehen darf: "However, minimalistic communication is the problem worldwide, not the solution." (Schröder 1993: 62). In diesem Zusammenhang könnten verkehrssprachliche Minimalanforderungen im Fremdsprachenunterricht gelehrt werden, um in „Notfällen" die Kommunikation aufrecht zu erhalten; ebenso könnte das Bewusstsein über die Problematik eines „kulturneutralen" Englisch geweckt werden, um die Lernenden auf reale Situationen vorzubereiten.

Daneben könnten Angebote aus dem Bereich der (außer)europäischen Migrantensprachen treten, um Zugang zu Kultur und Sprache dieser Mitbürger zu erhalten, das Zusammenleben zu erleichtern und dadurch Fenster zu öffnen hin zu gegenseitigem Verständnis und letztlich hin zu mehr Akzeptanz in den jeweiligen Gesellschaften. In diesem Zusammenhang geben Broadbent & Oriolo (1991) ein Beispiel für die Integration von Migranten, welches zwar den Umgang mit Migration auf nationaler britischer Ebene widerspiegelt, dennoch in gewissem Rahmen auf den Umgang mit Migranten in Europa übertragen werden kann: In den 80er Jahren gab es in Großbritannien ein EU-Pilotprojekt, in dem an Sekundarschulen Italienisch, Urdu und Punjabi gelehrt wurde. Dabei wurden *native speakers* dieser Sprachen als Lehrkräfte oder *assistant teachers* eingesetzt, um möglichst authentische Einblicke zu ermöglichen. Die Auswirkungen auf Migrantenkinder waren dahingehend positiv, dass sie in ihrer Herkunftssprache kommunizieren lernten, was nicht immer eine Selbstverständlichkeit ist. Auch Nicht-Migranten konnten davon profitieren, da dieser Unterricht Fenster öffnet auf die „nahe fremde" Kultur und Sprache. Dieses Projekt wurde auch auf den Sachfachunterricht übertragen – dort öffneten sich beispielsweise ganz neue Perspektiven im Geschichtsunterricht, als neben der „offiziellen" Geschichtsschreibung

[80] Vgl. dazu Schröder 1993.

auch die Sichtweisen der Migranten in deren Sprachen wahrgenommen werden konnten. Dieses Pilotprojekt könnte auch in anderen Ländern umgesetzt werden, natürlich angepasst auf die jeweiligen Bedürfnisse, um Mehrsprachigkeit nicht nur im Hinblick auf europäische Sprachen, sondern auch im Hinblick auf Migrantensprachen erfahrbar zu machen und die Vielfalt Europas auch und gerade aus den unterschiedlichen Perspektiven wahrzunehmen und zu akzeptieren. Denn ein geeintes Europa wird sich nicht durch „gemeinsame" Perspektiven kennzeichnen, sondern eher durch die Wahrnehmung und Akzeptanz der Vielfalt und der prinzipiellen Gleichwertigkeit der verschiedenen Perspektiven.

1.3.4 Begriffe des Lernens und Lehrens im GER

Nachdem in den vorangegangenen Abschnitten Grundbegriffe des Lernens und Lehrens von Sprache(n) und Kultur(en) erarbeitet wurden, sollen diese nun der Analyse des GER auf seine Lern- und Lehrbegriffe hin zugrunde gelegt werden. In den Unterkapiteln 1.3.4.1 mit 1.3.4.5 dieser Arbeit wird der GER auf folgende, oben in den Kapitel 1.3.1 mit 1.3.3 ausgeführte Konzepte des Lernens und Lehrens von Sprachen untersucht: *Erwerb und Lernen, Lernersprache, Lernprozesse und Lernprinzipien, Fremdsprache im Unterricht* und *Vermittlungskonzept*. Dies stellt aber keinesfalls eine abschließende Auflistung aller für Sprachenlernen und -lehren relevanten Aspekte dar – es handelt sich vielmehr um Schlüsselkonzepte im Rahmen der vorliegenden Arbeit, die deshalb als Analyserahmen dienen.

Um Aussagen zu angemessenen Beurteilungsverfahren treffen zu können, muss man neben der Klärung des Verständnisses von Sprache auch Prozesse und Konzepte des Lernens und Lehrens mit einbeziehen, um der Komplexität von Sprache, Sprachlernen und Sprachverwendung in der Testsituation gerecht zu werden; dies postuliert der GER auch auf S. 29:

> Es ist (...) klar, dass der Referenzrahmen sich nicht nur auf die Beschreibung von Kenntnissen, Fertigkeiten und Einstellungen beschränken kann, die Lernende erwerben müssen (...), sondern dass er sich auch mit den Prozessen des Spracherwerbs und des Sprachenlernens sowie mit Lehrmethoden befassen muss.

Doch vorher, gleich auf S. 8, stellen die Autoren des GER fest:

> Eines wollen wir aber von vornherein klarstellen: Wir wollen Praktikern NICHT sagen, was sie tun sollen oder wie sie etwas tun sollen. Wir stellen nur Fragen, wir geben keine Antworten. Es ist nicht die Aufgabe des *Gemeinsamen europäischen Referenzrahmens* festzulegen, welche Ziele die Benutzer anstreben oder welche Methoden sie dabei einsetzen sollen.

Es kann jedoch kein *Rahmen* gesteckt werden, wenn dabei nicht die Bandbreite didaktischer und spracherwerbstheoretischer Forschungspositionen und darauf aufbauender Empfehlungen beschrieben wird. Der GER bleibt bei Lernbegriff und Lehrkonzepten ebenso unverbindlich wie beim oben analysierten Sprachbegriff (GER 2001: 29):

> Die Rolle des Referenzrahmens in Bezug auf Aussagen über Spracherwerb und das Lernen und Lehren von Sprachen muss allerdings noch einmal klargestellt werden. (…) [Hier nun werden die Prinzipien einer pluralistischen Demokratie bemüht, Anm. d. V.] Deshalb kann er in der aktuellen Theorie-

> diskussion über das Verhältnis von Spracherwerb zu Sprachenlernen auch keine Position für die eine oder andere Seite beziehen. Er sollte auch keinen speziellen Ansatz zur Erklärung des Sprachenlernens darstellen, der andere Ansätze ausschließt. Die Rolle, die ihm zusteht, ist die, alle an Sprachlern- und -lehrprozessen als Partner Beteiligten zu ermutigen, ihre eigene theoretische Basis und ihre methodische Praxis so explizit und transparent wie möglich darzulegen.

Im GER wird zu wenig auf konkrete Lern- und Lehrprozesse und ihre Anforderungen eingegangen, es werden zu wenig Erkenntnisse verwendet, die in didaktischen Theorien oder Ansätzen verankert wären, es werden existierende Forschungsergebnisse nur marginal erwähnt und in keinem Fall belegt.[81] Dies ist kein fundiertes Vorgehen für ein Instrument solcher Tragweite. Da verwundert es nicht, dass Krumm (2003: 123) oder Vollmer (2003: 197) zu dem Schluss kommen, dass der GER kein lerntheoretisches Grundgerüst und keinen Lernbegriff vertrete.

Königs (2003: 114) bringt zwar den berechtigten Einwand, dass der GER nicht theoriegenerierend sein möchte und deshalb Befunde zum Lernen oder Lehren einer Sprache nicht darstellen, bewerten oder gar eigens erheben könne, sondern dass der GER „auf unterschiedlichen Abstraktionsebenen die Teilschritte dar[stellt], die zu einer – im europäischen Rahmen – vergleichbaren Planung, Zielbestimmung und Durchführung von Fremdsprachenunterricht entscheidend beitragen (können)." Doch auch ohne den Anspruch, zur (Weiter)Entwicklung einer umfassenden Fremdsprachentheorie beitragen zu wollen, wären ein Überblick über den Stand der Forschung und auf Theorien basierende Ableitungen hinsichtlich der Effektivität der GER-Vorschläge hilfreich für alle Nutzer des GER, um die eigene Position in Bezug auf den GER bestimmen und Praxiserfahrungen auf dem Hintergrund wissenschaftlicher Positionen reflektieren zu können.

Dazu kommt, dass die im GER, insbesondere in Abschnitt 6 *Fremdsprachenlernen und -lehren* verwendete Terminologie aus dem Bereich der Didaktik teils missverständlich, teils unscharf und teils auch falsch verwendet wird. Beispielsweise ist auf S. 150 des GER die Rede davon, mit welchen *Methoden* das Schriftsystem einer Sprache gelernt werden kann. Statt *Methoden* werden dann verschiedene *Übungsformen* aufgezählt. Es scheint auch, dass etwa die Begriffe *methodologische Optionen* und *Methoden* als austauschbar eingesetzt werden, wenn man sich die Aussage auf S. 140 des GER betrachtet: Der Unterabschnitt *methodologische Optionen* will „die *Methoden* für das Fremdsprachenlernen und –lehren umfassend beschreiben", doch wird dort nicht eine einzige *Methode* angesprochen, geschweige denn umfassend beschrieben – der GER verbleibt im unverbindlichen Aufzählen von unbeantworteten Fragen, die auch nicht in ein größeres methodisches Konzept eingebunden werden oder als *methodologische Optionen* betrachtet werden könnten. Denn bei Optionen würde man wenigsten erwarten, dass man begründete Entscheidungshilfen angeboten bekommt – dies ist aber nicht der Fall.

[81] Diese Aussagen werden im Verlauf dieses Kapitels konkretisiert und belegt. Hier sei nur vorwegnehmend auf Abschnitt 6.5 des GER verwiesen, in dem es zwei Auflistungen zu Einstellungen zu und Umgang mit Fehlern gibt; diese Auflistungen erwecken den Eindruck eines Maßnahmenkatalogs, bei dem alle Maßnahmen gleich effektiv wären. Intendiert war wohl eher, durch diese Liste die Benutzer des GER zum Nachdenken anzuregen – doch gerade im Bereich des Umgangs mit und der Bewertung von Fehlern gibt es Forschungsergebnisse, auf die man hätte zurückgreifen können, um diese irreführenden Auflistungen in ihrer Effektivität einzuschätzen und differenzierte Hinweise darauf zu geben, wann welche Arten von Fehlern wie zu bewerten sind.

Ebenfalls negativ fällt an GER-Abschnitt 6 auf, dass die Benutzerperspektiven vermischt werden: Abschnitt 6.1.4 *Unterschiedliche Lehr-/Lernziele in Bezug auf den Referenzrahmen* spricht auf S.134 des GER zunächst Curricula-Entwickler an, wohingegen er sich auf der nächsten Seite auch an Lernende und Lehrende wendet, ohne die unterschiedlichen Benutzer allerdings transparent zu benennen. Auf S. 139f des GER dann werden fünf Perspektiven (Prüfende, Behörden, Lehrwerkautoren, Lehrende und Lernende) beschrieben, die im Folgenden doch wieder vermengt werden: Beispielsweise spricht Abschnitt 6.4.6.4, S.146 die *Lehrenden* an, während sich der folgende Abschnitt zuerst mit der Perspektive der *Lernenden* beschäftigt, ehe dann zum unpersönlichen *man* gewechselt wird, das sich jedoch auf die *Lehrenden* bezieht. Auf der nächsten Seite (GER 2001: 147, unterer Kasten) werden „die Benutzer" angesprochen. Doch wer damit gemeint ist, Lehrende, Curriculumsplaner oder etwa Lerner, wird nicht verdeutlicht. Eine schärfere Trennung der Benutzer, an die sich bestimmte Abschnitte oder Aussagen des GER wenden, würde aber zur Transparenz des GER entscheidend beitragen.

Ein weiteres Manko ist die Strukturierung der didaktischen Aussagen: Wenn GER-Abschnitt 6 dem *Fremdsprachenlernen und -lehren* gewidmet ist, so wäre es wünschenswert, in diesem Abschnitt lern- und lehrbezogene Aussagen in didaktisch relevanten Kategorien zusammenzufassen und abzuhandeln. Stattdessen ist der Abschnitt selbst relativ unstrukturiert, wie gleich näher ausgeführt wird, und ist keinesfalls als umfassend zu betrachten; daneben finden sich weitere didaktische Aussagen über das gesamte Dokument verstreut und sind nur schwierig aufzufinden. Eventuell könnte ein Register bei der Auffindung bestimmter Konzepte und Begriffe helfen, doch kann dies terminologische oder strukturelle Inkonsistenzen nicht überwinden:

Die Unterabschnitte des GER-Abschnitts 6 zu den Bereichen der *Lernziele*, der *Prozesse des Sprachlernens*, zu *Möglichkeiten der Erleichterung des Sprachlernens*, zu *methodologischen Optionen* und zu *Fehlern* sind auf verschiedenen Ebenen anzusiedeln: Die Darstellung allgemeiner Aussagen zu Lernzielen auf derselben Ebene wie die der spezifischen Aussagen zum konkreten Umgang mit Fehlern leuchtet ebenso wenig ein wie die Beschreibung allgemeiner Prozesse des Sprachenlernens auf derselben Ebene wie die der konkreten Frage danach, wie die Benutzer des GER das Sprachenlernen erleichtern können. Innerhalb dieser Unterabschnitte zeigt sich, dass die genannten Bereiche wiederum relativ unstrukturiert abgehandelt werden: Unter der Überschrift *Lernziele* beispielsweise findet man zunächst allgemeine und relevante Überlegungen zu Zielbestimmung und den dabei zu bedenkenden Aspekten, ehe unvermittelt Aussagen zu Lernfortschritten getroffen werden; daraufhin werden konkrete Vermittlungsschwierigkeiten im Falle unterschiedlicher Begriffsfelder in L1 und L2 sowie konkrete Fragen bezüglich der Vermittlung und der Problemfelder bei der Aussprache dargestellt. Noch verwirrender wird es im darauf folgenden Abschnitt, der sich (auf derselben Gliederungsebene wie die Aussagen zu *Lernfortschritten*) mit *mehrsprachlicher und plurikultureller Kompetenz* (immer noch im Unterabschnitt zu *Lernzielen*) beschäftigt; innerhalb dieses Abschnittes jedoch sind Aussagen zu Merkmalen dieser Kompetenz zu finden (vgl. oben Kapitel 1.2.5.4 *Mehrsprachigkeit*

im GER), ebenso wie Aussagen zur Struktur und Entwicklung dieser Kompetenzen, doch wird relativ wenig über damit verbundene Lernziele ausgesagt. Lediglich am Schluss der Ausführungen auf wird S. 134 des GER die m. E. triviale Aussage getroffen, dass Lernenden geholfen werden soll, „ihre sprachliche und kulturelle Identität zu gestalten (…) und ihre Lernfähigkeit durch (…) vielfältige Erfahrung des Kontakts mit mehreren Sprachen und Kulturen zu verbessern."

An dieser Stelle sei nur noch auf ein weiteres Beispiel für die Unstrukturiertheit des GER verwiesen, denn die Aussagen des Abschnitts 6 werden im Folgenden genauer analysiert. Wenn man nach einer Aussage zum zentralen Prinzip der *Anordnung* von Vermittlungsinhalten sucht, so wird man an einer Stelle fündig, die nicht dafür spricht, dass die Autoren des GER diesen Punkt als wesentlich betrachtet hätten: In GER-Unterabschnitt 6.4 *Methodologische Optionen*, darin unter 6.4.7 *Linguistische Kompetenzen*, darin wiederum unter Punkt 6.4.7.4 finden sich tatsächlich Aussagen zur Anordnung nach dem „inhärenten Komplexitätsprinzip" (GER 2001: 148), die unter Punkt 6.4.7.5 erweitert werden um die Aspekte des kommunikativen Nutzens, der kontrastiven Faktoren, des Schwierigkeitsgrades von Texten und der natürlichen Erwerbssequenzen. Abschließens findet sich zur Anordnung folgende Aussage (ebd.):

> Der *Referenzrahmen* ersetzt keine Grammatikbücher und bietet keine strenge Reihenfolge an (obwohl das Skalieren eine Auswahl und somit einige globale Sequenzierungen beinhalten kann); er stellt jedoch einen Rahmen für die Entscheidungen der Praktiker dar, die sie anderen mitteilen wollen.

Auf die Implikation, dass die Skalen als Sequenzierungshilfe genutzt werden könnten, wird in dieser Arbeit in Kapitel 3 eingegangen – an dieser Stelle interessiert eher, wie solch ein Rahmen, der wenig Struktur zeigt und sich so unverbindlich gibt, denn bei der Positionsbestimmung helfen soll.

Will der GER einen Rahmen bieten, in dem sich alle am Sprachlern- und Lehrprozess Beteiligten in Europa wiederfinden können, so muss er Position beziehen und kann nicht im unverbindlichen Auflisten von unbewerteten Fragen oder Aussagen verbleiben. Das oben bei der Analyse der Sprachbegriffe im GER schon angedeutete prinzipielle Problem dieses Dokuments zeigt sich auch bei der Analyse des Lehr- und Lernbegriffs im GER: Es ist das Verharren in absoluter Unverbindlichkeit bei allen umstrittenen Fragen. Dies kann für ein Werkzeug zur Umsetzung europäischer Sprachenpolitik kein angemessenes Vorgehen sein. Zu einem bestimmten Aspekt oder einer Fragestellung, wie beispielsweise effektiver methodischer Ansätze im Fremdsprachenunterricht, müssten bedeutsame Eckpunkte und auch relevante Forschungsergebnisse dargestellt werden, um auf dieser Basis zu begründeten Empfehlungen zu kommen. Hinzu kommt die oben bereits kritisierte verwirrende und inkonsistente Terminologieverwendung nicht nur in GER-Abschnitt 6. Innerhalb des vom GER intendierten Rahmens muss eine kohärente Terminologie mit klaren Definitionen entwickelt werden, selbst wenn diese Definitionen nicht von allen Benutzern des GER, seien es Forscher oder Praktiker, geteilt werden – sie hätten dennoch ihre Gültigkeit im Rahmen dieses Instruments. Man kann sich nur dann *verorten*, wenn der Ort auch „lokalisierbar"

ist. Im Folgenden werden die negativen Folgen dieser Unverbindlichkeit und fehlenden Begriffsbestimmung am Beispiel der Begriffe des *Erwerbs* respektive *Lernens* aufgezeigt.

1.3.4.1 Erwerb und Lernen

Der lerntheoretische Ansatz des GER ist ein handlungsorientierter: „Der hier gewählte Ansatz ist im Großen und Ganzen *handlungsorientiert*, weil er Sprachverwendende und Sprachlernende vor allem als *sozial Handelnde* betrachtet …“ (GER 2001: 21). Er kann insofern als grobkörnig beschrieben werden, als dass keine Unterscheidung zwischen Lernen, Erwerb und Verwendung getroffen wird. Dies wird auf S. 21 des GER deutlich (Vgl. Zitat und Ausführungen oben unter Kapitel 1.2.5 dieser Arbeit), wie auch auf S.51 des GER: „Nach dem hier zugrunde liegenden handlungsorientierten Ansatz sind Sprachlernende angehende Sprachverwendende, so dass auf beide die gleichen Kategorien zutreffen.“

Es wird also einerseits nicht grundsätzlich differenziert zwischen Sprachlernen und Sprachverwendung, obwohl diese, wie oben beim Sprachbegriff erläutert, unterschiedlichen Bedingungen unterworfen sind, doch dazu unten mehr bei der Analyse des Konzepts der Lernersprache und des didaktischen Sprachbegriffs im GER. Andererseits wird auch nicht zwischen Erwerb und Lernen unterschieden, was mit der nicht standardisierten Terminologie und einem fehlenden Oberbegriff begründet wird (GER 2001: 137f).[82] Da der GER ein output-orientiertes Instrument ist, Erwerb und Lernen jedoch auch den Input betreffen, interessiert in diesem Zusammenhang, ob die fehlende Unterscheidung Auswirkungen auf die Reichweite und Verwendbarkeit des GER hat.

Inwiefern führen Erwerb und Lernen zu unterschiedlichen Kompetenzen? Wie oben unter Kapitel 1.3.1.1 dieser Arbeit erläutert, sind Erwerbs- und Lernprozesse nicht als Gegensätze zu verstehen, sondern vielmehr als komplementäre Enden eines Spektrums von Aneignungsprozessen. Erwerb, wie oben gesagt, betont die „natürlichen“ Prozesse, Lernen die gesteuerten. Insofern führt Erwerb eher zu Flüssigkeit und „natürlichem“, unbewusstem Sprachgebrauch, auf Basis der Erfahrungen im Umgang mit der Zielkultur. Dabei darf davon ausgegangen werden, dass affektive Erfahrungen und Kommunikationsstrategien die Folie der weiteren Aneignungsprozesse bilden. Lernen hingegen führt eher zu (formaler) Bewusstheit und Korrektheit, und zu Erfahrungen in einem eher geschützten Kommunikationsbereich, weshalb der Aspekt der Korrektheit im kommunikativen Ansatz nicht ohne Grund um Aspekte der Flüssigkeit und des „natürlichen“ Sprachgebrauchs erweitert wird.

Gerade in Bezug auf die Mehrsprachigkeit und den europäischen Normalzustand scheint es nicht angebracht, die so genannten natürlichen Erwerbsprozesse einer Zweitsprache überhaupt nicht zu thematisieren. Der (institutionalisierte) Weg zur mehrsprachigen Gesellschaft beinhaltet

[82] GER-Abschnitt 6.2.1 ist ein gutes Beispiel für die übertriebene Unverbindlichkeit des GER: Die Abgrenzung der Begrifflichkeiten *Erwerb* und *Lernen* ist so tautologisch, dass sie wenig bei Reflexion und Selbstverortung helfen kann.

auch Facetten des ungesteuerten Zweitspracherwerbs und dessen besonderer Bedingungen, doch im GER werden diese nirgends explizit thematisiert. Man könnte einwenden, dass der GER das institutionelle Fremdsprachenlernen fördern will und auf dieses ausgerichtet ist, doch fließen gerade bei (Arbeits-) Migranten in Zielsprachkursen deren ungesteuerte Erwerbserfahrungen mit in das gesteuerte Lernen ein, wie es beispielsweise – wie ebenfalls oben thematisiert – in der Didaktik des Deutschen als Zweitsprache anerkannt wird. Diesen Bedingungen kann der GER in seiner gegenwärtigen Form nicht gerecht werden. Um diese Kritik genauer zu belegen, sei an dieser Stelle ein Exkurs gestattet: Im Folgenden werden die GER-Darstellung des Umgangs mit Krisensituationen, auch die affektiven Aspekte betreffend, und der Vermittlung zwischen Kommunikationspartnern im interkulturellen Kontext beleuchtet, da es sich dabei um zwei zentrale Aspekte des Zweitspracherwerbs und der Mehrsprachigkeit handelt.

Potentiell kritische Situationen in der realen Kommunikation und Möglichkeiten zu deren Entschärfung finden im GER keine Erwähnung, wie oben in den Kapiteln 1.2.5.3 und 1.2.5.4 dieser Arbeit bereits erwähnt. Selbst wenn man sich im GER auf Situationen im Unterrichtskontext beschränkt hat, so müsste wenigstens der Einsatz von Kommunikationsstrategien angesprochen werden, seien es nun kommunikationsfördernde wie etwa Nachfragen oder Paraphrasierung, oder kommunikationsbehindernde wie zum Beispiel Themenwechsel; in diesem Zusammenhang hätte auch thematisiert werden müssen, auf welche Weise bestimmte Strategien im Unterricht gefördert werden können. Im GER finden sich Aussagen zu Strategien unter Abschnitt 4.4 *Kommunikative Aktivitäten und Strategien*, wobei sich Strategien dort nicht auf die erwähnten Strategien des Spracherwerbs generell beziehen, sondern lediglich auf „die Auswahl einer möglichst effektiven Handlungsweise“ (GER 2001: 63) – wiederum eine idealisierte Betrachtung kommunikativer Strategien. Konsequenterweise finden sich in den GER-Skalen auch keine Deskriptoren zum Einsatz von solchen Strategien, die eine kritische Situation entschärfen oder auflösen könnten. Wohl aber findet man auf S. 70f des GER Skalen zu *Planen*, *Kompensieren*, und zu *Kontrolle und Reparaturen*. Bei näherer Betrachtung allerdings bleiben die Deskriptoren eng an sprachliche Strategien angelehnt; kritische Kommunikationssituationen sind höchstens an Wendungen wie „Kann eigene Fehler korrigieren, wenn sie zu Missverständnissen geführt haben“ (GER 2001: 70) auszumachen. Wie aber solch ein Missverständnis erkannt wird oder welche pragmatischen und/oder nonverbalen Strategien in solch einer Situation noch zum Tragen kommen, darüber geben weder der GER-Text noch die GER-Skalen Aufschluss. Unter GER-Abschnitt 4.4.3.5 *Interaktionsstrategien* finden sich noch zwei Skalen, *Kooperieren* und *Um Klärung bitten*, die ebenfalls stark sprachlich orientiert sind und sich auf Gespräche beziehen, ohne jeden Hinweis auf interkulturelle Kontexte. Auf S. 155 des GER finden sich wiederum nur idealisierte Aussagen dazu, dass Sprachverwendende/Lernende die jeweils effektivsten Strategien einsetzen würden, um das kommunikative Ziel zu erreichen – sicherlich nicht der Realzustand in der (interkulturellen) Kommunikation.

Ebenfalls unter den Themenbereich der (interkulturellen) Kommunikation fällt der Umgang mit affektiven Aspekten: Welche Affekte kommen in realen (mehrsprachigen) Kommunikationssituationen zum Tragen? Welche Kompetenzen müssen herausgebildet werden, um in interkulturellen Situationen mit den eigenen Wertvorstellungen und Emotionen umgehen zu können? Wie oben in Kapitel 1.3.3 dieser Arbeit dargestellt, muss die affektive Komponente im Unterricht aufgefangen werden, sei es durch reflexive oder handlungsorientierte Phasen, in denen der Umgang mit eigenen und fremden Emotionen „geübt" werden kann, um in realen Situationen handlungsfähig zu bleiben – ein Ziel, das der GER sich ja, wie erwähnt, auch gesteckt hat. Allerdings finden sich dazu im GER-Dokument wenig hilfreiche Aussagen. Beispielsweise wird auf S. 19 des GER erwähnt, „…dass die Entwicklung kommunikativer Kompetenz auch andere Dimensionen umfasst als nur sprachliche im engeren Sinne (z. B. soziokulturelles Bewusstsein, Erfahrungen im Bereich der Imagination, affektive Beziehungen, das Lernen zu lernen usw.)", doch wird die affektive Domäne nicht konkret in Skalen umgesetzt. Man findet lediglich etwa auf S. 61 des GER bei Aussagen zur Aufgabenbeschreibung den Hinweis, dass man Informationen zu Aufgaben geben könne, wobei sprachliche Aktivitäten differenziert werden könnten hinsichtlich der Aspekte „kognitiv/affektiv". Hilfreicher sind die Ausführungen zur Rolle kommunikativer Aufgaben in GER-Abschnitt 7.3.1.2 *Affektive Faktoren*, die sich mit den Aspekten *Selbstwertgefühl*, *Engagement und Motivation*, *Befindlichkeit* sowie *Einstellung* beschäftigen. Doch auch hier fehlt die interkulturelle Dimension, denn die Aussagen beschränken sich auf die Auswirkungen dieser Faktoren bei der Ausführung kommunikativer Aufgaben.

Der Themenkomplex des Mittelns zwischen Sprachen und Kulturen wird im GER ebenfalls nur unzureichend thematisiert: Es wird beispielsweise nichts darüber ausgesagt, wie man „…die Bereitschaft, als 'kulturelle Mittler' zwischen der eigenen und der fremden Kultur zu fungieren und interkulturelle Missverständnisse und Konflikte zu lösen" (GER 2001: 157) im Unterricht fördern könnte noch wie man sie mittels einer Skala abbilden könnte. Unter GER-Abschnitt 4.4.4 *Aktivitäten und Strategien der Sprachmittlung (Übersetzen, Dolmetschen)* finden sich wiederum eher auf sprachliche Kompetenzen bezogene Ausführungen – die Perspektive des interkulturellen Vermittelns zwischen verschiedenen Gesprächspartnern fehlt. Auf S. 101 des GER findet sich eine schematische Darstellung zu den Prozessen des Übersetzens und Dolmetschens, die aber den größeren kulturellen Kontext ebenfalls nicht thematisiert. Lediglich folgende interkulturellen Fertigkeiten sind auf S. 106 des GER dargestellt:

> 5.1.2.2 – *Interkulturelle* Fertigkeiten umfassen:
> - die Fähigkeit, die Ausgangskultur und die fremde Kultur miteinander in Beziehung zu setzen;
> - kulturelle Sensibilität und die Fähigkeit, eine Reihe verschiedener Strategien für den Kontakt mit Angehörigen anderer Kulturen zu identifizieren und zu verwenden;
> - die Fähigkeit, als kultureller Mittler zwischen der eigenen und der fremden Kultur zu agieren und wirksam mit interkulturellen Missverständnissen und Konfliktsituationen umzugehen;
> - die Fähigkeit, stereotype Beziehungen zu überwinden.

Auf S. 120f des GER, unter 5.2.2.5 *Varietäten (sozial, regional, ethnisch* usw.) folgt eine Skala zu soziolinguistischer Angemessenheit, die wenigstens im obersten Bereich das Mitteln zwischen den Kulturen thematisiert (Auszug aus den Deskriptoren zu Niveau C2):

> Kann die soziolinguistischen und soziokulturellen Implikationen der sprachlichen Äußerungen von Muttersprachlern richtig einschätzen und entsprechend darauf reagieren.
> Kann als kompetenter Mittler zwischen Sprechern der Zielsprache und Sprechern aus seiner eigenen Sprachgemeinschaft wirken und dabei soziokulturelle und soziolinguistische Unterschiede berücksichtigen.

Wie sich jedoch Erwerb respektive Lernen auf die Herausbildung dieser Kompetenzen im interkulturellen Bereich auswirken, insbesondere wie diese Fertigkeiten durch Zweitsprachenkontakt beeinflusst werden, wie es in Europa gerade in mehrsprachigen Gesellschaften oder in Migrationssituationen der Fall ist, wird nicht näher ausgeführt. Man findet lediglich auf S. 134 des GER die Aussage, dass die Erfahrung von Mehrsprachigkeit und Plurikulturalismus „(…) bis zu einem gewissen Grad das spätere Lernen im sprachlichen und kulturellen Bereich beschleunigen“ kann – wie das jedoch geschehen kann, dazu wird nichts ausgesagt.

Auch wenn obige Beispiele schon tief in den unten anzusprechenden Bereich des Vermittelns von kommunikativen und akkulturativen Kompetenzen hineinragen, so sollen sie an dieser Stelle verdeutlichen, welche weitreichenden Folgen die fehlende Differenzierung des Lernbegriffs im GER haben kann. Vielleicht ist ein Grund für die Grobkörnigkeit des GER in Bezug auf den Lernbegriff im unter Kapitel 1.2.5.3 dieser Arbeit beschriebenen idealisierten Kommunikationsbegriff zu finden: Denn Sprachverwendung an sich wird sehr detailliert in ihrer idealisierten Form beschrieben, ohne allerdings zu differenzieren zwischen alltäglicher Kommunikation in der Muttersprache, alltäglicher Kommunikation in einer Zielsprache, die nicht Muttersprache ist, und didaktisch ausgerichteter Kommunikation in der Fremdsprache in Unterrichtssituationen. Ist diese Unterscheidung im GER aber überhaupt intendiert? Im Vorwort des GER (2001: 3) wird zunächst klargestellt, dass dieses Instrument „…jenes Wissen und jene Fertigkeiten [erfasst], mit denen Sprachlernende im öffentlichen, beruflichen und privaten Bereich sprachlich handlungsfähig werden.“ Diese Aussage ließe die Interpretation zu, dass sowohl der unterrichtliche Bereich als auch der öffentliche Bereich (sei es nun im Kontext der Mutter-, Fremd- oder Zweitsprache) Gegenstand des GER sein müssten. Die Aussagen auf S. 8ff des GER lassen allerdings darauf schließen, dass der GER im institutionellen Bereich des Lehren und Lernen angesiedelt ist und sich nur auf die Besonderheiten in diesem Bereich bezieht, denn er wendet sich konkret an „Lernende“ oder an „Angehörige einer der Berufsgruppen, die mit Sprachunterricht oder mit Beurteilen und Prüfen“ befasst sind. Insofern muss der GER die o. g. außerunterrichtlichen Bereiche nur insoweit mit einschließen, als dass diese die Ziele eines jeden Fremdsprachenunterrichts mitbestimmen. Doch selbst wenn der GER seinen Kommunikationsbegriff bewusst nicht in muttersprachliche, zweitsprachliche und unterrichtliche Kommunikation differenziert hat, so wirken auch im Bereich des institutionellen (unterrichtlichen) Lehrens und Lernens,

wie oben erläutert, Effekte des ungesteuerten Erwerbs ein, die nicht einfach ignoriert werden können.

1.3.4.2 Die Lernersprache

In diesem Bereich tritt ein weiteres, oben schon angedeutetes Problem ins Blickfeld: Wenn im GER nicht unterschieden wird zwischen Sprachverwendenden und Sprachlernenden, so dürfte dies auch Auswirkungen auf die Beschreibung dessen haben, was die beiden (m. E. nach unterscheidbaren) „Gruppen“ an Output produzieren: Bei Sprachverwendenden dürfte sich die Aufmerksamkeit auf die Beschreibung der Performanz als solche richten, bei Sprachlernenden dürfte auch der Lernfortschritt ein nicht unwesentlicher Beschreibungsgegenstand sein – an dieser Stelle soll nur auf dieses Problem verwiesen werden. Es wird entsprechend bei den nun folgenden Ausführungen zur Lernersprache und bei den Ausführungen zum Status der Skalen des GER in Kapitel 3 dieser Arbeit wieder aufgenommen.

Die Lernersprache im Sinne eines sich entwickelnden Interimsprachsystems scheint nicht Gegenstand des GER zu sein. Da Sprachverwendende und Sprachlernende, wie schon gesagt, gleich gesetzt werden, unterscheidet der GER auch konsequenterweise in seinen Beschreibungen der sprachlichen und außersprachlichen Kategorien in den GER-Abschnitten 4 und 5 nicht zwischen Muttersprache und Interimssprache. Lediglich findet sich der „wichtige Hinweis“ auf S.51 des GER: „Weder endet die muttersprachliche und kulturelle Kompetenz mit dem Erwerb einer zweiten oder fremden Sprache und Kultur, noch besteht die neue Kompetenz unabhängig von der alten.“ Es folgt aber keine Elaborierung dieses Hinweises, um das Konzept Lernersprache gegenüber der Muttersprache zu charakterisieren.

Die detailliert in den GER-Abschnitten 4 und 5 beschriebenen Kategorien der allgemeinen wie kommunikativen Wissensbestände fließen mit ein in die implizite Darstellung dessen, was die Lernersprache ausmacht, ebenso wie die in GER-Abschnitt 4 thematisierten Strategien der Kommunikation und des Lernens und die in Abschnitt 5 beschriebenen sprachlichen Kompetenzen – doch findet man nur eine Stelle im GER, wo eine marginale Aussage zur Lernersprache selbst getroffen wird – auf S.151, bei der Thematisierung von Fehlern: „Kompetenzfehler sind eine Erscheinung von ‚Lernersprachen’, d. h. von vereinfachten oder verzerrten Varianten der Zielsprache.“ Dies ist zu verkürzt dargestellt, da die Charakteristika der Interimssprache und deren Entwicklung eigentlich in den Mittelpunkt eines Dokuments zu Sprachlernen und -lehren gerückt gehören. Wenigstens sollte das Konzept der Interimsprache und sein theoretischer Hintergrund erläutert werden. Dazu sollten auch die oben unter Kapitel 1.3.1.2 dieser Arbeit beschriebenen Charakteristika des lernersprachlichen Systems genutzt werden: Die Variabilität und Veränderlichkeit der Lernersprache wird nur in Bezug auf die oben unter Kapitel 1.2.5.4 dieser Arbeit analysierten mehrsprachigen und plurikulturellen Kompetenzen anerkannt

(GER 2001: 134), es fehlt jedoch die Einbettung dieser Kompetenzen in ein größeres Bild der Lernersprache. Die Bedeutung von Fehlern als Kennzeichen der Interimsprache wird, wie gerade gesagt, zwar erkannt, doch sollten Fehler eher unter dem Oberbegriff der Interimsprachen eingeordnet und der effektive Umgang mit ihnen innerhalb eines größeren Sprachvermittlungskonzepts bewertet werden; stattdessen gibt es auf S. 151f des GER wie oben erwähnt eine unbewertete Auflistung von (didaktisch teils höchst fragwürdigen) Möglichkeiten des Umgangs mit Fehlern – wie diese Liste den Benutzern des GER helfen soll, ihre Einstellung zu Fehlern zu reflektieren, bleibt unklar.[83] Es sei noch ein weiteres Charakteristikum der Lernersprache genannt, der Einsatz von kommunikativen Strategien: Diese werden im GER zwar thematisiert, doch wie oben unter Kapitel 1.2.5.3 dieser Arbeit erläutert als Teil eines idealisierten Kommunikationsbegriffs: Die ineffektiven kommunikativen Strategien, die sowohl Kennzeichen des Spracherwerbs als auch des Fremdsprachenlernens sind, wie etwa (Themen- oder Gesprächs-) Abbrüche oder negativer Transfer, werden nicht angesprochen.

Vielmehr finden sich Skalen zu kommunikativen Strategien (GER 2001: 70, 78, 88, 89), als ob diese kennzeichnend und typisch für verschiedene *Niveaus* des Sprachgebrauchs und (aufgrund der Gleichsetzung von Sprachverwendung und Sprachlernen) des Sprachlernprozesses wären und bestimmten Kompetenzniveaus zugewiesen werden könnten. In der wissenschaftlichen Literatur ist keine Theorie zu finden, die solche Annahmen stützte – überhaupt sind Sprachlernprozess und Interimsprachenentwicklung so variabel, vielfältig und individuell geprägt, dass es nicht möglich sein dürfte, ihre relevanten und typischen Aspekte abgestuft zu skalieren. Die Skalen des GER, insbesondere die gerade erwähnten, könnten den Eindruck erwecken, dass der Lernprozess linear in den dort beschriebenen Abstufungen abliefe. Folgender Hinweis auf S. 9 des GER scheint deshalb gefährlich: „Die in Kapitel 3 vorgestellten Gemeinsamen Referenzniveaus bieten ein Mittel an, Fortschritte der Lernenden beim Aufbau ihrer Sprachkompetenz in den Kategorien des Beschreibungssystems abzubilden.“ Ist es denn tatsächlich der Fall, dass die Skalen die Entwicklung der Lernersprache beschreiben? Denn der GER stellt sie auf S. 46 als „Kompetenzskalen“ dar, also Skalen, die die angenommene Kompetenz in konstruierten Abstufungen darstellen, und eben nicht (empirisch abgesicherte) Lernfortschritte und daran beteiligte Strategien beschreiben. Skalen sollten aber nach Alderson (1991b) nicht für andere als die intendierten Funktionen eingesetzt werden – in diesem Fall könnte es zu Problemen kommen, da man unzulässigerweise von konstruierten Abstufungen, die eine Beurteilung erleichtern sollen, auf Lernfortschritte oder die Entwicklung der Lernersprache rückschließt und somit zu falschen Folgerungen wie beispielsweise der des stufenweisen, linearen Lernens kommen könnte. Hier wäre mehr Transparenz bei der Offenlegung der intendierten Funktionen der Skalen wünschenswert, doch mehr dazu in Kapitel 3 dieser Arbeit.

[83] Warum in einem Dokument, das sich dem Positivansatz in der Beurteilung verschrieben hat, wie in Kapitel 3 dieser Arbeit erläutert wird, dann aber keine Stellung genommen wird zu einer in diesem Ansatz notwendigen Einstellungsveränderung gegenüber Fehlern, ist nicht nachvollziehbar. Wenn das Prinzip der Positivbewertung und des *Rating*-Verfahrens in der Praxis angewandt werden soll, so muss den Benutzern des GER Hilfe angeboten werden, Fehler in der Beurteilung neu zu bewerten. Näheres dazu in dieser Arbeit unter Kapitel 3.3 *Rating*-Verfahren.

Sprache wird im GER lediglich, wie oben unter Kapitel 1.2.5.3 dieser Arbeit erläutert, als idealisiertes Kommunikationsmittel thematisiert, ohne dass beispielsweise die Rolle des Input, des Output und der Interaktion bei der Herausbildung der Lernersprache diskutiert würde – das Manko liegt zum einen im zu impräzisen Sprachbegriff, der sich auch durch Lern- und Lehransatz zieht und zu weiteren Impräzisionen führt; zum anderen dürfte es auch in der Output-Orientierung des GER liegen. Doch ehe man sich mit lernersprachlichem Output beschäftigt, muss man sich zuvor mit den Grundlagen des Outputs, namentlich den Lernprozessen und dem Input, auseinandersetzen.

1.3.4.3 Lernprozesse und Lernprinzipien

Da dem lernersprachlichem Output Lernprozesse und Input vorgeschaltet sind, soll zunächst auf das Verständnis des GER bezüglich dieser Lernprozesse eingegangen werden, ehe im Anschluss das Verständnis des GER bezüglich einer wichtigen Input-Quelle, der des *classroom discourse*, näher betrachtet wird.

Im GER finden sich immer wieder verstreut Aussagen zu den Prozessen des Lernens, wenngleich diese oft nur implizite Schlüsse auf den zugrunde liegenden Lernbegriff zulassen. Auf S.23 des GER beispielsweise wird festgestellt, dass neues Wissen „nicht einfach zum vorhandenen addiert“ wird, sondern „abhängig ist von der Beschaffenheit, dem Reichtum und der Struktur des bereits vorhandenen Wissens“. Auch wird die Strukturierung des neuen Wissens „zumindest teilweise“ auf schon vorhandene Strukturen zurückgeführt. Man könnte nun schlussfolgern, dass der GER von einem kontrastiven Lernbegriff ausgeht, doch dies würde zu kurz greifen, denn der GER hat hierzu noch mehr zu sagen.

Auf S. 24 des GER wird die Lernfähigkeit expliziert und deren Bedeutung im Lernprozess. Sie wird unter GER-Abschnitt 5.1.4, S.108f im Detail wieder aufgegriffen: Dort wird auf die Bedeutung des Sprach- und Kommunikationsbewusstseins, auf Lerntechniken und auf heuristische Fertigkeiten, wie beispielsweise dem Umgang mit Neuem, eingegangen. Diese Thematisierung ist zu begrüßen, da sie auch im Fremdsprachenunterricht Gegenstand sein sollte. Den interaktiven Aktivitäten und Strategien kommt im GER große Bedeutung zu (vgl. GER Abschnitt 4.4.3, S. 78) – ein weiterer Hinweis auf den impliziten Lernbegriff: Der authentischen Sprachverwendung, der Interaktion und der Bewusstheit über Sprache und kommunikative Abläufe wird Rechnung getragen, auch in einem eigenen Abschnitt 6.1.3.3 auf S. 133f.

Die Individualität der Lernprozesse wird im GER ebenfalls anerkannt: Auf S. 23 und in Abschnitt 5.1.3, S. 109 des GER werden die unterschiedlichen persönlichkeitsbezogenen Kompetenzen dargestellt und beschrieben, die neben anderen Faktoren die Lernprozesse individualisieren und denen im Fremdsprachenunterricht besondere Beachtung geschenkt werden muss, um die Lerner (gerade im schulischen Bereich) bei der Ausbildung einer mehrsprachigen und

damit interkulturellen „Persönlichkeit“ zu unterstützen. Auch wenn hierzu keine Skalen entwickelt werden konnten, so sind doch die wesentlichen Charakteristika im GER erfasst. Auf S. 28 des GER wird auf die Individualität allen Sprachenlernens im Kontext der Bedeutung der Niveaubeschreibungen eingegangen, um klarzustellen, dass Kompetenzbeschreibungen aufgrund dieser Individualität „bis zu einem gewissen Grad willkürlich sind“ und nicht auf eine allen Lernern gemeinsame Kompetenz verweisen.[84] Zur Variabilität des Lernens findet sich auf S.139 des GER die Aussage: „Es ist erforderlich, dass Lehrende die Vielfältigkeit der Lernprozesse verstehen.“

Die aktive und bewusste Teilnahme am Lernprozess wird auf S. 139f des GER thematisiert Die auf S. 109 des GER beschriebenen Lerntechniken lassen darauf schließen, dass der Lernerautonomie gewisse Bedeutung zukommt: Hinweise auf „die Fähigkeit, vorhandene Materialien für selbständiges Lernen zu nutzen“ oder auf „die Fähigkeit, die eigenen Bedürfnisse und Ziele zu identifizieren“ sind wichtig und hilfreich. Die Hinführung der Lernenden von einer meist reaktiven Haltung hin zu aktiver Selbstorganisation des Lernens kommt auf S. 140 des GER zur Sprache – auch im Hinblick auf autonomes Lernen, nachdem der Fremdsprachenunterricht geendet hat. Diese Ausdehnung des unterrichtlichen Sprachlernbegriffs auf ein lebenslanges Lernkonzept ist zu begrüßen.

Unter der Überschrift „Wie lernen Lernende“ (GER-Abschnitt 6.2.2, S. 138) findet sich zwar eine Kurzbeschreibung dieses Forschungsspektrums, doch (wieder einmal) ohne Verweise auf (oder Belege aus) Forschung und Literatur, so dass sich dieser Abschnitt eher als eine Reduktion auf Tatsachenbehauptungen herausstellt. In diesem Abschnitt des GER werden die gerade analysierten Aussagen zu Lernprozessen und -Prinzipien, die sich über den GER verstreut finden lassen, kurz zusammengestellt, doch man erhält auf dieser einen Seite (!) keine ausreichende Basis, auf der man seinen Lernbegriff reflektieren könnte.[85] So schließt denn auch dieser Abschnitt mit folgendem nichtssagenden Kasten (GER 2001: 139):

> Die Benutzer des *Referenzrahmens* sollten bedenken und, soweit sinnvoll, angeben,
> - mit welchen Annahmen über das Sprachenlernen sie arbeiten und welche methodischen Konsequenzen dies hat.

Zusammenfassend kann man feststellen, dass Im GER hinsichtlich der Lernprozesse durchaus positive Akzente gesetzt werden und dass ihm ein moderner Lernansatz zugrunde liegt: Dieser beinhaltet die Handlungsorientierung, die Bedeutung authentischer Sprachverwendung und Interaktion, und die Bedeutung von Sprachbewusstheit; er wird der Individualität und Vielfalt der

[84] Schon deswegen können die Skalen des GER keine Hinweise auf skalierbare Lernfortschritte geben, doch Näheres dazu unter Kapitel 3 dieser Arbeit.

[85] Beispielsweise wird dort (GER 2001: 138) die „Ansicht einiger Theoretiker“ dargestellt, die davon ausgehen, dass genügend Input den Spracherwerbsprozess in Gang setze und dieser nicht „durch bewusste Manipulation gefördert werden kann.“ Auf welche Theoretiker man sich hier bezieht, bleibt jedoch im Dunkeln; auch schlägt sich hier der nicht genügend differenzierte Lern-/Erwerbsbegriff negativ nieder, da es sich bei den hier genannten „Theoretikern“ höchstwahrscheinlich um Sprach*erwerbs*forscher handeln dürfte, die sich unter Umständen gar nicht mit den Prozessen des Sprach*lernens* beschäftigt haben. Im nächsten Abschnitt (ebd.) werden lediglich „Andere“ zitiert, die „der Ansicht“ seien, dass zu verständlichem Input noch „aktive Beteiligung an kommunikativer Interaktion“ treten müsse – eine Ansicht, die nachvollziehbar ist, doch um wen handelt es sich nun bei den „Anderen“?

Lernprozesse gerecht und anerkennt die Notwendigkeit von Lernerautonomie und entsprechender Strategien. Dennoch greifen die Aussagen zu kurz und werden nicht in ihre theoretischen und historischen Kontexte gebettet, so dass der GER seinen Benutzern auch bei der Reflexion über ihren Lernbegriff keine rechte Hilfe bietet.

1.3.4.4 Fremdsprache im Unterricht

Welche Bedeutung lassen die Autoren des GER der Fremdsprache im Unterricht zukommen? Aus welchen Perspektiven wird Fremdsprache im Unterricht betrachtet? Weder wird der unter Kapitel 1.3.2 dieser Arbeit erläuterte Dualismus „Fremdsprache als Lerngegenstand und Kommunikationsmittel zugleich“ explizit thematisiert noch wird er problematisiert. Auch dieses Manko lässt sich letztlich auf den unpräzisen Sprachbegriff zurückführen. Im GER werden die Konzepte „Sprache als soziales Gut in alltäglicher Kommunikation“ und „Lernersprache als individuelles Gut in der Entwicklung“ ebenso wenig voneinander unterschieden wie die Perspektiven „Sprachverwendung in realen Kommunikationssituationen“ und „Sprachverwendung in der Unterrichtssituation“.

Die Sprachlernenden sieht der GER – wie schon erwähnt – als denselben Bedingungen unterworfen wie die Sprachverwenden, doch sind – wie unter Kapitel 1.3.1 und 1.3.2 dieser Arbeit erläutert – die Kontexte der Verwendung der Lernersprache im Vergleich zur Verwendung der Muttersprache häufig nicht vergleichbar; dazu treten Unterschiede bei der Verwendung der Lernersprache im Unterricht und im zielsprachlichen Alltag. Man denke beispielsweise an die relativ geschützte unterrichtliche Situation eines simulierten *Critical Incident*, in welcher verbale wie nonverbale Verhaltensweisen erprobt und gegebenenfalls revidiert werden können – im Vergleich dazu ist die reale interkulturelle Situation geprägt durch das Fehlen dieser Erprobung von Handlungsspielräumen und führt im negativsten Fall zu Missverständnissen, Ängsten und Handlungsunfähigkeit.

Diese Unterscheidung der realen Sprachverwendung gegenüber der unterrichtlichen Verwendung hätte im GER explizit thematisiert werden müssen, da, wie oben erläutert, jeweils andere Voraussetzungen und Bedingungen gegeben sind. Es finden sich im GER ja durchaus Hinweise auf die Bedeutung authentischer Sprachverwendung im Unterricht, wie man sie aus dem kommunikativen Ansatz kennt; auch eine kognitiv orientierte Facette des Sprachbegriffs kann man ausmachen, da die Bedeutung der Bewusstheit im Sprachlernprozess anerkannt wird; selbst der Sprachbegriff des prozessorientierten Ansatzes zeigt sich an der Thematisierung der Lernprozesse. Der sprachvergleichende und interkulturelle Ansatz jedoch, namentlich die Funktion einer Fremdsprache und einer fremden Kultur als „Fenster“ zu anderen Sprachen und Kulturen im Unterricht wird nicht thematisiert, obwohl diese Funktion impliziert wird durch Aussagen wie beispielsweise die folgende (vgl. GER 2001: 18):

> Sprache ist nicht nur ein besonders wichtiger Aspekt einer Kultur, sondern auch ein Mittel des Zugangs zu kulturellen Erscheinungsformen und Produkten. (...) Die verschiedenen ... Kulturen, zu denen ein Mensch Zugang hat, (...) werden verglichen und kontrastiert (...).

Es findet sich dann aber keine Ausführung dazu, ob die Autoren des GER diese kulturelle „Zugangsfunktion" von Sprache als etwas Natürliches betrachten, das jeder Sprachverwender „automatisch" nutzen würde, oder ob es besonderer Vermittlungs- und Bewusstmachungsverfahren bedarf, um diese Funktion von Sprache nutzbar zu machen.

1.3.4.5 Vermittlungskonzept

Der Analyse des fremdsprachlichen Vermittlungskonzepts im GER werden die oben in Kapitel 1.3.3 dieser Arbeit erarbeiteten Kategorien zugrunde gelegt: Im Folgenden soll untersucht werden, welche Aussagen zu *methodischen Ansätzen*, zu *Auswahl und Anordnung*, zu *Darbietung*, zum *classroom discourse* und zur *europäischen Dimension im Fremdsprachenunterricht* im GER zu finden sind.

Der GER erwähnt auf S. 8, dass der Europarat solche *Lern- und Lehrmethoden* unterstütze, die das Ziel fördern, „Einstellungen, Kenntnisse und Fähigkeiten zu entwickeln, die notwendig sind, um im Denken und Handeln unabhängiger zu werden und in ... Beziehungen zu anderen Menschen verantwortungsbewusst und kooperativ zu handeln." Es geht also nicht mehr nur um die Vermittlung und Förderung reiner Sprachkenntnisse, sondern es soll übergeordnet auch um die „Förderung eines demokratischen, staatsbürgerlichen Bewusstseins" gehen (GER 2001: 8). Zur Förderung dieses Bewusstseins und der Sprachkenntnisse werden die schon erwähnten Empfehlungen des Ministerkomitees aufgelistet (ebd.: 14), namentlich Europas Vielfalt im kulturellen und sprachlichen Bereich zu bewahren und „als Quelle der Bereicherung" zu nutzen, die Kommunikation und Mobilität in Europa zu fördern über das Sprachenlernen und die Maßnahmen der Mitgliedsstaaten zu koordinieren. Dann erfolgt die Aufzählung allgemein gehaltener Maßnahmen: Beispielsweise soll allen Bevölkerungsgruppen Zugang zu Bildungsmöglichkeiten im sprachlichen Bereich ermöglicht werden (GER 2001: 15), oder es sollen „Methoden des modernen Sprachunterrichts" gefördert werden, „die die Unabhängigkeit des Denkens, des Urteilens und Handelns (...) stärken." (GER 2001: 16). Doch zur Frage, welche Methoden dazu geeignet sind, finden sich statt konkreter, begründeter Vorschläge eher generelle und wiederum unverbindliche Aussagen wie die folgende (ebd.: 10):

> Auch hier will der *Gemeinsame Referenzrahmen* keine bestimmte Methode vorschreiben oder empfehlen. Er präsentiert vielmehr Optionen und lädt Sie ein, Ihre eigene gegenwärtige Praxis zu reflektieren, Entscheidungen zu treffen und zu beschreiben, was genau Sie tun. (...) [D]er *Gemeinsame Referenzrahmen* will Ihnen vor allem bei Ihrer eigenen Entscheidungsfindung helfen.

Weiterführende Überlegungen zu methodischen Ansätzen finden sich im GER erst wieder in Abschnitt 6.4: Dort wird darauf hingewiesen, dass es „ein grundlegendes methodologisches

Prinzip des Europarats“ sei, diejenigen Methoden einzusetzen, „die als die effektivsten gelten, um die Ziele zu erreichen, auf die man sich in Hinblick auf die Bedürfnisse der einzelnen Lernenden in ihrem sozialen Kontext geeinigt hat.“ (GER 2001: 140). Doch wie bereits oben zu Beginn dieses Unterkapitels erläutert, finden sich im GER dann nur knappe Darstellungen einiger methodischer Ansätze, die noch dazu weder durch Literaturangaben noch durch Verweise auf die Forschung belegt sind. Auch in diesem Zusammenhang verwundert es daher nicht, in der Fachliteratur kritische Aussagen bezüglich der Tatsachenbehauptungen des GER zu finden.[86] Der GER stellt kein kohärentes Vermittlungskonzept vor und gibt auch keine Empfehlungen zu lernfördernden Merkmalen eines „guten“ Fremdsprachenunterrichts – er will ja keine *Summa Didactica* sein, doch zeigt er sich in seinem Vermittlungsansatz relativ konzeptionslos.

Der Überblick über allgemeine methodische Ansätze auf S.141f des GER ist an und für sich hilfreich, wenn diese Ansätze denn nur belegt wären. Die Thematisierung der Rollen von Lehrenden[87], Lernenden und Medien auf S.142f ist ebenfalls brauchbar und gibt einen Fragenkatalog an die Hand, den professionelle Sprachvermittler durchaus selbständig für ihr jeweiliges Konzept beantworten können. Auch die Thematisierung der Rolle von Texten (GER 2001: 143f) ist gerechtfertigt, ebenso wie die des Verhältnisses von Rezeption und Produktion in GER-Abschnitt 6.4.3.3; jedoch wird nichts zur Gewichtung dieser beiden Aspekte ausgesagt – es darf auf Bleyhl und Wandruszka[88] verwiesen werden, die der Rezeption naturgemäß ein weit größeres Gewicht zusprechen. Dies hätte an dieser Stelle des GER ebenfalls Gegenstand der Überlegungen sein müssen.[89] Daneben beleuchtet der GER die Rolle kommunikativer Aufgaben und Aktivitäten sowie kommunikativer Strategien (ebd.: 144f); dies impliziert wiederum den kommunikativen Ansatz, dem der GER verpflichtet ist. Der *Rolle kommunikativer Aufgaben* ist gar ein eigener Abschnitt 7 im GER gewidmet, der endlich unterscheidet zwischen kommunikativen Aufgaben im Alltag, die eher den Namen *Handlungen* oder *Aktivitäten* verdienen, und solchen im Unterricht, die durchaus den Namen *Aufgaben* verdienen, sofern sie eine *Aufgabe* repräsentieren und nicht etwa Teil der unterrichtlich bedingten Sprachverwendung sind (vgl. ebd.: 153). Auch in diesem GER-Abschnitt gibt es jedoch Terminologieprobleme: Der GER gibt Beispiele für kommunikative Aufgaben, die „sprachliche Aktivitäten in unterschiedlichem Umfang enthalten“ können (ebd.: 153): Worin allerdings bei den dann folgenden Beispielen wie etwa „malen“ oder „etwas reparieren oder zusammenbauen“ oder „Puzzles lösen“ die *kommunikative* Aufgabe oder gar die *sprachliche* Aktivität enthalten sein soll, bleibt zumindest der Verfasserin dieser Arbeit verschlossen. In GER-Abschnitt 7 werden, wiederum Bezug nehmend auf vorangegangene GER-Abschnitte, alle lernrelevanten Faktoren (kognitive, affektive und sprachliche) behandelt und schwierigkeitsbestimmende Merkmale hinsichtlich Interaktion/Produktion und

[86] Vgl. beispielsweise Schwerdtfeger 2003.
[87] Hier wird, bezogen auf die Perspektive der Lehrenden, der enge Ansatz von S. 139 des GER erweitert: Dort wird die Rolle der Lehrenden auf die Aufgabe beschränkt, „(…) offizielle Richtlinien zu befolgen, Lehrwerke und Unterrichtsmaterialien (...) zu benutzen, Tests zu entwickeln und durchzuführen und Schüler und Studierende auf Prüfungen vorzubereiten.“ (GER 2001: 139).
[88] Vgl. oben, Kapitel 1.3.3.2 dieser Arbeit.
[89] Dies wird auch thematisiert – leider an anderen Stellen: Beispielsweise werden die *partiellen Kompetenzen* unter 6.1.4 .1, S.135f angesprochen.

Rezeption beschrieben. Dieser Ansatz lässt vermuten, dass GER-Abschnitt 7 auf die Beschreibung und Erstellung von didaktischen Aufgaben im Rahmen des Unterrichts oder im Rahmen von Tests ausgerichtet ist.

Wenden wir uns nun der Behandlung der *Auswahl und Anordnung* der Unterrichtsinhalte im GER zu. Die Frage danach, was gelehrt und gelernt werden soll, hängt eng mit den jeweiligen Lehr- und Lern*zielen* zusammen. Diese werden im GER, wie oben schon angedeutet, unter Abschnitt 6.1.4 (S.134) thematisiert und ausführlich besprochen. Dort wird immer auch der Bezug zu den vorangehenden Abschnitten im GER hergestellt, so dass man sich in den entsprechenden Teilbereichen (beispielsweise *allgemeine Kompetenzen*, *Kommunikative Sprachkompetenzen* oder *Strategien*) die für die jeweilige Situation angemessenen Ziele stecken kann. Auch werden dort die partiellen oder rezeptiven Kompetenzen angesprochen, denen in Europa eine immer größere Bedeutung zukommt. Positiv zu bewerten ist das explizite Erwähnen der Vielfalt möglicher Ziele und der Notwendigkeit, diese immer auf konkrete Bedürfnisse auszulegen (vgl. GER 2001: 137). Ebenfalls positiv zu bewerten ist, dass der GER sich auf die bereits bestehenden Lernzielbestimmungen des Europarats bezieht und sie kohärent erweitert.[90]

Zur Frage der Auswahl der Themen, Inhalte und Gegenstände, um mehrsprachige und plurikulturelle Kompetenzen auszubilden und zu fördern, finden sich in GER-Abschnitt 6.1 einige allgemein gehaltene Aussagen zur Zielbestimmung, Bedarfsanalyse und Beschreibung dessen, was im jeweiligen Unterricht behandelt werden soll (vgl. ebd.: 131). Diese Aussagen sollen von den Lehrenden jeweils für ihre Kontexte gefüllt und bezogen werden auf die Kompetenzen (wie sie in GER-Abschnitt 5 beschrieben sind), auf deren Umsetzung in die Tat (vgl. GER-Abschnitt 4) und auf die dabei eingesetzten Strategien (ebenfalls GER-Abschnitt 4). Insofern zeigt sich der GER konsistent und transparent, als er sich bei den lern- und lehrbezogenen Aussagen auf die Ausführungen seiner vorangegangenen Abschnitte bezieht: Der GER gibt in den Abschnitten 4 und 5 viele Kategorien und Teilbereiche vor, die entsprechend des jeweiligen Unterrichtskonzepts ausgewählt und gewichtet werden müssen und können. In den GER-Abschnitten 6.4.6 respektive 6.4.7. finden sich auf diese Kategorien bezogene konkrete Anregungen zur Entwicklung der allgemeinen respektive der sprachlichen Kenntnisse, die die Basis der Auswahl des jeweiligen Unterrichtsstoffs bilden können. Da diese Auswahl immer von Zielen, Lernenden, Kursinhalten, Unterrichtsvorgaben etc. abhängt und gemeinsam zwischen Lehrenden und Lernenden bestimmt werden muss, können die Anregungen des GER in der Praxis eine durchaus hilfreiche Ausgangsbasis darstellen.

Doch das wichtige Ziel, mehrsprachige und plurikulturelle Kompetenzen zu entwickeln, wird nicht konsequent in didaktische Rahmenempfehlungen umgesetzt. Wohl finden sich einige Merkmale auf den Seiten 132ff beschrieben, und es findet sich die Aussage, dass „der Förderung

[90] Beispielsweise werden die Niveaus des Referenzsystems im GER auf diese Vorarbeiten des Europarats bezogen, vgl. dazu GER (2001: 33f und 42ff) und die Ausführungen in Kapitel 3.4 dieser Arbeit.

der Achtung sprachlicher Vielfalt und des Lernens von mehr als einer Fremdsprache in den Schulen große Bedeutung“ zukomme (GER 2001: 134), doch wie dies im Fremdsprachenunterricht umgesetzt werden kann, bleibt den Lehrenden überlassen. Zu diesen Kompetenzbereichen gibt es, wie oben bereits kritisiert, auch keine eigenen Skalen, die die Charakteristika mehrsprachiger oder plurikultureller Situationen und der dabei zum Einsatz kommenden Strategien oder kommunikativen Handlungen beschreiben und die in diesen Situationen benötigten Kompetenzen abstufen könnten. Dies steht in Widerspruch zu folgender Aussage (ebd.: 132):

> Der *Referenzrahmen* beschränkt sich nicht auf das überblicksartige Skalieren kommunikativer Fertigkeiten, sondern untergliedert globale Kategorien in ihre Komponenten und bietet für diese Skalen an. Dies spielt eine besondere Rolle für die Entwicklung der mehrsprachigen und der plurikulturellen Kompetenz.

Im darauf folgenden GER-Abschnitt 6.1.3.3 wird zwar der Entwicklung des Sprachbewusstseins eine wichtige Rolle im Sprachlernprozess zugeschrieben, doch wie beispielsweise die „Fähigkeit, sich auf andere Menschen und neue Situationen einzustellen“ (GER 2001: 133), zu entwickeln ist, bleibt wiederum den Benutzern überlassen. Es finden sich – im Gegensatz zu den gerade erwähnten Ausführungen zur Entwicklung der sprachlichen Kompetenzen – keine Anregungen, welche Themen, Inhalte und Teilbereiche zur Entwicklung der mehrsprachigen und plurikulturellen Kompetenzen empfehlenswert wären.

Im GER gibt es einige allgemein gültige Aussagen zu *Progressionsprinzipien*, wenn es dazu auch noch keinen übergreifenden Konsens in der Fremdsprachendidaktik geben mag. Doch heutzutage dürfte sich zumindest das oben genannte Konzept der Spiralprogression als konsensfähig erwiesen haben, wohingegen Hypothesen bezüglich des Vorhandenseins so genannter natürlicher Erwerbssequenzen als Basis für unterrichtliche Progression in einer starken Form nicht bestätigt werden konnten.[91] Wie oben schon angedeutet, finden sich im GER lediglich unter Abschnitt 6.4.7 *Linguistische Kompetenzen* Aussagen zu Ordnungsprinzipien wie inhärente Komplexität, kommunikative Funktionen, Verwendbarkeit und Häufigkeit der sprachlichen Phänomene oder natürliche Erwerbssequenzen, allerdings ohne weitere Belege oder Ausführungen.

Positiv fällt die Anregung auf, dass sich der Wortschatz „organisch“ entwickeln könnte „in Reaktion auf die Bedürfnisse der Lernenden bei kommunikativen Aufgaben“ (ebd.: 148). Es fehlt jedoch ein umfassendes und wissenschaftlich fundiertes Progressionskonzept jenseits der grammatischen Progression.

[91] Vgl. beispielsweise Edmondson & House (1993, insb. Kap. 9.1): Hier sind u. a. Studien von Brown 1973, Dulay & Burt 1974 oder Krashen 1982 dargestellt, in denen „natürliche“ Erwerbssequenzen im Unterricht sprachübergreifend nicht bestätigt werden konnten; es gibt aber Hinweise auf von L1 beeinflusste Zwischenstrukturen beim Ausbau einer L2, welche vergleichbar sind mit Erscheinungen beim Erstspracherwerb. Die Entwicklung dieser Zwischenstrukturen deckt sich weitgehend mit allgemeinen Prinzipien der Progression, beispielsweise den Prinzipien des Voranschreitens von Bekanntem zu Neuem, von Einfachem zu Komplexem oder vom Häufigen zum Seltenen.

Aussagen zur *Darbietung* des Unterrichtsstoffs im GER sind eigentlich nur implizit erschließbar über etwa die methodologischen Anregungen (GER-Abschnitt 6.4) oder über Hinweise, wie etwa grammatische Kompetenz entwickelt werden kann (GER: 149), doch werden solche didaktischen Fragen nicht unter didaktischen Gesichtspunkten angeordnet und behandelt, sondern wohl aufgrund pragmatischer Überlegungen immer gleich bei denjenigen Teilkompetenzen abgehandelt, bei denen sie relevant erscheinen. Dies macht das Auffinden von Hinweisen zur Darbietung etwas mühsam: Beispielsweise finden sich die Hinweise auf Darbietung des Wortschatzes unter GER-Abschnitt 6.4.7.1, während die grammatische Darbietung auf unter GER-Abschnitt 6.4.7.7 thematisiert wird; das kontrastive Herangehen, um den Einbau neuen Wissens in bestehende Systeme zu erleichtern, findet sich unter Abschnitt 6.4.7.5 respektive ist implizit aus Abschnitt 6.4.6.1 zu erschließen. Man muss allerdings wieder an anderer Stelle suchen, um (implizite) Aussagen bezüglich der generellen Darbietung von Sprache etwa in ganzheitlichen, sinnvollen Zusammenhängen zu finden. Beispielsweise geht aus GER: 51 hervor, dass „Sprachlernende angehende Sprachverwender (sind), so dass auf beide die gleichen Kategorien zutreffen." Auch wenn oben unter Kapitel 1.3.1 und 1.3.4 dieser Arbeit gezeigt wurde, dass diese Aussage so nicht korrekt ist, lässt sie doch die Schlussfolgerung zu, dass die Autoren des GER Sprache beim Lernen möglichst denselben (kontextuellen) Bedingungen unterwerfen wollen wie bei der Verwenden in realen Kontexten: Sprache sollte in möglichst unterschiedlichen und möglichst authentischen Kontexten dargeboten werden und von den Lernenden selbst aktiv verwendet werden.

Es ist auffällig, dass sich im gesamten GER keine Aussagen zu den Besonderheiten des *classroom discourse* finden lassen, wie der GER ja auch die Besonderheit des Fremdsprachenunterrichts (Sprache als Unterrichtsgegenstand und Kommunikationsmittel zugleich) nicht thematisiert. Wohl ist der GER dem kommunikativen und teils auch dem interaktiven Ansatz verpflichtet, doch gibt es keine Beschreibung der lernfördernden Merkmale des *classroom discourse*. Wünschenswert wäre auch eine Skala zu den Kompetenzen der Lehrenden in diesem Bereich. Ein wichtiger Bereich des *classroom discourse* und der Lernersprache ist allerdings thematisiert: der der Fehler und der *Feedback*-Möglichkeiten in GER-Abschnitt 6.5 *Kompetenz- und Performanzfehler*. Dort finden sich, wie oben in Kapitel 1.3.4.2 dieser Arbeit bereits kritisiert, Aussagen, die den Eindruck von Tatsachenbehauptungen erwecken, obwohl sie vermutlich als Denkanstöße gedacht sind. Bedauerlicherweise werden auch in diesem Zusammenhang die durchaus vorhandenen konkreten Ergebnisse aus Diskurs- und Lernforschung, wie sie in Kapitel 1.3.3.4 dieser Arbeit dargestellt sind, nicht erwähnt. Die Anregungen auf S.152 des GER zur Beobachtung und Analyse von Fehlern jedoch sind als Denkanstoß durchaus sinnvoll.

Ebenso auffällig ist, dass sich im GER sich keine konkreten Umsetzungsvorschläge finden, wie die *europäische Dimension* beim Sprachenlernen und -lehren berücksichtigen werden kann. Nun könnte man argumentieren, dass diese Umsetzung in den Bereich der Sprachenpolitik reicht, somit den einzelnen europäischen Staaten obliegt und nicht zentralisiert geregelt werden kann. Dennoch sollte gerade ein europäisches Instrument zur Förderung der Mehrsprachigkeit in diesem Bereich einen konsensfähigen Rahmen stecken und Optionen vorstellen, wie sie etwa oben unter Kapitel 1.3.3.5 dieser Arbeit erörtert wurden. Wieso sich im GER außer den erwähnten sehr generellen Empfehlungen des Europarats keine konkreten Aussagen dazu finden lassen, ist nicht nachvollziehbar.

Hingegen findet man in GER-Abschnitt 8 Überlegungen zu den *Curricula* hinsichtlich der Diversifizierungsmöglichkeiten, der verschiedenen methodischen Ansätze und der je nach Kontext und Zielgruppe variierenden Ziele. Es werden in GER-Abschnitt 8.1 und 8.2 Überlegungen angestellt, wie die *Curricula* durch den GER verbessert und erweitert werden können. Auf S. 163f des GER ist von drei Grundsätzen die Rede, die bei diesem Prozess wirksam seien: Einmal der Grundsatz der „Förderung der Mehrsprachigkeit und der Sprachenvielfalt“, dem Curricula gerecht werden sollten; zum Zweiten der Grundsatz, dass Diversifizierung ökonomisch gestaltet werden muss (beispielsweise durch Nutzung von Synergieeffekten); und schließlich der Grundsatz, sich bei curricularen Überlegungen nicht auf eine Sprache zu beschränken, sondern an ein „integriertes Curriculum für mehrere (einzelne) Sprachen“ (GER 2001: 164), so dass grundsätzlich die Möglichkeit gefördert werde, Kenntnisse und Fähigkeiten auf verschiedene Sprachen zu übertragen. In GER-Abschnitt 8.3 finden sich „Entwürfe für Curriculumsszenarien“ (ebd.: 165), die obige Überlegungen konkretisieren. Im Anschluss wird auf S.169 auf lebenslanges Lernen auch nach der Schulzeit kurz eingegangen, ohne allerdings konkret zu werden. Auch die Bedeutung des Sprachenportfolios und der Profilbildung wird herausgestellt, um die Lernenden beim lebenslangen Lernprozess begleiten zu können, und um auch Teilkompetenzen im Sinne von unterschiedlichen Profilen anerkennen zu können. Die Bedeutung der Mehrdimensionalität des Sprachenlernens und der Modularisierung des Lernprozesses wird auf S.169f betont. Ob der Fragekasten am Ende des GER-Abschnitts 8 Curriculumsentwicklern helfen kann, ihre Entscheidungen zu treffen und zu begründen, muss die Praxiserfahrung zeigen.

1.3.4.6 Fazit

In GER-Abschnitt 6 (und nicht nur dort) werden Tatsachenbehauptungen aufgestellt, die nicht durch entsprechende Verweise auf Fachliteratur oder Forschungsergebnisse gestützt werden. Die in GER-Abschnitt 6 angebotenen Fragenkataloge können Denkanstöße geben und die Beteiligten zu Reflexion ermuntern. Doch eine bloße Auflistung von Denkanstößen ohne

wissenschaftliche Absicherung scheint für solch ein politisch bedeutsames Instrument zu oberflächlich. Gerade ein europäischen Instrument, das auch der Umsetzung der europäischen Sprachenpolitik dienen soll, sollte sich bemühen, wissenschaftlichen Ansprüchen zu genügen.

Im GER fehlen kohärente Lern- und Lehrkonzepte, die durchaus offen sein dürfen, doch bezüglich existierender didaktischer Forschungsergebnisse Position beziehen müssten. Auch wenn der GER kein wissenschaftliches Dokument ist, sollte er als politisches Instrument dennoch transparent darstellen, woher die Erkenntnisse stammen, die dort als Tatsachen präsentiert werden. Auch sollte die Auswahl des im GER Dargebotenen begründet werden.

Das Fehlen wichtiger didaktischer Kernbereiche wie etwa der Charakteristika der Lernersprache, der lernfördernden Merkmale eines „guten" Fremdsprachenunterrichts oder des *classroom discourse* trägt nicht zu einer soliden didaktischen Basis bei. Unzureichende Thematisierungen wie die der Rolle der Muttersprache oder die eines Mehrsprachigkeitskonzepts (vgl. Neuner 2003: 141) stellen ebenso ein Manko dar wie die Nicht-Thematisierung interkultureller Aspekte und deren Didaktik – ein Manko, das gerade im europäischen mehrsprachigen und plurikulturellen Kontext nicht nachzuvollziehen ist.

Die Strukturierung der lern- und lehrrelevanten Abschnitte des GER in Anlehnung an die Kategorien der GER-Abschnitte 4 und 5 mag von einem pragmatischen Gesichtspunkt aus Sinn machen, doch sollte dabei der didaktische Standpunkt nicht völlig in den Hintergrund geraten. Zur Erhöhung der Benutzerfreundlichkeit würde schon ein Register entscheidend beitragen, ebenso wie eine transparentere Differenzierung der Benutzerperspektiven.

Positiv fällt auf, dass die Bedeutung der Lernerautonomie (im Sinne eines *life-long learning*), die Bedeutung authentischen Inputs und möglichst authentischer, interaktiver Sprachverwendung und auch die Bedeutung von Bewusstheit über Sprache und Lernen anerkannt werden. Auch gibt es ausführliche Darstellungen zu kommunikativen Aktivitäten und zur Entwicklung der allgemeinen und der sprachlichen Kompetenzen, sowie einige hilfreiche methodische und didaktische Hinweise. Das Hervorheben der Bedeutung von Mehrdimensionalität und Modularität für die sprachliche Diversifizierung ist ebenfalls positiv zu bewerten.

Doch der GER kann und will eine solide Ausbildung in der Fremdsprachendidaktik keinesfalls ersetzen. Schon im Vorwort wird ausgesagt, dass sich der GER an „professionell im Bildungsbereich Tätige" (GER 2001: 3) wendet – und diese sollten Grundlagenwissen mitbringen. Aufbauend auf diese Grundlagen können die Anregungen im GER schon hilfreich sein, doch man darf sich keinesfalls eine umfassende Konzeption oder Darstellung erwarten; viele Lücken müssen von den Benutzern (noch) selbst geschlossen werden – dies lässt zwar Freiräume für eigene Akzentuierungen, doch ob dies einem *Referenzrahmen* gerecht wird, der „wirklich alles" bieten möchte, was man benötigt, um seine „Ziele, Methoden und Produkte zu beschreiben" (GER 2001: 9), ist fraglich.

2 Das Testen des Sprachvermögens

Auch wenn diese Arbeit im Feld der Beurteilung von Sprachvermögen angesiedelt ist und sich im Praxisteil konkret auf die Schulleistungsstudie DESI und ein innerhalb dieser Studie entwickeltes Testmodul bezieht, so soll Gegenstand dieses Kapitels nicht der größere Kontext der Beurteilung von Schülerleistungen und deren Rahmenbedingungen sein; Gegenstand dieses Kapitels ist vielmehr die Ebene der Beurteilungsinstrumente und besonders der Sprachtests als einer konkreten Möglichkeit, das Sprachvermögen von fremdsprachlichen Lernenden zu bewerten.

In den vorangegangenen Kapiteln wurde ein sprachlicher und didaktischer Rahmen gesteckt, der zunächst als Analysegrundlage für den Sprachbegriff und das Lern- und Lehrkonzept des GER gedient hat. An dieser Stelle sollen diese sprachlichen und didaktischen Grundlagen nun herangezogen werden, um Aussagen in Bezug auf Sprachtests zu treffen: Was erfassen welche Arten von Tests? Welche Testformate wurden in welchem Paradigma entwickelt und zu welchen Zwecken eingesetzt? Wie können die verschiedenen Formate ausgewertet werden? Welche Möglichkeiten gibt es, valide und reliable Testformate zu entwickeln, die darüber hinaus den Bedingungen von Sprachverwendung und Sprachlernen gerecht werden? Nach Betrachtung dieser Fragen wird wiederum der GER analysiert: Welchen Beitrag können GER und der *User's Guide for Examiners*[92] (vgl. Council of Europe 1996b respektive 2002[2], im Folgenden kurz UGE genannt) auf dem Gebiet „Testen von Sprachvermögen" leisten? Nicht hier abgehandelt werden Skalenansätze in der Beurteilung, konkrete Fragen der Bewertungsmöglichkeiten von Tests mittels *Rating*-Verfahren und die Analyse des Skalenansatzes des GER, denn diesem Thema soll aufgrund seiner Bedeutsamkeit ein eigenes Kapitel 3 *Der Skalenansatz in der Beurteilung des Sprachvermögens* gewidmet werden.

Wenn die innersprachliche Organisation nach dem Prototypenmodel (wie in Kapitel 1.2.1 dieser Arbeit dargestellt) beschrieben werden kann, und wenn Lerner sprachliche Elemente in „Gestalten" oder Schemata abspeichern, die wiederum bestimmte Assoziationen auslösen bezüglich der Funktionen und Kategorien der jeweiligen sprachlichen Elemente, so ist es für das Lernen, Lehren und Beurteilen von Sprachvermögen ratsam, dieser Organisation gerecht zu werden. Dies dürfte sich erreichen lassen, wenn man Sprache – wie schon lange in der Fremdsprachendidaktik gefordert – in natürliche Kommunikations- und Verwendungszusammenhänge bettet, sie in diesen präsentiert und von den Lernern in solchen Kontexten verwenden lässt. Dann kann man davon ausgehen, dass Sprachlernen zu kommunikativer Handlungskompetenz führt und nicht zu einem unverbundenen Ansammeln isolierter Fakten und Einheiten, die in realen kommunikativen Verwendungssituationen nicht zur Verfügung stehen. Beispielsweise führt ein Erlernen von Regeln und Ausnahmen dazu, dass die Lerner alle Ausnahmen als „sprachliche

[92] Die aktuellste Version ist im Internet herunterladbar unter:
http://culture2.coe.int/portfolio/documents/Guide%20October%202002%20revised%20version1.doc, Zugriff am 13.09.2005.

Einzeltatsachen“ (Frey 2002: 26) speichern müssen – keine effektive Memorierungsstrategie. Leichter dürften Merkmalsbündel zu memorieren sein, die in Form von Assoziationen und Schemata gespeichert werden. Wenn man die Erkenntnisse der Gehirnforschung[93] mit einbezieht, die u. a. besagen, dass sich im Gehirn Repräsentationsstrukturen dieser Art finden lassen, so können diese Erkenntnisse nicht nur im Fremdsprachenunterricht genutzt und umgesetzt werden, sondern auch auf das Testen von Sprache übertragen werden: Tests sollten so entwickelt und eingesetzt werden, dass sie diesen natürlichen Prinzipien gerecht werden und Sprachverwendung in möglichst authentischen Kontexten erfassen.

Was sagen beispielsweise traditionelle Formen der Leistungsmessung wie etwa *discrete-point tests* aus, in denen dekontextualisierte, isolierte Einheiten bearbeitet werden? Man kann nicht auf das Kommunikationsvermögen der Probanden im Alltag rückschließen, da traditionelle Testformate diese gar nicht erfassen. Die vorliegende Arbeit rückt integrative und kommunikative Ansätze in den Mittelpunkt, um zu zeigen, wie die Komplexität von Lernersprachentwicklung, Sprachverarbeitung und Sprachproduktion mit entsprechend komplexen Formaten erfasst werden könnte. Natürlich können nicht einfach alle Anforderungen der klassischen Testtheorie außer Kraft gesetzt werden, doch müssen Psychometrie und Testtheorie auf neue linguistische und lerntheoretische Modelle mit neuen Rechenmodellen antworten, die der Realität gerecht werden. Da die Psychometrie jedoch nicht Gegenstand dieser Arbeit ist, kann deshalb nur ganz am Rande ein Blick darauf geworfen werden.

Zu Beginn sollten einige terminologische Fragen geklärt werden: Was soll mit Sprachtests erfasst werden? Geht es um das Können, die Kompetenz, die Qualität der Performanz, die Leistung? Geht es um das generelle Sprachvermögen oder um einzelne Teilbereiche? Geht es um Wissensbestände oder um Anwendbarkeit dieses Wissens? Wie kann man diese Termini definieren und voneinander abgrenzen? Da im Bereich des Sprachtestens Termini oft im angelsächsischen Raum geprägt wurden, lohnt es, die englischen und deutschen Begrifflichkeiten im Vergleich zu betrachten.

Wenden wir uns zunächst dem komplementären Begriffspaar Kompetenz – Performanz zu: Während Performanz das empirisch zu beobachtende „tatsächliche“ Sprachhandeln bezeichnet, bezieht sich Kompetenz auf das theoretische Konstrukt, das der zu beobachtenden Performanz zugrunde liegt und selbst nicht direkt zu beobachten ist. Diese beiden Begriffe stammen ursprünglich aus der Linguistik und wurden von Chomsky (1965) geprägt. Neben diesem kognitiv-linguistischen Zugang zum Begriff der Kompetenz gibt es auch ein verhaltensbasiertes Konzept der Kompetenz, welches Wissen und Fertigkeiten umfasst und Performanz als Teil der kommunikativen Kompetenz begreift.[94] Dabei wird die Beherrschung verschiedener Fertigkeiten oder *skills*

[93] Vgl. beispielsweise Neville & Bavelier 1998.
[94] Vgl. beispielsweise die Ausführungen zur kommunikativen Kompetenz in North 2000, 43ff. Hier wird eine übersichtliche Darstellung der verschiedenen Ansätze und Modelle von Kompetenz und *Proficiency* gegeben.

als Voraussetzung für kompetentes Sprachhandeln betrachtet. Diese Beherrschung könnte man mit dem englischen Terminus *proficiency* wiedergeben. Auch die Fähigkeit zur Anwendung von Wissen spielt mit herein; beispielsweise hat Hymes[95] neben anderen psychologischen Faktoren diese Fähigkeit mit in sein Kompetenzkonzept aufgenommen. Ein dritter Zugang zur kommunikativen Kompetenz lässt sich in sozio-kulturellen und kommunikationstheoretischen Konzepten ausmachen: Dabei ergibt sich die Angemessenheit und Bedeutung sprachlicher Handlungen erst aus der Einbettung der Kommunikation in sozio-kulturelle Kontexte, so dass bei diesem Konzept auch pragmatische und sozio-kulturelle Kompetenzen zur kommunikativen Kompetenz beitragen. Bei der Vielzahl der möglichen Ursprünge, Zugänge und Definitionen des Kompetenzbegriffs ist es ratsam, das eigene Verständnis des Kompetenzbegriffs offen zu legen, wie es in dieser Arbeit gleich im Anschluss auch geschieht.

Ein weiterer Begriff, der der Klärung bedarf, ist der englische Terminus der s*kills*: Diese entsprechen in der didaktischen Auffassung den Fertigkeiten, die erlernt und geübt werden können, traditionellerweise Hörverstehen, Sprechen, Leseverstehen, Schreiben und in jüngerer Zeit auch die Fertigkeit des Übersetzens, welche erweitert werden sollte um die Fertigkeit des Vermittelns im interkulturellen Kontext. Fähigkeiten dagegen könnte man auf Englisch mit *abilities* wiedergeben; Fähigkeiten bezeichnen im Gegensatz zu Fertigkeiten in der Didaktik etwas, das die Lerner „mitbringen“, mit dem sie bereits ausgestattet sind. *Knowledge* im Sinne von Wissen bezieht sich auf erlernte oder erfahrene Wissensbestände, die man unterteilen könnte in prozedurales oder handlungsbezogenes Wissen einerseits und deklaratives Wissen andererseits. Fertigkeiten, Fähigkeiten und Wissensbestände sind so genannte „Subkompetenzen“, sie stellen der kommunikativen Kompetenz untergeordnete Teilkompetenzen dar.

CLA, communicative language ability, bezeichnet die kommunikative Sprachfähigkeit, das Sprachvermögen: das Können, Sprache kommunikativ (wirksam) zu verwenden. Der Begriff der *proficiency* ist dem der *CLA* ähnlich, er ist ebenfalls holistisch ausgerichtet, wurde aber in der Zeit des integrativen Testens geprägt: Man suchte ja Tests, die den globalen Sprachstand – den *global language proficiency factor* – erfassten. *Proficiency* bezieht sich auf den Grad der generellen Sprachbeherrschung, der beispielsweise in einem Test beurteilt werden soll: “The term ’language proficiency’ - something in between competence and performance – is conventionally used to describe what gets evaluated in language assessment.“[96] Der Begriff *achievement* dagegen bezieht sich auf das Erreichen eines bestimmten (Lern-)Ziels. Dieser Begriff im Sinne einer gerichteten Leistung kann im Deutschen mit *Leistung* wiedergegeben werden, doch *Leistung* trägt im Deutschen auch eine weitere Bedeutung: Man kann beispielsweise neutral von den *Leistungs*dimensionen sprechen, die ein bestimmter Test erfasst. Dieses Konzept könnte im Englischen mit *performance* wiedergegeben werden, soweit der Test Performanz erfasst, oder mit *abilities* bzw. *knowledge*, je nachdem, was der Test erfasst. Falls man sich auf die in

[95] Vgl. Hymes 1972b, zitiert in North (2000: 45).
[96] *Statement* von Brian North, *posted* in einem informellen Austausch im Rahmen der EALTA-Diskussion (E-mail Forum) um den GER im April 2004.

einem Test zu beobachtende sprachliche Handlung oder Produktion beziehen möchte, sollte man im Deutschen jedoch nicht von *Leistung* allgemein, sondern zutreffender von *Performanz* sprechen. Es gibt, wie man sieht, eine Vielzahl von Konzepten und Termini, die den jeweiligen Testgegenstand, die zu erfassende „Leistungsteildimension" beschreiben und bezeichnen, und die, teils bedingt durch die Traditionen, in denen sie geprägt wurden, nicht immer leicht voneinander zu differenzieren sind, da sie teils denselben Aspekt aus unterschiedlichen Perspektiven beschreiben.

Um der häufig englischsprachigen Testliteratur gerecht zu werden, sei es an dieser Stelle gestattet, englische und deutsche Termini zu kontrastieren und folgende Übersetzungen vorzuschlagen:

competence	Kompetenz (i. S. eines kommunikationstheoretischen Konstrukts)
performance	(Leistung i. S. v.) Performanz (neutral, im Gegensatz zu *achievement*)
achievement	Leistung i. S. des Erreichens eines (Lern-)Ziels
proficiency	Beherrschung von Fertigkeiten; Kenntnisse und deren Anwendbarkeit; Können, Fähigkeit (angesiedelt zwischen Kompetenz und Performanz: das, was üblicherweise in Sprachtests erfasst wird)
CLA: communicative language ability	kommunikative Sprachfähigkeit, kommunikatives Sprachvermögen
skills	Fertigkeiten (als der kommunikativen Kompetenz untergeordnete Subkompetenzen)
knowledge: declarative vs. procedural	Wissen, Wissensbestände: deklaratives vs. prozedurales Wissen (als Bestandteile der Kompetenz)

Tabelle 2: Gegenüberstellung englischer und deutscher Termini

Der vorliegenden Arbeit liegt ein kommunikationstheoretischer Begriff der Kompetenz zugrunde, welcher kommunikative Kompetenz als umfassendes Konzept versteht, in das relevante Fähigkeiten und Fertigkeiten, alle relevanten Wissensbestände und deren Anwendbarkeit sowie angemessenes Verhalten in verschiedenen Kontexten mit einfließen.[97] Kompetenz (als theoretisches Konzept) ist nicht direkt beobachtbar oder messbar – man kann sich ihr jedoch nähern über die Performanz, den tatsächlichen Sprachgebrauch in den produktiven Fertigkeiten, respektive über Indikatoren bezüglich der Kompetenzen, die sich einer direkten Beobachtung entziehen. Der Grad der Beherrschung der sprachlichen Fertigkeiten und der Grad der Anwendbarkeit des zugrunde liegenden Wissens kann mit dem Begriff der *proficiency* bezeichnet werden. Wissen und angemessene Verhaltensweisen werden durch entsprechenden Gebrauch und sprachliches Handeln eingebettet in sozio-kulturelle Kontexte erworben. Wenn man nun die kommunikative Kompetenz erfassen möchte, so muss man allen genannten Facetten dieses komplexen Gegenstandes gerecht werden: Man muss für Anlässe sorgen, die die fraglichen Wissens- und Verhaltensbestände aktivieren, so dass man über den Sprachgebrauch, die Performanz, auf den Grad der Beherrschung, die *proficiency*, rückschließen kann. Dieser Grad der Beherrschung muss für möglichst viele Aspekte (seien es kommunikative Aktivitäten,

[97] An dieser Stelle sei auf die Ausführungen zur kommunikativen Kompetenz in Kap. 1.2.3, S.11 dieser Arbeit verwiesen.

Sprechanlässe, Fertigkeiten, Themengebiete oder verschiedene Wissensbestände, um nur einige zu nennen) erfasst werden, um wiederum Rückschlüsse auf die kommunikative Kompetenz ziehen zu können.

Was im Einzelfall von welchem Testformat mit welchen Instrumenten zu welchem Zweck gemessen wird, hängt natürlich von den jeweiligen Gegebenheiten und Kontexten ab. Ob etwa ein bestimmtes Ziel im Test zu erreichen ist (im Sinne von *achievement*) oder ob die *overall proficiency,* der Grad an Sprachbeherrschung insgesamt festgestellt werden soll, kommt ebenso auf den Einzelfall an wie die Beantwortung der Frage, welche Fertigkeiten der Test erfassen soll: Geht es eher um deklaratives oder prozedurales Wissen, um Hörverstehen oder um die Schreibfertigkeit, oder soll der Test das kommunikative Sprachvermögen auf globaler Ebene erfassen? Diese Fragen sind nicht generell zu beantworten, sondern müssen immer im jeweiligen Kontext betrachtet werden, weshalb im Folgenden nur allgemeine Grundsätze angesprochen werden können, ehe sich dann Kapitel 4 dieser Arbeit mit einem konkreten Testbeispiel aus der DESI-Praxis beschäftigen wird.

Im vorliegenden Kapitel 2 dieser Arbeit sollen einige grundsätzliche Fragen geklärt werden, die zyklisch miteinander verbunden sind: Die Frage nach geeigneten Testformaten hängt beispielsweise eng mit der Frage zusammen, was Gegenstand des Testens sein soll (sollen etwa deklarative Wissensbestände getestet werden oder sprachliches Handeln) und zu welchen Zwecken der jeweilige Test in welcher Zielgruppe eingesetzt werden soll. Auch die Zielgruppe muss spezifiziert werden, um ihrem Hintergrund und ihrem Vorwissen gerecht zu werden, denn es gibt Hinweise darauf, dass beispielsweise Unterschiede in den Testgewohnheiten verschiedener Kulturen die Performanz in Tests unterschiedlicher Formate durchaus beeinflussen können.[98] Nur wenn diese grundsätzlichen Fragen geklärt sind, lassen sich davon abhängige Fragen wie etwa die der Gütekriterien eines Tests beantworten. Beginnen wir mit der Erörterung dessen, was ein Test messen soll, ehe wir uns dann der Frage nach geeigneten Testformaten zuwenden. Gegen Ende dieses Kapitels schließt sich der Kreis der relevantesten Fragen mit der Erörterung der wichtigsten Testgütekriterien und Testziele, so dass im Anschluss der GER auf seinen Testbegriff hin untersucht werden kann.

[98] Vgl. hierzu Farhady (1979: 354).

2.1 Auswahl der Leistungsdimensionen[99]

Wenden wir uns zuerst dem Gegenstand zu, der mit einem bestimmten Test erfasst werden soll. Sicherlich sind in diesem Zusammenhang konkrete Fragen nach Zweck und Ziel des Tests, der Probandengruppe und Ähnlichem zu beantworten. Doch diese Auswahl wird vom jeweils vorherrschenden Paradigma beeinflusst: Ob man nun „diskret" eine Teilfertigkeit neben der anderen erfassen will, sprachliche Wissensbestände mittels isolierter Wortschatz- und Grammatiktests abprüft, ob man verschiedene Teilfertigkeiten verschränkt betrachten will, oder ob man Sprache in all ihrer Komplexität integrativ so erfassen will, wie sie auch im Leben verwendet wird – letztlich bestimmen die zugrunde gelegten Modelle von Sprache und Kompetenz, aus welchem „sprachlichen Fundus" man schöpfen kann. Gerade die Diskussion um die (teils notgedrungene) Erfassung von Intelligenz, sei es nun in generellerer Form oder explizit bezogen auf verbale Intelligenz, zeigt wie bedeutsam es ist, noch vor Beginn der Testkonstruktion zu definieren und zu spezifizieren, was man erfassen will, um valide Instrumente entwickeln zu können. Gehen wir in Gedanken nochmals zurück[100] zu Bachmanns Modell der kommunikativen Kompetenz: Wenn die dort aufgeführten Komponenten Teil der kommunikativen Kompetenz sind, so können und sollen sie folgerichtig auch als Testgegenstand fungieren. Jeder Test muss begründet offen legen können, welche dieser Komponenten jeweils in welchen Kontexten durch welche Testformate erfasst werden sollen. Generell lässt sich sagen, dass kommunikative Tests sowohl sprachliche Kompetenzen (die sich wiederum aus Wissensbeständen und Handlungskompetenzen zusammensetzen) als auch strategische Kompetenzen erfassen sollen, sowie zu einem Teil auch Wissensbestände, die außerhalb des sprachlich-kommunikativen Wissens anzusiedeln sind. Nicht zu vergessen sind in diesem Zusammenhang die oben erwähnten Schlüsselqualifikationen, vor allem wenn es sich um Schulabgangsprüfungen oder Aufnahmetests handelt: Dabei muss natürlich der Erfassung der Schlüsselkompetenzen eine sorgfältige Analyse der jeweiligen Ansprüche an die Zielgruppe vorausgehen. Dies gilt auch für den Fall, dass der Test die oben beim Vermittlungskonzept erläuterten interkulturellen Fertigkeiten wie beispielsweise das Vermitteln im interkulturellen Kontext erfassen soll, wobei sich hierbei das Problem der Entwicklung neuer, valider Formate stellt. In jedem Fall wird die Testspezifizierung umso transparenter, je genauer der Testgegenstand im Konstrukt umrissen und in einem Modell

[99] Zur Terminologie „Dimension" darf auf die Ausführungen in North & Schneider (1998: 232f) verwiesen werden: Konstrukte und Dimensionalitäten verhalten sich relativ zueinander. Dimensionalität ist ein relatives Konstrukt, das entweder dem Verständnis komplexer Phänomene oder der Erleichterung von Entscheidungen dient. Die Frage der Ein- oder Mehrdimensionalität hängt auch immer von der Art des psychometrischen Dimensionalitätstests ab. Henning (1992: 8) nimmt an, dass die den meisten Tests zugrunde liegenden psychologischen Dimensionen so hoch korrelieren, dass sie dazu tendieren, ein psychometrisch eindimensionales Konstrukt zu bilden. Diese Annahme dürfte für alle Tests und Fragebögen gelten – McNamara (1996: 217-81) sagt, dass die Rasch-Analyse als Test eingesetzt werden könne um herauszufinden, inwieweit diese Annahme zutreffe. Carroll (1983: 83, 93) stellt fest, dass Eindimensionalität lediglich bedeute, dass die Ergebnisse interkorrelieren, da Menschen sich in den verschiedenen Teilfertigkeiten doch eher „gleichmäßig" weiter entwickeln. Demnach müssen Ergebnisse statistischer Rasch-Analysen im Licht von Theorien und Modellen psychologischer Konstrukte (die den Tests zugrunde liegen) interpretiert werden. Um es mit den Worten von North & Schneider (1998: 232f) auszudrücken: "(…) tests can exhibit sufficient *psychometric* unidimensionality without justifying assumptions about *psychological* unidimensionality. In other words, dimensionality in, for example, Rasch, has a technical meaning related to the technical meaning of reliability as separability."

[100] Vgl. Kapitel 1.2.3 dieser Arbeit

kommunikativer Kompetenz verankert wird. In solch einem Fall spricht man von modell-basiertem Testen. Diese Verankerung ist der Ausgangspunkt für die Testentwicklung, denn die Operationalisierung eines Testkonstrukts führt nur dann zu einem validen Test, wenn das Instrument letztlich auch erfasst, was es erfassen soll.

Es versteht sich von selbst, dass nicht alle Teilbereiche der kommunikativen Kompetenz in einem einzigen Test oder auch in einer Testbatterie erfasst werden können, da in der Regel zeitliche, organisatorische und finanzielle Grenzen gesetzt sind. Wenn man aber notgedrungenerweise Einschränkungen vornehmen muss, so sollten diese möglichst begründet erfolgen. Dabei spielen Testzweck und Probandengruppe die entscheidende Rolle: Was soll mit diesem Test bezweckt werden? Stellt er eine Qualifikationsprüfung „für das Leben“ dar, so wird sie umfangreicher ausfallen müssen als beispielsweise eine Abschlussprüfung am Ende eines akademischen Schreibkurses in einer Fremdsprache – bei letzterem Test wird es genügen, sich auf die Schreibfertigkeit in akademischen Kontexten zu beschränken, während man im ersteren Fall gut daran tut, möglichst alle Fertigkeiten und Kombinationen verschiedener Fertigkeiten in möglichst unterschiedlichen, alle relevanten Situationen abdeckenden Kontexten zu erfassen; in diesem Fall bietet es sich an, sich dem generellen Sprachstand auch über integrative Formate zu nähern. In jedem Fall aber wird es sich bei dem, was ein bestimmter Test erfasst, um ein *sampling* handeln, das nur in gewissem Rahmen im Hinblick auf die Spezifika des Testkonstrukts und der Probandengruppe verallgemeinert werden kann. Mehr zur Problematik der Generalisierbarkeit von Testergebnissen unter Kapitel 2.3 *Testgütekriterien*.

Zur Erfassung sprachlicher Teilbereiche und Fertigkeiten gibt beispielsweise Carroll (1972) eine zweidimensionale Matrix vor, die eine Auswahl von relevanten Bereichen und Fertigkeiten erlaubt – selbstverständlich muss das Vorgehen begründet werden und es muss genau dokumentiert werden, warum eine bestimmte Auswahl getroffen wurde.

Cooper (1972) stellt eine dreidimensionale Matrix vor, die neben Fertigkeiten und Wissensbereichen auch noch sprachlichen Varietäten gerecht werden will:
- *Skills* beziehen sich auf Hörverstehen, Leseverstehen, Sprechen und Schreiben;
- *Knowledge* umschließt bei Cooper Kompetenzen im phonologischen, syntaktischen und semantischen Bereich sowie eine globale Ebene;
- *Variety* bezieht sich auf Unterschiede im Dialekt, Register und Stil.

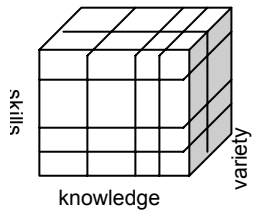


Abb. 6 Testmatrix nach Cooper

Cooper stellt ausdrücklich fest, dass seine Matrix zwar einen logischen Überblick über Kombinationsmöglichkeiten gäbe, doch mitnichten die Komplexität aller Möglichkeiten abdecken könne. Man denke beispielsweise an Situationen, Aktivitäten oder Themen, die man ebenfalls der Übersichtlichkeit halber in diese Matrix aufnehmen müsste, ganz zu schweigen von sprachlichen Funktionen und kommunikativen Absichten. Auch ein Integrieren verschiedener Teilfertigkeiten und sprachlicher Teilbereiche oder die Einbindung authentischer Formen der

Sprachverarbeitung dürfte sich nur schwer in solch einer Matrix darstellen lassen. Um solch eine ganzheitliche Herangehensweise an den "total communicative effect of an utterance" (Carroll 1972: 318) ging es aber seit den 70er Jahren einem Teil der Testforscher: Sie wollten nicht nur „eine Sache zu einer Zeit" prüfen, sondern möglichst authentischen Sprachgebrauch auf einer holistischen Ebene betrachten. Dazu wurden neue, integrative Formate benötigt: Auf den augenscheinlichen Gegensatz *diskretes versus integratives Testen* soll gleich im Anschluss unter Kapitel 2.2 dieser Arbeit eingegangen werden.

Hier schließt sich die Überlegung an, wie man die ausgewählten Leistungsdimensionen angemessen erfassen kann: Es bieten sich offene versus geschlossene Verfahren an, in direkten versus indirekten Tests, um die Enden der möglichen Spektren zu benennen, je nach dem welche Dimensionen von Wissen, Kenntnissen, Fertigkeiten, Varietäten, Aktivitäten oder Sprachfunktionen man in welchen Kombinationen erfassen will. Offene Aufgabenstellungen geben den Probanden den Raum, ihre Antworten in freier Form zu versprachlichen, also mit einer „echten" Sprachhandlung auf einen Test zu reagieren. Sie sind jedoch sehr aufwändig in ihrer Auswertung und im strengen Sinn nicht objektiv. Geschlossene Tests hingegen sind zwar sehr objektiv auszuwerten, doch reagieren die Probanden oft nicht mit sprachlichen Handlungen, sondern etwa in *Multiple-Choice*-Aufgaben mit bloßem Ankreuzen. Was genau also mit den jeweiligen Testtypen erfasst wird, muss sorgfältig beschrieben und empirisch abgesichert werden.

Die Auswahl der Leistungsdimensionen bestimmt auch die Wahl des indirekten oder direkten Testens: Sollen Kenntnisse und Wissensbestände erfasst werden oder rezeptive Fertigkeiten, so bieten sich indirekte Formate an, die die zu elizitierende Leistung über Indikatoren erfassen. Denn im sprachlichen Bereich ist es häufig nicht der Fall, dass sich Testperformanz und zu messende Fertigkeit decken würden.[101] Man denke nur an Leseverstehen – wie sollte der Grad des Leseverstehens erfasst werden, wenn nicht über Indikatoren: Seien es nun MC-Formate oder Fragen, die in offener Form beantwortet werden – in beiden Fällen ist die Testperformanz (Ankreuzen respektive Schreiben) nicht deckungsgleich mit der zu messenden Fertigkeit. Wo immer es aber möglich ist, sprachliche Handlungen und Sprachproduktion direkt zu beobachten, sollte das auch getan werden, um möglichst realitätsnah auf die zugrunde liegende Kompetenz schließen zu können.[102]

An dieser Stelle sei ein kurzer Ausblick auf die Komplexität des Testentwicklungsprozesses gestattet, um die Bedeutsamkeit eines klar definierten Testkonstrukts als Basis aller testbezogenen Überlegungen und Entscheidungen aufzuzeigen: Schon bei der Festlegung der Leistungsdimensionen muss man sich darüber klar werden, wie diese erfasst werden können und welche Testformate die für diesen Fall geeigneten sind. Diese Überlegungen müssen ebenso wie die

[101] Anders als im sprachlichen Bereich ist beispielsweise beim Sport die zu messende Fertigkeit häufig deckungsgleich mit der Testperformanz: Derjenige ist der beste *Läufer*, der am schnellsten *läuft*.
[102] Vgl. hierzu beispielsweise Hughes 1989, Bachmann 1991a, Brown 1993, u. a..

Begründung der Leistungsdimensionsauswahl im Textkonstrukt spezifiziert werden. Auch die Integration verschiedener Fertigkeiten sollte im Testkonstrukt schon genau definiert werden: Wenn beispielsweise Leseverstehen mittels eines offenen Antwortverfahrens erfasst werden soll, so muss das theoretische Verständnis dessen, was Leseverstehen kennzeichnet, im Konstrukt offen gelegt werden: Zielt der Test nur auf die Erfassung des „reinen" Leseverstehens als rezeptive Fertigkeit ab oder liegt dem Test ein integratives Verständnis von Leseverstehen zugrunde und wird deswegen eine integrierte Erfassung von Leseverstehen und Schreibfertigkeit (als ein möglicher Indikator des Leseverstehens[103]) angestrebt? Im ersten Fall kann man darauf verzichten, etwa die sprachliche Korrektheit der Antworten zu betrachten; vielmehr kann man sich dann auf die Bewertung der inhaltlichen Korrektheit und Verständlichkeit der Antworten konzentrieren. Im letzteren Fall jedoch ist es legitim, die inhaltliche und sprachliche Angemessenheit der Antworten in die Bewertung mit einzubeziehen. Diese integrative Erfassung von Fertigkeiten muss jedoch wie gesagt im Testkonstrukt begründet werden, um auf dieser Basis valide Tests entwickeln zu können: Beispielsweise sollten Fragen zu einem Lesetext als Indikatoren des Leseverstehens die im Testkonstrukt definierten theoretischen Annahmen über Charakteristika des Leseverstehens widerspiegeln, ebenso wie sich die im Konstrukt definierten Merkmale der Textproduktion in den Bewertungskriterien der Antworten zum Text niederschlagen sollten. Zu Beginn der Konstruktion etwa eines Leseverstehenstests müssen also Fragen wie die folgenden beantwortet werden: „Welche Merkmale des Leseverstehens sollen durch welche Testaufgaben operationalisiert werden? Welche Merkmale des Leseverstehens sollen mittels eines offenen Antwortformats überprüft werden und müssen somit in die Bewertungskriterien der Antworten einfließen?" Nur durch eine validen Operationalisierung des Konstrukts kann man dann auf die Kompetenzen rückschließen, die mit diesem Test erfasst werden sollen. Hierbei zeigt sich deutlich, dass das Testkonstrukt der Dreh- und Angelpunkt im Testkonstruktionsprozess ist (Näheres dazu unter Kapitel 2.6 dieser Arbeit).

2.2 Testformate und Auswertungsmöglichkeiten

Ähnlich den Präferenzen für bestimmte Sprachmodelle oder Lehrmethoden sind auch Präferenzen für bestimmte Testformate im jeweils gültigen Paradigma begründet. In der von Spolsky[104] postulierten *pre-scientific period* gab es beispielsweise keine allgemein anerkannten Testgütekriterien oder statistische Kontrollen: Man testete informell, meist mittels Übersetzungen und Schreibaufgaben, die mangels fehlender Gütekriterien oft von fraglicher Qualität waren. Im sich

[103] Beispielsweise könnte man argumentieren, dass sich erst bei der schriftlichen Beantwortung von Fragen zu einem Text zeigt, inwieweit dieser verstanden wurde, denn bei den MC-Formaten bleibt meist ein gewisser „Rate"-Spielraum. Auch finden sich für diese sprachliche Handlung im Schnittbereich der Rezeption und Produktion Beispiele im realen Sprachgebrauch, wenn etwa Anfragen beantwortet werden müssen oder ein Leserbrief zu einem Zeitungsartikel verfasst wird.
[104] Vgl. Kapitel 1.1 und vgl. Farhady (1979: 347ff).

anschließenden Paradigma der *psychometric-structuralistic period* testete man nach dem zugrunde gelegten *Item-and-Process*-Modell, vorwiegend mit *discrete-point tests*, im Glauben, man könne die mittels des strukturalistischen Ansatzes identifizierten sprachlichen Phänomene am validesten mithilfe isolierter Tests erfassen, die nur einen klar umrissenen sprachlichen Teilbereich testen und sich dabei auf formale Aspekte konzentrieren. Man glaubte auf Basis eines additiven Sprachmodells, sich der sprachlichen Kompetenz als Ganzes dadurch nähern zu können, dass man die isoliert erfassten Teilkompetenzen am Ende wieder aufsummierte. Nun wurde aber in der vorliegenden Arbeit in den Kapiteln 1.2.1 *Innersprachliche Organisation* und 1.2.3 *Modell der kommunikativen Kompetenz* im Detail erläutert, warum dieses Postulat im Zeitalter des kommunikativen Paradigmas nicht mehr haltbar ist. Daher werden im Folgenden nur knapp die Charakteristika diskreter Formate und die wichtigsten Argumente aufgeführt, die seit nunmehr 50 Jahren gegen rein diskret ausgerichtetes Testen vorgebracht werden:[105]

- *Additives Modell der Sprache und der Kompetenz*: Das Postulat, dass „das Ganze" in diskrete Teile zerlegt werden könne, die bei ihrer Addition wieder das Ganze ergeben, kann der sprachlichen Organisation nicht gerecht werden: Das Ganze ist etwas anderes als die Summe der Teile; Sprache und ihre innersprachliche Organisation können zutreffender mithilfe des Prototypenansatzes beschrieben werden, welcher dann allerdings verbietet, das Ganze in klar getrennte Einzelteile zu zerlegen.

- *Produktion von dekontextualisierter, isolierter Sprache ohne sozio-pragmatische Funktion*: In der Regel erfassen *discrete-point tests* formale Aspekte der Sprache. CLA und die zugrunde liegende kommunikative Kompetenz können jedoch nicht in möglichst authentischen Kontexten erfasst werden, wenn die entscheidende Rolle von Sprache nicht thematisiert wird, nämlich ihr Auftreten in bestimmten Kontexten und zu bestimmten Funktionen, wobei die Auftrittsbedingungen wiederum die sprachliche Form mitbestimmen, die letztlich über die kommunikative Wirksamkeit entscheidet.

- *Mangelnde Generalisierbarkeit*: Aus isoliert gelösten *discrete-point items*, die isolierte Teilaspekte innerhalb einer Teilfertigkeit erfassen, kann nicht auf die Anwendbarkeit des zugrunde liegenden Wissens oder gar auf die Handlungsfähigkeit im realen Leben geschlossen werden. Die so gewonnenen Einsichten können nicht reliabel und valide auf eine generellere, übergreifende Kompetenz – und sei es nur innerhalb eines Fertigkeitsbereichs – verallgemeinert werden. In diesem Zusammenhang fordern folgende Probleme nähere Betrachtung:

  • *Problem der validen Operationalisierung*: Das den *discrete-point items* zugrunde liegende Konstrukt repräsentiert die kommunikative Realität nicht in ausreichender Weise. Deshalb kann von *discrete-point items* nicht auf authentische Sprachproduktion geschlossen werden, da diese meist gar nicht gefordert ist. Beispielsweise können so

[105] Vgl. hierzu beispielsweise Farhady 1979, Oller 1975, Carroll 1961 u. a..

genannte *sentence transformations* nicht als Indikatoren einer generellen Schreibfertigkeit genutzt werden, da sie nur einen minimalen Bereich der komplexen Leistungsdimension der Textproduktion abdecken und nicht authentischen Sprachgebrauch widerspiegeln.

• *Problem der repräsentativen Auswahl*: Die Auswahl spezifischer Phänomene in isolierten Teilbereichen wird immer gewisse Einschränkungen mit sich bringen: *Discrete-point items* können aufgrund ihres engen Gegenstandsbereichs die Leistungsdimensionen eines Tests nur begrenzt repräsentieren; die Abdeckung aller relevanten Facetten der jeweiligen Leistungsdimensionen impliziert jedoch in der Regel eine nicht mehr administrierbare Anzahl an *items*.

Da diese Probleme des Testens mittels *discrete-point items* für ein valides Erfassen von Sprachvermögen überwunden werden mussten, wurden neue Formate unabdingbar. Deren Entwicklung wird im Folgenden dargestellt.

2.2.1 Integrative Formate

Um den genannten Einwänden gegen diskrete Testformate Rechnung zu tragen, wurden ab den 70er Jahren integrative Formate entwickelt, die das neue Paradigma des kommunikativen Zeitalters widerspiegelten: Gesucht wurden Tests, "… that focus on the total communicative effect of an utterance rather than its discrete linguistic components" (Farhady 1979: 348), oder, wie Oller (1979: 37) es ausdrückt: "The concept of an integrative test was born in contrast with the definition of a discrete point test. If discrete items take language skill apart, integrative tests put it back together." Auch wenn dort von Kontrast die Rede ist, so dürfte es zutreffender sein, von einem Kontinuum zu sprechen, das an einem Ende das Extrem absolut diskreter Tests aufweist, wohingegen sich am anderen Ende des Spektrums hochintegrative Formate befinden, die sich dem globalen Sprachstand nähern wollen. Dazwischen sind alle Arten von Abstufungen denkbar. Beispiele für integrative Testformate sind das Diktat, aber auch der *cloze*-Test oder der C-Test. Wichtig ist zu bedenken, dass im historischen Ablauf integrative Formate nicht einfach an die Stelle der diskreten traten, sondern dass man sie als Erweiterung und Ergänzung zu bisherigen Formaten sah, wie Ingram (1978: 12) schreibt:

> It is, in any case, quite unnecessary to suppose that one has to make an either/or choice, that if one approves of integrative tests, one should therefore disapprove of discrete-point ones. This 'disjunctive fallacy,' as Carroll calls it, stems, it seems to me, from misunderstanding about the nature of language command.

Die Wahl des geeigneten Testformats hängt demnach vom jeweiligen Testkontext ab, davon, was der Test erfassen soll, in welcher Art und Weise, zu welchem Zweck und in welcher Zielgruppe.

Nun sind aber integrative Formate denkbar, die auf mehr als einen eng begrenzten Teilbereich der Sprache oder Grammatik abzielen, dennoch aber nicht den sozio-pragmatischen

Auftrittsbedingungen von Sprache und den Forderungen nach Authentizität und Lebensnähe gerecht werden. Deshalb trat neben die Forderung nach Integration mehrerer sprachlicher Teilbereiche die Forderung nach Einbettung der Tests in reale sozio-pragmatische Bedingungen und sinnstiftende kommunikative Kontexte, die der Funktionalität von Sprachgebrauch Rechnung tragen. Oller entwickelte zwei Kriterien, *naturalness criteria* genannt, die forderten, dass die so genannten pragmatischen Tests in bedeutungsvollen Kontexten angesiedelt werden und unter authentischen Bedingungen ablaufen (Oller 1979: 38):

> It [i.e. a pragmatic test] is any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language, and which requires the learner to relate sequences of linguistic elements … to extralinguistic context.

Im Gegensatz zu den diskreten Tests sind integrativ-pragmatische Tests darauf ausgelegt Situationen vorzugeben, die der Realität von Sprachverwendung möglichst nahe kommen, Sprachverarbeitung in bedeutungsvoller Weise verlangen und möglichst mehrere Fertigkeiten oder Teilaspekte dieser Fertigkeiten gleichzeitig ansprechen, innerhalb welcher möglichst mehrere Komponenten abgedeckt werden sollten. Denn die entscheidende Frage beim Testen lautet, wie man von den Testergebnissen auf eine valide und reliable Generalisierung für die jeweilige Zielgruppe kommen kann – wie sollen beispielsweise Satztransformationen vom Aktiv ins Passiv oder umgekehrt verallgemeinert werden, wo sich Aktiv und Passiv gerade nicht entsprechen, sondern jede der beiden *voices* ihre ganz eigenen Auftrittsbedingungen besitzt? Und wie könnte man das Übersetzen isolierter Wortlisten[106] verallgemeinern auf die Anwendbarkeit dieser Wörter in echten Kommunikationssituationen, in denen Wortschatzelemente nicht in Isolation auftreten, sondern in denen sich die jeweilige Bedeutung eines Wortes und somit auch die Wortwahl immer auch aus dem sprachlichen wie außersprachlichen Kontext ergibt, dem wiederum das kommunikative Ziel inhärent ist?

Die Generalisierbarkeit von Testergebnissen hängt unter anderem[107] davon ab, ob ein gegebener Test den Gegenstandsbereich so valide und authentisch erfasst, dass die Ergebnisse auf das Sprach- und Kommunikationsvermögen der Probanden im realen Sprachgebrauch hin verallgemeinert werden können. Dazu müssen Tests jedoch den realen Auftrittsbedingungen von kommunikativen Sprachhandlungen Rechnung tragen, wie es die kommunikativen Formate tun.

[106] Wie es etwa in informellen Lehrervokabeltests noch manchmal der Fall ist. Ein etwas anders gelagerter Fall findet sich beim Einstufungstest des DIALANG-Projekts: In der Pilotphase wurden dabei isolierte Listen von realen und erfundenen Wörtern vorgegeben; die Probanden sollten zwischen diesen beiden Wortgruppen unterscheiden (vgl. http://www.ling.lancs.ac.uk/dialang/pilotsite, Zugriff am 13.05.2002). Hier zeigt sich wieder, dass *discrete-point tests* zum Beispiel als grobe Einstufungstests, die leicht administrierbar und sogar computergestützt auswertbar sind, ihre Berechtigung haben – solange sie nicht auf Aussagen verallgemeinert werden, die mit diesem Format nicht erfasst werden können.

[107] Auf die Bedingungen der Validität und Reliabilität, die ebenfalls Auswirkungen auf die Generalisierbarkeit haben, wird unter Kapitel 2.3 dieser Arbeit eingegangen.

2.2.2 Kommunikative Formate

Neben den gerade beschriebenen Anforderungen der pragmatisch ausgerichteten Tests, namentlich der Berücksichtigung der Erfordernisse und Einschränkungen authentischer Kommunikationssituationen und der bedeutungsvollen Bezugsetzung von Sprache zum außersprachlichen Kontext, muss auch die kommunikative Funktion von Sprache, die Wirkung beachtet werden.[108] Bei den kommunikativ ausgerichteten Testformaten liegt das Hauptaugenmerk nicht mehr auf der Erfassung eines „undifferenzierten" *global language factor*, sondern es geht darum, sich der Komplexität der kommunikativen Kompetenz zu nähern. Deshalb steht die Konstruktion solcher Testformate im Mittelpunkt, die möglichst viele Aspekte der kommunikativen Kompetenz erfassen können, die in bedeutungsvollen Kontexten möglichst authentische Sprachverarbeitung fordern, indem sie ein kommunikatives Ziel vorgeben, das die Probanden erreichen sollen. Bachmann stellt folgende vier Kriterien für kommunikative Tests auf:

> First, such tests create an "information gap," requiring test takers to process complementary information through the use of multiple sources of input. Test takers, for example, might be required to perform a writing task that is based on input from both a short recorded lecture and a reading passage on the same topic. A second characteristic is that of task dependency, with tasks in one section of the test building upon the content of earlier sections, including the test taker's answers to those sections. Third, communicative tests can be characterized by their integration of test tasks and content within a given domain of discourse. Finally, communicative tests attempt to measure a much broader range of language abilities – including knowledge of cohesion, functions, and sociolinguistic appropriateness – than did earlier tests, which tended to focus on the formal aspects of language – grammar, vocabulary, and pronunciation. (Bachmann 1991b: 678)

Während die Forderung nach einer Informationslücke, die es zu schließen gilt, aufgrund ihrer Lebensnähe einleuchtet und operationalisierbar ist, stellt die zweite Forderung aus testtheoretischer Sicht ein Problem dar: Wenn *tasks* voneinander abhängig sind und gar auf Antworten der Probanden in vorangehenden Testteilen aufbauen, so ist die statistisch geforderte Unabhängigkeit der Testitems nicht mehr gegeben – dies könnte im negativsten Fall zu einem verzerrten Bild der Leistung führen, im positiven Fall jedoch zu einem realitätsnahen Bild des Probandenkönnens, denn auch in der Realität bauen kommunikative Handlungen aufeinander auf: Beispielsweise liest man einen Zeitungsartikel, ehe man sich in einem Leserbrief auf den Artikel bezieht – das Verstehen ist dabei der Produktion vorgeschaltet, weshalb letztere von ersterem abhängig ist. Wesentlich ist gerade bei voneinander abhängigen Testitems, im Testkonstrukt zu definieren, welche Gegenstandsbereiche auf welche Weise erfasst werden sollen, um zu einer validen Generalisierung zu kommen. Das dritte Bachmannsche Kriterium zeugt von der Verankerung des kommunikativen Testens u. a. im linguistischen Ansatz der Diskursanalyse, der bestimmten Domänen bestimmte Diskurstypen zuordnet. Indem man diese Domänen und die damit verbundenen unterschiedlichen Diskurtypen und Sprachfunktionen bei der Konstruktion von Testitems berücksichtigt, bildet man reale Kommunikationssituation ab und kommt somit der Forderung nach authentischen Tests näher. Das Kriterium der Breite der abgedeckten Gegenstandsbereiche versteht sich von selbst – je

[108] Vgl. beispielsweise Bachmann 1991a u. a..

breiter ein Test angelegt ist, je umfassender er die zugrunde liegende kommunikative Kompetenz „anzapft“ und je mehr „Fenster“ er auf die zu testenden Fähigkeiten und/oder Fertigkeiten öffnet, desto umfassender wird auch das Bild, das sich von den Probanden, ihrem kommunikativen Sprachvermögen und dessen Ausprägung ergibt.

Ziel der Beurteilung sollte im Idealfall ein detailliertes Profil der Kompetenzen der Probanden sein. Um sich diesem Ziel zu nähern, sollten alle adäquaten Formate genutzt werden, denn weder rein diskrete noch rein integrative Formate lassen solch eine Profilbildung zu: Wie oben gezeigt, sind *discrete-point items* nur schwer auf den generellen Leistungsstand hin generalisierbar; dennoch lassen sie einen konkreten Blick auf einzelne sprachliche oder fertigkeitsbezogene Teildimensionen zu, wie etwa die Beherrschung der *tenses* im Englischen, die über Lückenformate[109] erfasst werden könnte. Hingegen lassen integrative Formate wie etwa der C-Test nur in bedingtem Maß Rückschlüsse auf einzelne Fertigkeiten zu, eben auf die bei der Lösung des C-Tests zum Einsatz kommenden; dafür geben sie Aufschluss über den generellen Leistungsstand. Beispielsweise könnten rezeptive Fertigkeiten in einem groß angelegten Test aus Praktikabilitätsgründen im *Multiple Choice*-Format erfasst werden, Schreiben und Sprechen über offene, handlungsorientierte Formate wie etwa dem semikreativen Schreiben, und der generelle Sprachstand etwa mittels eines integrativen C-Tests, wie es im DESI-Projekt der Fall ist. Swain (1990: 403) etwa gibt eine übersichtliche Tabelle, aus der Testformate und Testgegenstände eines *proficiency tests* hervorgehen. So kann detailliert dokumentiert werden, auf welcher theoretischen Basis und mit welcher Methode man sich welchen Teilbereichen nähern will.

Nimmt man zum Beispiel Bachmanns Modell der kommunikativen Kompetenz als Ausgangsbasis, so zeigt sich, dass in einer Testsituation neben den Sprachkompetenzen auch strategische Kompetenzen wirksam sind, in Interaktion mit den Gebrauchsbedingungen der jeweils vorgegebenen Situation, um den Test möglichst erfolgreich zu meistern. Dies verdeutlicht erneut, warum kommunikative Kompetenz nicht zuverlässig mittels diskreter Formate allein erfasst werden kann. Abbildung 3 in Kapitel 1.2.3 dieser Arbeit illustriert den Prozess einer sprachlichen Äußerung und die dabei wirksamen Faktoren, sei die Äußerung nun in der Realität oder in einem Test angesiedelt. Bachmann (1991a: 98ff) bemerkt in diesem Zusammenhang, dass bei authentischer Sprachverarbeitung nicht nur isolierte sprachliche Wissensbestände „angezapft“ werden, sondern dass immer auch komplexere Strategien zum Einsatz kommen, wie etwa das Interpretieren des kommunikativen Ziels, das Einschätzen der Situation, das Planen der Zielrealisierung und die damit verbundene Auswahl der geeigneten sprachlichen Mittel, sowie der Einsatz kommunikativer und natürlich testrelevanter Strategien, um nur einige zu nennen. Deshalb muss vor Beginn der

[109] Dazu werden in einem gegeben Text oder in einzelnen Sätzen die Verben elidiert und in Grundform vorgegeben. Diese Verben müssen dann in der korrekten *tense* in die Lücken eingesetzt werden.

eigentlichen Testentwicklung im Testkonstrukt deutlich gemacht werden, was erfasst werden soll, damit man nicht Gefahr läuft, statt Sprache beispielsweise Intelligenz[110] zu testen.

2.2.3 Auswertungsmöglichkeiten der verschiedenen Formate

Jedes Testformat verlangt aufgrund seiner Besonderheiten nach einem je anders gearteten Auswertungsschema, denn nur wenn Testkonstrukt, Testformat und Auswertungsschema aufeinander abgestimmt sind, wird der Test zu validen Ergebnissen führen. Die Entscheidung über das Auswertungsschema hängt vom Testgegenstand, vom Testformat und von den Testzielen ab; sie hat Auswirkung auf die Generalisierbarkeit und die Rückmeldungsmöglichkeiten der Testergebnisse.

In diesem Zusammenhang können grundsätzlich zwei Typen der Beurteilung unterschieden werden, die das Auswertungsschema beeinflussen: die des normorientierten und die des kriterienorientierten Testens. Das normorientierte Testen nimmt die Probandengruppe als Bezugspunkt – wo steht das Individuum in Bezug zu einer bestimmten Gruppe? Dabei können keine Aussagen darüber getroffen werden, wie „gut" das Individuum ist oder wo es sich in Bezug auf ein bestimmtes Kriterium befindet, denn bei der Auswertung des Tests geht es darum, die Probanden innerhalb der Gruppe in eine Reihenfolge zu bringen, sie zu *ranken*. Dieses *Ranking* erfolgt oft gemäß einer angesetzten Normalverteilung. Die Frage danach, was die Probanden tatsächlich in der Fremdsprache „können", ist bei ausschließlich normorientierter Herangehensweise nicht zu beantworten. Ein Vergleich mehrerer Gruppen wird dabei nahezu unmöglich, eben da es an vergleichbaren Kriterien mangelt. Auch wenn beispielsweise Proband A in Gruppe X der beste ist, so weiß man nichts darüber, wie gut er in Gruppe Y abschneiden würde. Das kriteriumsorientierte Herangehen dagegen bietet den Vorteil, dass in Bezug auf ein bestimmtes Kriterium (das natürlich in die Testkonstruktion einfließen muss) festgestellt werden kann, inwieweit dieses Kriterium von den jeweiligen Probanden erfüllt wird, ohne dabei die Gruppe als Bezugsnorm anzusetzen. Dieses Herantreten bietet die Möglichkeit Vergleiche zu ziehen, da man mithilfe der besagten Kriterien einen gemeinsamen Referenzpunkt hat. Gerade bei Vergleichsuntersuchungen und Schulleistungsstudien wird kriteriumsorientiert an die Auswertung herangegangen, um feststellen zu können, welche Kriterien die Lernenden erfüllen und was sie schon können. Die beiden Herangehensweisen schließen sich nicht gegenseitig aus und sollten je nach Bedarf und Angemessenheit eingesetzt und kombiniert werden, um den jeweiligen Testzwecken und Zielen so gerecht wie möglich zu werden.

Neben Norm- und Kriterienorientierung kann man Bewertungsverfahren für geschlossene von solchen für offene Testformate unterscheiden. Auf Verfahren im Rahmen geschlossener Formate

[110] Es ist seit langem umstritten, ob die so genannte *language proficiency* Teil der generellen Intelligenz ist oder ob es sich hierbei um eine eigenständige Fertigkeit handelt (vgl. etwa Carroll 1972 und Oller 1976). Da ich persönlich auch hier zu einem systemischen Ansatz neige um die Struktur von Intelligenz zu beschreiben, sehe ich keinen Widerspruch zwischen den beiden hier angerissenen Positionen: Sprachlich-kommunikative Intelligenz ist eingebettet in die übergeordnete generelle Intelligenz, doch kann sie auf der nächstuntergeordneten Ebene als eigenständiges System beschrieben werden.

soll hier nur am Rand eingegangen werden, denn wenn es nur eine korrekte Lösung für ein Testitem gibt, so ist diese unkompliziert zu bewerten; mittels Zählen der korrekten Lösungen kann der Leistungsstand dann über den ganzen Test hinweg ermittelt werden. Solch quantitative Zählverfahren kommen meist bei *items* eines *discrete-point tests* zum Einsatz oder bei der Erfassung rezeptiver Fertigkeiten über Indikatoren. Aber auch integrative geschlossene Formate wie beispielsweise der C-Test können über quantitative Verfahren ausgewertet werden. Ob dann Rohpunkte rückgemeldet werden oder ob die Testitems einer Skala zugeordnet werden, die mittels ihrer Deskriptoren inhaltliche Rückschlüsse auf den Leistungsstand in Bezug auf die vom Test erfassten Leistungsdimension zulässt, hängt vom jeweiligen Testkonzept ab. Skalierungen von Testitems beruhen auf psychometrischen Rechenmodellen wie beispielsweise dem Rasch-Modell[111]. Dabei werden die Schwierigkeiten der Testitems aufgrund ihrer Lösungshäufigkeiten mittels probabilistischer Rechenmodelle abgeschätzt und die *items* auf einer Skala in aufsteigender Schwierigkeit angeordnet. Die Probandenfähigkeiten werden auf Basis der so ermittelten Schwierigkeiten der Testitems und des Verhaltens der jeweiligen Probanden abgeschätzt und auf derselben Skala dargestellt. Auf konkretere Fragen der Skalierung kann im Rahmen dieser Arbeit jedoch nicht näher eingegangen werden, da Messmodelle der Psychometrie wie gesagt nicht Gegenstand dieser Arbeit sind.

Gegenstand dieser Arbeit ist dagegen die valide Bewertung offener Aufgabenstellungen, bei der es nicht um die Entscheidung „korrekte Antwort oder nicht?“ geht, sondern bei der es um die Bewertung der qualitativen Seite der Antwort geht. Bei rezeptiven Aufgaben mag es sinnvoll sein, diese in ihren Charakteristika zu beschreiben und nach ihren Schwierigkeiten einzustufen, so dass die korrekte Bearbeitung eines Testitems schon Rückschlüsse auf den Leistungsstand geben kann, doch bei offenen Aufgabenstellungen ist das in dieser Art nicht möglich.[112] Dabei ist es ratsamer, sich zur Ermittlung des Leistungsstands auf die jeweilige Performanz zu konzentrieren statt auf den Stimulus. Dazu bedarf es valider Bewertungsinstrumente, die das jeweilige Testkonstrukt widerspiegeln und den Bewertungsprozess unterstützen und erleichtern. Hierbei kommen qualitative Verfahren zum Einsatz, die jedoch nicht über das Maß an Objektivität verfügen, wie es quantitative Verfahren tun. Folgende Gegenüberstellung der extremen Ausprägungen quantitativer und qualitativer Auswertungsverfahren (nach Pollitt 1991a: 52 und Pollitt & Murray 1996) macht die jeweiligen Charakteristika deutlich:

[111] Für eine knappe Erläuterung des Rasch-Modells darf auf das Glossar dieser Arbeit verwiesen werden. In statistischen Zusammenhängen wird von Probanden*fähigkeiten* gesprochen, welche dem Begriff des kommunikativen Sprach*vermögens* entsprechen.
[112] Vgl. hierzu beispielsweise Pollitt (1991a: 88) oder Alderson 1991a.

	Counting	**Judging**
Procedure	add up scores: counting strategies	rate a performance: judging strategies
Technical Concern	difficulty of item	precision of rating criteria
Assumption	all acceptable performances are equally good	all tasks are equally difficult
Criteria	quantity of acceptable performance	quality of acceptable performance
Focus	stimulus; ranked by difficulty	response, performance; described in rating scale
Advantages	objectiveness of scoring	qualitative description of observable performance or behaviour
Disadvantages	ignoring qualitative side of performance; scores without direct relation to assessment or performance	subjectiveness of assessment; ignoring quantity and task difficulty

Tabelle 3: Gegenüberstellung quantitativer und qualitativer Auswertungsverfahren

In der Realität zeigen sich die Ausprägungen nicht so extrem: Beispielsweise können die Schwierigkeiten offener Aufgaben nicht eindeutig bestimmt werden, so dass man in der Regel nicht davon ausgeht, dass alle Aufgaben gleich schwer sind. Es gibt Messmodelle, die der Aufgabenschwierigkeit, der Strenge der Bewerter und den Schülerleistungen Rechnung tragen.[113] Zudem werden in den meisten Beurteilungen sowohl die quantitative als auch die qualitative Seite der Leistung berücksichtigt: Beispielsweise geben rein quantitative Fehlerzählverfahren weder Rückschlüsse auf das Können noch bieten sie Lernanreize, so dass es sich anbietet, solche Verfahren durch qualitative Fehleranalysen zu ergänzen. Schon aus den wenigen, hier genannten Gründen bietet sich eine Kombination verschiedener Auswertungsverfahren an, um die genannten Nachteile möglichst zu minimieren und die Vorteile nutzen zu können.

Pollitt & Murray (1996: 75f) geben zwei Vorschläge, wie die quantitative und die qualitative Seite bei der Auswertung zusammengeführt werden können:

(a) Man kann genau spezifizierte und in ihren Schwierigkeiten definierte Testaufgaben, die durch ein Zählverfahren ausgewertet wurden, skalieren, so dass der Zusammenhang zwischen Probandenfähigkeit und Aufgabenschwierigkeit sichtbar wird. Durch die Spezifizierung der Aufgaben werden die mit dem Test erfassten Leistungsdimensionen in Zusammenhang gebracht mit den Probandenfähigkeiten, so dass auch qualitative Aussagen möglich werden. Dies ist ein Verfahren, das sich bei geschlossenen Aufgabenstellungen anbietet.[114]

(b) Es bietet sich bei offenen, produktiven Aufgaben an, neue Wege der Testentwicklung zu beschreiten: Man administriert in der Prätestphase einen Test, der die zu testenden Fertigkeiten elizitiert. Die so gewonnenen Performanzbeispiele werden in Deskriptoren beschrieben, die wiederum in den Testkonstruktionsprozess rückfließen und die bei diesem Test zu erwartende Performanz beschreiben. Auf dieser Basis muss ein Interpretationsschema entwickelt werden, das den Leistungen in diesem Test Performanzniveaus zuweist, welche dann mit Außenkriterien oder existierenden Standards in Zusammenhang gebracht werden können, um von der

[113] Vgl. beispielsweise Lumley (2002: 251), der auf die *multi-faceted Rasch analysis* mit der Software *FACETS* verweist, die von Linacre & Wright (1992-1996: *FACETS*. Chicago, IL: MESA Press) entwickelt wurde. Lumley verweist auch auf McNamara 1996 und Weigle 1998, die die Anwendung der Rasch- Analyse in der Sprachbeurteilung beschreiben.

[114] Dieser Weg wird bei den geschlossenen Aufgabenstellungen in DESI gegangen. Auch die Vorgehensweise (b) bei offenen Aufgabenstellungen wird in DESI genutzt und wird hier in Kapitel 4 dokumentiert.

Testperformanz auf Fähigkeiten im realen Leben rückschließen zu können. So können quantitative Einstufungen mit qualitativen Beschreibungen zusammengebracht werden.

Offene Formate können traditionell quantitativ nach Fehlern beurteilt werden, doch in jüngerer Zeit gewinnt der Ansatz der qualitativen Positivbewertung an Bedeutung. Traditionell beurteilte man eine Leistung nach Fehlerhäufigkeit und eventuell nach Schwere der Fehler. Es gibt bestimmte Situationen, in denen dieses Herantreten Sinn macht – man denke etwa an diagnostische Lehrertests, die Aufschluss geben sollen, ob bestimmte Strukturen erlernt worden sind beziehungsweise wo es noch Defizite gibt, um das weitere Vorgehen zu planen. Dennoch ist es aus mehreren Gründen[115] ratsam, die Negativkorrektur um Positivansätze zu ergänzen: Reines Fehleranstreichen wirkt demotivierend, da es den Lernenden nicht zeigt, was sie schon können. Es kann keine Aufschlüsse über Lernfortschritte oder das Erreichen bestimmter (positiv formulierter) Kriterien geben und ist somit als alleiniges Vorgehen bei Sprachstandstests unangemessen. Angemessenere Verfahren sind im Ansatz der Positivbewertung zu finden: Durch das positive Herantreten an eine Leistung wird die Qualität derselbigen und das Können der Lernenden in den Mittelpunkt gerückt, ohne dass das Fehlende aus dem Blickfeld gerät. Der Positivansatz gibt Aufschluss über das Erreichen bestimmter Kriterien oder Standards und über den Sprachstand, so dass den Lehrenden qualitative Einblicke und den Lernenden motivierende Lernanreize geboten werden können.[116]

Welcher Weg letztlich eingeschlagen wird, hängt vom jeweiligen Testkonzept ab und muss im Einzelfall aus dem Testkonstrukt begründet und entschieden werden. Wichtig ist, dass man ein valides Auswertungsschema entwickelt, das die Leistungsdimensionen erfasst, welche der Test messen will. Letztlich wird der Testzweck mit entscheiden, welche Dimensionen man auf welche Weise auswerten und rückmelden will.

2.3 Testgütekriterien

Spätestens seit der o. g. psychometrisch-strukturalistischen Ära gibt es klar definierte Anforderungen, denen Tests genügen müssen, um einen bestimmten Qualitätsstandard beim Testen zu gewährleisten. Die Kriterien der Durchführbarkeit und Praktikabilität, Objektivität, Reliabilität, Validität und Generalisierbarkeit sind dabei die relevantesten.[117] Diese Kriterien spielen schon in der ersten Planungsphase der Testerstellung eine wichtige Rolle, denn nur wenn sie mit in den Prozess der Testerstellung einfließen und fortlaufend überprüft werden, wird der Test am Ende den an ihn gestellten Anforderungen genügen können. Bei der empirischen Qualitätskontrolle

[115] Vgl. hierzu beispielsweise Alderson 1991a, Bleyhl 2003, Börner 1989, Hamp-Lyons & Kroll 1996, Pollitt 1991a, u. a..
[116] Im Rahmen des Positivansatzes werden in Kapitel 3.3 *Rating-Verfahren* vorgestellt. In Kapitel 4 wird an einem konkreten Beispiel ein dem Positivansatz verpflichtetes Bewertungsverfahren begründet und entwickelt.
[117] Für einen Überblick vgl. beispielsweise Brown 1994.

kommen neben inhaltlichen und theoretischen Überlegungen Verfahren der Psychometrie zum Einsatz. Auf diese soll nur insoweit eingegangen werden, wie sie zur Erläuterung relevanter Konzepte benötigt werden. Diese Arbeit erhebt nicht den Anspruch, Aussagen über die Angemessenheit des einen oder anderen statistischen Verfahrens zu treffen.

Ein Test muss, um überhaupt stattfinden zu können, durchführbar und praktikabel sein, sei es hinsichtlich seiner Machbarkeit oder hinsichtlich seiner Administration: Wenn die zur Verfügung stehenden Ressourcen nicht korrekt bedacht wurden oder der finanzielle Rahmen nicht eingehalten wurde, wird es nicht zum Testlauf kommen. Ebenso wenig durchführbar ist ein Test, wenn er „Unmögliches" von den Probanden verlangt oder sie vor unlösbare Fragen stellt – die Durchführbarkeit stellt einen Aspekt der Testvalidität dar, welche im Anschluss unter Kapitel 2.3.1 *Aspekte der Validität* besprochen wird.

Die Forderung nach objektiver Leistungsmessung dürfte sich von selbst verstehen, denn eine rein subjektive Beurteilung ist nicht reproduzierbar und damit auch nicht verallgemeinerbar, ganz abgesehen davon, dass sie keine akzeptable Art des Testens darstellt.[118] Gerade im Testentwicklungsprozess gibt es viele Entscheidungen, die letztlich nur subjektiv getroffen werden können; dennoch ist eine Objektivierung anzustreben, und sei es beispielsweise nur mittels des Offenlegens von Entscheidungsgründen. Entscheidungen über Inhalte, Formate, Verfahrensweisen u. Ä. sollten deshalb von Anfang an transparent dokumentiert werden. Beim direkten Testen mittels offener Aufgabenstellung nimmt die Frage der Auswertungsobjektivität eine zentrale Rolle ein: Wenn keine gemeinsamen Kriterien und standardisierten Verfahren festgelegt werden, so wird die Auswertung so subjektiv ausfallen, dass der Test nicht mehr zuverlässig misst.

An dieser Stelle zeigt sich, dass das Kriterium der Objektivität eng mit dem der Reliabilität verbunden ist, denn ein Test gilt dann als reliabel, wenn er unter gleichen Bedingungen zuverlässig vergleichbare Ergebnisse liefert, also reproduzierbar ist. Wenn ein Test zu subjektiv misst oder zu subjektiv ausgewertet wird, so wird er keine vergleichbaren, zuverlässigen Ergebnisse bringen. Ein Test jedoch, der bei jeder Durchführung zu anderen Ergebnissen kommt, hat nicht nur Mängel in seiner Reliabilität, er besitzt auch keine Validität. Denn wenn mehrere Testdurchführungen jeweils zu anderen Ergebnissen führen, lässt dies darauf schließen, dass bei jedem Testlauf etwas anderes gemessen wird.

Das Reliabilitätskriterium kann auch auf die einzelnen Testitems innerhalb eines Tests bezogen werden. Das bedeutet, dass die *items* ein gewisses Maß an Homogenität aufweisen müssen, um innerhalb des gegebenen Testkonstrukts zuverlässig zu messen. Dieses Kriterium wird mittels statistischer Verfahren – beispielsweise der Berechnung von Cronbachs Alpha,

[118] In diesem Zusammenhang sei auf die Testfairness verwiesen: Wenn die Probanden nach subjektiven Maßstäben bewertet werden, kann es zu unterschiedlichen und damit unfairen Bewertungen kommen, wenn etwa persönliches Missfallen die Bewertung beeinflusst.

einem Korrelationsindex – geprüft.[119] Auf die Bedeutsamkeit von Korrelationen und den Zusammenhang zwischen Reliabilität und Validität wird bei den folgenden Ausführungen zur Validität integrativer und kommunikativer Formate noch näher eingegangen.

Wie man zu einem reliablen und möglichst objektiven Bewertungsschema bei offenen Schreibaufgaben kommen kann, wird in Kapitel 4 dieser Arbeit beschrieben.

2.3.1 Aspekte der Validität

Das Kriterium der Validität eines Tests bezieht sich darauf, ob der Test auch das misst, was er messen soll. Dieses Kriterium hängt eng mit den gerade erwähnten Kriterien der Objektivität und Reliabilität zusammen und setzt teils deren Erfüllung voraus. Beispielsweise kann ein Test immer nur so valide sein wie er reliabel ist, denn nur wenn er zuverlässig misst, kann man prüfen, inwieweit der Test auch das erfasst, was intendiert war; ein nicht reliabler Test wird gar nicht erst zur Validitätsprüfung kommen. Validität ist ein komplexes Konzept und umfasst verschiedene Aspekte, die in der Testsituation wirksam sind. Es spielen Testziel und Testzweck, Testkonzept, Testgegenstand (Inhalte und abgedeckte Bereiche), Format, Aufgabentyp, Bekanntheitsgrad der Instrumente, Zusammensetzung und Hintergrund der Probanden und ggf. die Art des Unterrichts (falls der Test in einem unterrichtlichen Kontext situiert ist) eine Rolle, um nur einige Facetten zu nennen.

Deshalb hat man traditionell verschiedene Aspekte der Validität unterschieden und sie mit jeweils angemessenen Methoden untersucht. Beispielsweise wurde die Inhaltsvalidität von der Konstruktvalidität unterschieden: Erstere bezieht sich auf das, was der Test inhaltlich abdecken soll, welche Teilbereiche er erfassen soll und mit welchen Formaten und Typen dies erreicht werden soll; letztere bezieht sich darauf, ob der Test in seinen Ergebnissen das ihm zugrunde liegende theoretische Konstrukt widerspiegelt, ob die Testergebnisse angemessen und bedeutungsvoll interpretiert werden können.[120] Letzterer Aspekt wird in der Regel mittels empirischer Methoden ermittelt, immer in Bezug auf den Zweck des jeweiligen Tests. Beispielsweise sollten zwei Tests, die ein ähnliches Konstrukt der Schreibfertigkeit auf jeweils unterschiedlichem Weg zu messen behaupten, innerhalb einer Probandengruppe auch ähnliche Ergebnisse erzielen, selbst bei unterschiedlichen Formaten, während ein Hörverstehenstest in dieser Probandengruppe zu deutlich anderen Ergebnissen kommen sollte. Konstruktvalidität ist also ein Indikator dafür, inwieweit die zu messenden Fähigkeiten auch vom Test abgedeckt werden, inwieweit

[119] Es darf auf die Fußnote 97 zum Begriff der Dimensionalität verwiesen werden: *Items* innerhalb eines Tests sind dann reliabel, wenn sie dasselbe zugrunde liegende Konstrukt erfassen. Dies zeigt sich daran, dass die *items* miteinander so hoch korrelieren, dass sie eine statistische Dimension bilden. Man könnte auch sagen, dass diese Dimension dem Konstrukt der Leistungsdimension entspricht, das der Test erfassen soll. In einem *discrete-point test* unterscheiden sich psychometrische und psycho-linguistische Dimension deshalb nicht, weil der Test darauf ausgelegt ist, genau einen spezifischen Teilaspekt zu erfassen. Wenn der Test jedoch integrativ ausgerichtet ist, so kann er psychometrisch dennoch eindimensional sein, obwohl oder gerade weil er mehrere interagierende Teildimensionen integrativ erfasst.

[120] Vgl. hierzu beispielsweise Bachmann (1996: 23ff).

also die Testergebnisse verallgemeinert werden können. Im Rahmen der Validitätsprüfung sollte auch erfasst werden, ob und inwieweit andere als die vom Test intendierten Wissensbestände oder Strategien bei der Testbearbeitung zum Ziel führen. Dies kann durch Prätests oder Prozessanalysen während der Testentwicklung geschehen.

Um das Kriterium der Inhaltsvalidität zu prüfen, werden inhaltliche Analysen der Tests durchgeführt: Beispielsweise wird empirisch untersucht (etwa durch Prozessanalysen) und beschrieben, welche Teilkompetenzen in welcher Art und Weise zur Lösung der jeweiligen Testitems benötigt werden; auch die Testitems können durch Merkmale beschrieben werden, die die Schwierigkeiten und Anforderungen der jeweiligen *items* charakterisieren. Die Anmerkungen oben bei den zu erfassenden Leistungsdimensionen spielen hier wieder herein: Nur auf der Basis einer Spezifizierung der Dimensionen, Inhalte und geforderten Sprachhandlungen des Tests, und auf Basis der genauen Beschreibung der Merkmale, die den betreffenden Test charakterisieren, können valide Tests entwickelt werden. Man denke dabei auch an den oben erwähnten Einfluss der strategischen Kompetenzen auf die Testperformanz: Nur wenn dieser, wie andere „Störvariablen“ auch, hinreichend erkannt und kontrolliert wird, können Testergebnisse auf die Dimensionen hin verallgemeinert werden, die sie erfassen. Eine sorgfältige Dokumentation aller Entscheidungsgründe, insbesondere bei der Auswahl der Leistungsdimensionen, trägt zur Transparenz der inhaltlichen Validität bei. Denn jeder Test wird zwangsläufig eine Auswahl treffen müssen hinsichtlich dessen, was er im jeweils gegebenen Rahmen erfassen kann. Nicht nur der Testgegenstand, auch die Testitems stellen letztlich eine Auswahl dar. Nur wenn diese begründet erfolgt und eine für den Test und seinen Kontext repräsentative Auswahl darstellt, wird der Test dem Kriterium der Inhaltsvalidität standhalten können. Das *sampling*, die Auswahl repräsentativer *items*, findet aber nicht nur nach inhaltsvaliden Gesichtspunkten statt, sondern auch aufgrund der schon erwähnten Reliabilitätsprüfungen der *items* innerhalb eines Tests. Das *sampling* wird letztlich auch über die Generalisierbarkeit des betreffenden Tests entscheiden – seine Bedeutsamkeit darf nicht unterschätzt werden. Hierbei darf man das theoretische Modell, das die Basis eines Tests bildet, nicht aus den Augen verlieren: Soll ein Test beispielsweise die kommunikative Kompetenz messen, so tut man gut daran, auch alle relevanten Teildimensionen mit zu bedenken und entsprechend des jeweiligen Testkonzepts zu operationalisieren. Anderenfalls wäre die Validität und Generalisierbarkeit der Testergebnisse fraglich.

Das traditionelle Validitätskonzept wurde immer wieder erweitert, beispielsweise von Oller (1979: 50-69), nicht zuletzt, um es auf integrative Formate anzuwenden. Ausgehend von seinem Begriff des Sprachgebrauchs als einem Prozess interagierender Pläne und Hypothesen bezüglich des Verknüpfens von linguistischem mit außersprachlichem Kontext, und ausgehend von der *Interlanguage*-Hypothese des Spracherwerbs schlussfolgerte er, dass valide Sprachtests eben diese *Interlanguage* herausfordern und elizitieren müssen, um Aussagen bezüglich Stand und Effektivität der *Interlanguage* bzw. bezüglich der Kompetenzen der Lernenden machen zu können. Er setzte die Kriterien der Inhalts- und Konstruktvalidität an, erweiterte aber die inhaltliche

Validität um den Aspekt, ob der Test auch das sprachliche Verhalten elizitiere, das im normalen Sprachgebrauch mit der betreffenden Fertigkeit einhergeht: Führen die Probanden die entsprechende Tätigkeit auch im Test aus? Streng genommen hätten etwa Leseverstehensaufgaben im *Multiple Choice*-Format bei Oller keine Inhaltsvalidität. Dennoch wird man nicht an indirekten Testverfahren vorbeikommen, gerade bei den rezeptiven Fertigkeiten, die sich per se einer direkten Beobachtung entziehen, weshalb in der Testkonstruktionsphase besondere Sorgfalt auf inhaltliche Aspekte gelegt werden muss.

Bachmann et al. (1996: 23ff) verstehen zwei weitere Konzepte als wesentlich für die Gütebestimmung von Tests: Authentizität und Interaktivität. Authentizität bezieht sich dabei auf den Grad der Übereinstimmung zwischen tatsächlichem Sprachgebrauch in der Zielsprache und dem Sprachgebrauch in der jeweiligen Testsituation – nur wenn sich die reale Welt in der Testsituation widerspiegelt, lassen sich in diesem Rahmen auch Generalisierungen ableiten. Bachmann (ebd.: 29) sieht Authentizität deshalb als mit der Inhaltsvalidität verbunden. Das Kriterium der Interaktivität hingegen bezieht sich darauf, inwieweit bestimmte Charakteristika der Probanden zur Lösung der Testaufgabe involviert sind: “The interactiveness of a given language test can […] be characterized in terms of the ways in which the test taker’s areas of language knowledge, metacognitive strategies, topical knowledge, and affective schemata are engaged by the test task.“ (ebd.: 25). Da sowohl im realen Leben als auch in der Testsituation diese Interaktivität zwischen Testaufgabe und Wissensbeständen variieren könne, müsse diese von der Authentizität unterschieden und in Beziehung zur Konstruktvalidität betrachtet werden: Elizitiert eine bestimmte Aufgabe die sprachlichen Verhaltensweisen, aktiviert sie die sprachlichen, strategischen und außersprachlichen Wissensbestände, die durch das Testkonstrukt auch intendiert sind? Dabei seien Authentizität und Interaktivität immer nur relativ betrachtbar in Bezug auf die Probanden, die reale Sprachgebrauchsituation und die Testsituation (ebd.: 29).

Das Konzept der *face validity*, der augenscheinlichen Validität aus der Perspektive der Probanden und Außenstehender, trägt im weitesten Sinn ebenfalls zur Validität bei: Ist nachvollziehbar, was der Test erfassen soll? Dieses Kriterium wird auch *test appeal*[121] genannt und lässt sich nicht objektiv überprüfen. Es spielt nach Oller (1979: 52) auch nur dann eine Rolle, wenn sich eine augenscheinlich nicht gegebene *face validity* negativ auf die Testleistung auswirken sollte. Ein Beispiel für ein Testformat, dem immer wieder mangelnde *face validity* vorgeworfen wird, ist der C-Test. Seine äußere Form mag Anlass zur Kritik geben, denn wo im Leben würde man Texten begegnen, in denen Wörter teils elidiert sind. Doch betrachtet man die Fertigkeiten näher, die der Test erfordert, so findet man sehr wohl Parallelen zum authentischen Sprachgebrauch: Antizipatorische Sprachverarbeitung und Nutzung von Redundanzen machen einen großen Teil natürlicher Sprachverarbeitung aus.[122] Auch sind reale Situationen denkbar, in denen fehlende Wörter erschlossen werden müssen, wie etwa Lautsprecherdurchsagen an Bahnhöfen oder Flughäfen. Das

[121] Vgl. beispielsweise UGE (1997: 43).
[122] Vgl. etwa Lehmann et al. (1999: 23).

C-Test-Format hat spätestens seit dem Bundeswettbewerb Fremdsprachen in die Schulen Einzug gehalten.[123] Folgt man dem Grundsatz, dass Testformate den Probanden auch bekannt sein sollten um sicherzustellen, dass man nicht etwa rein teststrategische Kompetenzen erfasst, und folgt man dem Grundsatz der Authentizität, so dürfte die *face validity* das geringste Problem darstellen.

Des Weiteren gibt es das Konzept der *concurrent validity*, des Maßes an statistischer Korrelation zwischen den Ergebnissen eines zu validierenden Tests und eines bereits validierten Tests als Außenkriteriums, welcher dieselben Gegenstandsbereiche misst wie der zu prüfende Test. Dies ist ein gängiges Verfahren bei der Testvalidierung. Oller (1979) weist in diesem Zusammenhang auf die enge Beziehung zwischen Reliabilität und Validität hin: Er sieht Reliabilität als Spezialfall der *concurrent validity*, verständlich, wenn man Reliabilität als Maß der Homogenität von *items* innerhalb eines Tests begreift und *concurrent validity* als Maß der Homogenität zwischen zwei oder mehreren Tests, die man sich innerhalb eines übergeordneten Tests denken kann. In beiden Fällen werden Korrelationen als Messverfahren genutzt und als Homogenitätsindikator interpretiert.[124] Das Konzept der *convergent validity* hingegen untersucht, inwieweit zwei Tests in ihren Ergebnissen korrelieren, die beide dasselbe Konstrukt zu messen vorgeben, unabhängig von der jeweils gewählten Testmethode oder dem tatsächlichen Testgegenstand. *Convergent validity* ist somit ein Unteraspekt der Konstruktvalidität.

Ein weiterer Unteraspekt der Validität findet sich im Konzept des *instructional value*, des instruktiven Effekt, den ein Test in der getesteten Zielgruppe beabsichtigt oder unabsichtlich erzielt. Auf diesen Effekt wird in Kapitel 2.4 *Testziele und Zwecke* noch näher eingegangen, da sich dahinter ein testtheoretisch wie sprachen- und bildungspolitisch bedeutsames Konzept verbirgt. Bei Camp (1996: 138) wird dieses Konzept als „systemische Validität“ bezeichnet: “For an assessment to demonstrate ’systemic validity’ (...) it must support instruction and learning in the cognitive skills it is intended to measure”. Diese Anforderungen an Beurteilungen im Bildungsbereich sollten eine zentrale Zielsetzung sein, denn nur wenn Testergebnisse wieder in die betroffenen Institutionen rückfließen, ist *large scale assessment* auch gerechtfertigt.

All diese hier genannten Aspekte werden, zusammen mit den Kriterien der Praktikabilität bei Oller als *true* oder *valid validity* bezeichnet und als ein integratives Konzept begriffen. Camp versteht Validität ebenfalls als ein umfassendes und integratives Konzept, als “single, unified concept in which the *construct* [Herv. d. V.] to be measured – the theoretical understanding of the knowledge and skills targeted in the assessment – is central to all other considerations“ (Camp 1996: 136). Wieder einmal rückt das theoretische Konstrukt in den Mittelpunkt, das jedem Test zugrunde liegen sollte: Die wesentlichen Entscheidungen, die die Testvalidität beeinflussen, werden in der Testentwicklung relativ früh getroffen – sie müssen auf einer soliden Basis fußen, will man valide Tests entwickeln. Camp fordert, dass alle „Beweise“ der Validität

[123] Vgl. hierzu etwa Harsch & Schröder 2005b.
[124] Auf die Bedeutung von Korrelationen kann hier nur am Rande eingegangen werden. Dennoch ist es lohnenswert zu betrachten, wie Korrelationen interpretiert werden können und sollen. Doch dazu Näheres unter Kapitel 2.3.2 in Bezug auf die Konstruktvalidität integrativer Formate.

interpretiert werden müssen in Bezug auf das theoretische Konstrukt, in Bezug auf Ziele und Zwecke des Tests und damit auch in Bezug auf die aus den Testergebnissen gezogenen Schlussfolgerungen: die sozialen Konsequenzen des jeweiligen Tests. Denn Tests werden zu ganz bestimmten Zwecken eingesetzt und ihre Validität muss auch dahingehend geprüft werden, ob sie für die jeweiligen Zwecke angemessen sind. Man denke beispielsweise an die Konsequenzen eines Zulassungstests zu einem bestimmten Bildungsgang oder an Sprachprüfungen für Immigranten: Falls der Test keine Validität hinsichtlich seiner sozialen Konsequenzen besäße, so würden Entscheidungen über zukünftige Lebenswege ohne solide und valide Basis getroffen – ein Umstand, der nicht nur politisch unhaltbar wäre.

2.3.2 Exkurs: Sprachstruktur – Testformat – Validität

An dieser Stelle sei ein Exkurs zur in den 60er und 70er Jahren geführten Diskussion[125] um die Validität integrativer Formate gestattet. Einerseits soll damit der oben angedeutete zirkuläre Zusammenhang verdeutlicht werden, der zwischen den Strukturmodellen von Sprache und Kompetenzen, die im Testkonstrukt definiert werden, und den vom Test elizitierten Leistungen existiert. Denn wenn die im Testkonstrukt zugrunde gelegte Struktur dieser Leistungen wiederum genutzt werden soll, um die theoretischen Modelle der Kompetenz zu validieren, die dem betreffenden Test zugrunde gelegt wurden, könnte es sich bei dieser Art der Validierung um einen Zirkelschluss handeln, wie im Folgenden gezeigt wird. Andererseits soll gezeigt werden, wie vorsichtig Korrelationsanalysen interpretiert werden müssen.

Die Wahl des geeigneten Testformats hängt wie schon erwähnt eng mit dem Testgegenstand zusammen. Wie oben erläutert, kann man einzelne Teildimensionen oder auch übergreifende sprachliche Fertigkeiten in unterschiedlichen Kombinationen erfassen. Annahmen über die Struktur der zu erfassenden Dimensionen sollten auf einem Modell von Sprache und kommunikativen Kompetenzen basieren. (Solche Modelle sind in Kapitel 1.2.1 mit 1.2.3 dieser Arbeit erläutert). Diese Annahmen werden sich bei valider Operationalisierung nicht nur in den Testitems widerspiegeln, sondern sie werden sich vermutlich auch in der dadurch elizitierten Sprache niederschlagen. Nun ist die Frage gerechtfertigt, ob Sprachtests helfen können, die Annahmen über die Struktur der zu beurteilenden Kompetenzen zu validieren, wenn ihrer Konstruktion eben diese angenommene Struktur zugrunde gelegt wird.

Bei der Untersuchung der Struktur der kommunikativen Kompetenz mittels Sprachtests ist Vorsicht geboten, nicht einem Zirkelschluss zu unterliegen, denn wenn das Testformat die Struktur der elizitierten Sprache beeinflusst, so lässt dieses Testformat keine eindeutigen Rückschlüsse auf die Struktur der zu erfassenden Sprachproduktion und der ihr zugrunde liegenden

[125] Die Ausführungen hier beziehen sich im Wesentlichen auf die Darstellungen in Farhady (1979) und Oller (1979), da diese die Thematik gut illustrieren.

Kompetenzen zu. Wenn beispielsweise Testergebnisse von *discrete-point items* aufsummiert werden und somit ein additives Modell der Kompetenzen angesetzt wird, so verwundert es nicht, wenn dieses Modell durch die Testergebnisse bestätigt wird. In Bezug auf die integrativen Formate hat beispielsweise Oller (1979) untersucht, ob es einen *global language factor*, eine *overall proficiency* gibt, oder ob sich Sprachkönnen aus separaten Bestandteilen zusammensetzt. Doch auch dabei ist Vorsicht geboten, da integrative Formate ja gerade versuchen, das Sprachkönnen ganzheitlich zu erfassen. Wenn ein kommunikatives Testformat jedoch authentische Sprachproduktion elizitiert, so müsste die in diesem Test erfasste Performanz Hinweise auf Strukturen der Interimsprache und der zugrunde liegenden Kompetenzen zulassen.

Den Zusammenhang zwischen Testformat und Struktur von Sprache und Kompetenzen hat man versucht, mittels Korrelations-, Regressions- und Faktorenanalysen[126] zu untersuchen, doch bis heute ist er nicht abschließend geklärt. Psychometrische Rechnungen alleine werden nicht zum Ziel führen, ebenso wenig wie sprachwissenschaftliche oder didaktische Theorien für sich genommen. Vielmehr scheint ein interdisziplinäres Herangehen notwendig, um Strukturmodelle auf linguistischer wie didaktischer Basis mithilfe psychometrischer Methoden zu überprüfen. An dieser Stelle darf auf die Arbeiten zu Strukturmodellen im DESI-Projekt verwiesen werden, die derzeit am DIPF gerechnet werden.[127]

Nicht nur Strukturmodelle von Sprache, auch die darauf basierenden Testformate müssen validiert werden. In den 70er Jahren wurden oft *discrete-point tests* zur Validierung der neu entwickelten integrativen Formate eingesetzt, mit teils sehr hohen Korrelationen. Integrative Formate scheinen untereinander höher zu korrelieren als mit *discrete-point tests*; *discrete-point tests* zu ein und demselben Gegenstand wiederum zeigen untereinander geringere Korrelationen als mit integrativen Formaten.[128] Diese Korrelationen können nun auf mindestens zwei Arten[129] interpretiert werden: Entweder messen die Tests dasselbe oder die Ursache der hohen Korrelation liegt in einem ähnlichen Testkonstrukt. Letzteres wäre verwunderlich, denn gerade im Konstrukt sollten sich „diskrete“ und integrative Formate doch deutlich unterscheiden, wie oben erläutert. Die erstgenannte Interpretation ist so ebenfalls nicht haltbar, wie Farhady (1979: 352)

[126] Korrelation bezieht sich auf das Maß des Zusammenhangs zwischen zwei (oder mehr) Testergebnissen, angegeben im Bereich von -1 bis 1, wobei der Wert -1 bedeutet, dass es einen negativen Zusammenhang gibt (dort wo etwa ein Proband sehr gut in Test A abschneidet, schneidet er sehr schlecht in Text B ab), wohingegen der Wert 1 bedeutet, dass ein Proband in beiden Tests dasselbe Verhalten zeigt. Dieser Zusammenhang darf aber dennoch nicht als „die beiden Tests messen dasselbe“ interpretiert werden. Nach Oller (1979: 54ff) kann ein Zusammenhang im Bereich von 0.95 als Hinweis auf einen akzeptablen Zusammenhang zwischen zwei Tests interpretiert werden, während eine Korrelation im Bereich von 0,60 nicht mehr akzeptabel wäre. Dabei sind niedrige Korrelationen immer schwieriger zu interpretieren als hohe: Hohe Korrelationen können nach Oller (ebd.) mit dem Fündigwerden bei der Goldsuche verglichen werden, während niedrige Korrelationen einem „nicht Fündigwerden“ entsprechen, für welches sich die Gründe sehr viel schwieriger interpretieren lassen: Beispielsweise könnte eine niedrige Korrelation an den unterschiedlichen Schwierigkeiten der beiden Tests liegen, oder einer der Tests könnte unreliabel messen bzw. nicht valide sein.
Bei den beiden letztgenannten Analysen werden Korrelationsmuster zwischen verschiedenen Testvariablen untersucht, um zu sehen, welche Teildimensionen in wie weit zusammenhängen respektive ob unterschiedliche Teildimensionen durch dieselben zugrunde gelegten Faktoren erklärt werden können. Im Übrigen darf auf das Glossar verwiesen werden.

[127] Vgl. Klieme, Eichler et al. (2003: 187ff) und Jude & Klieme 2005.

[128] Vgl. beispielsweise Farhady 1979, Oller 1979 u. a.. Allerdings besteht bei dieser Validierungspraxis die Gefahr von Zirkelschlüssen, denn ausgehend von korrelativen Validierungsversuchen wurden oft diejenigen integrativen Formate als valide betrachtet, die hohe Korrelationen mit diskreten Formaten zeigten. Dann ist es nur logisch, dass die beiden Formate bei späteren Untersuchungen wieder hohe Korrelationen zeigen, wenn sie schon als Validierungskriterium herangezogen wurden.

[129] Eine weitere Interpretationsmöglichkeit sei hier noch erwähnt: die Leistungsunterschiede innerhalb der Probandengruppe, welche sich in beiden Testformaten ähnlich niederschlagen müssten. Ein fortgeschrittener Sprachlerner wird in diskreten wie integrativen Tests besser abschneiden als ein „Anfänger“

schreibt: "… a high correlation between two given tests should not be interpreted as though they tested the same thing". Denn eine hohe Korrelation sei ein Maß für die "go-togetherness" zweier Tests oder *items*, könne aber von ganz unabhängigen Faktoren „zufällig" bestimmt sein. Deshalb schlug Farhady (ebd.: 352ff) Faktorenanalysen vor als angemessenere Methode, herauszufinden ob zwei Tests dieselben zugrunde liegenden Kompetenzen erfassten. Er meinte, man könne zeigen, dass zwei Tests zwar hoch korrelierten, aufgrund beispielsweise dreier (hypothetischer) Faktoren, auf die sie jeweils laden – doch der springende Punkt sei, dass die beiden Formate ganz unterschiedliche Ladungen auf die Faktoren zeigen könnten, was sich jedoch nicht in den Korrelationen zeigen würde, was aber bedeute, dass sie eben doch nicht dieselben zugrunde liegenden Faktoren testeten. Er folgerte daraus: Nur wenn eine Faktorenanalyse ergäbe, dass zwei Tests zu ähnlichen Anteilen auf dieselben Faktoren laden, käme eine Aussage wie „die Tests messen dieselben zugrunde liegenden Komponenten" in Betracht.

Oller beispielsweise sah aufgrund der oben erwähnten Korrelationen zwischen *scores* von „diskreten" und integrativen Tests und aufgrund ähnlicher Ladungen auf dieselben Faktoren einen Hinweis auf einen *overall language proficiency factor* – doch es konnten keine eindeutigen, empirisch schlüssigen Hinweise auf diesen gefunden werden. Es wurden viele Erklärungen angeboten und bemüht, doch welche ist die zutreffendste? Beispielsweise schlug Oller (1979: 61) vor, dass die hohen Korrelationen zwischen integrativen Tests ein Beweis ihrer Validität seien: "The results of correlation studies can be easily understood or at least straightforwardly interpreted as evidence of the fundamental validity of the variety of language tests that have been shown to correlate at such remarkably high levels." Er betrachtete die annähernd gleichen Reliabilitäts- und Validitätsindizes vieler pragmatischer Tests als Hinweis auf den postulierten *unitary language factor*. Als weitere mögliche Erklärung bot Oller (ebd.: 61f) das grammatikalische System der Lerner an: Dieses sei für große Teile der Varianz bei unterschiedlichsten Sprachtests verantwortlich (vgl. die Erklärung oben, Fußnote 127). In diesem Zusammenhang wurden auch Fehleranalysen der Lernersprache durchgeführt, die ebenfalls keine eindeutigen Ergebnisse brachten. Wie aber kann die verbleibende Varianz, die sich nicht mittels der *Interlanguage* erklären lässt, gedeutet werden?

Es könnte sein, dass sich Korrelationen zwischen Tests mittels eines Modells der systemischen Organisation von Sprache, wie es in Kapitel 1.2.1 dieser Arbeit besprochen wurde, erklären lassen. Es ist wahrscheinlich, dass es Testformate gibt, die auf verschiedenen systemischen Ebenen angesiedelt sind: Solche, die nur einen ganz bestimmten Bereich erfassen, sind wahrscheinlich auf unteren Systemebenen angesiedelt; je komplexer die von einem Test erfassten Bereiche sind, desto mehr Subsysteme dürfte er erfassen, und je mehr Subsysteme erfasst werden, auf einer umso höheren Ebene dürfte der Test angesiedelt sein. Hinweise auf solch ein systemisches Modell lassen sich beispielsweise aus einer Studie Bachmanns und Palmers (1987) ableiten: Sie versuchten ebenfalls mittels Faktorenanalysen einige Komponenten der "c*ommunicative proficiency*" zu validieren. Dazu setzten sie Interviews, verschiedene schriftliche Aufgaben und MC-Formate ein. Interview und schriftliche Produktion wurden im *Rating-*

Verfahren (Näheres dazu in Kapitel 3.3 dieser Arbeit) bewertet, wobei drei "*main traits*" eingeschätzt wurden, namentlich pragmatische Kompetenz mit den "*subtraits*" Vokabular, Kohäsion und Organisation, grammatikalische Kompetenz mit den *subtraits* Umfang und Korrektheit sowie soziolinguistische Kompetenz mit den *subtraits* Register, Nähe zu Muttersprachlern (*nativeness*) und kulturelle Referenzen. Bachmann und Palmer fanden einen generellen Faktor, der alle Messungen ihrer Studie beeinflusste, sowie zwei *primary-trait* Faktoren der grammatisch-pragmatischen und der soziolinguistischen Kompetenz. Sie konnten das ihren Tests zugrunde gelegte Modell der kommunikativen Kompetenz (vgl. oben, Kapitel 1.2.3) teilweise bestätigen.

2.4 Testziele und Zwecke

Wenden wir uns nun den Testzielen zu, die in der Praxis natürlich ganz am Anfang des Testentwicklungsprozesses festgelegt werden. Tests sind dann gerechtfertigt und sinnvoll, wenn die damit erzielten Ergebnisse in der jeweiligen Institution, in deren Kontext der Test durchgeführt wurde, auch genutzt werden, sei es um den Unterricht in konkreten Punkten zu verbessern, um anstehende Entscheidungen zu treffen oder um die Bildungsqualität insgesamt zu entwickeln. Um Testergebnisse nutzen zu können, müssen die Tests Informationen liefern, welche relevant für die Unterrichtsverbesserung, für die zu treffenden Entscheidungen oder die Qualitätsentwicklung sind.[130] In dieser Nutzung von Testergebnissen und Informationen liegt der Testzweck, liegen die Testziele begründet: Nur wenn diese Ziele über den gesamten Testentwicklungsprozess hin beachtet werden und in die Testerstellung einfließen, kann der Test dem Anspruch auf Validität genügen, denn eigentlich alle weiteren Entscheidungen hängen vom jeweiligen Ziel ab: die Auswahl der Testbereiche, *skills* und Wissensbestände; die Bestimmung geeigneter Formate und Typen; das Bewertungsverfahren; die Interpretation der Daten; und nicht zuletzt die Art des Feedbacks. Jede dieser Entscheidungen muss am Testzweck, an den zu erreichenden Zielen ausgerichtet werden und in ihren potentiellen Vor- und Nachteilen sorgfältig abgewogen werden.

Im Hinblick auf Schulleistungsstudien wird diese Forderung nach systemischer Validität noch deutlicher: Solche Leistungsstudien sind im Spannungsfeld zwischen edukativer und psychometrischer Leistungsmessung angesiedelt und müssen beiden genügen. Die mit großen Studien verbundenen hohen Kosten lassen sich nur rechtfertigen, wenn sich das Kosten-Nutzen Verhältnis in einem akzeptablen Rahmen bewegt, wenn folglich die getestete Institution einen Erkenntnisgewinn erzielt, der dem Aufwand entspricht. Dazu ist es notwendig, in der Studie nicht nur den Ansprüchen der Psychometrie und Statistik gerecht zu werden, sondern auch denjenigen der Didaktik, Pädagogik und Sprachwissenschaft, die die Ausgangsbasis für jede psychometrische Analyse bilden. Torrance beschreibt die Tendenz, dass sich die Testformate

[130] Vgl. hierzu beispielsweise Cooper 1972, Elbow 1996, Bachmann 1996, u. a..

gerade in *large scale assessments* wegbewegen von “integrated extended tasks over a period of weeks“ hin zu “short paper-and-pencil tests“ (1998: 34). Bleyhl (2003: 39f) sieht eine ähnliche Entwicklung und warnt davor, dass die Forderung nach „Qualitätssicherung“ zu leicht durchzuführenden Tests führen könne, die oft von ihren Formaten her die Handlungsfähigkeit gar nicht erfassen könnten – wohl eine zum Teil notgedrungene Entwicklung, denn die finanziellen Ressourcen genügen in den meisten Fällen nicht, aufwändigere Untersuchungen durchzuführen.

Trotzdem ist Torrances (1998: 34f) und Shales (1996: 95f) Einwand gerechtfertigt, dass das „Heft“ im Bereich der Sprachtests nicht alleine in den Händen der Psychometriker liegen dürfe und man das Bewerten von Sprache nicht auf rein quantitative Verfahren reduzieren dürfe, gleichwohl diese auf große Akzeptanz in der Bevölkerung stießen, nicht zuletzt da sie bekannt seien und die Öffentlichkeit an sie gewöhnt sei. Denn auch die Psychometrie arbeitet mit Modellen, die die Komplexität von Sprachverarbeitung noch nicht adäquat widerspiegeln können und – wie alle Modelle – nur Annäherungen an die Realität darstellen. Dennoch sollte die Entwicklung darauf hinauslaufen, dass nicht im Extremfall die sprachliche Testrealität auf die mathematischen Modelle hin konstruiert und an diese angepasst wird, sondern dass sich Rechenmodelle an die Komplexität der Sprache annähern. Hierzu müssen die qualitativen Ansprüche und Bedürfnisse aus Didaktik, Pädagogik und Linguistik gleichberechtigt neben die quantitativen Ansprüche der Psychometrie treten. Testentwicklung ist ein interdisziplinäres Unterfangen und sollte an Standards ausgerichtet werden, die von allen Beteiligten gemeinsam erarbeitet werden. Die Frage bliebt, ob Schulleistungsstudien diesen Ansprüchen je gerecht werden können, denn aus logistischen und finanziellen Gründen verbietet sich oft der Einsatz geeigneter Testformate aufgrund ihrer zu aufwändigen Durchführung. Man denke beispielsweise an die Bewertung der mündlichen Kommunikationsfähigkeit und die dazu nötigen aufwändigen Bewertungsverfahren – diese sind oft aus praktischen Gründen nicht realisierbar.

Um dennoch Informationen über den Leistungsstand auf breiter Basis zu erhalten, bietet es sich an, begründete und ausgewählte Stichproben zu ziehen, wie es Elbow (1996: 129) fordert:

> Testing on a massive scale is not the best way to measure every student – and it gives no substantive feedback. Selective but more trustworthy testing would serve the goals that are reasonable for such huge tests: identifying schools that need extra resources to bring more students up to par and providing samples of unsatisfactory and exemplary portfolios for teachers – samples that could be used for local assessment at school or regional levels to give some genuine feedback to every student.

Für diese ausgesuchte Zielgruppe müssten im nächsten Schritt Tests entwickelt werden, die den Anforderungen an kommunikative, systemvalide Tests weitgehend genügen, aber noch in großem Stil durchführbar sind. Beispielsweise spricht nichts dagegen, das Hör- oder Leseverstehen im Multiple Choice-Format zu erfassen, wenn die Texte und Fragen dazu in einem kommunikativ-handlungsorientierten Rahmenkonzept verankert sind, die Aufgaben der Probandengruppe, den Testzielen und den Rahmenbedingungen entsprechen und die Distraktoren sorgfältig entwickelt wurden.

Die Ergebnisse großer Schulleistungstests sollten in der jeweils beurteilten Institution interpretiert und verarbeitet werden, und zwar auch diesbezüglich, dass quantitative Daten wiederum in qualitativ-beschreibende Ergebnisse transformiert werden, um beispielsweise auf Klassenebene genutzt werden zu können. Die Tests einer Studie könnten etwa im Unterricht wieder aufgegriffen und um Bewertungsverfahren ergänzt werden, die sich Aspekten widmen können, die in der Studie selbst nicht erfasst werden konnten, wie zum Beispiel um ein Portfolio-*Assessment* im Bereich des Schreibens, das sich an einen nationalen Schreibtest anschließen könnte (vgl. dazu Kapitel 4.7 dieser Arbeit). In diesem Rahmen wäre es denkbar, die im Test geschriebenen Aufsätze im Unterricht auf bestimmte Aspekte hin zu bewerten und zu überarbeiten, die jeweils für die einzelnen Klassen relevant sind. Diese Überarbeitungen könnten in einem Portfolio dokumentiert werden. Dies ist nur ein Beispiel, wie externe Leistungsmessung und interne Bedürfnisse miteinander verknüpft werden könnten.

In diesem Zusammenhang fordert Torrance (1998) die Beachtung des größeren Kontexts des Lehrens und Lernens der jeweils in einem Test erfassten Leistungsbereiche. Er schlägt vor, die Bewertung auszurichten hin auf die Lernkontexte, in denen und für die die Bewertung auch stattfindet. Das traditionelle Vermessen mit dem Ziel der Selektion oder Zertifizierung müsste um Evaluationsverfahren erweitert werden, die der Ganzheitlichkeit von Lehren und Lernen Rechnung tragen, wobei Messverfahren zum Einsatz kommen sollten mit dem Zweck, Bildung und Lernen zu fördern und bereits erreichte (Teil-) Kompetenzen anzuerkennen. Schulen haben Torrances Meinung nach eine neue Aufgabe bekommen: “(...) [T]he emphasis of school systems (...) must surely be on the intellectual development of all, to the highest possible standards, rather than the selection of a few by methods which arguably lower standards” (Torrance 1998: 35). Solch umfassende Evaluation, mit der die erwähnte Entwicklung aller Lernenden überprüft werden muss, sollte auch einen Beitrag leisten zur Verknüpfung von Außen- und Innenperspektive und von Fremd- und Selbstevaluation. Sie sollte den Rückfluss von Ergebnissen äußerer Bewertung in die schulische Entwicklung ermöglichen, Lehrende und Lernende in der Beurteilung von Lernfortschritten unterstützen, den Unterricht verbessern helfen und die Eigenverantwortung aller Beteiligten fördern. Des Weiteren sollte sie einen Beitrag leisten zur Curriculums- und Bildungsstandardentwicklung.

Wenn Schulleistungsstudien nicht eingebettet werden in die größere Diskussion um die Sicherung der Bildungsqualität, so fehlt ihnen ihre Berechtigung. Allerdings muss an dieser Stelle darauf hingewiesen werden, dass bei Schulleistungsstudien die Anonymität aller Beteiligten aus Gründen des Datenschutzes gewährleistet werden muss – dies erschwert ein Rückfließen des Tests und seiner Ergebnisse in die Schule und macht eine Weiterverarbeitung der Schulleistungstests im Unterricht (etwa durch ein sich an eine Studie anschließendes Portfolio-Projekt wie oben angedeutet) meist unmöglich, da die Schülerinnen und Schüler die eigenen Tests (beispielsweise die Aufsätze aus einem Test, die man nun im Unterricht weiter verwenden könnte)

in der Regel nicht zurückerhalten. Solch äußere, aus Perspektive des Datenschutzes auch wichtige und sinnvolle Beschränkungen verhindern leider manch aussichtsreiche Möglichkeiten.

Neue Testformate, die der Komplexität des Bildungsprozesses Rechnung tragen können, müssen sich natürlich, wie alle anderen Testformate auch, an oben genannten Gütekriterien messen lassen und ihre systemische Validität erst unter Beweis stellen. Auch wenn auf diesem Gebiet die Forschung erst am Anfang steht, so lohnt es doch, einen kurzen Blick darauf zu werfen: Die oben angedeutete Möglichkeit, beispielsweise Schreibfertigkeit mittels des Portfolio-Ansatzes zu bewerten, wird vielerorts diskutiert.[131] Auch bieten sich Formen des *extended assessment*[132] an, die sich über längere Zeiträume erstrecken. Inwieweit sich solche Formen in großen Leistungsstudien einsetzen lassen, müsste erst empirisch untersucht werden. Dennoch scheint es sinnvoll, den traditionell engen Testkontext zu erweitern – am Beispiel des Schreibens etwa um Ansätze zur Erforschung von Schreibprozessen, Schreibentwicklung und Schreibbewertung – und ihn zu ergänzen um Hintergrunderhebungen, etwa in Form von Befragungen[133] aller Beteiligten zum Unterricht, zur Lernsituation und auch zur Wahrnehmung, Validität und Durchführbarkeit des betroffenen Tests.

Es geht um die Frage, welchen Beitrag Beurteilung im weiteren Sinn und Schulleistungsstudien und Sprachtests im engeren Sinn leisten können, um die Bildungsqualität im Fremdsprachenunterricht zu verbessern und wie solch eine Beurteilung im Idealfall aussehen müsste, um *instructional value* zu besitzen. Dazu müssten auch die Lehrkräfte vorbereitet und in die Testentwicklung mit einbezogen werden, um in der Praxis neue Methoden und Wege auszuprobieren. Ebenso müssten Wege des *performance assessment* in der Praxis empirisch erforscht werden, wie Torrance (1998: 36) beschreibt:

> Research in the fields of cognitive science, student motivation and attribution, and classroom interaction, as well as assessment, all have their contribution to make to our understanding of the impact of new approaches to assessment on teaching and learning, but the improvement of practice depends on teachers recognizing the relevance of such research and interrogating it and acting on it in the context of practice. Thus what is perhaps needed most of all are well conducted and well reported action research studies of performance assessment in practice, including detailed evidence from the students' perspective, as well as good accounts of the practice and problems of teachers.

Basierend auf solcher Forschung könnten dann empirisch fundierte Aussagen zu validen neuen Formaten getroffen werden, die sich für Schulleistungsstudien eignen.

Die vorliegende Arbeit muss sich jedoch auf Aspekte schon erfolgter Forschung beschränken – in diesem Fall auf den Aspekt des positiven *Washback*-Effekts, der beim Kriterium des *instructional value* eines Tests eine nicht zu unterschätzende Rolle spielt. *Washback* bezieht sich auf den Tatbestand, dass Tests immer Auswirkungen auf die Institution haben, in der sie stattfinden. Während ein *teaching to the test* im Allgemeinen nicht wünschenswert ist und die

[131] Vgl. beispielsweise Torrance 1998, Larson 1996, Murphy & Grant 1996, u. a..
In Kapitel 4.7.2 dieser Arbeit wird anhand des erwähnten Praxisbeispiels skizziert, welchen Beitrag ein Portfolio-*Assessment* zur Einbindung eines nationalen Schreibtests in den Schulalltag leisten könnte, um die Testergebnisse möglichst sinnvoll in die getestete Institution rückfließen zu lassen.
[132] Vgl. hierzu Torrance (1998: 35).
[133] Dieser Weg wurde u. a. in der DESI-Studie eingeschlagen.

unterrichtliche Praxis negativ beeinflusst, können Tests auch positiv rückwirken und Effekte erzielen, die erwünscht sind. Im Idealfall kann man sich Tests denken, die schon bei ihrer Konstruktion auf solch erwünschte Effekte hin ausgelegt wurden. Damit könnte dann die beispielsweise von Cohen (1994) angesprochene Kluft zwischen Lehren und Bewerten überwunden werden: Einerseits soll das im Unterricht Gelehrte angemessen überprüft werden und andererseits soll diese Überprüfung sinnvoll in den Lehrbetrieb rückfließen. Eine denkbare Lösung sieht Cohen im Vorschlag von Paris et al. (1991), die empfehlen, Lehrende und Lernende gemeinsam komplexe Tests entwickeln zu lassen: "They recommend a *developmental approach* to testing, whereby teachers and pupils would work collaboratively on authentic testing tasks that were *longitudinal* and *multidimensional*." (Cohen 1994: 3). Auch wenn dieser Weg nicht immer gangbar sein wird, so stellt er doch eine interessante Alternative zur gängigen Testentwicklung dar, die empirisch erforscht werden müsste.

Ganz unabhängig von der Art der Testentwicklung sollten folgende Charakteristika beachtet werden, um zu praktikablen, reliablen und validen Tests zu kommen, die den Sprachlehr- und Lernprozess unterstützen und seinen Bedingungen gerecht werden:[134]

- Leistungsbeurteilung erweitern: Alle angemessenen Wege und Techniken der Beurteilung nutzen, nicht nur auf den Einsatz von Tests beschränken; Beurteilung als Gelegenheit für bedeutungsvolle Interaktion zwischen Lehrenden und Lernenden betrachten.

- Alle Fähigkeiten, Fertigkeiten und Kenntnisse in die Beurteilung einfließen lassen, die in Schule oder Unterricht gefördert werden sollen, in ausgewogenem und den gegebenen Bedingungen angemessenem Verhältnis; alle in der Beurteilung erfassten Gegenstandsbereiche durch möglichst viele verschiedene *items*, Indikatoren oder Aufgaben erfassen, nicht nur durch eine einzige Messung oder Beurteilung.

- Lernende auf Basis dessen beurteilen, was sie schon können, und nicht allein auf Basis ihrer Fehlleistungen, wobei die Beurteilungsergebnisse die Performanz der Probanden in möglichst vielen Gebieten und mittels möglichst unterschiedlicher Beurteilungsmethoden widerspiegeln sollten.

- Wo immer möglich, die Performanz der Lernenden beurteilen und direkte Beurteilungsmethoden einsetzen; solche Testformate wählen, die zu ihrer Lösung diejenigen Fertigkeiten verlangen, die in diesem Test auch erfasst werden sollen; solche Beurteilungsmethoden wählen, die darauf ausgelegt sind, den Lernenden zu helfen ihre Fertigkeiten zu verbessern.

- Wo angebracht und möglich, kriterienorientierte Tests einsetzen, die es den Probanden ermöglichen vor der Beurteilung herauszufinden, zu was sie in der Lage sein sollten, um den Test mit einem bestimmten Erfolgsgrad zu bestehen; Bewertungskriterien und Erwartungen an die Probanden vorher klarstellen; Strategien zur Meisterung der Beurteilungssituation und neue Beurteilungsformate vorher üben.

[134] vgl. hierzu Alderson et al. 1996, Bachmann & Palmer 1995, Brown 1993, Cohen 1994, Hughes 1998 u. a..

- Unterstützung der Lehrerschaft, gerade bei neuen Beurteilungswegen oder unbekannten Testformaten, um sie etwa auf den Umgang mit neuen Formaten in einer anstehenden Leistungsstudie vorzubereiten, so dass wiederum die Schülerschaft angemessen vorbereitet werden kann und die Ergebnisse dann auch im Unterricht sinnvoll genutzt werden können.

- Ergebnisse und Feedback möglichst zeitnah in den beurteilten Kontext rückfließen lassen, um sie unter den Beteiligten diskutieren und bestmöglich nutzen zu können.

- Abwägen der Kosten und Nutzen einer anstehenden Beurteilung: Auch wenn beispielsweise indirekte Formate leichter durchzuführen und auszuwerten sind und damit auch kostengünstiger implementiert werden können als etwa ein direktes Bewertungsverfahren, das aufwändige *Rating*-Prozesse mit sich bringt, so darf man dennoch die langfristig entstehenden Kosten nicht vergessen, wie Hughes (1989: 47) zu bedenken gibt:

> When we compare the cost of the test with the waste of effort and time on the part of teachers and students in activities quite inappropriate to their true learning goals (and in some circumstances, with the potential loss to the national economy of not having more people competent in foreign languages), we are likely to decide that we cannot afford *not* to introduce a test with a powerful beneficial backwash effect.

2.5 Konzepte der Sprachbeurteilung und des Sprachtestens im GER

Auf Basis der bisher in Kapitel 2 erarbeiteten Grundlagen wird nun der GER auf seinen Beurteilungs- und Testansatz hin untersucht. Der Titel des GER lautet „Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen“. Nachdem Sprachbegriff und Lehr-/Lernkonzepte des GER in den Kapiteln 1.2.5 respektive 1.3.4 dieser Arbeit erörtert wurden, wendet sich das vorliegende Kapitel 2.5 den Beurteilungs- und Bewertungskonzepten des GER zu.

Der GER gibt keine umfassende Darstellung zum Beurteilen und Testen des Sprachvermögens, er will kein Beurteilungs- oder Testerstellungsleitfaden sein, sondern er stellt in seinem Abschnitt 9 diejenigen Aspekte dar, die für den im GER gewählten handlungsorientierten kommunikativen Ansatz relevant sind. Zum Aspekt der Testerstellung gibt es wie erwähnt als Ergänzung zum GER einen eigenen Leitfaden zur Testentwicklung, den *User's Guide for Examiners* (im Folgenden mit UGE abgekürzt). Dieser Leitfaden wird in Kapitel 2.6 dieser Arbeit vorgestellt, da er den Beurteilungs- und Bewertungsansatz des GER in den größeren Kontext der Testerstellung rückt. Zudem wird auf ihn im Praxisteil in Kapitel 4 dieser Arbeit rekurriert.

Bei der Analyse des GER-Abschnitts 9, in welchem neben allgemeinen Testgütekriterien verschiedene Typen von Beurteilung und Bewertung und Möglichkeiten der Verwendung des GER in der Sprachbeurteilung vorgestellt werden, fällt auf, dass das GER-Skalensystem beim Beurteilen und Bewerten im GER eine zentrale Stellung einnimmt: Beispielsweise ist ihm ein eigener Unterabschnitt 9.2 gewidmet mit dem Titel „Der Referenzrahmen als Hilfsmittel beim

Beurteilen und Bewerten“. Auch werden die Skalen in GER-Abschnitt 9.3 den unterschiedlichen Bewertungstypen zugeordnet und ihre jeweiligen Verwendungsmöglichkeiten benannt. Um also die GER-Aussagen bezüglich der Einsatzmöglichkeiten seines Skalensystems bei der Sprachbeurteilung einschätzen zu können, muss zunächst das GER-Skalensystem in seiner Tragweite analysiert werden. Daher wird dem Skalenansatz in der Sprachbeurteilung ein eigenes Kapitel 3 in dieser Arbeit gewidmet. Dort wird auf wichtige Aspekte wie beispielsweise auf die Grundlagen von Skalen, auf Skalenentwicklung, auf die Bedeutung von Skalen bei der Beurteilung sprachlicher Leistung, bei der Testentwicklung und bei der Verortung von Tests in einem gegebenen Referenzsystem eingegangen. Aufbauend darauf werden das Skalensystem des GER, seine Konstruktion und die ihm zugrunde gelegten Modelle analysiert, ehe dann die Verwendbarkeit der Skalen betrachtet wird. Zum genannten Aspekt der Anbindung von Tests an das Referenzsystem des GER wurde im September 2003 das *Manual: Relating Language Examinations to the Common European Framework of Reference for Languages* (vgl. Council of Europe 2003a) veröffentlicht, das entsprechend in Kapitel 3.5 der vorliegenden Arbeit angesprochen wird.

Im Folgenden werden grundlegende Fragen im Zusammenhang mit dem Beurteilungs- und Testbegriff im GER untersucht: Welche Schlüsselkonzepte im Bereich des Beurteilens und Testens des Sprachvermögens lassen sich ausmachen? Welchen Kompetenzbegriff liegt der Sprachbeurteilung im GER zugrunde? Durch welche Charakteristika ist der Testbegriff im GER gekennzeichnet? Wie betrachtet der GER seine Verwendungsmöglichkeiten in der Sprachbeurteilung?

2.5.1 Schlüsselkonzepte des Beurteilens und Bewertens im GER

Positiv fällt die Einleitung zum neunten Abschnitt des GER auf; sie enthält neben einer übersichtlichen Gliederung des Abschnitts eine klare Begriffsbestimmung, indem sie allgemein verständlich die Termini Evaluation, Beurteilung und Bewertung unterscheidet:

Der GER versteht den Begriff der Evaluation als Oberbegriff: Evaluation greife am weitesten und erfasse neben der tatsächlichen Leistungsmessung auch die Rahmenbedingungen und die Institutionen, innerhalb derer die Leistungsmessung stattfindet (GER: 172). Die Termini *Beurteilung* und *Bewertung* siedelt der GER innerhalb der Evaluation an: Während sich Beurteilung (engl. *assessment*) auf einen generellen Prozess innerhalb der breiter angelegten Evaluation bezieht – beispielsweise auf die auch informelle Beurteilung von Lernfortschritten – , sieht der GER Bewertung als „[S]ynonym für ‚Leistungsmessung’, ‚Prüfung’, ‚in eine Reihenfolge bringen’ usw.“ (GER: 172 Fußnote). Der GER grenzt sich klar ab von einer evaluativen Ausrichtung und beschränkt sich auf Aspekte der Beurteilung und Bewertung. Solch eindeutige Stellungnahme trägt zur Transparenz des Instruments bei.

Der GER legt der Beurteilung drei Schlüsselkonzepte zugrunde: Validität, Reliabilität und Durchführbarkeit (GER: 172). Hauptaugenmerk liegt dabei auf der Validität, wohingegen Reliabilität ein „technischer Begriff“ sei, auf den der GER nicht weiter eingeht. Zur Validität vermerkt der GER, dass die Validität der Niveaueinteilungen, die Validität der Kriterien und die Validität der Kriterienfindung entscheidend zur Testvalidität und damit zur „…Genauigkeit der Entscheidungen, die unter Bezug auf einen Standard getroffen werden…“ beitrage (GER: 172). Entscheidend sei, was gemessen werde und wie die Leistung interpretiert werde (ebd.). Wie allerdings ein valider Sprachtest entwickelt oder auf welche Weise die oben erörterten unterschiedlichen Aspekte der Validität geprüft werden können, wird im GER nicht angesprochen. Auf den Aspekt der Durchführbarkeit wird nur noch kurz auf S. 173 des GER eingegangen: Praktikabilität der Bewertung, Zeitdruck, Momentaufnahme der Performanz und begrenzte Anzahl von Bewertungskriterien aus Gründen der Handhabbarkeit werden dort als restringierende Faktoren der Testdurchführung erwähnt. Andere wichtige Faktoren, wie beispielsweise die oben erläuterte Machbarkeit eines Tests, von der die Testvalidität ebenfalls abhängt, werden im GER nicht erwähnt.

Der Aspekt der Praktikabilität wird im GER noch einmal aufgenommen, dabei jedoch nur auf die flexible Adaption des GER-Kategoriensystems bezogen (vgl. GER: 187f). Dieses Kategoriensystem (vgl. die Kategorien der GER-Abschnitte 4 und 5) ist so umfangreich, dass die Autoren des GER es weder für möglich noch machbar halten, alle Kategorien oder Skalen des GER bei einer einzigen Beurteilung anzuwenden (ebd.: 187) – dies versteht sich eigentlich von selbst. Stattdessen wird im GER vorgeschlagen, jeweils angemessene und relevante Kategorien zu analysieren und auf eine handhabbare Menge zu reduzieren, um die Praktikabilität zu gewährleisten. Denn das GER- „System soll (…) als Mittel zur Herstellung von Bezügen dienen“ (GER: 187) und könne aufgrund seines Umfangs nicht „eins zu eins“ zur Bewertung übernommen werden. Die erforderliche Reduktion der Bewertungskriterien könne entweder durch Auswahl der geeigneten Kategorien für den jeweiligen Kontext erfolgen oder durch das Zusammenlegen von Merkmalen mehrerer Kategorien zu einem Kriterium. Wie dieser Vorgang in der Praxis aussehen könnte, wird in Kapitel 4 dieser Arbeit dokumentiert. Den Abschluss der Praktikabilitätsüberlegungen im GER bilden vier Beispiele aus der Praxis (ebd.: 188ff), die Wege aufzeigen, auf Basis des GER-Systems zu einem praktikablen Bewertungssystem zu kommen – eine anschauliche Darstellung, die den Nutzern des GER Anregungen gibt, wie das eigene Bewertungssystem zu dem Kategoriensystem des GER in Beziehung gesetzt werden kann und wie das GER-Referenzsystem an eigene Verhältnisse angepasst werden kann. Allerdings werden neben dem Aspekt der Anpassung des GER-Systems keine weiteren Faktoren der Praktikabilität angesprochen; wichtige Aspekte wie etwa der oben erwähnte Zusammenhang zwischen Testgegenstand und praktikablem Testformat oder der Aspekt der praktikablen Administration eines breit angelegten *proficiency tests* werden nicht thematisiert.

Der GER macht in seinem Abschnitt 9 deutlich, dass er lediglich einen Rahmen stecken kann, den die Nutzer selbst füllen und in die Praxis umsetzen müssen: „Der *Referenzrahmen* versucht, Bezugspunkte zur Verfügung zu stellen, nicht aber praktische Beurteilungsinstrumente. Der *Referenzrahmen* muss nämlich umfassend sein, seine Benutzer hingegen müssen auswählen“ (GER: 173). Dieser Anspruch auf „umfassend sein“ kann jedoch auch in Abschnitt 9 – wie beispielsweise schon beim Vermittlungskonzept des GER-Abschnitts 6 – nicht eingelöst werden, da wesentliche Punkte wie die verschiedenen Konzepte der Validität oder Reliabilität nur gestreift und andere, ebenso wesentliche Testgütekriterien (vgl. die Ausführungen in Kapitel 2.3 oben) jedoch gar nicht erst erwähnt werden.

2.5.2 Kompetenzkonzept des GER in der Sprachbeurteilung

Wie oben in Kapitel 2.1 erläutert, sollte einem Sprachtest ein Modell dessen, was er erfassen soll, zugrunde liegen, um die Testvalidität zu gewährleisten. Auch der GER wählt einen modellbasierten Ansatz: Wie in Kapitel 1.2.5.3 dieser Arbeit erläutert, beruht das GER-Referenzsystem auf Modellen einer handlungsbezogenen kommunikativen Kompetenz, ohne dass diese Modelle jedoch im GER benannt würden. Da sie die Basis der horizontalen Einteilung des GER-Skalensystems in seine Kategorien bilden, werden sie bei den entsprechenden Ausführungen in Kapitel 3.4.1.1 dieser Arbeit konkretisiert.[135] Hier wird untersucht, welche theoretische Konzeptionalisierung des Kompetenzbegriffs der GER der Sprachbeurteilung zugrunde legt und welche Möglichkeiten der GER vorschlägt, diese kommunikative Kompetenz in Sprachtests zu erfassen.

Bei der Analyse des der Sprachbeurteilung zugrunde gelegten Kompetenzbegriffs in GER-Abschnitt 9 fällt zunächst auf, dass manche Formulierungen im deutschen Dokument nicht recht verständlich sind. Erst das englischsprachige Original (vgl. *Council of Europe* 1996a: *Common European Framework of Reference*, im Folgenden mit CEF abgekürzt) erhellt die teils missverständlichen Formulierungen, die sich meist auf Übersetzungsprobleme zurückführen lassen. Deshalb werden vorab diese Probleme betrachtet, ehe wir uns dem Kompetenzkonzept zuwenden, das der GER der Sprachbeurteilung zugrunde legt.

2.5.2.1 Übersetzungsproblematik in GER-Abschnitt 9

Die erwähnten Übersetzungsprobleme[136] (in der Regel inkonsistente Übersetzungen von Fachtermini) sollen an folgenden Beispielen verdeutlicht werden:

[135] Diese Modelle werden in North (2000) bei der Dokumentation der GER-Skalenkonstruktion erörtert.
[136] Eine Analyse der europaweiten Übersetzungen des CEF wäre ein eigenes Forschungsprojekt, würde jedoch den Rahmen der vorliegenden Arbeit sprengen. Deshalb wendet sich diese der Übersetzungsthematik nur insofern zu, als es für die Analyse des GER-Kompetenzbegriffs, den er der Beurteilung zugrunde legt, unmittelbar notwendig ist.

- Beispiel (a) (GER: 174f, CEF: 180, Fettdruck d. V.):

> Es ist jedoch normalerweise *nicht* ratsam, Deskriptoren für 'kommunikative Aktivitäten' in den Kriterienkatalog für Prüfende bzw. Korrektoren aufzunehmen, wenn diese die **Leistung** in einem bestimmten Test zum Sprechen bzw. Schreiben auf dem jeweils erreichten **'Kompetenzniveau'** einstufen sollen. Der Grund dafür ist, dass sich eine Beurteilung der **Sprachkompetenz** nicht vorwiegend auf irgendeine spezielle **einzelne Leistung** beziehen, sondern vielmehr versuchen sollte, eine generalisierbare **Kompetenz** zu beurteilen, für die die vorliegende einzelne Leistung nur ein Indiz ist.
>
> However, it is *not* usually advisable to include descriptors of communicative activities in the criteria for an assessor to rate **performance** in a particular speaking or writing test if one is interested in reporting results in terms of a **level of proficiency** attained. This is because to report on **proficiency**, the assessment should not be primarily concerned with any one particular **performance**, but should rather seek to judge the generalisable **competences** evidenced by that performance.

- Beispiel (b) (GER: 179, CEF: 184, Fettdruck d. V.):

> Die Skalen der Beispieldeskriptoren beziehen sich auf die Beurteilung der **Sprachkompetenz** (also auf Qualifikationsprüfungen): das Kontinuum der **Fähigkeiten bei der Anwendung der Sprache** in realen Kontexten.
>
> The scales of illustrative descriptors relate to **proficiency** assessment: the continuum of real world **ability**.

- Beispiel (c) (GER: 186f, CEF: 192, Fettdruck d. V.):

> Die Benutzer des Referenzrahmens sollten bedenken und, soweit sinnvoll, angeben, […]
> - auf welche Art und Weise in ihrem jeweiligen System die Beurteilung von **Leistung** (schulorientiert; lernorientiert) und die Beurteilung von **Sprachkompetenz** (bezogen auf die reale Welt; ergebnisorientiert) ausbalanciert gehandhabt und ergänzt werden; in welchem Umfang neben **Sprachkompetenz** auch die **Performanz** beurteilt wird;
>
> Users of the Framework may wish to consider and where appropriate state: […]
> • the way in which the assessment of **achievement** (school-oriented; learning-oriented) and the assessment of **proficiency** (real world-oriented; outcome-oriented) are balanced and complemented in their system, and the extent to which **communicative performance** is assessed as well as **linguistic knowledge**.

- Beispiel (d) (GER: 178, CEF: 183, Fettdruck d. V.):

> Ein ***Sprachstandstest*** (auch: Leistungstest, *achievement test*) überprüft, ob bestimmte Ziele erreicht wurden; er überprüft also, was unterrichtet worden ist. Er bezieht sich somit auf die Arbeit einer Woche, eines Semesters, das Lehrbuch oder den Lehrplan. Ein Sprachstandstest orientiert sich am Kurs und stellt somit eine Binnenperspektive dar.
>
> Eine ***Qualifikationsprüfung*** (auch: Feststellungsprüfung, *proficiency test*) hingegen überprüft, was jemand kann oder weiß, wenn er/sie einen Lerngegenstand im "wirklichen Leben" anwendet. Diese Art der Beurteilung stellt eine Außenperspektive dar.
>
> ***Achievement assessment*** is the assessment of the achievement of specific objectives – assessment of what has been taught. It therefore relates to the week's/term's work, the course book, the syllabus. Achievement assessment is oriented to the course. It represents an internal perspective.
>
> ***Proficiency assessment*** on the other hand is assessment of what someone can do/knows in relation to the application of the subject in the real world. It represents an external perspective.

Wie wirken sich diese Übersetzungen auf die Konzeptionalisierung der Begriffe *Kompetenz*, *Performanz* und *Leistung* im GER aus?

Der Terminus **Kompetenz** im GER umfasst in Beispiel (a) die Begriffe der *proficiency* und der *competence* im englischen Original: *level of proficiency* wird mit '*Kompetenzniveau'*, *proficiency* mit *Sprachkompetenz* und *competences* mit *Kompetenz* wiedergegeben. In Beispiel (b)

wird *Sprachkompetenz* gleichgesetzt mit *proficiency*, in Beispiel (c) mit *proficiency* und mit *linguistic knowledge*. Inwiefern diese inkonsistenten Übersetzungen den Kompetenzbegriff in der Sprachbeurteilung des GER verschleiern, wird unter Kapitel 2.5.2.2 näher ausgeführt.

Der Begriff der **Performanz** im GER bezieht sich in der Regel auf den Begriff der *performance* im CEF (vgl. Beispiel b); doch das Konzept der *performance* im CEF wird im GER nicht immer mit *Performanz* wiedergegeben (vgl. Beispiel (a), bei dem *performance* mit *Leistung* wiedergegeben wird, und s. die folgenden Ausführungen zum Leistungsbegriff).

Dem Konzept der **Leistung** im GER entsprechen mehrere Konzepte im CEF: Im obigen Beispiel (a) bezieht sich *Leistung* auf den Begriff *performance*, in den Beispielen (b) und (d) auf den Begriff *achievement*. Allerdings wird im letzten Beispiel der Begriff des *achievement* im deutschen Dokument gleichgesetzt mit den Begriffen *Sprachstand* und *Leistung* – doch der Begriff des globalen *Sprachstands* entspricht eher dem englischen Begriff der *proficiency*. So ist auch der *Leistungs*begriff, ähnlich wie der Kompetenzbegriff im GER durch problematische Übersetzungen undurchsichtig.

Eine Folge dieser Intransparenz macht sich beispielsweise an den Termini ***Sprachstandstest*** und ***Qualifikationsprüfung*** des Beispiels (d) bemerkbar: Die deutschen Termini sind unglücklich gewählt, impliziert *Sprachstand* doch das generelle Sprachvermögen, die *proficiency* also, während eine *Qualifikationsprüfung* konkrete Ziele impliziert, für die man sich in einer entsprechenden Prüfung qualifiziert. Wäre es da nicht zutreffender, *achievement test* mit *Lernzielkontrolle*, und *proficiency test* mit *Sprachstandstest* zu übersetzen?

Für eine Möglichkeit, die (deutsche und englische) Terminologie zu systematisieren, darf auf die Ausführungen zu Beginn dieses Kapitels 2 und die dort in Tabelle 2 vorgestellten Termini verwiesen werden.

2.5.2.2 Der Begriff der Kompetenz in GER-Abschnitt 9

Der Kompetenzbegriff, den der GER dem Testen von Sprachvermögen zugrunde legt, und die Möglichkeiten, die verschiedenen Teildimensionen dieser Kompetenz zu erfassen, werden durch die gerade dargestellten Übersetzungsprobleme verschleiert: Im deutschen Dokument umfasst der Kompetenzbegriff die im englischen Dokument differenzierten Begriffe der *competence* und der *proficiency*, enthält also neben einem theoriebasierten Konzept auch das Konzept der Beherrschung von Fertigkeiten. Wie oben zu Beginn dieses Kapitels erläutert, stammen diese beiden Begriffe aus unterschiedlichen Kontexten: *Proficiency* bezeichnet in der Regel das, was in Sprachtests beurteilt wird, wohingegen *competence* sich auf das dem Sprachvermögen zugrunde liegende theoretische Konstrukt bezieht. Generell kann auf Kompetenzen geschlossen werden über die Performanz oder über Testitems, die Wissensbestände erfassen; die

proficiency hingegen, der Grad der Beherrschung und Anwendbarkeit der Kompetenzen, kann im Sprachgebrauch beobachtet und beurteilt werden.

Wie wirkt sich nun die Subsumierung dieser beiden Konzepte unter dem Begriff der *Kompetenz* im deutschsprachigen Dokument aus? Zur Erfassung von Kompetenzen findet sich im GER folgende Aussage (ebd.: 182, Herv. d. V.):

> Leider kann man Kompetenzen niemals direkt testen. Man kann sich nur auf ein Spektrum von Beispielen der Performanz stützen, von dem aus man verallgemeinernde Schlüsse auf die Kompetenz zu ziehen versucht: *Kompetenz zeigt sich nämlich im Gebrauch*. In diesem Sinne *beurteilen alle Tests lediglich die Performanz*, obwohl man darüber hinaus zugleich Schlüsse auf die zugrunde liegende Kompetenz zu ziehen versucht.

Die Aussage, dass Kompetenzen nicht direkt beobachtbar sind, sondern über Indikatoren auf sie rückgeschlossen werden muss, deckt sich mit dem Verständnis von Kompetenz, das dieser Arbeit zugrunde liegt. Doch dass sich „Kompetenz im Sprachgebrauch" zeige, wäre bei einer theoriebasierten Konzeptionalisierung des Kompetenzbegriffs nicht nachvollziehbar. Es ist doch vielmehr die *proficiency*, der Grad der Beherrschung und Anwendbarkeit der Kompetenzen, der sich im Sprachgebrauch zeigt.

Ausgehend davon, dass der Begriff *Kompetenz* im GER auch den Aspekt der *proficiency* umfasst, werden Kompetenzen nach oben zitierter GER-Aussage über Performanzbeispiele erfasst. Doch es ist nicht zutreffend, dass „alle Tests lediglich die Performanz" beurteilen, denn beispielsweise erfassen indirekte Tests sprachbezogene Kenntnisse und rezeptive Fertigkeiten, die sich nicht in einer Performanz zeigen können – dies anerkennt der GER auch an anderer Stelle (ebd.: 182): „Bei der *Beurteilung von Kenntnissen* verlangt man von den Lernenden die Lösung von Aufgaben, (…) die einen Nachweis für den Umfang ihrer sprachlichen Kenntnisse und ihrer Beherrschung sprachlicher Mittel liefern sollen." Eine Möglichkeit, diese zu erfassen, sieht der GER in indirekten Testformaten (ebd.: 181): „*Indirekte Beurteilung* (…) benutzt einen Test, üblicherweise einen schriftlichen, der Kompetenzen und Fertigkeiten prüft, die einer Leistung zugrunde liegen." Klar davon unterschieden werden Performanzbeurteilungen durch direkte Formate (ebd.): „Bei der *Beurteilung der Performanz* verlangt man von den Lernenden, dass sie in einem direkten Test mündliche oder schriftliche Beispiele ihrer Sprachproduktion vorlegen." Wieso es auf dem Hintergrund dieser differenzierten Aussagen dennoch zur o. g. widersprüchlichen Schlussfolgerung im GER kommt, dass „… alle Tests lediglich die Performanz beurteilen", wird erst deutlich, wenn man sich die Konzeptionalisierung des Begriffs der *Performanz* im GER betrachtet:

Bei den gerade zitierten GER-Aussagen zur Beurteilung von Kenntnissen respektive Performanzen und bei den o. g. Aussagen zu direktem und indirektem Testen lässt sich eine konzeptionelle Unterscheidung von Performanzen und zugrunde liegenden Kompetenzen im GER ausmachen. Performanz im GER wird jedoch wiederum nicht transparent definiert; stattdessen werden mögliche Definitionen des Begriffs angeboten: Beispielsweise könne Performanz allgemein als „Bezeichnung von 'Sprachproduktion'" genutzt werden (hierunter wird auch das

Ankreuzen bei einer *Multiple-Choice*-Aufgabe verstanden, vgl. GER: 182); Performanz könne aber auch konkret definiert werden als „… relevante Sprachproduktion in einem (relativ) authentischen und für gewöhnlich berufs- oder ausbildungsbezogenen Kontext“ (GER: 182). Die Autoren des GER lassen jedoch offen, wie der Begriff der Performanz im GER definiert und genutzt wird. Dieser Begriff wird im GER zusätzlich getrübt, da die Konzepte *Performanz* und *Leistung* nicht stringent voneinander unterschieden werden, wie oben bei den Übersetzungsproblemen gezeigt.

Abschließend soll oben gegebenes Beispiel (a) die inhaltliche „Unverständlichkeit“ mancher GER-Textpassagen in Bezug auf die Erfassung von Kompetenzen illustrieren, die durch die Vermischung der Konzepte *Kompetenz*, *Performanz*, *Leistung* und *proficiency* bedingt ist (GER: 174f, Fettdruck d. Verf.):

> Es ist jedoch normalerweise *nicht* ratsam, Deskriptoren für 'kommunikative Aktivitäten' in den Kriterienkatalog für Prüfende bzw. Korrektoren aufzunehmen, wenn diese die **Leistung** in einem bestimmten Test zum Sprechen bzw. Schreiben auf dem jeweils erreichten '**Kompetenz**niveau' einstufen sollen. Der Grund dafür ist, dass sich eine Beurteilung der **Sprachkompetenz** nicht vorwiegend auf irgendeine spezielle einzelne **Leistung** beziehen, sondern vielmehr versuchen sollte, eine generalisierbare **Kompetenz** zu beurteilen, für die die vorliegende einzelne **Leistung** nur ein Indiz ist.

Es ist nicht nachvollziehbar, weshalb man die Deskriptoren für kommunikative Aktivitäten nicht zur Beurteilung derselbigen heranziehen sollte: Selbstverständlich kann man nur auf Kompetenzen rückschließen, wenn man genügend (Performanz-)Beispiele hat, um solche Generalisierungen vorzunehmen – und wieso sollte das gerade auf kommunikative Aktivitäten nicht zutreffen? Auch das im Zitat erwähnte Vorgehen der Einstufung von Performanzen auf ein bestimmtes Kompetenzniveau ist bedenklich – man sollte zur Einstufung tunlichst eigens dafür konstruierte Bewertungsskalen nutzen, die dann erst auf Kompetenzniveaus hin verallgemeinert werden müssen – eine direkte Einstufung auf ein Kompetenzniveau, das man ja gerade nicht direkt beobachten kann, sondern auf das man rückschließen muss, macht keinen Sinn und widerspricht dem auf S. 182 des GER dargelegten Kompetenzbegriff. Der zweite Satz des obigen Zitats trägt zudem wenig semantischen Gehalt. Man betrachte folgende Aussage, die sich durch einfaches Umstellen der Satzglieder ergibt: „Eine Beurteilung der Sprachkompetenz bezieht sich nicht vorwiegend auf irgendeine spezielle einzelne Leistung, sondern [eine Beurteilung der Sprachkompetenz] sollte eine generalisierbare Kompetenz beurteilen, für die die vorliegende einzelne Leistung nur ein Indiz ist.“ Wie aber soll eine Beurteilung der generellen Sprachkompetenz eine generalisierbare Kompetenz beurteilen? Die Übersetzer drehen sich im Kreis, da sie die Schlüsselkonzepte Kompetenz, Performanz und Leistung unzutreffend übersetzt beziehungsweise konzeptionalisiert haben.

Man kann ausgehend von der deutschen Fassung des GER nur schwer analysieren, welcher Kompetenzbegriff in GER-Abschnitt 9 mit dem Terminus *Kompetenz* gemeint ist: Die Übersetzer lassen den Rezipienten die Wahl zwischen dem theoretisch-linguistischen Konzept der Kompetenz,

welches sich bezieht auf Wissensbestände und Kenntnisse, die komplementär zur Performanz treten; einem anwendungsbezogenen Kompetenzbegriff, der sich auf die *proficiency*, den konkreten Grad an Sprachbeherrschung, auf Wissensbestände und deren Anwendbarkeit bezieht; und einem kommunikationsorientierten Kompetenzbegriff, der Sprachvermögen und Kommunikation eingebettet in deren sozio-pragmatische und kulturelle Auftrittsbedingungen betrachtet.

Es scheint geboten, dass Autoren und Übersetzter des GER ihre jeweils hinter den Termini stehenden Konzepte vergleichen und ein europäisches Termini-System schaffen, das der postulierten Transparenz des GER auch gerecht wird. Auch bietet es sich an, die Konzepte der *competence, performance* und *proficiency* in der englischen Originalausgabe zu analysieren, denn im Gegensatz zum GER werden diese drei Konzepte im CEF auch klar unterschieden. Diese CEF-Konzepte könnten dann mit den GER-Konzepten der *Kompetenz*, *Performanz* und *Leistung* vergleichen werden, doch das würde den Rahmen der vorliegenden Arbeit sprengen. Daneben wäre es ratsam, die Quellen offen zu legen, aus denen sich der Begriff der Kompetenz im GER respektive die Begriffe der *competence* und der *proficiency* im CEF speisen. In diesem Zusammenhang darf auf die aufschlussreichen Ausführungen in North (2000: 41ff) verwiesen werden: Dort trifft North die grundlegenden Unterscheidungen zwischen den Konzepten *performance*, *proficiency* und *competence*, die die theoretische Basis des Referenzsystems im CEF bilden. Die Ausführungen in North (2000) sind klar, konsistent und mit Quellen belegt. Sie werden in Kapitel 3 dieser Arbeit wieder aufgenommen, wenn die Basis der horizontalen Einteilung des GER-Skalensystems (d.h. die Einteilung in die Kategorien des GER) analysiert wird.

2.5.3 Der Testbegriff des GER

Der Testbegriff des GER zeigt sich im Großen und Ganzen konsistent mit dem kommunikativ-handlungsorientierten Ansatz, den der GER in seinen Abschnitten 4 und 5 entwickelt. Folgende Prinzipien zeichnen sich in GER-Abschnitt 9 ab:

- *Kommunikatives Testen*: Die kommunikativ-handlungsorientierten Kategorien und Deskriptoren des GER-Abschnitts 4 „Kommunikative Aktivitäten“ werden in GER-Abschnitt 9 wieder aufgenommen; dort wird konkretisiert, wie die Deskriptoren der Kategorien der kommunikativen Aktivitäten beim Testen des Sprachvermögens verwendet werden können (vgl. dazu auch die Ausführungen unten). Der Testansatz im GER kann deshalb als kommunikativ-handlungsorientiert bezeichnet werden. Folgende Charakteristika dieses Ansatzes lassen sich im GER (GER 2001: 178) ausmachen: Sprachtests sind im GER situiert in einem „bedürfnisorientierten Lehr-/Lernkontext“, sie prüfen „die praktische Sprachverwendung in relevanten Situationen“, und sie bestehen idealiter „aus Sprach- und Kommunikationsaufgaben ..., die den Lernenden Gelegenheit geben zu zeigen, was sie erreicht haben“. Damit sind Ollers oben erwähnte

naturalness criteria[137], namentlich die Situierung von Test*items* in bedeutungsvollen Kontexten und authentischen Bedingungen, erfüllt. Bachmanns o. g. Kriterien für kommunikative Tests sind zum Teil erfüllt: Die Forderung nach einer Informationslücke, die es zu schließen gilt, und die Forderung, dass *tasks* aufeinander aufbauen sollen, spielen bei der generellen Natur des Referenzrahmens keine Rolle – sie müssen von den jeweils für die Testerstellung Verantwortlichen umgesetzt werden. Jedoch wird die von Bachmann geforderte Integration von *tasks*, Inhalten und Diskursdomänen auch im GER-Testansatz gefordert: Aufgaben und deren Inhalte werden in bestimmten Domänen verankert (ebd. 153):

> Kommunikative Aufgaben sind ein Merkmal des alltäglichen Lebens im privaten, öffentlichen und beruflichen sowie im Bildungsbereich. Die Bewältigung einer kommunikativen Aufgabe beinhaltet die strategische Aktivierung spezieller Kompetenzen, um innerhalb eines bestimmten Lebensbereichs eine Gruppe zielgerichteter Handlungen mit einem klar definierten Ziel und einem speziellen Ergebnis auszuführen (vgl. Abschnitt 4.1).

Bachmanns letztes Kriterium für kommunikative Tests, die Forderung nach möglichst breiter Erfassung vielfältiger sprachrelevanter Bereiche, ist ebenfalls erfüllt, wenn man folgendes Charakteristikum des GER-Testansatzes betrachtet:

- *Verknüpfung aller Wissens- und Kompetenzbereiche*: Dieser Verknüpfung liegt der in Kapitel 1.2.5.3 dieser Arbeit analysierte idealisierte Kompetenzbegriff der GER-Abschnitte 4 und 5 zugrunde, der sprachliche Wissensbestände, kommunikative Aktivitäten, Strategien, deklaratives und prozedurales Wissen, persönlichkeitsbezogene Kompetenzen, Lernfähigkeit, Handlungsfähigkeit und sprachlich-kommunikative Fertigkeiten als miteinander verbunden betrachtet (ebd.: 22):

> Unterstellt man, dass die (…) verschiedenen Dimensionen bei der Verwendung von Sprache und beim Sprachenlernen miteinander verknüpft sind, dann ist jede Tätigkeit beim Sprachenlernen und -lehren auf die eine oder andere Weise mit jeder dieser Dimensionen verbunden: mit den Strategien, den Aufgaben, den Texten, mit den allgemeinen Kompetenzen und der kommunikativen Sprachkompetenz, mit kommunikativen Sprachaktivitäten und Sprachprozessen, mit Kontexten und mit Lebensbereichen.

Dieser Verbundenheit der Dimensionen können Tests laut GER (ebd.: 178) gerecht werden, indem sie darauf abzielen, „ein ausgewogenes Bild einer sich entwickelnden Kompetenz zu bieten“. Diese Ausgewogenheit könne erzielt werden über das Ansetzen jeweils relevanter Bereiche, die abgestuft auf einem Kontinuum der Sprachkompetenz dargestellt werden (ebd.: 179).

- *Kriteriumsorientiertes Testen*: Die Referenzniveaus stellen nach Ansicht der GER-Autoren ein „System allgemeiner Standards“ (GER: 179) dar. Die Skalen und Beispieldeskriptoren sind „aus kriteriumsorientierten Aussagen für Kategorien des Bescheibungssystems“ (ebd.) gewonnen. Relevante Kriterien werden durch die Kategorien der GER-Abschnitte 4 und 5 zur Verfügung gestellt. Die Kriterien können gemäß GER-Aussagen (ebd.: 179f) benutzt werden, um Lernfortschritte auf einem Kontinuum (den Skalen und Niveaus des GER) darzustellen oder um diese am Erreichen eines bestimmten Lernziels (etwa eines äußeren Kriteriums oder eines

[137] Zu den Kriterien von Oller vgl. Kapitel 2.2.1 dieser Arbeit, zu denen von Bachmann Kapitel 2.2.2.

Kriteriums aus dem Katalog des GER) zu messen. Durch diesen Ansatz wird beispielsweise die Überprüfung von Standards erst ermöglicht: Eine normorientierte Leistungsmessung, bei der die Gruppe der Lernenden als Bezugspunkt genommen wird, lässt Aussagen darüber zu, wo in Relation zur Lernergruppe sich ein Individuum befindet; sie lässt jedoch keine Aussagen bezüglich des Erreichens eines bestimmten Standards oder Kriteriums unabhängig vom Leistungsstand der Gruppe zu. Daher können Vergleiche über verschiedene Institutionen oder Qualifikationen hinweg nur schwerlich gezogen werden (vgl. dazu auch die Ausführungen oben unter Kapitel 2.2.3). Setzt man aber als Bezugspunkt ein Kriterium an (sei es nun ein konkretes Lernziel oder bestimmte Standards, die erreicht werden sollen), können die individuellen Lernenden in Bezug auf dieses Kriterium eingeschätzt werden, unabhängig davon, wie „gut" oder „schlecht" sie in Bezug auf den Rest der Lernergruppe sind. Diese Einschätzungen können dann in verschiedenen Kontexten genutzt werden, etwa zu einem Vergleich verschiedener Prüfungen (vgl. GER: 176f).

- *Direkter Testansatz*: Bei der Beurteilung von Performanz in den produktiven Fertigkeiten erwähnt der GER das direkte Testen als angemessene Vorgehensweise: Da rezeptive Aktivitäten selbstverständlich nicht direkt beobachtet werden können, beschränkt der GER das direkte Testen „im Grunde … auf das Sprechen, Schreiben und Zuhören bei Interaktion" (GER: 181). Der GER will „…einen Fundus für die Entwicklung *genau definierter, spezifischer Kriterien* für direkte Tests zur Verfügung stellen" (ebd.: 183). Dieser direkte Testansatz erfordert aber auch subjektive Bewertungsweisen:

- *Positive und subjektive Bewertung*: Der GER schlägt subjektive Bewertung bei direkten Formaten vor, da diese der Komplexität von Sprache und Kommunikation gerecht werde (GER: 182f). Er zeigt Möglichkeiten auf, wie ein subjektives Urteil so weit objektiviert werden kann, dass unter verschiedenen Bewertern hinreichender Konsens erzielt wird (ebd.: 183):

  - Man entwickelt *inhaltliche Vorgaben* für die Beurteilung, z. B. basierend auf einem Referenzrahmen für den betreffenden Kontext;
  - man stützt sich bei der Auswahl von Inhalten und/oder der Beurteilung der Leistungen auf *gemeinsame Entscheidungen*;
  - man verwendet *Standardverfahren*, die festlegen, wie geprüft wird;
  - man stellt *verbindliche Bewertungsschlüssel* für indirekte Tests zu Verfügung und stützt die Urteile in direkten Tests auf *spezifische, klar definierte Kriterien*;
  - man fordert *mehrfache Beurteilung* und/oder die *Gewichtung verschiedener Faktoren*;
  - man bietet entsprechendes *Training* in Bezug auf die *Beurteilungsrichtlinien* an;
  - man kontrolliert die Qualität von Leistungsbeurteilungen (Validität, Reliabilität) durch eine *Analyse der Prüfungsdaten*.

  Wie am Anfang dieses Abschnitts bereits dargelegt, bestehen erste Schritte in Richtung auf eine Verminderung der Subjektivität auf allen Stufen eines Beurteilungsverfahrens darin, ein gemeinsames Verständnis vom betreffenden Konstrukt herzustellen, d. h. einen gemeinsamen Bezugsrahmen. Der *Referenzrahmen* versucht, eine solche Basis für die *Beschreibung der Inhalte* und einen Fundus für die Entwicklung *genau definierter, spezifischer Kriterien* für direkte Tests zur Verfügung zu stellen.

Ob der GER diesem Anspruch gerecht werden kann, soll in Kapitel 4 dieser Arbeit untersucht werden: Dort wird am Beispiel einer offenen, direkten Schreibaufgabe gezeigt, inwieweit der GER bei der Testentwicklung und Auswertung helfen kann und wo seine Grenzen liegen.

Der Ansatz der Positivbewertung, namentlich die KANN-Beschreibungen in den Deskriptoren der Skalen sind das Novum und die Stärke des GER: Es ist an der Zeit, bei der Beurteilung und Bewertung zuerst nach dem zu sehen, was die Lernenden schon können, ehe man sich mit dem beschäftigt, was durch die Lernsituation bedingt noch fehlt. Nur wenn man sich loslöst von der Tradition des „Fehler-Anstreichens", welche demotiviert und wenig Aufschlüsse über Lernfortschritte gibt, wird man die motivierende und lernfördernde Wirkung einer Positivbewertung wahrnehmen können. In diesem Zusammenhang kann der GER einen wertvollen Beitrag zur „Neuorientierung" in der Beurteilung des Sprachvermögens leisten. Der GER (ebd.: 184f) bietet an, seine Skalen zur Erstellung und Abstufung von Bewertungskriterien zu nutzen, sie bei der Bewertung einzusetzen um zu gelenkten Urteilen statt rein subjektiver Eindrücke zu kommen und um holistische wie analytische Bewertungsverfahren zu kombinieren. Die Referenzniveaus bietet der GER (ebd.: 183) als Möglichkeit, zu einem breiten Konsens über das Verständnis bestimmter Kompetenzniveaus zu kommen. Ob die Skalen all diesen Funktionen gerecht werden können, wird in Kapitel 3 dieser Arbeit untersucht.

- *Selbstbeurteilung*: Der GER lässt an verschiedenen Stellen die Bedeutsamkeit der Selbstbeurteilung erkennen: Schon zu Beginn wird festgestellt, dass der GER u. a. folgendem Zweck dienen soll (ebd: 18):

> (…) der Planung von selbst bestimmtem Lernen, was mit einschließt,
> - das Bewusstsein der Lernenden für den Kenntnisstand, den sie erreicht haben, zu entwickeln;
> - dass erreichbare und sinnvolle Lernziele durch die Lernenden selbst festgelegt werden;
> - die Auswahl von Lernmaterialien;
> - die Anwendung von Instrumenten der Selbstbeurteilung.

Einer der Verwendungszwecke des GER wird auf S. 30 wie folgt beschrieben:

> Er [der GER] kann Kriterien bereitstellen, mit deren Hilfe man bei der Beurteilung einer bestimmten mündlichen oder schriftlichen Leistung feststellen kann, ob ein Lernziel erreicht wurde oder nicht, und er kann dies sowohl für die kontinuierliche Beurteilung durch Lehrende oder die Lerngruppe tun als auch für die Selbstbeurteilung.

Der GER betrachtet Selbstbeurteilung als Teil der Selbstbestimmung, welcher im Sprachlernprozess eine nicht zu unterschätzende Bedeutung zukommt (vgl. auch die Ausführungen unter Kapitel 1.3.1 respektive 1.3.4 dieser Arbeit): Um *feedback* aus Beurteilungen umsetzen zu können, brauche es die Wahrnehmungs- und Aufnahmefähigkeit der Lernenden (vgl. GER 2001: 181): Sie müssen Lücken in ihrem Wissen identifizieren und interpretieren können, um neue Informationen ins bestehende Wissenssystem zu integrieren und den weiteren Lernweg festzulegen. Dabei könne das Selbstbeurteilungsraster aus GER-Abschnitt 3 (ebd.: 36) helfen.[138] Darauf müssten die Lernenden aber vorbereitet werden, etwa in Form eines „bewusstseinsbildenden Trainings", welches den eigenen „subjektiven Eindruck (…) mit der Realität zu vergleichen" sucht (ebd.: 181). Selbstbeurteilung stellt der GER in gewissen Kontexten als „eine wirkungsvolle

[138] Auch die Entwicklung des Sprachenportfolios in Anlehnung an den GER ist ein Beleg für die zunehmende Bedeutung der Selbstbeurteilung im europäischen Kontext. Da Selbstbeurteilung jedoch nicht Thema der vorliegenden Arbeit ist, kann das Sprachenportfolio im Rahmen dieser Arbeit nicht analysiert werden.

Ergänzung für Tests“ (ebd.: 186) dar, doch die wichtigste Funktion sieht er in der Motivation (ebd.):

> Die größte Bedeutung hat die Selbstbeurteilung aber als ein Instrument für die Motivation und für ein bewussteres Lernen: So kann sie den Lernenden helfen, ihre Stärken richtig einschätzen zu lernen, ihre Schwächen zu erkennen und ihr Lernen effektiver zu gestalten.

2.5.4 GER-Aussagen bezüglich seiner Verwendungsmöglichkeiten bei der Beurteilung des Sprachvermögens

Der GER stellt die „drei wichtigsten Verwendungszwecke“ (GER 2001: 30) des Instruments im Kontext der Sprachbeurteilung[139] in seinem Abschnitt 9.2 vor (GER: 173-177): Der GER könne verwendet werden zur inhaltlichen Beschreibung von Tests oder Prüfungen, zur Festlegung der Kriterien der Beurteilung oder Bewertung und nicht zuletzt zur Beschreibung von Kompetenzniveaus, um verschiedene Qualifikationssysteme vergleichen zu können. Dabei könnten zur inhaltlichen Beschreibung die in GER-Abschnitt 4 dargestellten kommunikativen Aktivitäten und Strategien beitragen, genau wie die in GER-Abschnitt 5 beschriebenen sprachlichen Teilkompetenzen und die Aufgabenmerkmale aus GER-Abschnitt 7. Um relevante Beurteilungs- bzw. Bewertungskriterien zu finden, könnten neben theoretischen Analysen wiederum die GER-Abschnitte 4 und 5 genutzt werden, um Kriterien für „kommunikative Aktivitäten“ und für die dabei involvierten „Aspekte der Sprachbeherrschung“ zu finden (GER: 174). Zum Vergleich verschiedener Qualifikationssysteme böten sich die Skalen aus den Abschnitten 4 und 5 sowie die Lernzielniveaus des Europarates an.

Die eindeutigen Verweise auf die Verwendungsmöglichkeiten des GER-Systems scheinen benutzerfreundlich, doch die Möglichkeiten beziehen sich letztlich alle auf den Einsatz der GER-Skalen. Ob diese direkte Verwendung der GER-Skalen in der in GER-Abschnitt 9.2 vorgeschlagenen Weise jedoch tatsächlich möglich ist, kann, wie oben schon angedeutet, erst nach Analyse des GER-Skalenansatzes und des Status der GER-Skalen abschließend beurteilt werden: Deshalb ist den GER-Skalen das Kapitel 3 dieser Arbeit gewidmet, welches auf theoretischer Basis die Verwendungsmöglichkeiten der GER-Skalen untersucht, während Kapitel 4 die Verwendbarkeit der GER-Skalen anhand eines Praxisbeispiels beurteilt. Hier nun werden im Folgenden die GER-Aussagen zur Verwendbarkeit seines Referenzsystems für einen ersten Überblick zusammengefasst. In Kapitel 3.4.4.4 dieser Arbeit werden die betreffenden GER-Aussagen nach der GER-Skalenanalyse wieder aufgenommen und abschließend beurteilt.

Der GER gibt auf S. 173f als groben Rahmen zur *Beschreibung von Testaufgaben* – der ersten Verwendungsmöglichkeit – die Taxonomie der kommunikativen Aktivitäten (Abschnitt 4.4) „Produktion vs. Interaktion im mündlichen vs. schriftlichen Bereich“ vor. Aufbauend darauf

[139] Weitere Nutzungsmöglichkeiten des GER-Systems jenseits derer im Kontext der Sprachbeurteilung werden in Kapitel 3.4.4.3 dieser Arbeit vorgestellt und beurteilt.

seien die Aspekte „Kontext der Sprachverwendung“ (Abschnitt 4.1), „Texte“ (Abschnitt 4.6) und „Schwierigkeitsgrade kommunikativer Aufgaben“ (Abschnitt 7.3) zu beachten. Die inhaltliche Basis bei der Testkonstruktion sei in den „kommunikativen Sprachkompetenzen“ (Abschnitt 5.2) zu finden. Dieser Rahmen mag hilfreiche Überlegungen bei der Testentwicklung bieten, ist jedoch – wie der GER explizit erwähnt – immer von den jeweiligen Nutzern im jeweiligen Fall zu füllen. Die Inhalte und Aspekte des GER alleine werden nicht ausreichen, Testinhalte in einem theoretischen Testkonzept zu verankern. Der GER kann durchaus hilfreiche Rahmenpunkte und Zusatzinformationen liefern, doch kann er keine theoretische Basis ersetzen, aufgrund derer eine Beschreibung relevanter Merkmale der Testaufgaben, welche aus einem gegebenen Testkonzept entwickelt wurden, möglich wäre. [140]

Auf S. 174ff gibt der GER eine Kurzdarstellung der zweiten Nutzungsmöglichkeit seiner Skalen: der *Entwicklung von Bewertungskriterien*. Der GER sieht seine Skalen als „Quelle zur Entwicklung von Bewertungsskalen“ und die Deskriptoren als „Hilfe bei der Formulierung von Kriterien“ (GER: 174). Dabei wird unterschieden zwischen den Skalen für kommunikative Aktivitäten des GER-Abschnitts 4 und jenen für Aspekte der Sprachbeherrschung des GER-Abschnitts 5: Die Skalen des GER-Abschnitts 4 könnten helfen bei den (sehr unterschiedlichen) Funktionen der Aufgabenerstellung, der Rückmeldungen oder der Beurteilung selbst (GER: 174), die Skalen der Sprachbeherrschung hingegen seien nützlich zur Selbst- und Fremdbeurteilung und zur Beurteilung der Performanz (GER: 175f). Im Zusammenhang mit der Entwicklung von Beurteilungsskalen werden im GER die Begriffe *Sprachkompetenzskala* und *Bewertungsskala für Prüfungen*[141] unterschieden (ebd.: 176), wobei es scheint, dass der GER unterstellt, man könne diese beiden ganz unterschiedlichen Skalentypen aus denselben Deskriptoren des GER entwickeln. Diese Aussage scheint gewagt, denn der Status der Deskriptoren und Skalen des GER ist undurchsichtig.[142] Es handelt sich, wie später noch gezeigt wird, um *proficiency scales*, doch ob diese Skalen tatsächlich zur direkten Bewertung einer Performanz hergenommen werden können, sei momentan dahingestellt – diese und damit verwandte Fragen werden im Detail in Kapitel 3 dieser Arbeit erörtert. In jedem Fall muss der Entwicklung von Bewertungskriterien nicht nur eine GER-Skalen- und Deskriptorenanalyse vorangehen, sondern valide Bewertungskriterien müssen bei einem gegebenen Test aus dem konkreten Testkonstrukt abgeleitet werden. Erst auf dieser Basis können die aus dem Testkonstrukt abgeleiteten Bewertungskriterien mit relevanten Aussagen aus dem GER abgeglichen werden. Ansonsten

[140] Zur Beschreibung von Testaufgaben vgl. auch das Kriterienraster des so genannten *Dutch Grid*, der im September 2004 veröffentlicht wurde und in den Kapiteln 3.4.4 und 3.5 dieser Arbeit angeführt wird. Es zeigt Lücken und Inkonsistenzen in den Kategorien *Hörverstehen* und *Leseverstehen* im System des GER auf und versucht diese zu schließen, um ein Raster zu entwickeln, mit dessen Hilfe Testitems auf die Niveaus des GER eingestuft werden können (vgl. Alderson et al. 2004).

[141] An dieser Stelle darf auf die Termini des englischen Originals verwiesen werden: *Sprachkompetenzskala* wird dort als *Proficiency Scale* bezeichnet, und *Bewertungsskala für Prüfungen* als *Examination Rating Scale* (CEF: 181).

[142] Der Status einer Skala hängt von ihrem Beschreibungsgegenstand und ihrer Ausrichtung ab und bestimmt die Funktionen, die die betreffende Skala übernehmen kann. Beispielsweise gibt es Skalen, die der Testaufgabenkonstruktion dienen und deshalb konkrete Aufgabenmerkmale beschreiben. Dieser Typus Skalen unterscheidet sich stark etwa vom Typ der Kompetenzskalen, die in der Regel auf Rückmeldungen von Testergebnissen ausgelegt sind und deshalb generalisierte Kompetenzen der Lernenden beschreiben. Bezüglich des Status der GER-Skalen darf auf die Ausführungen in Kapitel 3.1 dieser Arbeit zu den verschiedenen Typen von Skalen und auf die Ausführungen in Kapitel 3.4.4 dieser Arbeit zu Status und Verwendbarkeit der GER-Skalen verwiesen werden.

dürfte den Bewertungskriterien eine valide Grundlage fehlen – diese auf konkrete Kontexte bezogenen Grundlagen kann der GER jedoch aufgrund seines Rahmencharakters nicht bieten.

Die dritte Anwendungsmöglichkeit sieht der GER in der „Beschreibung von Kompetenzniveaus“, die es dann erlaube, verschiedene Prüfungen miteinander zu vergleichen. Der GER erwähnt klassische Verfahren des Vergleichs (GER: 176): Neben Gleichung, Kalibrierung und statistischer Prüfung[143] schlägt er *benchmarking* und gemeinsame Standardfindung durch Konsens vor. Dabei sollen die Skalen des GER dazu beitragen, Referenzniveaus zu schaffen, die von allen Betroffenen geteilt und in vergleichbarer Weise interpretiert und verstanden werden können. Wenn die GER-Skalen die Beschreibung solcher Standards unterstützen sollen, dann weist dies wiederum auf den Status der Skalen als „Kompetenzskalen“ hin, was im Widerspruch zur o. g. GER-Postulation der Verwendungsmöglichkeit als *rating scales* steht. Der GER sieht seine Skalen als „Begriffsraster“ (ebd.: 177), um verschiedene Systeme und Prüfungsziele aufeinander zu beziehen – ein weiterer Hinweis auf den Status der Deskriptoren und Skalen, welche nicht die Performanz beschreiben, sondern eher Standards und generalisierte Aspekte der Sprachbeherrschung. Folgende Aussage beleuchtet die Hauptaufgabe des GER (ebd.):

> Es ist eines der Ziele des Referenzrahmens, [den] Prozess des Aufbaus eines gemeinsamen Verständnisses [im Hinblick auf Standards bei der Beschreibung von Kompetenzniveaus] zu unterstützen. (...)
>
> Man ist sich darüber einig, dass die Entwicklung eines standard-orientierten Ansatzes zeitaufwändig ist, weil die am Verfahren Beteiligten nur durch einen Prozess der Veranschaulichung und des Austauschs von Meinungen ein Gefühl dafür erwerben, was bestimmte Standards bedeuten. (...)
>
> Der Referenzrahmen bietet einen fundierten Lösungsversuch für dieses vorrangige und grundlegende Problem des Lernens moderner Sprachen [bezogen auf die Vergleichbarkeit der Referenzniveaus] in einem europäischen Kontext. In den Kapiteln 4 bis 7 wird ein Beschreibungssystem dargestellt, das auf praxisorientierte Weise Sprachverwendung, Kompetenzen und die Prozesse des Lehren und Lernens begrifflich zu klären versucht.

Wiederum wird deutlich, dass die Skalen den Kern des Referenzrahmens bilden und dass es das Hauptanliegen des GER ist, Konsens bezüglich der gemeinsamen Referenzniveaus aufzubauen, auch wenn in obigem Zitat postuliert wird, dass „Prozesse des Lehrens und Lernens begrifflich“ geklärt würden – wie in Kapitel 1.3 dieser Arbeit jedoch analysiert, ist dies in umfassender Form nicht der Fall. Die Bedeutsamkeit der erwähnten Konsensfindung und der Vergleichbarkeit von Qualifikationsprüfungen steht gerade im europäischen Kontext außer Frage. Offenbar war den Autoren bewusst, dass der GER dabei nur in unzureichender Form einen Rahmen stecken kann, da es das bereits erwähnte Zusatzdokument zur Anbindung von Tests an das Niveausystem des GER gibt (das so genannte *Manual*, vgl. Council of Europe 2003a), welches in Kapitel 3.5 dieser Arbeit untersucht wird.

Die Frage, ob die GER-Skalen und Deskriptoren tatsächlich für die drei hier grob umrissenen Zwecke im Zusammenhang mit der Beurteilung von Sprachvermögen genutzt werden können, die der GER in seinem Abschnitt 9.2 auflistet, kann wie gesagt abschließend erst in

[143] Für Erläuterungen dieser Konzepte vgl. das Glossar.

Kapitel 3.4.4.4 dieser Arbeit beurteilt werden, aufbauend auf detaillierten Skalen-Analysen, die eine Beurteilung des Status und damit der Verwendbarkeit der betreffenden GER-Skalen zulassen.

2.5.5 Fazit

Das Hauptanliegen des GER wird durch seinen Untertitel „Sprachen: lernen, lehren, beurteilen" eher verdeckt: Es drängt sich der Eindruck auf, dass es dem GER nicht primär darum geht, die Begriffe *Sprache*, *Lernen* und *Lehren* näher zu beleuchten und Konzepte des Lernens und Lehrens zu entwickeln, sondern vorwiegend scheint der GER auf die *Beurteilung* von Sprachvermögen ausgelegt zu sein. Der GER ist nicht so sehr bemüht um die postulierte Förderung der Mehrsprachigkeit, die Diversifizierung von Lernangeboten, die Entwicklung eines (europäischen) Rahmens für das Lehren und Lernen von Sprachen, sondern eher bemüht um die Herausbildung eines gemeinsamen Referenzsystems zur Beurteilung von Sprache und zum Vergleich verschiedener Qualifikationssysteme. Hierfür lassen sich viele Belege im GER finden: Gleich zu Beginn etwa werden die Niveaus zusammen mit „Messverfahren" und „Prüfungen" erwähnt (GER 2001: 10, Herv. d. V.):

> (…)Niveaustufen sollten – wie übrigens alle *Messeinheiten* – nicht unnötig vermehrt werden! (…)
>
> Das System mit sechs Niveaustufen, das wir durchgängig benutzen, entspricht dem, was bei einer Reihe von *Prüfungsanbietern* üblich ist. (…)
>
> Es hat sich jedoch schon gezeigt, dass ein System von Referenzniveaus als Kalibrierungsinstrument besonders von Praktikern aller Art begrüßt wird, die es, wie in vielen anderen Arbeitsbereichen auch, vorteilhaft finden, mit *stabilen und anerkannten Mess- und Formatstandards* zu arbeiten.

Dann finden sich in GER-Abschnitt 2.2 Aussagen zu Zwecken und Aufgaben eines Referenzsystems: Neben der Beschreibung von Lernzielen und der Entwicklung von Lernprogrammen dienten die Niveaubeschreibungen der Vergleichbarmachung von Qualifikationen und der Beurteilung von Lernfortschritten (GER: 27f, Herv. d. V.):

> - Die Definition von Kompetenzbeschreibungen mit Hilfe der Kategorien des ‚Referenzrahmens' könnte dazu beitragen zu konkretisieren, was sinnvollerweise auf den einzelnen Niveaus erwartet werden kann. Dies wiederum könnte zur Entwicklung von transparenten und realistischen Beschreibungen von globalen Lernzielen beitragen.
> - Wenn sich Lernen über einen längeren Zeitraum erstreckt, muss es in Einheiten unterteilt werden, die Progressionen bzw. Lernfortschritte berücksichtigen und Kontinuität gewährleisten. Lernzielbeschreibungen und Materialien müssen aufeinander bezogen werden. Ein Referenzrahmen mit Niveaustufen wäre für diesen Prozess äußerst hilfreich.
> - *Lernleistungen* im Rahmen solcher Ziele und Einheiten müssen ebenfalls auf der vertikalen Achse eingeordnet, d. h. als Fortschritte in der Sprachkompetenz *beurteilt* werden. Sprachkompetenzbeschreibungen könnten dabei hilfreich sein.
> - Eine solche *Beurteilung* sollte auch zufälliges Lernen, außerschulische Erfahrungen und "nebenbei" erworbene Bereicherungen der Sprache berücksichtigen, wie wir sie weiter oben erwähnt haben. Die Bereitstellung von Sprachkompetenzbeschreibungen, die über den Bereich spezifischer Lehrpläne hinausgehen, wäre für diesen Zweck nützlich.
> - Ein System von gemeinsamen Sprachkompetenzbeschreibungen erleichtert den *Vergleich* von Lernzielen, Niveaustufen, Materialien, *Tests* und von Lernerfolgen in unterschiedlichen Systemen und Situationen. (…)

Der enge Zusammenhang zwischen GER und Beurteilung wird schon bei den einführenden Erläuterungen zum Ansatz des Referenzrahmens deutlich (ebd.: 30, Herv. d. V.):

> Dieses Dokument ist ein Gemeinsamer Europäischer Referenzrahmen für Sprachen: Lernen, lehren und beurteilen. Bis zu diesem Punkt lag der Schwerpunkt unserer Ausführungen auf dem Wesen der Sprachverwendung und des Sprachverwendenden sowie auf den Implikationen für das Lernen und Lehren. In Kapitel 9, dem letzten Kapitel, richten wir unsere Aufmerksamkeit auf die *Funktionen, die der Referenzrahmen für die Beurteilung und die Bewertung von Sprachkompetenz* hat. Das Kapitel umreißt die wichtigsten drei Verwendungszwecke des Referenzrahmens (…)

Wenngleich dort zwar die Rede ist von „bisherigen Schwerpunkten" und von Abschnitt 9 als „letztem Kapitel", vermag der GER jedoch bis zu besagtem Abschnitt 9 den Eindruck eines beurteilungslastigen Instrumentes nicht widerlegen. Folgende Aussage stellt klar, zu welchem Zweck die Niveaus entwickelt worden sind (ebd.: 32, Herv. d. V.):

> Eines der Ziele des *Referenzrahmens* ist es, allen beteiligten Partnern bei der Beschreibung der *Kompetenzniveaus zu helfen, die gemäß den Standards ihrer Tests und Prüfungen erwartet werden*. Dies soll den *Vergleich zwischen verschiedenen Qualifikationssystemen erleichtern*. *Zu diesem Zweck sind ein Beschreibungssystem und die Gemeinsamen Referenzniveaus entwickelt worden*.

Die Referenzniveaus und das Skalensystem sind das „Herzstück" des GER. Allerdings werden die GER-Skalen in ihrer Einsetzbarkeit bei der Beurteilung und Bewertung überschätzt. Hier gilt es, mit Vorsicht heranzugehen, denn ein Einsatz zu anderen als den bei der Konstruktion intendierten Zwecken kann zu einem Validitätsverlust der Skalen führen (Näheres dazu in Kapitel 3 dieser Arbeit). Positiv fällt auf, dass in GER-Abschnitt 9 Stellung zu den Verwendungsmöglichkeiten der Skalen und Deskriptoren in der Beurteilung des Sprachvermögens genommen wird, wenn auch – wie noch gezeigt wird – nicht alle Aussagen nachvollziehbar oder in sich schlüssig sind. Für eine Analyse des GER-Skalenansatzes und seiner Verwendbarkeit darf wie gesagt auf Kapitel 3 dieser Arbeit verwiesen werden.

Die GER-Ausführungen zur Beurteilung des Sprachvermögens sind im Grunde nur auf das eigene Referenzsystem und auf die Einsatzmöglichkeiten der Kategorien, Skalen und Deskriptoren des GER ausgelegt. Jenseits dieser Bereiche werden wesentliche Dinge im GER nicht erwähnt: Beispielsweise wird die Problematik der unterschiedlichen Testformate (vgl. oben die Ausführungen zu den diskreten, integrativen und kommunikativen Formaten in Kapitel 2.2 dieser Arbeit) ignoriert, ebenso wie die Bedeutsamkeit der in Kapitel 2.3 dieser Arbeit erwähnten systemischen Validität, wo diese von einem Instrument, das sich auch mit Qualifikationsvergleichen im europäischen Bildungsbereich beschäftigt, nicht unterschätzt werden dürfte. Auch wird der in Kapitel 2.4 dieser Arbeit erläuterte *instructional value* nicht in dem Maße thematisiert, wie sich das für ein Instrument zur Förderung des Sprachlernangebots verstehen würde. Man findet zwar einige relevante Aussagen in diesem Zusammenhang, beispielsweise die Ausführungen zu *Feedback* (GER: 181) oder der motivierenden Wirkung, die von Selbstbeurteilung ausgehen kann (ebd.: 186); es finden sich jedoch keine Hinweise, wie Tests mit *instructional value* entwickelt werden könnten.

Der GER kann aufgrund seines auf sein eigenes Referenzsystem ausgelegten Testbegriffs kein theoretisches Rahmenkonzept, kein in wissenschaftlichen Theorien verortetes Testkonstrukt ersetzen – er kann lediglich Eckpunkte geben und als ein möglicher Referenzpunkt dienen, nicht aber als Basis oder Ausgangspunkt der Testentwicklung. Professionelles Wissen um den Prozess der Testentwicklung, die Konsultation von Fachliteratur, die Anwendung wissenschaftlicher Theorien und Modelle u. v. m. sind unabdingbare Voraussetzungen für eine valide Testentwicklung.

2.6 Testentwicklungsprozess und der UGE

In Kapitel 2 der vorliegenden Arbeit wurden wesentliche und relevante Konzepte im Zusammenhang mit Sprachtests erörtert, auf deren Basis der Testansatz im GER analysiert wurde. Die bisher angesprochenen Aspekte der Kompetenzmodelle und Leistungsdimensionen, der Testformate, der Gütekriterien, der Testziele und der Charakteristika valider Tests hängen, wie zu Beginn dieses Kapitels dargestellt, zyklisch miteinander zusammen, weshalb sie nun eingebettet werden sollen in den größeren Kontext des Testentwicklungsprozesses. In diesem Zusammenhang interessiert, welche Aspekte in welcher Entwicklungsphase realisiert werden – wann beispielsweise wird das Testformat bestimmt, wann wird etwa die Testvalidität geprüft, oder wann sollten Testziele festgelegt werden?

Testentwicklung ist ein komplexer Prozess, der in bestimmten Phasen abläuft und an dem in Idealfall ein interdisziplinäres Team von Experten teilnimmt. Dieser Prozess ist vielbeschrieben[144], weshalb sich die vorliegende Arbeit auf einen generellen Überblick beschränkt. Denn jede Testentwicklung ist anders angelegt und unterliegt anderen Bedingungen; somit müssen alle Entscheidungen jeweils neu begründet und getroffen werden. Was für eine Situation eine angemessene Entscheidung darstellt, muss nicht automatisch in einer anderen Situation tragbar sein. Dennoch kann im Folgenden ein Rahmen gesteckt werden, der im Einzelfall von den Verantwortlichen jeweils adäquat umgesetzt werden muss.

Als Testkonstruktionsleitfaden kann der GER, wie oben gezeigt, nicht dienen, wenn er auch mit seinen drei Hauptverwendungsmöglichkeiten die groben Phasen der Testerstellung, Auswertung und Rückmeldung umreißt. In diesem Zusammenhang erweist sich das erwähnte Zusatzdokument, der *User's Guide for Examiners* (abgekürzt mit UGE, vgl. Council of Europe 2002[2]), als wesentlich informativer. In ihm finden sich in übersichtlicher Form alle relevanten Aspekte die Testentwicklung betreffend, weshalb er bei der hier folgenden Beschreibung des Testentwicklungsprozesses mit einbezogen wird. Nachstehend wird deshalb einerseits der UGE vorgestellt,

[144] Vgl. beispielsweise Alderson et al. 1995, Bachmann et al. 1996, Lienert & Raatz 1994, oder den hier besprochenen UGE, in dem ein umfassender Überblick gegeben wird.

andererseits werden grundlegende Aspekte des Testentwicklungsprozesses erörtert, und zugleich wird – zunächst auf theoretischer Basis – die Bedeutsamkeit des UGE bei diesem Prozess beurteilt. Die praktische Verwendbarkeit von GER und UGE bei der Entwicklung, Auswertung und Rückmeldung eines konkreten Tests wird wiederum in Kapitel 4 dieser Arbeit beurteilt, am erwähnten Beispiel der Entwicklung des Moduls *Schreiben Englisch* im DESI-Projekt.

2.6.1 Grundlagen und Zielsetzung des UGE

Der UGE wendet sich an alle an der Testentwicklung Beteiligten, insbesondere an GER-Benutzer: “This guide is designed to help anyone involved with the preparation of language tests, and particularly those wishing to make use of the Council of Europe's ‘Common European Framework of reference for language learning and teaching’." (UGE: 1) Die Nutzer werden aufgefordert, die Aussagen des UGE jeweils auf die eigenen Kontexte hin anzuwenden. Der UGE will keine Produkte beschreiben, sondern Prinzipien des Testentwicklungsprozesses aufstellen, die zu den gewünschten Ergebnissen führen sollen. Der UGE will darüber hinaus zur Bildung eines gemeinsamen Sprachtestsystems in Europa beitragen, wobei lokale Traditionen durchaus fortgeführt werden sollen (UGE: 3). Den gemeinsamen Bezugspunkt dieses Testsystems sieht der UGE im GER: Daher will der UGE seinen Beitrag dazu leisten, Tests zu entwickeln, die im GER-System verankert sind und die den üblichen Standards der Testentwicklung genügen (ebd.). Allerdings muss schon hier eingewandt werden, dass diese Verankerung alleine mit den Instrumenten GER und UGE nicht möglich ist, wie in Kapitel 3.4 dieser Arbeit belegt wird. Wohl auch deshalb wurde das bereits erwähnte *Manual* entwickelt, welches in Kapitel 3.5 dieser Arbeit in seiner Bedeutung näher beleuchtet wird.

In der Einleitung des UGE wird deutlich gemacht, dass der GER die notwendigen theoretischen Überlegungen zum Testdesign und zur Testentwicklung anstelle und es im UGE um die praxisorientierte Seite der Testentwicklung auf Basis der Rahmenpunkte des GER gehe (vgl. UGE: 3f) – diese Aussage kann wiederum als Hinweis auf den GER als „Testinstrument“ gedeutet werden.

Der UGE baut also auf den Grundlagen des GER auf und knüpft an Vorarbeiten des Europarats an. Beispielsweise folgt der Europarat der oben beschriebenen Paradigmenentwicklung von Sprachmodellen des Strukturalismus hin zu Modellen der kommunikativen Kompetenz und der Betrachtung von Sprache im Gebrauch: Er betrachtete schon in früheren Veröffentlichungen[145], was man können muss, um in einem fremden Land „funktionstüchtig“ und unabhängig zu sein. Diesem funktionalen Ansatz sehen sich GER und UGE ebenfalls verpflichtet:

> The emphasis is firmly on language as a social instrument, or a way of enabling people to interact with one another. The starting point is the range of situations in which language learners commonly find themselves in a foreign country; the goal is to be able to use language to do whatever is necessary in order to act appropriately in those situations. (UGE: 2).

[145] Vgl. etwa van Ek 1975, van Ek 1980, van Ek & Trim 1990.

Dem modell-basierten Ansatz des Sprachtestens, der im GER gewählt wurde, ist auch der UGE verbunden. Im Gegensatz zu den Darstellungen im GER werden im UGE die dem GER zugrunde gelegten Modelle der kommunikativen Kompetenz mit Quellenangaben belegt – diese Art der Transparenz müsste auch im GER selbst erzielt werden. Im UGE wird die Entwicklung nachgezeichnet ausgehend von dem Modell, das Canale & Swain (1981) mit vier Komponenten – Grammatik, Soziolinguistik, Diskurs, Strategien – vorstellten, bis hin zu Bachmanns darauf aufbauendem Modell der kommunikativen Kompetenz von 1990 (resp. 1991[2]a), welche sich aus sprachlichen Kompetenzen einerseits und aus der Ausübung dieser Kompetenzen im angemessenen Sprachgebrauch andererseits zusammensetzt. Ein solcher modell-basierter Ansatz hat, wie schon erwähnt, den Vorteil, dass man im Testkonstrukt leichter und kohärenter definieren kann, welche Gebiete der Kompetenz der Test erfassen soll; diese Konstruktdefinition sieht nicht nur der UGE (vgl. ebd., insb. S. 2) als Basis der Inhaltsvalidität von zu entwickelnden Tests, als Basis für Testkonstrukteure und nicht zuletzt als Basis für aussagekräftige *samples* der kommunikativen Kompetenz.

Die Autoren des UGE anerkennen die Vielfalt des Testentwicklungsprozesses, stellen jedoch folgende Grundsatzfragen vor, die bei jeder Testentwicklung auf rationaler Basis beantwortet und aus Objektivitätsgründen transparent belegt werden müssen (UGE: 3):

- is the test a test of general proficiency or does it test mainly what has been learnt in a course?
- how much time is available for the test?
- what level of performance is expected?
- is the aim to spread and rank students?
- are the results to be used diagnostically?

2.6.2 Testentwicklungsprozess

Der UGE betrachtet, ebenso wie die Verfasserin dieser Arbeit, die Testentwicklung als zyklischen Prozess (UGE: 5): “It is important and useful to think of the process of test development as cyclical and iterative. This involves feeding back the knowledge and experience gained at different stages of the process into a continuous re-assessment of a given test and each administration of it.”

Das folgende Schaubild (UGE: 5) gibt einen ersten Überblick über den generellen Prozess der Testerstellung und die unterschiedlichen Phasen. Daran schließen sich allgemein gültige Überlegungen der Verfasserin bezüglich der Testerstellungsphasen an, ehe jeweils die Aussagen im UGE zu den verschiedenen Phasen beurteilt werden:

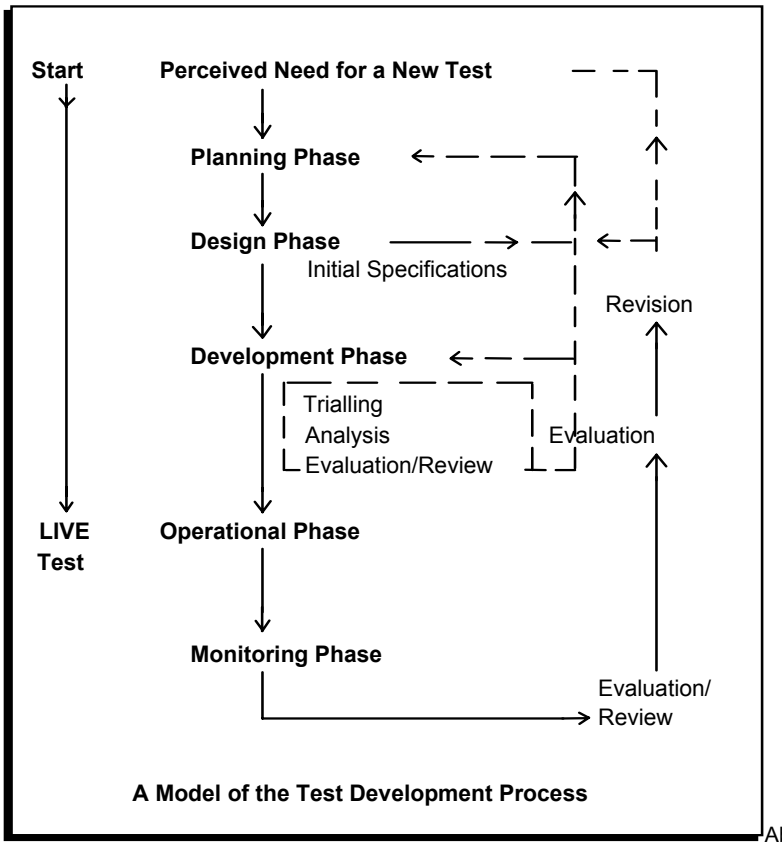


Abb. 7: Modell des Testentwicklungsprozesses

Das Bedürfnis nach einem Test entsteht, bevor man sich an die generelle Planung macht. Ausgehend von dieser wird ein Konzept entworfen, das dann in der Entwicklungsphase umgesetzt werden muss. Die so entwickelten Tests und ggf. Subtests werden prägetestet und analysiert, um ihre Qualität zu gewährleisten. Alle Ergebnisse und Erkenntnisse dieses Prozesses werden fortlaufend genutzt, um die Tests zu optimieren. Erst wenn dies erreicht ist und die Tests den Qualitätsstandards genügen, kommt es zum eigentlichen Testlauf. Die Ergebnisse werden analysiert und interpretiert. Der Umgang mit der dann anstehenden Rückmeldung kann in seiner Bedeutung gar nicht unterschätzt werden, denn daraus ziehen gerade große Leistungsstudien ihre Berechtigung. Nur wenn Testergebnisse den Institutionen, in denen sie gewonnen wurden, auch wieder zur Verfügung stehen, können sie einen sinnvollen Beitrag zur Qualitätsentwicklung der jeweiligen Institutionen leisten.

Bezogen auf die Planung kann ein Fragenkatalog nach Cohen (1994: 6) helfen, den Entscheidungsrahmen bei der Planung und Entwicklung von validen und reliablen Tests zu stecken:

> **WHAT** abilities – language, sociocultural and lingual, grammatical, strategic? Which skill area? For what purpose – administrative, instructional or research?
> **WHY** these abilities? What approaches and techniques with what rationale for using them?
> **HOW** to develop new instruments? How to validate them?
> **WHEN** and how often will assessment take place?
> **WHERE** – relating to the physical environment will assessment take place?
> **WHO** are the testees? personal characteristics, socio-cultural background, cognitive styles, test-wiseness?
> **THROUGH WHAT** processes are the answers derived at?
> **FOR WHOM** are the results intended? Which method for reporting the results and for giving feedback that benefits all concerned?

Zuallererst bietet es sich an, die Bedingungen zu betrachten, die in einer gegebenen Testsituation wirksam sind. Denn nur wenn die Variablen, die die Testsituation beeinflussen, bekannt sind, können sie auch kontrolliert werden. Fragen nach Auftraggeber, Finanzierung, Umfang und zeitlichem Rahmen werden wohl zuerst auftreten. Dieser Rahmen bestimmt die weitere Testentwicklung und den späteren Testverlauf. Die Zielgruppe, ihr Kontext und Hintergrund müssen bedacht werden, ebenso wie der Testzweck. Aus letzterem begründet bestimmt sich der jeweils adäquate Testtyp: Je nach Auslegung kann es sich etwa um *achievement tests* oder *proficiency tests* handeln, je nachdem, ob beispielsweise eine Beurteilung eher auf das Erreichen eines vorgegebenen Ziels ausgerichtet ist oder ob der generelle Sprachstand ermittelt werden soll. Die Auswahl der Leistungsdimensionen muss in Anlehnung an den Zweck und die Ziele der Leistungsbeurteilung erfolgen: Im schulischen Kontext bieten sich Curricula-, Lehrbuch- und Unterrichtsanalysen neben didaktischen Überlegungen an, um die relevantesten Bereiche abzudecken. Schon in der Planungsphase werden sich beschränkende Variablen wie beispielsweise die zur Verfügung stehende Zeit bemerkbar machen und die Auswahl der zu testenden Bereiche mit bestimmen, so dass man in der Praxis fortlaufend Kompromisse finden muss, um Tests in einem gegebenen Rahmen in angemessener Weise entwickeln zu können. Beispielsweise können nicht alle Schüler einer Jahrgangsstufe getestet werden, so dass repräsentative Stichproben von Probanden gezogen werden müssen, auch um den Zielen der Studie möglichst gerecht zu werden.

In der Phase des Testdesigns werden dann die geeigneten Testformate für die jeweils zu testenden Bereiche bestimmt: Welche Fertigkeit wird direkt, welche indirekt erfasst? Wo bieten sich integrative Formate an, die einen Blick auf die generelle Sprachfertigkeit ermöglichen? Wo bieten sich kommunikative Formate an, die die Performanz erfassen? Möchte man neben der Performanz auch Wissenssysteme und deklaratives Wissen prüfen? Dazu müssen Testspezifikationen erstellt werden und die Tests in ihren Formen, Inhalten, Dimensionen und schwierigkeitsbestimmenden Merkmalen genauestens charakterisiert werden als Basis der Operationalisierung, der Umsetzung dieser Spezifikationen in konkrete Testitems.

Diese und weitere Aspekte, die am Beginn der Testentwicklung stehen, finden sich im Detail im UGE in den Abschnitten 2.1 (Beschreibung der einzelnen Phasen) und 2.2 (Entwicklung von Testspezifikationen). Dort wird auch explizit auf die jeweiligen Verwendungsmöglichkeiten des GER verwiesen. Der UGE ist in seinem Herangehen so umfassend und transparent, wie es der GER zwar von sich postuliert, doch nicht erreicht. Nutzer werden hilfreiche und umfassende Anregungen aus den Katalogen und Übersichten ziehen können, die der UGE zur Verfügung stellt und die ihnen in der Praxis beim Testdesign gute Dienste leisten können.

In der sich der Designphase anschließenden Entwicklungsphase gilt es, die Spezifikationen, die in der Anfangsphase erstellt wurden, zu operationalisieren, sie also in Testitems umzusetzen. Diese *items* werden in informellen Prätests und/oder repräsentativeren Pilotierungen getestet und einer *Item*-Analyse mittels statistischer Analyseverfahren unterzogen, um sie auf Reliabilität, Validität und Machbarkeit hin zu untersuchen. Auch Fragen der statistischen wie linguistischen Dimensionalität des erfassten Konstrukts, der angemessenen Schwierigkeit und der Verteilung der *items* wie der Probanden müssen in dieser Phase beantwortet werden, immer unter Zuhilfenahme angemessener statistischer Verfahren. Es ist denkbar Fragebögen einzusetzen, um die Perspektiven der Probanden als auch der in der jeweiligen Institution Tätigen zu erfassen. Alle so gewonnenen Einsichten müssen fortlaufend in den Testentwicklungsprozess rücklaufen, um zu qualitativ hochwertigen Testitems, zu angemessenen Bewertungsschemata und damit zu einer aussagekräftigen Beurteilung zu kommen. In dieser Phase können noch wichtige Informationen einbezogen werden und Grundsätzliches kann noch geändert werden, ehe die Tests dann zum Einsatz kommen.

Die Entwicklungsphase wird im UGE ebenfalls ausführlichst beschrieben: In UGE-Abschnitt 2.3.1 werden umfassende Informationen gegeben zu den Stadien des eigentlichen Schreibens und Konstruierens von Testitems und zur Rekrutierung von Mitarbeitern und deren Trainingsbedarf, um die Testspezifikationen auch in valides Testmaterial umsetzen zu können; es werden ausführliche Ratschläge gegeben zur Textauswahl, zur Präsentation des Testmaterials und zur Erstellung von Testanweisungen. Illustriert werden diese Erläuterungen anhand eines konkreten Praxisbeispiels und einer Checkliste (vgl. UGE: 18ff), welche direkt in der Praxis benutzt werden kann. UGE-Abschnitt 2.3.2 beschäftigt sich mit konkreten Fragen der Materialauswahl und -Überarbeitung. Auch diese Ausführungen sind äußerst praxisnah und hilfreich.

Auf die Methoden des Prätestens und Pilotierens wird unter UGE-Abschnitt 2.4 im Detail eingegangen: Es werden Empfehlungen zur Stichprobengröße (UGE: 22) gegeben neben Ausführungen zum Vorgehen bei verschiedenen Testtypen und Bewertungsschemata. Ziel dieser Prätests ist das Sammeln von Daten und Informationen zu den Testitems selbst, zum Verhalten der Probanden und zu praktischen Aspekten der Administration. Auch auf statistische Analysen der so gewonnenen Informationen wird eingegangen: Im Anhang 1 (UGE: 39 mit 46) findet sich ein Beispiel einer solchen *Item*-Analyse mit genauen Erläuterungen der Fachtermini, relevanter Rechenformeln und eines Beispielausdrucks des Statistikprogramms MicroCAT, der in seiner Bedeutung genau erläutert wird.

In UGE-Abschnitt 2.5 *Test construction* (UGE: 23ff) wird auf konkrete Fragen bezüglich der einzuhaltenden Standards hinsichtlich der Schwierigkeiten, des Abdeckens der verschiedenen Bereiche und des Inhalts bei der Testkonstruktion eingegangen, ebenso wie auf weitere kontingente Merkmale, die es bei der Konstruktion zu beachten gilt. Ausdrücklich wird auf die

Notwendigkeit der Balance und Ausgeglichenheit aller Bereiche gerade in großen Untersuchungen hingewiesen.

Relativ umfangreich ist UGE-Abschnitt 2.6 *Issues in Item Writing* (UGE: 25ff), welcher sich ausführlich mit allen Fragen des Testitem-Schreibens – also dem Kern der Testkonstruktion – beschäftigt und *Guidelines* für folgende Aspekte gibt (ebd.: 25):

- task design;
- text selection (authenticity, difficulty, etc.);
- choice of item types;
- rubrics;
- keys, mark schemes and rating scales.

Die Erläuterungen zu diesen Kernpunkten sind so umfangreich und die Prinzipien des Vorgehens werden so übersichtlich dargestellt, dass die geneigten Leserinnen und Leser auf den UGE verwiesen werden dürfen, denn diesen Ausführungen ist nichts mehr hinzuzufügen. Auch an dieser Stelle verweist der UGE auf relevante Aussagen im GER und wie sie beim *Item*-Schreiben genutzt werden können. Zudem werden sie immer noch durch Checklisten ergänzt, die in der Praxis wertvolle Dienste leisten können. Es werden viele Problembereiche angesprochen, offene von geschlossenen Testaufgaben unterschieden und in ihren unterschiedlichen Konstruktionsanforderungen charakterisiert und es wird auf die verschiedensten Bewertungs- und Kodierschemata eingegangen. Da diese Konstruktionsprinzipien in Kapitel 4 dieser Arbeit anhand eines Praxisbeispiels erläutert und illustriert werden, kann an dieser Stelle auf eine Wiedergabe des UGE-Inhalts verzichtet werden.

2.6.3 Testevaluation

Im dritten und letzten UGE-Abschnitt (UGE: 36ff) geht es um die Evaluation der entwickelten Tests selbst. Diese Qualitätskontrolle findet während des gesamten Entwicklungs- und Pilotierungsprozesses statt. Damit soll fortlaufend überprüfen werden, ob sich Testzweck, Effekte und Konsequenzen im intendierten Bereich bewegen oder ob es zu unerwünschten Nebenwirkungen oder Störeffekten kommt. Folgende Maßnahmen werden vorgeschlagen (UGE: 36):

- validate the test;
- evaluate the impact of the test;
- provide relevant information to test users;
- ensure that a high quality of service is maintained.

Im UGE (ebd.: 36) wird die nicht zu unterschätzende Feststellung getroffen, dass Tests im Allgemeinen Auswirkungen auf den Bildungsprozess sowie auf die Gesellschaft haben, und dass es prinzipiell erstrebenswert ist, Tests so zu planen und zu konstruieren, dass sie einen positiven Effekt nicht nur auf Gesellschaft und Bildungsprozess generell, sondern auch direkt auf die Beteiligten im Lehr- und Lernprozess haben – ein Anerkennen der Notwendigkeit systemischer

Validität von Sprachtests. Es werden folgende Maßnahmen für die Testkonstruktion vorgeschlagen, um einen möglichst positiven *Washback*-Effekt zu erzielen (UGE: 36):

- the identification of suitable experts within any given field to work on all aspects of test development;
- the training and employment of suitable experts to act as question/item writers in test production;
- the training and employment of suitable experts to act as examiners.

Daneben wird vorgeschlagen folgende Fragen zu bedenken, um die Auswirkungen des Tests auf den Bildungsprozess kontrollieren zu können (UGE: 36f):

- who is taking the test (i.e. profile of the candidates);
- who is using the test results and for what purpose;
- who is teaching towards the test and under what circumstances;
- what kinds of courses and materials are being designed and used to prepare candidates;
- what effect the test has on public perceptions generally (e.g. regarding educational standards generally);
- how the test is viewed by those directly involved in educational processes (e.g. by students, test-takers, teachers, parents, etc.);
- how the test is viewed by members of society outside education (e.g. politicians, businessmen, etc.).

Es schließt sich eine Literaturliste mit Vorschlägen zur vertiefenden Lektüre an. Im Anhang findet sich neben dem oben erwähnten Anhang 1 ein Glossar als Anhang 2 (UGE: 48ff), in dem alle Fachtermini klar und verständlich erläutert werden.

Die umfangreichen Darstellungen, praxisorientierten Erläuterungen und informativen Anhänge machen den UGE zu einem umfassenden Instrument, das auch die nötige Transparenz besitzt, selbst „Neulingen“ einen relevanten und verständlichen Überblick über die wichtigsten Phasen und Aspekte der Testkonstruktion zu geben.

3 Der Skalenansatz bei der Beurteilung des Sprachvermögens

Skalen kann man als einen Versuch betrachten, eine lineare Ordnung in ein gegebenes System zu bringen und diese Ordnung gleichzeitig inhaltlich ansteigend zu beschreiben. Skalen stellen ein Modell eines bestimmten Ausschnittes der Realität dar; in diesem abgestuften Modell können Lernende, Aufgaben oder Tests eingeordnet werden. Während Skalen schon in den 50er Jahren in der Psychologie zum Einsatz kamen, hielten sie erst in den 70er Jahren Einzug in den Fremdsprachenunterricht.[146] Dort wurden sie vorwiegend bei der Beurteilung produktiver und integrativer *skills* genutzt. Alderson (1991b) gibt einen knappen Überblick über die Entwicklung einiger wichtiger Skalen, von der *Proficiency Scale* des *Foreign Service Institute* der Vereinigten Staaten über die IRT-Skalen des *InterAgency Round Table*, die ACTFL-*Guidelines* des *American Council for the Teaching of Foreign Languages* bis hin zu den *Australian Second Language Proficiency Ratings*, den IELTS-Skalen oder den Skalen der ALTE. Auch in großen Prüfungen, wie den *Cambridge ESOL-Tests,* werden Skalen eingesetzt, teils unter Anbindung an bereits existierende Skalensysteme.[147]

Je nach Beschreibungsgegenstand können Skalen vielfältige Funktionen in der Beurteilung erfüllen, wie im Folgenden gezeigt wird. Generell bieten Skalen einige Vorteile[148] gegenüber zählenden Verfahren: Sie treten im Allgemeinen positiv an zu bewertende Leistungen heran, statt traditionell Fehler und Defizite zu zählen, so dass von Skalen eine motivierende Wirkung ausgehen kann; sie beschreiben prototypisches Verhalten, das Lernenden bei der Selbsteinschätzung helfen kann; sie geben detaillierte Informationen statt reiner Punktwerte und können damit einen Beitrag zur Transparenz in einem gegebenen Bildungssystem leisten; sie können die Reliabilität einer Beurteilung durch die Vorgabe gemeinsamer Standards und Konzepte erhöhen; sie schaffen die Möglichkeit, verschiedene Prüfungen oder Bildungssysteme zu vergleichen mittels Niveaubeschreibungen, die gemeinsam interpretiert werden können. Skalen können zu einem Skalensystem zusammengestellt werden, wie etwa im GER geschehen: Sein Skalensystem hat Referenzcharakter und stellt einen Bezugsrahmen dar, in dem u. a. existierende oder neu zu konstruierende Skalen verortet werden können. In diesem Kapitel der vorliegenden Arbeit soll u. a. untersucht werden, ob und in wie weit der GER dem Anspruch eines solchen Referenzsystems gerecht werden kann.

Um die nötigen Grundlagen dazu zu schaffen, werden im ersten Teil dieses Kapitels zunächst Funktionen und Typen von Skalen mit ihren jeweiligen Beschreibungsgegenständen erörtert, ehe auf generelle Aspekte der Skalenentwicklung eingegangen wird, die dann am Beispiel von *rating scales* konkretisiert werden. Denn die Hauptaufgabe von Skalen im Fremdsprachenunterricht lag und liegt in der Bewertung komplexer Leistungen, für die rein quantitative

[146] Vgl. hierzu beispielsweise Upshur & Turner 1995 u. a.. Zu Ursprung und Entwicklung von Skalen vgl. auch North (2000: 13ff).
[147] Nähere Informationen zu diesen Skalen finden sich beispielsweise im Internet:
Zum *International English Language Testing System*: http://www.ielts.org, Zugriff am 28.03.2003.
Zu den *Cambridge Exams*: http://www.cambridgeesol.org/exam/5level.cfm, Zugriff am 28.03.2003.
[148] Vgl. hierzu beispielsweise Alderson 1991b, Brindley 1998, North & Schneider (1998: 219), u. a..

Zählverfahren unangemessen scheinen. Die in Kapitel 2.2.3 dieser Arbeit erwähnten *Counting*-Verfahren mögen bei rezeptiven Aufgaben sinnvoll sein, doch die Qualität einer Sprachproduktion lässt sich angemessener mittels *Judging*-Verfahren erfassen und wiedergeben.[149] Aufgrund der Bedeutsamkeit dieses Bewertungsverfahrens wendet sich Kapitel 3.3 den *Rating*-Verfahren als einer Verwendungsmöglichkeit von Skalen zu.

Aufbauend auf diesen Grundlagen wird im zweiten Teil dieses Kapitels der Status der GER-Skalen analysiert, um die Verwendbarkeit dieses Skalensystems beurteilen zu können. Im Anschluss wird das schon erwähnte *Manual* (vgl. Council of Europe 2003a) untersucht im Hinblick auf die dort vorgestellten Möglichkeiten der Testanbindung an das Referenzsystem des GER.

3.1 Funktionen – Beschreibungsgegenstand – Typen von Skalen

Grundsätzlich können Skalen nach ihrer Funktion typologisiert werden. Handelt es sich beispielsweise um eine Skala, die zur Beurteilung einer in einem Test elizitierten Performanz benutzt werden soll, wird sie *rating scale*[150] genannt; eine Skala, die ein Beurteilungsergebnis kommunizieren soll, wird entsprechend *reporting scale* genannt; eine Skala zur Konstruktion von Testaufgaben kann entsprechend *construction scale* benannt werden. Je nach Funktion einer Skala wird etwas anderes beschrieben, so dass man Skalen auch nach ihrem Beschreibungsgegenstand klassifizieren könnte. Dazu bieten sich Begriffe wie Kompetenzskala, Performanzskala oder *proficiency scale* an. Die Differenzierung von *proficiency scales* in *real-life approach* und in *interactive-ability approach* findet sich bei Bachmann (1991a: 325-330) und bezieht sich darauf, ob solch eine *proficiency scale* ausgelegt ist auf das, was ein Proband mit seiner Sprache im Leben anfangen kann oder ausgelegt ist auf das Können eines Probanden in einem bestimmten Test. Brindley (1998) unterscheidet die Typen „verhaltensbasierte Skalen“ und „konstruktbasierte Skalen“: Erstere Skalen beschreiben (sprachliches) Verhalten, der Beschreibung des letzteren Skalentypus hingegen liegen theoretische Modelle und Konstrukte zugrunde.

Eine weitere Einteilungsmöglichkeit findet sich in der Unterscheidung derer, für die die Skala bestimmt ist: Ist sie konstruiert für Lernende und Lehrende zur Rückmeldung; konstruiert für Beurteilende in einer Bewertungssituation; konstruiert für Testautoren zur Testerstellung? Da die Aspekte Funktion, Gegenstand und Verwendungsbereich so eng zusammenhängen, sollen sie hier gemeinsam betrachtet werden. Ähnlich der Kategorisierung sprachlicher Phänomene, wie sie in Kapitel 1.2.1 dieser Arbeit diskutiert wird, kann auch in diesem Bereich keine scharfe Trennung der identifizierbaren Typen erfolgen, sondern eher eine Einteilung in Prototypen, die aber gemeinsame Überschneidungs- und Berührungsbereiche besitzen. Die gängigste

[149] Zur Gegenüberstellung der Konzepte des *counting* und des *judging* darf auf Alderson (1991a) verwiesen werden, auf die Bemerkungen in Kapitel 2.2.3 dieser Arbeit und auf die folgenden Ausführungen unter Kapitel 3.3 *Rating-Verfahren*.
[150] Die Terminologie in diesem Bereich wird auf Englisch belassen, wann immer es keinen adäquaten Terminus im Deutschen gibt oder die Übersetzung missverständlich wäre.

Einteilung findet sich bei Alderson (1991b), der drei Typen unterscheidet (benutzerorientierte, beurteilungsorientierte und aufgabenorientierte Skalen), zu denen nach Pollitt & Murray (1996) noch ein vierter Typus tritt, die diagnoseorientierten Skalen.[151] Im Folgenden werden diese vier Typen beschrieben.

3.1.1 Benutzerorientierte Skalen

Dieser Typus dient der Berichtsfunktion und beschreibt (typisches oder wahrscheinliches) Verhalten oder (typische oder wahrscheinliche) Wissensbestände, meist in Form von positiven Kann-Aussagen. Die Beschreibung dessen, was auf einem bestimmten Niveau gekonnt wird, ist meist einfach gehalten und eher holistisch ausgerichtet, doch findet man auch benutzerorientierte Skalen, die auf die Beherrschungsgrade in den einzelnen Teilfertigkeiten hin ausgerichtet sind.[152] Ob eine benutzerorientierte Skala eher Kompetenzen im Sinne des unter Kapitel 1.2.3 erläuterten Konzepts der kommunikativen Kompetenz oder den Grad der Anwendbarkeit und Beherrschung des Wissenssystems, die *proficiency*, beschreibt, kommt auf die jeweilige Ausrichtung der Skala an: Wenn über die Herausbildung eines Wissenssystems berichtet werden soll, ist eine auf Kompetenzen ausgerichtete Skala angemessen. Wenn jedoch das Augenmerk eher auf die Anwendbarkeit des zugrunde liegenden Wissens im realen Leben gerichtet ist, ist eine Skala des erwähnten *real-life approach* angemessener. Wenn dagegen über Lernfortschritte oder Erwerbshierarchien berichtet werden soll, müssten solche Skalen typische Erwerbssequenzen beschreiben.[153]

Benutzerorientierte Skalen können bei der allgemeinen Feststellung des generellen Leistungsstands helfen, sei es nun im Rahmen von Selbst- oder Fremdbeurteilung. Sie lassen Rückschlüsse auf den allgemeinen Leistungsstand zu, sind jedoch nicht auf die Bewertung einer einzelnen Performanz ausgerichtet – sie beschreiben nicht eine erwartete oder tatsächlich beobachtete Performanz in einem bestimmten Kontext, sondern geben in generalisierter Form Beschreibungen des Könnens wieder. Diese Könnens-Beschreibungen sind in der Regel auf das reale Leben bezogen und nicht auf eine spezifische Beurteilungssituation. Allerdings sind auch benutzerorientierte Skalen denkbar, die einen Bezug zu konkreten Beurteilungssituationen aufweisen: Wenn über Testergebnisse in möglichst generalisierter Form berichtet werden soll, so wird sich die Skala auf ausgewählte Testaspekte, wie etwa Testanforderungen, und auf konkrete Beschreibungen der Testleistungen (etwa der Performanzen in direkten Tests) als Basis

[151] Vgl. auch North (2000: 17ff), der Aldersons Einteilung erweitert und als Basis der GER-Skalen nutzt.
[152] Teils findet sich in der Literatur (vgl. z. B. GER: 48 oder North 2000: 19) die Aussage, dass benutzerorientierte Skalen beschreiben, WAS ein Lernender schon kann, und nicht, WIE GUT er es schon könne. Doch die Frage, wie gut man etwas schon kann, spielt bei benutzerorientierten Skalen ebenfalls - etwa bei der Selbstbeurteilung - mit herein, so dass ich persönlich dieser Aussage nicht folge.
[153] Wie schon erwähnt, gibt es jedoch noch keine Theorie aus der Didaktik oder Spracherwerbsforschung, die Erwerbssequenzen schlüssig in lineare Abfolge bringen könnte, so dass eine solche Skala noch nicht entwickelt wurde. Aufgrund der in Kapitel 1.3.1 geschilderten Besonderheiten des Spracherwerbs und des Sprachenlernens bleibt es fraglich, ob solch eine Skala überhaupt konstruiert werden kann. Hingegen gibt es Theorien und Modelle, die typische Erwerbsfolgen für einzelne Teilfertigkeiten beschreiben, wie etwa das Modell zur Entwicklung der Schreibfertigkeit von Bereiter (1980), auf das in Kapitel 4 dieser Arbeit näher eingegangen wird.

beziehen, die dann in solch einer *reporting scale* auf Kompetenzen oder auf den Grad der Sprachbeherrschung hin generalisiert werden müssen.

Als Benutzer werden in der Regel die Lernenden genannt, doch können m. E. auch Lehrende Benutzer solcher Skalen sein und sich dort wertvolle Informationen über den Stand ihrer Lerner einholen.

3.1.2 Beurteilungsorientierte Skalen

Wie der Name schon sagt wird dieser Skalentypus bei der Beurteilung eingesetzt, in der Funktion, den Bewertungsprozess zu lenken und zu erleichtern. Solche *rating scales* beschreiben entweder konkrete Merkmale der zu beobachtenden oder erwarteten Performanz auf einen bestimmten Stimulus hin, sind also dem o. g. *interactive-ability approach* zuzuordnen; oder sie beschreiben theoretische Modelle und Konstrukte, die mit dieser Aufgabe erfasst werden sollen, sind also nicht auf das reale Leben hin ausgerichtet, sondern sie beschreiben abgestuft bestimmte Merkmale eines theoretischen Modells, das dem Test zugrunde liegt.[154] Sie können holistisch ausgerichtet sein oder detailliert und analytisch verschiedene Beurteilungsaspekte beschreiben. Dabei sollten sie konkrete Merkmale der zu bewertenden Leistung oder des Testkonstrukts auf ein bestimmtes Niveau einordnen, um bei der Einstufung der jeweiligen Performanz helfen zu können. Denkbar sind aber auch Skalen, die ein Lernziel abgestuft beschreiben, dessen Erfüllung in einem konkreten Test überprüft werden soll.

Beurteilungsorientierte Skalen dienen natürlich der Beurteilung; sie können beispielsweise ausgelegt sein auf die Bewertung eines spezifischen Tests oder aber auf die Bewertung eines bestimmten Kriteriums über verschiedene Tests hinweg. Sie bilden gemeinsame Standards ab, die einem Test zugrunde liegen, und dienen somit auch der Validität und Reliabilität der Bewertung. Sie bilden darüber hinaus eine der Grundlagen für *Rater*-Schulungen, die der Nutzung dieser Skalen vorangehen müssen. In solchen Schulungen wird ein gemeinsames Verständnis der Niveaubeschreibungen erarbeitet, so dass die Niveaus in der Bewertung selbst möglichst vergleichbar interpretiert werden.[155]

Beurteilungsorientierte Skalen richten sich an Beurteiler, seien es nun Lehrende im Fremdsprachenunterricht, Lernende bei der Selbstbewertung einer spezifischen Leistung oder externe Beurteiler in einer Testsituation.

---

[154] Vgl. hierzu Brindley 1998. Für eine genauere Analyse dessen, was eine *rating scale* im Idealfall beschreiben sollte, vgl. unten Kapitel 3.3 *Rating*-Verfahren.
Auch bei diesem Skalentyp ist die Frage berechtigt, ob die Basis der Beschreibung in Modellen des Spracherwerbs, des Sprachgebrauchs oder der Sprachkompetenz liegt. Denn grundlegend muss bei der Konstruktion von *rating scales* offen gelegt werden, ob sie beschreiben, was jemand auf einem bestimmten Niveau können SOLLTE, oder ob sie beschreiben, was jemand tatsächlich TUT

[155] Den *rating scales* wird aufgrund ihrer Bedeutsamkeit (nicht nur) in dieser Arbeit ein eigenes Kapitel 3.3 *Rating*-Verfahren gewidmet. Die Grundlagen für eine *Rater*-Schulung werden in Kapitel 3.3.3 dieser Arbeit diskutiert.

3.1.3 Aufgabenorientierte Skalen

Diese Skalen dienen der Testkonstruktion und sollen bei der Testerstellung helfen, zu validen Instrumenten zu kommen. Sie spezifizieren in der Regel die Aufgaben, Inhalte, Kontexte, Aktivitäten, Schwierigkeiten oder Texte, die der Testkonstruktion zugrunde liegen. Sie beschreiben in der Regel, WAS Lernende in Bezug auf bestimmte Aufgabenstellungen können (sollten) und nicht, WIE GUT sie es können, denn letzteres kann erst bei der Bewertung festgestellt werden.

Aufgabenorientierte Skalen sollen helfen, Testitems entsprechend bestimmter Vorgaben oder Erwartungen zu konstruieren und sie in einem größeren Beurteilungsrahmen vergleichbat zu machen. Solche Skalen können auch helfen, Testitems inhaltlich zu spezifizieren als Basis von beispielsweise Validitätsprüfungen. Sie schaffen Transparenz von der Testerstellung bis hin zur Bewertung und Rückmeldung, wenn ihre Inhalte sich in den Bewertungs- und Rückmeldeskalen widerspiegeln.[156]

Dieser Typus Skalen wendet sich an Testkonstrukteure, seien es nun die Lehrenden im Bildungssystem oder externe Testkonstrukteure.

3.1.4 Diagnoseorientierte Skalen

Dieser vierte Funktionstypus umfasst Skalen, die Lehrenden wie Lernenden diagnostische Informationen zu Lernstand, Lernzielen, Schwächen und Stärken geben sollen. Meist sind sie analytisch ausgerichtet, um beispielsweise Profile oder Lernerfolge im Detail kommunizieren zu können. Kennzeichnend für diesen Typ sind sehr detaillierte Beschreibungen, die sich je nach Diagnosezweck und Ausrichtung auf verschiedene Beschreibungsgegenstände beziehen. Bei diesen Skalen geht es vorrangig darum, WIE GUT etwas gekonnt wird, doch auch WAS gekonnt wird, darf nicht vernachlässigt werden.

Diagnoseorientierte Skalen hängen sehr eng mit den drei zuvor erwähnten Typen zusammen, weshalb im Folgenden die Zusammenhänge zwischen den verschiedenen Skalentypen beleuchtet werden sollen.

3.1.5 Zusammenhänge zwischen den Funktionen, Gegenständen und Typen

Wenn man aufgabenorientierte Skalen als Ausgangspunkt in einem gegebenen Testkonstruktionsprozess betrachtet, so liefern diese die Spezifikationen für den zu konstruierenden Test. Die Beurteilungsskalen sind insofern mit den aufgabenorientierten Skalen verwandt, als dass sie deren Spezifikationen – etwa bezüglich der *tasks* oder Testitems – umsetzen in Bewertungskriterien der in diesem Test zu erwartenden Performanz. Die erstgenannten benutzerorientierten

[156] Näheres zum Zusammenhang der verschiedenen Skalentypen unter Kapitel 3.1.5 dieser Arbeit.

Skalen hängen insofern mit diesen beiden Typen zusammen, insoweit sie Testergebnisse auf die *proficiency* oder Kompetenzen hin generalisieren und dem Berichten dieser Testergebnisse dienen. Die letztgenannten diagnoseorientierten Skalen können grundsätzlich von den vorgenannten Skalen abgeleitet werden: Beispielsweise können aufgabenorientierte Skalen diagnostische Hinweise darauf geben, welche Schwierigkeiten in welcher Jahrgangsstufe gemeistert werden können; Bewertungsskalen können, wenn sie analytisch und detailliert sind, diagnostische Profile liefern; benutzerorientierte Skalen geben den Nutzern die gewünschte diagnostische Information, wenn sie nicht zu global gehalten sind, sondern beispielsweise analytische Profile kommunizieren oder auf das Erreichen eines bestimmten Lernziels ausgelegt sind.

Die verschiedenen Skalentypen lassen sich bei angemessenen Adaptionsmethoden ineinander umformulieren, um etwa den Prozess der Testerstellung von der Konstruktionsphase bis zur Rückmeldungsphase transparent und kohärent zu dokumentieren. Beispielsweise lässt sich, wie Alderson (1991b: 80f) beschreibt, eine *rating scale*, die konkrete und detaillierte Merkmale zur Bewertung aufweist und deshalb auf eine bestimmte Aufgabe ausgerichtet ist, in eine *reporting scale* umformulieren. Dazu muss erstgenannte Skala in dem Umfang, in dem es die Aufgabe und der Test zulassen, verallgemeinert werden. Die konkreten *task*-bezogenen, detaillierten Beschreibungen müssen in auf die Benutzer hin abgestimmte generalisierte Beschreibungen überführt werden, die im Idealfall noch einen gewissen Aufgabenbezug haben sollten. Doch ist dabei Vorsicht geboten, denn Skalen dürfen nicht zu anderen als den bei ihrer Konstruktion intendierten Funktionen eingesetzt werden; Näheres dazu wird unter Kapitel 3.2.4 *Validitätsaspekte* erläutert.

Auch die Nutzer der verschiedenen Skalentypen stehen in gewissen Verbindungen, von denen hier nur einige exemplarisch genannt seien: Oft sind Testkonstrukteure und Bewerter ein und dieselbe Person; falls nicht, müssen Konstrukteur und Bewerter gemeinsame Erwartungen an den Test und die zu elizitierende Leistung haben – dazu müssen die Testersteller klare Vorgaben an die Bewerter geben und klare Aufgabenstellungen an die Probanden. Probanden müssen ebenfalls um die Testerwartungen wissen, um die gewünschte Performanz zu zeigen. Testkonstrukteure müssen auch um den Hintergrund der Probanden wissen, um valide Tests konstruieren zu können. Bewerter müssen ebenfalls um den Horizont der Probanden wissen, um zu angemessenen und validen Bewertungen zu kommen.[157] Skalen sind dann sinnvoll, wenn sie dem Kommunikationsdreieck *Exam Setter – Rater – Testee* gerecht werden. Folgendes Schaubild zeigt die Charakteristika dieses Dreiecks nach Cohen (1994: 307f):

[157] Vgl. hierzu beispielsweise Cohen (1994: 307f) oder Pollitt (1991b: 87f).

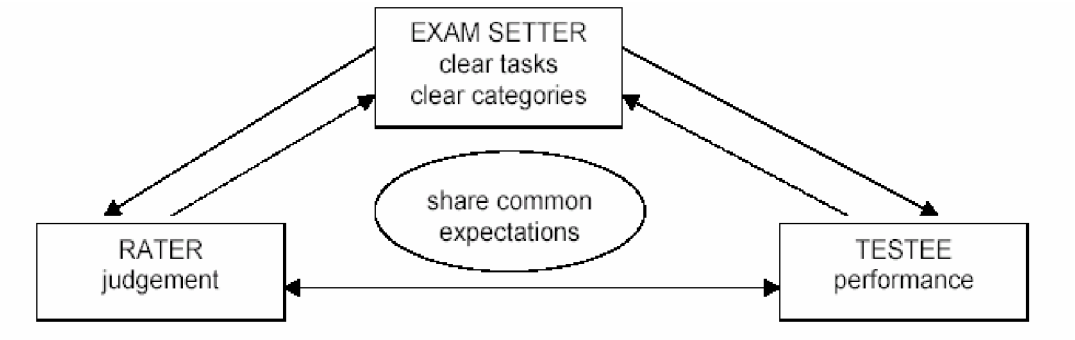


Abb. 8 Kommunikationsdreieck

3.2 Konstruktion von Skalen

Bei der Skalenkonstruktion kommt dem Offenlegen der Basis und der Entstehung der Skalen gewichtige Bedeutung zu: Welche Bereiche muss die Skala beschreiben, um ihrer intendierten Funktion auch gerecht zu werden? Wie kommt man zu den horizontalen Teildimensionen, die mit je einer Skala beschrieben werden sollen? Wie viele solcher Teilbereiche sollen in welchen Kategorien angesetzt werden? Welche Modelle, seien es nun Modelle der kommunikativen Kompetenz, des Sprachgebrauchs oder der Sprachentwicklung, liegen diesen Kategorien zugrunde? Wie kommt man dann zu den vertikalen Abstufungen? Wie viele davon sind praktikabel? Auf welche Modelle und Theorien, seien es Theorien des Spracherwerbs, der Fertigkeitsentwicklung oder psychometrische Messmodelle, bezieht man sich dabei? Wie können Gegenstandsbereiche und Abstufungen validiert werden und in welchen Kontexten können die so entwickelten Skalen valide genutzt werden?

Diese Entscheidungen können nur konkret auf die jeweilige Situation bezogen getroffen werden, so dass im Folgenden allgemein gültige Aussagen[158] angesprochen werden, ehe in Kapitel 4 solch ein konkreter Fall geschildert wird.

3.2.1 Dimensionen

Bei der Skalenkonstruktion wird zunächst der zu beschreibende Gegenstandsbereich festgelegt, der zugleich den Einsatzbereich der zu konstruierenden Skala mitbestimmt, denn wie oben erläutert hängen Gegenstand, Verwendungszweck und Benutzerkreis eng zusammen. Die Basis der horizontalen Einteilung von Skalen in einzelne Teilbereiche muss in Theorien und Modellen gefunden werden, die die Grundlage für eine valide Beschreibung darstellen.[159] Denn nicht jeder Gegenstandsbereich kann durch eine einzige Skala und damit eine Beschreibungskategorie

[158] Die Aussagen des Kapitels 3.2 basieren auf den Ausführungen in Alderson 1991a+b, Brindley 1998, Lumley 2002, North 2000 und North & Schneider 1998.
[159] Vgl. beispielsweise North (2000: 29).

dargestellt werden. Je nach Zweck der Skala muss man unterschiedlich detaillierte Kategorien ansetzen. Selbst wenn man, an einem Ende des möglichen Spektrums, den Sprachstand global durch eine Skala beschreiben will, ist es fraglich, ob dies für alle Teilaspekte der Sprachbeherrschung in einer einzigen Kategorie erreicht werden kann, ohne dass die Deskriptoren sehr vage gehalten werden und dadurch wenig Information transportieren. Am anderen Ende des denkbaren Spektrums stehen detailliert aufgeschlüsselte Teildimensionen, die durch je eine (Sub-) Skala repräsentiert sind. Nun gibt es keine Ideallösung, wie man zu angemessener Kategorisierung sprachlicher Teilbereiche oder kommunikativer Fertigkeiten kommt, noch wie viele Kriterien in wie vielen Kategorien repräsentiert sein sollen, noch wie die Kategorien klar voneinander abgegrenzt werden können. Deshalb muss die horizontale Einteilung von Skalen in Einzeldimensionen immer begründet werden im jeweiligen Kontext, für den die Skala konstruiert wird, und verankert werden in Theorien und Modellen, die für diesen Kontext zutreffen.

Will man beispielsweise eine beurteilungsorientierte Skala entwickeln, so sollten die Bewertungskriterien einerseits in Modellen der kommunikativen Kompetenz und des Sprachgebrauchs verankert werden; andererseits ist es in diesem Bereich ratsam, relevante Merkmale zusätzlich aus empirischen Beschreibungen der zu beurteilenden Leistungen abzuleiten. Aufgabenorientierte Skalen hingegen müssen Charakteristika, Anforderungen und Schwierigkeiten der Aufgaben möglichst kohärent beschreiben.[160] Skalen wiederum, die den Grad der Sprachbeherrschung in Bezug auf bestimmte Aufgaben, die *proficiency* im *interactive-ability approach*, beschreiben sollen, müssen mangels einer empirisch validierten Theorie der *proficiency*[161] alle Aspekte der *proficiency* unter Zuhilfenahme des gesunden Menschenverstands und pragmatisch gehaltener Beschreibungen beachten, so zum Beispiel den Zusammenhang zwischen den Aufgabenanforderungen und der Problemlösefähigkeit und Sprachverarbeitungskapazität der Probanden, denn "(…) proficiency is a function of the processing skills required by a task" (Brindley 1998: 125).

Generell weisen Skalen umso detaillierte Kategorien auf, je mehr spezifische Informationen sie tragen. Beurteilungsorientierte Skalen etwa müssen mehr Details aufweisen, die eine Beurteilung erst ermöglichen, als solche Skalen, die die Ergebnisse dieser Beurteilung in allgemeiner Form mitteilen sollen. Bei letzteren könnten – je nach Kontext – beispielsweise mehrere detaillierte Beurteilungsaspekte in einer Kategorie zusammengefasst werden, um lediglich das generalisierte Ergebnis zu transportieren.[162] Wenn jedoch nicht global über den Sprachstand informiert werden soll, sondern auch in der Beurteilungsrückmeldung Profile kommuniziert werden sollen, so müssen auch die Rückmeldungsskalen über detaillierte Kategorien verfügen.

Welche Kriterien letztlich als relevant angesetzt und zu welchen Beschreibungskategorien zusammengefasst werden, muss im jeweiligen Kontext begründet und entschieden werden.

[160] Für solch einen Rahmen vgl. beispielsweise Bachmann (1991a, 70ff und 116ff, insbesondere Abb. 5.1: 119).
[161] Vgl. North (2000: 32).
[162] Vgl. etwa Alderson (1991b: 80f).

Wichtig dabei ist, dass die Anzahl der Kategorien beziehungsweise der (Sub-)Skalen, die diese Kategorien repräsentieren, handhabbar sein muss. Man denke etwa an eine Beurteilung der Sprechfertigkeit – dabei sind nur wenige Kategorien bewertbar, wenn die Sprechfertigkeit in einem direkten Interview beurteilt werden soll. Hingegen können mehr Kategorien bewertet werden, wenn es sich um die Beurteilung beispielsweise der Schreibfertigkeit handelt, da hierbei das Performanzbeispiel von „dauerhafter Natur" ist.

Bei dieser Kategorisierung gelten die in Kapitel 1.2.1 dieser Arbeit getroffenen Feststellungen bezüglich des Prototypenmodells, das Kategorien kennzeichnet durch prototypische Merkmale und fließende Übergänge, und das die Zugehörigkeit zu einer gegebenen Kategorie über das (Nicht-)Vorhandensein von prototypischen Merkmalen beschreibt. Um gegebene Kategorien respektive Subskalen voneinander abgrenzen zu können, sollte der Gegenstandsbereich der Kategorien respektive der sie repräsentierenden Skalen in seinen Charakteristika definiert werden. Letztlich ist die Dimensionsbestimmung eine Ermessensentscheidung, die über die oben angesprochenen Wege der Theoriebasierung, der Entwicklungsdokumentation und der Offenlegung von oft pragmatischen Entscheidungsgründen transparent gehalten werden muss.

3.2.2 Abstufungen

Sind die Dimensionen und damit die Beschreibungskategorien gefunden und ist entschieden, welche Kriterien und Aspekte in welchen Kategorien zusammen gefasst werden, so müssen Abstufungen geschaffen werden, die als Deskriptoren versprachlicht eine valide hierarchische Ordnung des Gegenstandsbereichs der jeweiligen Skala beschreiben. Solche Abstufungen müssen einerseits, wo möglich, aus Theorien und Modellen abgeleitet werden, die sich mit solchen hierarchischen Ordnungen beschäftigen. Man denke beispielsweise an Modelle der Entwicklung einzelner Teilfertigkeiten. Andererseits kommt psychometrischen Messmodellen bei der Skalierung der Deskriptoren eine große Bedeutung zu, die hier jedoch nur erwähnt, nicht jedoch diskutiert werden können, da dies nicht in den Bereich der vorliegenden Arbeit fällt.

Grundsätzlich ist es eine arbiträre Entscheidung, in wie viele Niveaus eine Skala eingeteilt wird, doch wird diese Entscheidung beeinflusst durch pragmatische und empirische Gesichtspunkte: Handelt es sich beispielsweise um eine *reporting scale* um Beurteilungsergebnisse zu kommunizieren, so sollte die Bandbreite der Leistungen mit in Betracht gezogen werden – streuen die Leistungen nicht stark, so genügen weniger Abstufungen; bei breiter streuenden Leistungen können auch entsprechend mehr Niveaus beschreiben werden. Bei einer *rating scale* wird die Handhabbarkeit über die Anzahl der Niveaus entscheiden: allgemein gelten fünf bis sieben Niveaus noch als benutzbar.[163] Die Anzahl der Niveaus hängt also davon ab, zu welchen Zwecken die Beschreibung genutzt werden soll und wie stark der Beschreibungsgegenstand

[163] Vgl. beispielsweise Lehmann 1990.

vertikal differenziert werden kann. Man kann sich dieser Frage auch empirisch nähern und aufgrund statistischer Reliabilitäts- und Separabilitätsanalysen entscheiden, wie viele Abstufungen eine gegebene Skala zulässt.[164] Nach Lehmann (1990) erhöht sich die Reliabilität, je größer der Wertebereich einer Skala ist; es haben sich zwischen 4 und 6 Stufen bewährt.[165]

Bei North & Schneider (1998: 221ff) und im GER (2001: 202ff) finden sich Übersichten über Methoden der vertikalen Skalenentwicklung. Ausgangspunkt können entweder empirische Analysen etwa auf der Basis von Performanzbeispielen sein oder aber existierende Deskriptoren, die als Grundlage der Konstruktion dienen. Es sind dabei drei Ansätze auszumachen: intuitive, qualitative und quantitative Methoden. Erstere beziehen sich auf Experteneinschätzungen, die entweder auf Basis von Bedarfsanalysen oder aufgrund empirischer Analysen etwa von Performanzbeispielen oder Aufgaben abgegeben werden, sei es in Form eines ersten Skalenentwurfs durch einen Experten oder in Form von Diskussionen bereits existierender Skalen in Expertengruppen, die zu dem nötigen Konsens hinsichtlich der Abstufungen führen sollen.

Qualitative Ansätze beschäftigen sich mit der Analyse der Merkmale – seien es nun Schlüsselkonzepte oder primäre Eigenschaften – des Beschreibungsgegenstands, also etwa der Leistungen, Kompetenzen, Aufgaben oder Lernziele, je nach Zweck und Gegenstand der Skala. Die Analyse soll helfen, die für die jeweils angesetzten Kategorien relevanten Merkmale in ihren Abstufungen zu identifizieren und charakteristische Merkmale bestimmten Niveaus zuzuweisen. Beispielsweise fällt hierunter die Methode, eine existierende Skala in ihre Deskriptoren zu zerlegen und diese von einer Gruppe von Experten in eine hierarchische Ordnung bringen zu lassen. Eine weitere qualitative Methode besteht darin, Stichproben von Leistungen zu vergleichen und/oder zu analysieren, um charakteristische Merkmale auf bestimmten Niveaus oder in bestimmten Ausprägungen zu identifizieren. Letztere Methode bietet sich gerade bei beurteilungsorientierten Skalen an, da diese einen direkten Bezug zur zu bewertenden Performanz oder Leistung haben sollten; auf Details in diesem Zusammenhang wird unter Kapitel 3.3 *Rating-Verfahren* eingegangen.

Bei den qualitativen Ansätzen muss die grundlegende Entscheidung getroffen werden, ob die Grenzen zwischen zwei Niveaus identifiziert und beschrieben werden sollen oder ob es darum geht, das Prototypische eines Niveaus zu beschreiben. In Kapitel 3.3.1.2 *Die Rolle der Deskriptoren* wird u. a. auf eine binäre Skala nach Upshur & Turner (1995) eingegangen, um zu zeigen zu welchen Schwierigkeiten das Fokussieren auf Übergänge zwischen konstruierten und damit bis zu einem gewissen Maß auch arbiträren Niveaus führen kann. Denn wie in Kapitel 1.2.1 erläutert sind die Grenzen zwischen gegebenen oder konstruierten Kategorien meist fließend, so dass das Verhältnis von gegebenen Kategorien zueinander durch den Prototypenansatz angemessen beschrieben werden kann: Auch bei konstruierten Niveaus einer Skala

[164] Vgl. hierzu beispielsweise North (2000: 38f) oder North & Schneider (1998: 231); nach Pollitt (1991a) gibt es einen Zusammenhang zwischen der Reliabilität einer Datensammlung und der Anzahl von *Levels*, in die sie unterteilt werden kann.
[165] Studien hierzu finden sich beispielsweise in Coffman (1971).

handelt es sich um eine mehr oder minder arbiträre Einteilung einer linearen Skala in bestimmte Abschnitte, die durch fließende Übergänge und prototypische Mittelbereiche gekennzeichnet sind. Deshalb macht es Sinn, prototypische Merkmale zu beschreiben statt der Grenzen – die Beschreibung der Grenzen würde auch implizieren, dass ein Merkmal erst ab einer bestimmten „Stufe“ auftritt beziehungsweise dass das Auftreten eines isolierten Merkmals das Erreichen eines bestimmten Niveaus zwingend nahe legt, während man aus der Praxis weiß, dass es durchaus der Fall sein kann, dass nicht alle Merkmale eines Niveaus auch immer dort auftreten müssen. Vielmehr liegt es nahe, dass ein Niveau umso wahrscheinlicher erreicht ist, je mehr der dort beschriebenen prototypischen Merkmale vorhanden sind.

Sind die Merkmale einer Skala auf verschiedenen Niveaus intuitiv und/oder qualitativ identifiziert und beschrieben, so muss die so konstruierte Skala im Idealfall noch quantitativ skaliert werden. Hierbei finden psychometrische Messmodelle[166] Anwendung, wie beispielsweise die Rasch-Skalierung von Deskriptoreneinschätzungen oder die Skalierung von Testitems, die aufgrund einer intuitiv oder qualitativ entwickelten Skala konstruiert und administriert worden sind. Auch finden Diskriminanzanalysen in der Bewertung von Leistungen Anwendung. Sie helfen bei der Bestimmung von Merkmalen, die sich in der Bewertung als relevant erwiesen haben. Mithilfe von Faktorenanalysen lassen sich darüber hinaus Schlüsselmerkmale einer gegebenen Performanz identifizieren.

Die quantitativen Methoden helfen, die Subjektivität der intuitiven und qualitativen Methoden zu minimieren und sie tragen dazu bei, zu Deskriptoren zu kommen, die unabhängig von Testitems und Expertengruppen allgemein verständlich sind. Beispielsweise kann der Grad an Übereinstimmung des Verständnisses der Deskriptoren in verschiedenen Kontexten überprüft werden, indem die Deskriptoren von unterschiedlichen Gruppen eingestuft werden und diese Einstufungen wiederum mittels psychometrischer Verfahren analysiert werden. Auch können auf diese Weise missverständlich formulierte Deskriptoren identifiziert und überarbeitet werden, denn der Versprachlichung kommt eine nicht zu unterschätzende Bedeutung in der Skalenentwicklung zu.

3.2.3 Aspekte der Beschreibung[167]

Bei der Skalenkonstruktion sind Aspekte der Beschreibung von Bedeutung, Aspekte der Semantisierung der Deskriptoren also, die eine Skala bilden. Die Forderung nach Benutzerfreundlichkeit der Skalen nimmt dabei eine zentrale Rolle ein: Die Sprache der Deskriptoren muss auf den Verwendungszweck und die Zielgruppe ausgelegt sein, damit die Nutzer der Skala auch zu einem gemeinsamen Verständnis der Beschreibungskategorien und Abstufungen kommen können. Dazu müssen relevante Merkmale der in einer Skala zu beschreibenden Kategorie so

[166] Für einen knappen Überblick über Modelle aus der Messtheorie vgl. beispielsweise Hamp-Lyons (1996: 235) oder North & Schneider (1998: 221f, 225).
[167] Vgl. hierzu die Ausführungen in Alderson 1991b, Brindley 1998, Lumley 2002, North 2000, North & Schneider 1998.

präzise beschrieben werden, dass die Skalennutzer die Skaleninhalte auch mit dem Beschreibungsgegenstand in Zusammenhang bringen können. Dabei sollte die Sprache so gehalten werden, dass möglichst wenig Rekurs auf Spezialvokabular genommen werden muss; dieser Grundsatz ist jedoch abhängig vom Benutzerkreis nicht immer durchsetzbar. Eine *reporting scale* beispielsweise, die Schülerinnen, Schülern und Eltern Rückmeldung über einen Leistungstest geben soll, muss möglichst frei von Fachtermini gehalten sein, wohingegen eine *rating scale*, die von geschulten Experten benutzt werden soll, durchaus solche Fachtermini enthalten kann, deren Verständnis in der voranzustellenden Schulung gesichert wird. Auch eine aufgabenorientierte Skala für Testentwickler wird nicht ganz ohne Fachjargon auskommen – Voraussetzung ist jedoch, dass Fachjargon und Benutzerkreis aufeinander abgestimmt sind.

Ein weiterer Grundsatz der Beschreibung ist die Positivformulierung: In der Regel werden Aspekte in einer Skala positiv beschrieben, um nicht Defizite zu kommunizieren, sondern um Vorhandenes abgestuft zu beschreiben, seien es nun Anforderungen, Ziele, Wissensbestände oder das Können allgemein. Gerade in Bezug auf benutzerorientierte Skalen spielt dabei der Aspekt der Motivation eine Rolle. Traditionell hat man (nicht nur) im deutschen Bildungssystem eher Defizite und Fehler wahrgenommen als das, was schon vorhanden ist. Der Blick auf das Vorhandene jedoch zeigt Lernerfolge auf und motiviert dadurch zum Weiterlernen. Allerdings wird man sich nicht immer an diesen Grundsatz halten können, beispielsweise wenn sich die Performanz gerade auf unterem Niveau eher durch Defizite als durch schon Vorhandenes auszeichnet. Auf die Notwendigkeit der Verknüpfung von Positiv- und Negativansatz wird unter Kapitel 3.3 *Rating-Verfahren* näher eingegangen.

Bei der Versprachlichung spielt ein weiterer Aspekt mit herein: der der Kontextfreiheit respektive der Kontextualisierung von Skalen. Während eine *proficiency scale* eines unabhängigen Referenzsystems naturgemäß möglichst über alle denkbaren Kontexte hinweg einsetzbar sein sollte, kann man sich etwa bei *rating scales* eine solche Dekontextualisierung nicht denken, wenn die *rating scales* für die Bewertung spezifischer Aufgaben konstruiert wurden – dazu Näheres unter Kapitel 3.3 *Rating-Verfahren*. Der Grad der Kontextfreiheit hängt, wie andere Aspekte auch, vom Verwendungszweck und dem Einsatzbereich der betreffenden Skalen ab.

Die Niveaus einer Skala sollten nicht rein verbal abgestuft werden, da sonst die Bedeutung eines Niveaus nicht aus sich heraus verständlich wird. Denn eine Skala verliert an Handhabbarkeit, wenn man zum Verständnis eines Niveaus immer auch die benachbarten Niveaus heranziehen muss. Deshalb sollten relevante prototypische Merkmale eines Niveaus inhaltlich und qualitativ charakterisiert werden, um Beschreibungen der prototypischen Mitte der Niveaus zu erhalten. Basieren sollten diese abgestuften Beschreibungen im Idealfall auf Analysen relevanter Kontexte, Situationen und Aufgaben; auf Analysen relevanten Verhaltens, das unter Umständen empirisch beobachtet und beschreiben werden kann; sowie auf Analysen von Lernbedingungen und Curricula. Um jedoch Redundanzen und Tautologien über die Skalenniveaus hinweg zu vermeiden, die ihrerseits der Handhabbarkeit abträglich sind, sollten nicht alle

Merkmale auf allen Niveaus wiederholt werden – Merkmale sollten wie gesagt auf dem Niveau beschrieben werden, wo sie auch prototypisch beobachtet werden. Auf den Niveaus darunter treten sie in der Regel noch nicht auf, auf den höher gelegenen Niveaus werden sie als vorhanden betrachtet und müssen deshalb nicht immer wieder genannt werden: Obere Niveaus schließen in der Regel die unteren mit ein. Das mag auf den ersten Blick dem gerade genannten Grundsatz der Unabhängigkeit der einzelnen Niveaus widersprechen, doch nur unter Beachtung beider Grundsätze, der Unabhängigkeit der Niveaus und der Redundanzvermeidung, lassen sich Skaleninhalte und Abstufungen benutzerfreundlich beschreiben.

Generell werden an die Sprache einer Skala folgende Anforderungen gestellt: Die Formulierungen sollen in klarer, präziser und verständlicher Sprache verfasst werden, positiv formuliert und möglichst kurz gehalten sein. Dabei sollen die jeweiligen Bereiche konsistent und kohärent beschrieben werden, frei von Widersprüchen oder Ambiguitäten. Die Beschreibungskategorien und die Deskriptoren der verschiedenen Niveaus der Skala müssen den ihnen zugrunde gelegten Theorien gerecht werden und vom Zielpublikum gemeinsam interpretiert und verstanden werden können.

3.2.4 Validitätsaspekte[168]

An dieser Stelle muss vermerkt werden, dass es – anders als beispielsweise in den Naturwissenschaften – kein „hartes“; objektives Außenkriterium gibt, an dem Skalen validiert werden könnten. Denn das jeweilige Modell, nach welchem Sprache beschrieben wird, die Merkmale, welche als relevant für bestimmte Aspekte und Niveaus betrachtet werden, oder auch die Sprache der Deskriptoren – all diese Aspekte haben keinen Absolutheitsanspruch, sondern ihre Gültigkeit ist relativ: relativ in Bezug auf das jeweils gültige Paradigma, relativ bezüglich der jeweils vorherrschenden Vorstellung von dem, was beispielsweise Sprachvermögen ausmacht, relativ bezogen etwa auf den Sprachgebrauch der Skalenkonstrukteure.

Dennoch gibt es Möglichkeiten, valide Skalen zu entwickeln, die sich nicht nur an den Vorstellungen der an der Konstruktion Beteiligten messen lassen. Dazu sollten bei der Skalenkonstruktion die bisher angesprochenen Grundsätze beachtet werden. Eine Skala kann beispielsweise nur in dem Maße valide sein, in dem sie den Gegenstandsbereich, den sie abbilden soll, auch tatsächlich beschreibt. Eine Skala muss demnach hinsichtlich ihres Beschreibungsgegenstands, ihrer Einteilung in Kategorien und Abstufungen und hinsichtlich der verwendeten Sprache validiert werden. Hält man sich nicht an die o. g. Grundsätze oder setzt man Skalen zu anderen als den intendierten Funktionen ein, so kann es zu Konflikten kommen, die im Folgenden erläutert werden:

[168] Vgl. hierzu die Ausführungen in Alderson 1991b, Brindley 1998, Lumley 2002, North 2000, North & Schneider 1998, Pollitt & Murray 1996.

Das der Skala zugrunde gelegte Konstrukt muss in wissenschaftlichen Theorien verankert werden, ebenso wie die angesetzten Teildimensionen, in die der Skalenbereich gegebenenfalls zerlegt werden soll. Nur wenn sich theoretisch begründen lässt, warum welches Konstrukt durch welche Merkmale in der Skala repräsentiert ist, und in welchen Theorien oder Modellen die Aufteilung des Gegenstandsbereichs verortet ist, kann die Subjektivität solcher teils intuitiv, teils pragmatisch bedingten Entscheidungen limitiert werden. Wenn diese Verankerung nicht erfolgt, kann es u. U. schwer sein, die Skaleninhalte zu kommunizieren und zu einem gemeinsamen Verständnis unter den Benutzern einer Skala zu kommen. Wenn jedoch jeder Nutzer eine gegebene Skala anders interpretiert, so dürfte eine reliable Nutzung unmöglich sein und somit auch die Validität der betreffenden Skala nicht gegeben sein.

Eine weitere Ursache für Validitätsprobleme neben der gerade erwähnten Aufteilung eines Bereichs in mehrere Beschreibungskategorien kann im Zusammenfassen mehrerer Kriterien in einer Beschreibungskategorie liegen. Oft sind aus pragmatischen Gründen nicht so viele Kriterien in je einzelnen Skalen beschreibbar, wie es etwa ein komplexes Konstrukt verlangen würde. Gerade bei Bewertungsskalen können sich nach Lumley (2002) dadurch Probleme der inneren Gewichtung ergeben – etwa dass eines der Kriterien die anderen überwiegt und demnach die Kriterien in einer Kategorie nicht gleichwertig wahrgenommen werden können. Wie man bei *rating scales* diesen Problemen vorbeugen kann, wird in Kapitel 3.3 erläutert. Generell muss auch hier die Entscheidung des Zusammenlegens gut begründet werden und auf die Zielgruppe hin abgestimmt werden.

Auch die Abstufungen einer Skala müssen validiert[169] werden, denn gerade die o. g. intuitiven und qualitativen Methoden unterliegen der Subjektivität. Ansonsten läuft die betreffende Skala Gefahr, eine Hierarchie zu postulieren, die so in der realen Welt nicht existiert. Wie nun können die Abstufungen validiert werden? Zum einen müssen die Abstufungen selbst ebenso wie die Annahme, dass obere Niveaus die unteren mit einschließen, wie gesagt nachvollziehbar aus Theorien und Modellen abgeleitet werden; zum anderen kann durch Diskussionen in Expertengruppen ein gewisser Konsens hinsichtlich der vertikalen Abstufungen erzielt werden; empirische Beschreibungen und Analysen von Aufgaben, Anforderungen, Fertigkeiten oder Performanzen stellen einen weiteren Weg der Validierung dar. Doch letztlich sollten solche meinungsbasierten Skalierungen, wie North & Schneider (1998: 221f) sie nennen, durch datenbasierte Skalierungen empirisch validiert werden. Psychometriker müssen hierbei die jeweils adäquaten Messmodelle ansetzen.

Das gemeinsame Verständnis der Abstufungen und Kategorien innerhalb der Zielgruppe ist dennoch nicht automatisch gegeben, selbst wenn alle Aspekte einer Skala validiert worden sind. Das mag einerseits daran liegen, dass es sich oft um heterogene Benutzer handelt, die erst auf den Umgang mit einer gegebenen Skala vorbereitet werden müssen; andererseits kann es aber

[169] Vgl. Brindley (1998) für einen Überblick über Studien, die sich mit der empirischen Validierung von Abstufungen beschäftigen.

auch an der Versprachlichung der Deskriptoren liegen. Lumley (2002: 258ff) beschreibt die Probleme, die auftreten können, wenn die Beschreibungen der Merkmale und Abstufungen der Interpretation bedürfen: Gerade bei *rating scales* kann dies zu einer unreliablen und nicht validen Bewertungssituation führen – wie dem vorgebeugt werden kann, wird im nächsten Kapitel erläutert. Man muss allerdings die Grenzen einer klaren und verständlichen Beschreibung, die nicht in Fachjargon gleiten soll, akzeptieren: Jedes Wort trägt in der Regel mehrere Bedeutungen, so etwa auch eine „Alltagsbedeutung", die neben der in einer Skala erwünschten spezifischen Bedeutung immer mitschwingt. Interpretationsprobleme werden sich auch dann ergeben, wenn man auf rein verbale Abstufungen wie (*sehr*) *wenig*, *durchschnittlich*, (*sehr*) *viel* verzichtet und statt dessen versucht, qualitativ beschreibende Adjektive zu verwenden, denn diese müssen in ihrer Bedeutung erst von den Skalennutzern erschlossen werden. Im Extremfall könnte man sich vorstellen, dass zu jedem beschreibenden Terminus eine Bedeutungsdefinition gegeben wird – natürlich ist dies weder machbar noch praktikabel, so dass gerade bei der Versprachlichung den Skalennutzern und deren Hintergrund verstärkte Aufmerksamkeit zukommen muss. Ratsam sind Schulungen im Umgang mit den betreffenden Skalen, um einen möglichst breiten Konsens in der jeweiligen Benutzergruppe zu erzielen. Sollte es sich um mehrere zusammengehörige Skalen oder gar um ein Skalensystem handeln, so wäre eine Systematisierung der Begrifflichkeiten um der Kohärenz und Verständlichkeit Willen wünschenswert.

Zusätzlich können so genannte *benchmarks* die Bedeutung des jeweiligen Niveaus illustrieren. *Benchmarks* sind prototypische Performanzbeispiele oder typische Aufgaben eines Skalenniveaus, die zu einem gemeinsamen Verständnis der Niveaus beitragen. In Kapitel 4 werden solche *Benchmark*-Texte vorgestellt, die im Modul *Textproduktion Englisch* in der DESI-Studie bei der *Rater*-Schulung und bei der Bewertung der offenen Schreibaufgaben eingesetzt wurden.

Im Rahmen der Validierung von Skalen müssen auch Fragen der Generalisierung und Simplifizierung geprüft werden: Skalen sind Modelle der Realität, doch in der Regel ist diese wesentlich komplexer als man sie in Modellen abbilden könnte. Deshalb geben Skalen Beobachtungen oder Beschreibungen der Realität in vereinfachter oder verallgemeinerter Form wieder. Dabei muss jedoch darauf geachtet werden, dass es nicht zu Übergeneralisierungen kommt, in deren Folge die Skala Dinge beschreibt, die so nicht empirisch belegbar sind. Brindley (1998: 133f) warnt vor Übergeneralisierungen und zu starker Vereinfachung, da die Gefahr besteht, dass beispielsweise eine zu simplifizierende Bewertungsskala den hochkomplexen, multidimensionalen und variablen Sprachproduktionsprozessen nicht gerecht wird. Er rät in der Beurteilung deshalb zu komplexem Testen möglichst vieler Fertigkeiten und zur Bewertung mithilfe von kontextualisierten Skalen, um zu einem Lernerprofil zu kommen statt zu einem wenig aussagekräftigen, da zu generell gehaltenen Gesamtergebnis. Denn auch eine Kompetenzskala, die Bewertungsergebnisse in generalisierter Form rückmelden soll, darf nur auf die durch die Bewertung tatsächlich elizitierten Prozesse, Fertigkeiten oder Wissensbestände hin verallgemeinert werden, wie Alderson (1991b: 80) und Pollitt (1991b: 91) bemerken. Daher schlägt

Brindley konkrete, auf Empirie basierende, kontext- wie *task*-bezogene Skalen vor, deren Status, Zweck und „Reichweite“ klar definiert werden muss.

Die Validität einer *rating scale* etwa und ihrer Abstufungen kann aber nicht dadurch überprüft werden, indem man auf Basis der fraglichen Skala Testitems konstruiert und diese wiederum mithilfe der fraglichen Skala bewertet. Das wäre ein Zirkelschluss, der nach Brindley (1998: 120f) weder die Deskriptoren noch die Testitems validieren kann. Er schlägt vielmehr vor (ebd.: 125ff), folgende Gebiete näher zu analysieren: Inhaltsanalysen von Rasch-skalierten Testitems, um zu validen schwierigkeitsbestimmenden Charakteristika zu kommen; intro- wie retrospektive Erforschung der Prozesse, die ein bestimmter *task* verlangt, um zu validen Beschreibungen der Strategien und Prozesse zu gelangen, die im Test auch tatsächlich eingesetzt werden; *Task*-Analysen, um die Effekte der *tasks* auf die Testperformanz kontrollieren zu können; Textanalysen bei textbasierten Testitems, um empirisch validierte Texthierarchien zu erhalten; und nicht zuletzt die Erforschung der Perspektive des Spracherwerbs.

Auch wenn Skalen in der Regel nicht den Spracherwerb selbst beschreiben, so bildet er doch die Basis jeden Sprachlernfortschritts, wie de Jong (1988: 74)[170] schreibt:

> What we need to know if we want to develop good scales is not linguistic knowledge of how language is structured, what all the features of language are; we need to know how somebody acquires language, that is, what the developmental stages in language acquisition are.

Dennoch ist Vorsicht geboten bei der Interpretation der hierarchischen Abstufungen als Spracherwerbshierarchien. Auch wenn solch eine ansteigende Beschreibung des sprachlichen Verhaltens oder des Könnens etwa in einer *proficiency scale* einen Lernfortschritt impliziert, so darf dieser nicht als Erwerbshierarchie interpretiert werden, wenn der abgestuften Beschreibung nicht der Spracherwerb einer Probandengruppe über eine bestimmte Zeit hinweg zugrunde liegt, sondern eine Momentaufnahme der unterschiedlichen Leistungsstände innerhalb der Probandengruppe. Der Zusammenhang zwischen Querschnitt (beobachtete Unterschiede in der Lernentwicklung) und Längsschnitt (beobachtete Lernentwicklung) ist insofern gegeben, als dass Unterschiede im Querschnitt als unterschiedlicher „Stand“ des Spracherwerbs interpretiert werden könnten, doch diese Interpretation müsste durch empirische Längsschnittuntersuchungen abgesichert werden, ebenso wie implizierte Erwerbssequenzen empirisch validiert werden müssten, ehe sie als solche interpretiert werden können. Deshalb scheint es bei der Entwicklung von Skalen ratsam, zwischen Momentaufnahmen einer Fähigkeit und Langzeitbeobachtungen von Lernfortschritten zu unterscheiden. Im Fall von *proficiency scales* werden vermutlich beide Perspektiven ihren Beitrag zu einer validen Beschreibung leisten können, doch dabei ist es wesentlich, die jeweilige Basis der Deskriptoren offen zu legen.

Probleme hinsichtlich der Validität von Skalen können sich auch ergeben, wenn sie zu Funktionen verwendet werden, zu denen sie nicht konstruiert wurden.[171] Alderson (1991b: 74f)

[170] Zitiert in Brindley (1998: 130).
[171] Vgl. beispielsweise Alderson 1991b, Brindley (1998: 133f).

beispielsweise beschreibt, dass die Performanz in einem bestimmten *task* nicht unbedingt Rückschlüsse auf die sprachliche Kompetenz insgesamt zulässt, so dass die für die Bewertung des *tasks* konstruierte Skala nicht einfach als Kompetenzskala zur Rückmeldung genutzt werden kann. Auch kann es Validitätsprobleme geben, wenn etwa die Funktionen von aufgabenorientierten und benutzerorientierten Skalen verquickt werden: Wenn in einer *reporting scale* Fähigkeiten aus einer aufgabenorientierten Skala übernommen und beschrieben werden, die der der Rückmeldung zugrunde gelegte *task* gar nicht elizitiert hat, ist die Validität der Rückmeldeskala in Frage gestellt. Eine Bewertungsskala darf aus den gerade genannten Gründen nicht einfach als diagnostische Information zum Entwicklungsstand der Interimsprache interpretiert werden, wenn sie (auf Basis einer Momentaufnahme) bestimmte Kriterien einer Leistung beschreibt, welche auf bestimmten Niveaus erwartet werden.

Skalen sind nach Pollitt (1991b: 87) dann valide konstruiert, wenn die o. g. Kommunikation zwischen Testkonstrukteuren, Probanden und Testbewertern gegeben ist:

> It is worth reminding ourselves that grade descriptors are themselves an exercise in communication, intended to convey some message to some audience or some purpose. If assessment is to be valid there must be no breakdown in the Triangle of Communication between Exam Setter, Candidate and Rater.

Auch wenn Validität und Reliabilität in der Regel eng zusammen hängen, werden Fragen der Reliabilität von Skalen hier nicht abgehandelt, da sie sich in der Regel in Bezug auf die reliable Nutzung von Skalen in der Bewertung stellen, weshalb sie erst im Kapitel 3.3 *Rating*-Verfahren behandelt werden.

3.2.5 Ein Metasystem zur Vergleichbarkeit von Skalen

Sind erst einmal valide Skalen zu bestimmten Zwecken konstruiert worden und sind ihre Niveaus valide definiert und beschrieben, so stehen solche Skalen dennoch isoliert und können nicht einfach mit anderen Skalen respektive deren Niveaus in Verbindung gebracht werden. Dies wäre jedoch zu Zwecken der Vergleichbarkeit äußerst wünschenswert. Wie aber können verschiedene Skalen, die valide nach obigen Gesichtspunkten entwickelt worden sind – etwa Rückmeldeskalen verschiedener Tests – aufeinander bezogen werden? Was sagt beispielsweise das Erreichen eines bestimmten Kompetenzniveaus in der PISA-Studie aus im Vergleich etwa zu den *levels of proficiency* des GER oder zu den Kompetenzniveaus in der DESI-Studie? Kann ein Metasystem helfen, verschiedene Skalen aufeinander zu beziehen? Wie müsste solch ein Metasystem beschaffen sein?

Wenn Skalen Beschreibungen liefern sollen, die jenseits des konkreten Skalenkontextes interpretierbar sind, so muss es einen Referenzpunkt geben, auf den man sich beziehen kann – sonst verbleiben die Aussagen einer Skala isoliert auf den jeweiligen Kontext genau dieser Skala und können nicht verallgemeinert werden. Dies wird jedoch in Zeiten der Debatte um

Bildungsstandard in einem europäischen Kontext immer dringlicher, um über Bildungs- und Ländergrenzen hinweg Inhalte und Informationen transportieren zu können. Dabei spielt das Verständnis der Dimensionen und Niveaus die entscheidende Rolle – nur wenn es gelingt, gemeinsame Dimensionen und Niveaus über existierende oder zu konstruierende Skalen hinweg zu schaffen, werden Aussagen gegebener Skalen auch vergleichbar.

Hilfreich bei solch einem Vergleich ist ein Skalensystem, auf das die einzelnen Skalen bezogen werden können. Solch ein Metasystem müsste über Dimensionen verfügen, die auf einem konsensfähigen Kompetenzmodell von sprachlicher Handlungsfähigkeit und Kommunikationsvermögen basieren und so ausdifferenziert sind, dass sie möglichst vielen Kontexten gerecht werden. Die vertikalen Abstufungen müssten ebenfalls auf konsensfähigen Modellen fußen und empirisch abgesichert werden. Die den Skalen des Systems zugrunde gelegten Beschreibungen müssten alle relevanten Aspekte von Sprache und Kommunikation abdecken und ihre Beschreibungsbasis[172] offen legen, so dass ihr Status deutlich wird und damit auch ihr Verwendungsbereich. Dann könnte solch ein Skalensystem als Referenzsystem dienen, auf das neu zu konstruierende oder existente Skalen bezogen werden können.[173]

An dieser Stelle wird ein exemplarischer Ausblick gegeben, wie unterschiedliche Skalen auf solch ein gemeinsames Niveausystem bezogen werden könnten; dieser wird in Kapitel 4 anhand des Moduls *Textproduktion Englisch* des DESI-Projekts konkretisiert und dokumentiert.

Als Beispiel dienen an dieser Stelle zunächst geschlossene Testaufgaben, für die eine *reporting scale* oder Kompetenzskala, wie sie in PISA und DESI genannt wird, konstruiert werden soll. Dazu bietet es sich neben der Verankerung des Testkonstrukts in entsprechenden Theorien und Modellen an, *Task*-Analysen und wenn möglich Analysen der bei dieser Aufgabenstellung elizitierten Prozesse, Fertigkeiten und Wissensbestände durchzuführen. So kommt man zu den Anforderungen, die eine Aufgabe an die Probanden stellt, und zugleich zu empirisch fundierten Merkmalen, die die Aufgabenschwierigkeiten bestimmen. Diese Anforderungen und Merkmale können nun abgestuft beschrieben werden und die einzelnen Testitems können anhand dieser Merkmale eingestuft werden. Diese Einstufungen können dann wiederum mit den empirisch ermittelten Aufgabenschwierigkeiten, die in der Regel über Rasch-Skalierungen abgeschätzt werden, in Verbindung gebracht werden, beispielsweise über Korrelationsberechnungen – hohe Korrelationen legen einen Zusammenhang zwischen schwierigkeitsbestimmendem Merkmal und der tatsächlichen Schwierigkeit nahe. So können diese Merkmale einerseits genutzt werden, die auf Basis der Rasch-Skala zu konstruierende *reporting scale* inhaltlich valide zu beschreiben, und andererseits können diese Merkmale bei der Bestimmung von Niveaugrenzen

[172] Der Terminus „Beschreibungsbasis“ bezeichnet die Basis der Skalenbeschreibungen; beispielsweise können Beschreibungen auf Erwartungen, auf theoretischen Annahmen oder auf empirisch beobachteten Tatsachen basieren.
[173] Der GER will solch einen Referenzrahmen bieten; ob sein Skalensystem den Ansprüchen eines Referenzsystems genügt, wird im Anschluss unter Kapitel 3.4 dieser Arbeit erörtert. Mögliche Wege der Anbindung an die Niveaus des GER werden im erwähnten *Manual* beschrieben, welches unter Kapitel 3.5 dieser Arbeit vorgestellt wird.

dieser *reporting scale* mittels Regressionsanalysen genutzt werden.[174] Die so erhaltenen Niveaubeschreibungen können dann inhaltlich mit einem Referenzniveausystem abgeglichen werden, um zu sehen, wo es zwischen den Referenzniveaus und den Niveaus der fraglichen Skala zu Berührungspunkten kommt. Zusätzlich kann die Anbindung mittels quantitativer Methoden empirisch validiert werden.

Bei offenen Aufgaben wie etwa Schreibaufgaben ist das Vorgehen ein etwas anderes: Dabei liegt das Augenmerk nicht auf der Skalierung der Aufgaben selbst, die wie oben in Kapitel 2.2.3 erläutert nicht exakt in ihrer Schwierigkeit zu bestimmen sind, sondern es liegt auf der Konstruktion valider *rating scales*. Dazu bietet es sich neben *Task*- und Prozessanalysen und der Verankerung der Skalen in Theorien und Modellen an, Performanzbeispiele qualitativ zu analysieren und auf diesem Weg zu abgestuften Merkmalen zu kommen, die charakteristisch für ein bestimmtes Niveau sind. Die so entworfenen Niveaus der *rating scales* können zur Validierung einerseits inhaltlich mit den Niveaus eines gegebenen Referenzsystems abgeglichen werden; andererseits können die Deskriptoren zusätzlich durch Experten skaliert werden. Mithilfe der so konstruierten *rating scales* wird dann durch geschulte *raters* die Performanz bewertet. Im nächsten Schritt werden die *raters* skaliert; durch das Ansetzen eines probabilistischen Messmodells kann den Eigenarten der *raters* Rechnung getragen werden, so dass diese Skalierung letztlich zu einer validen Einschätzung der Fähigkeiten der Probanden führt. Im letzten Schritt müssen die *rating scales* in Kompetenzskalen überführt werden; letztere können wiederum qualitativ-inhaltlich mit den Niveaus eines gegebenen Referenzsystems abgeglichen werden und dieser Abgleich wiederum über quantitative Methoden abgesichert werden, so dass man Kompetenzskalen erhält, die auch Aussagen jenseits des konkreten Tests zulassen.

Bei dem hier skizzierten Vorgehen werden nicht gegebene Referenzniveaus (wie etwa die des GER) als Ausgangspunkt genommen, sondern die zu konstruierenden Skalen werden auf theoretischen wie empirischen Grundlagen entwickelt und somit auf eine eigene, dem Kontext der betreffenden Skala gerecht werdende Basis gestellt. Erst am Ende des Entwicklungsprozesses werden sie mit dem Referenzsystem in Verbindung gebracht, um Berührungspunkte und Anbindungsmöglichkeiten zu eruieren. Dies stellt eine Möglichkeit dar, die Einteilungen und Niveaus eines gegebenen Referenzsystems und diejenigen der neu zu konstruierenden Skala gegenseitig zu validieren.

[174] Dieser Weg wurde etwa im DESI-Projekt bei den Testmodulen *Hörverstehen*, *Leseverstehen* und *C-Test* eingeschlagen. Für nähere Informationen zu Vorgehen und zum angesetzten Regressionsmodell vgl. Hartig 2005; für einen Kurzüberblick über das Vorgehen beim Modul *C-Test* vgl. Kapitel 3.5.6 dieser Arbeit.

3.3 *Rating*-Verfahren

In jüngster Zeit wird der Positivbewertung mehr und mehr Beachtung geschenkt, da die traditionelle Negativ- oder Fehlerkorrektur wie erwähnt einige Nachteile[175] mit sich bringt: Das Fehleranstreichen oder Fehlerzählen betrachtet vorwiegend die quantitative Seite einer Performanz, vernachlässigt die Qualität der Leistung, ermöglicht wenig Einblicke in das Können der Lernenden, bietet keine oder nur wenig Lernanreize und wirkt daher eher demotivierend. Der Positivansatz hingegen stellt die Qualität einer Performanz in den Mittelpunkt und betrachtet, was vorhanden ist und was gekonnt wird. Damit gibt er qualitative Informationen über den Leistungsstand, bietet Lernanreize und motiviert zum Weiterlernen.

Während etwa Fehlerquotienten objektiv auszählbar sind, dabei aber die Qualität einer Leistung außer Betracht lassen, haben Positivansätze den Nachteil, dass sie subjektive Einschätzungen der Qualität einer Leistung mit sich bringen. Gerade die Bewertung produktiver offener Aufgabenstellungen, deren Bearbeitung ein komplexes Zusammenspiel von Prozessen, Strategien und Wissensbeständen verlangt, stellt besondere Anforderungen an das Bewertungsinstrument: Es muss so beschaffen sein, dass es allen Aspekten der zu bewertenden Performanz gerecht wird, valide erstellt und reliabel benutzbar ist. Die Entscheidung für den einen oder anderen Bewertungsansatz oder für eine Kombination beider Verfahren, um ein für das jeweilige Testkonstrukt angemessenes Bewertungsinstrument zu entwickeln, muss im jeweiligen Kontext begründet werden. Pollitt (1991b: 88) beispielsweise schlägt bei allen produktiven *tasks judging-based assessment* vor, um sich auf Performanzen und darauf bezogene Bewertungskriterien zu konzentrieren und so den Leistungen auch gerecht zu werden, während Lehmann (1990) analytische Verfahren gestützt auf wenige objektivierbare Bewertungsaspekte vorschlägt.

An dieser Stelle werden die Grundlagen von *Rating*-Verfahren diskutiert, nicht aber das Für und Wider dieses Ansatzes, denn dieses lässt sich nur bezogen auf konkrete Kontexte erörtern.[176] Ausgehend vom Beschreibungsgegenstand und den Abstufungen werden in den folgenden Unterkapiteln Charakteristika valider *rating scales* beschrieben, ehe die verschiedenen Arten von *rating scales* und ihre Verwendungsmöglichkeiten betrachtet werden. Im Anschluss daran werden die Prozesse, die beim *rating* ablaufen, beleuchtet, um Anforderungen eines *Rater*-Trainings abzuleiten. Dabei kommt Aspekten der Bewertung der Schreibfertigkeit vermehrte Aufmerksamkeit zu, da das Praxisbeispiel dieser Arbeit in diesem Feld angesiedelt ist. Dennoch gelten die hier getroffenen Aussagen in gewissem Rahmen auch für die Bewertung der Sprechfertigkeit, auf die an dieser Stelle jedoch nicht näher eingegangen werden kann.

[175] Vgl. hierzu beispielsweise Alderson 1991a, Bleyhl 2003, Börner 1989, Hamp-Lyons & Kroll 1996, Pollitt 1991a, u. a..
[176] Das Für und Wider des *Rating*-Ansatzes und die Konstruktion eines solchen Bewertungsinstruments werden in Kapitel 4 dieser Arbeit am erwähnten Praxisbeispiel aus dem DESI-Projekt konkretisiert.

3.3.1 *Rating Scales*

Wie oben erläutert muss der Gegenstand einer Skala valide beschrieben werden, immer bezogen auf den jeweiligen Zweck der Skala. Bei *rating scales*, deren Verwendungszweck in der Bewertung produktiver Aufgabenstellungen liegt, bietet es sich an, nach den oben skizzierten Methoden die dabei elizitierten Performanzen selbst zu beschreiben, ebenso wie die dabei zum Einsatz kommenden Fertigkeiten und Wissensbestände. Dazu müssen die Beschreibungen auf den jeweiligen *task* bezogen sein, um die Reaktionsmöglichkeiten auf diesen *task* abzudecken, denn *rating scales* sollen u. a. als Werkzeug dienen, die Komplexität der Antworten in möglichst vergleichbarer Form darzustellen. Das Bewertungsinstrument muss aus Validitätsgründen im Testkonstrukt verankert werden, denn wenn ein Test nicht valide bewertet wird, so kann er auch nicht valide testen. *Task*-Analysen, Prozessanalysen und Performanzanalysen bilden die Basis der Konstruktion des Bewertungsinstruments.

Wenn entschieden ist, welche Kriterien in welchen Bewertungskategorien angesetzt werden, so muss die Frage nach Abgrenzung und Gewichtung der angesetzten Bewertungskriterien respektive der Gewichtung der Kriterien innerhalb einer Kategorie beantwortet werden. Denn bei letzterem Fall muss den *raters* mit an die Hand gegeben werden, welche Rolle die einzelnen Kriterien innerhalb einer Kategorie spielen sollen. Im ersteren Fall hingegen muss entschieden werden, ob die Kategorien unabhängig als Profil rückgemeldet werden oder ob darüber hinaus ein Gesamtergebnis rückgemeldet werden soll, dessen Zusammensetzung transparent dargestellt werden muss. In der Literatur, insbesondere zur Bewertung von Schreibaufgaben, finden sich keine schlüssigen Hinweise, welche Facetten wie zu gewichten sind. Lehmann (1990) etwa nennt einige Untersuchungen, die durch das Ansetzten von Messmodellen versucht haben, relevante Bewertungsfaktoren zu analysieren. Es scheinen sich drei Faktoren *Inhalt, Aufbau* und *Stil* zu manifestieren, doch lassen sich bezüglich deren Gewichtung keine eindeutigen Hinweise finden. Untersuchungen darüber, wie *raters* Kriterien ohne nähere Vorgaben abgrenzen und gewichten, ergeben ebenfalls keine eindeutigen Befunde. Milanovic, Saville & Shuhong (1996) beispielsweise fanden heraus, dass die Gewichtung sehr heterogen gesehen wurde und dass es nicht möglich war, komplexe Kriterien absolut unabhängig von anderen Kriterien zu bewerten: “This criterion [i. e. communicative effectiveness, Anm. d. V.] is obviously related to grammar, vocabulary and structure, all factors which assist coherence and comprehension.“ (ebd.: 101).[177] Des Weiteren schreibt er (ebd.: 103): “It seems that the majority of raters sought to balance grammatical accuracy with communicative competence, giving the latter a slight priority“.

Aus Mangel an eindeutigen Forschungsergebnissen müssen im Einzelfall Testkonstrukt und Datenlage darüber entscheiden, wie welche Kategorien voneinander abgegrenzt werden

[177] Diese Befunde kann man interpretieren als Beleg für das oben erwähnte Prototypenmodell und die systemische Organisation von Sprache auf verschiedenen hierarchischen Ebenen: Das komplexe Kriterium der kommunikativen Wirksamkeit kann als den Kriterien Grammatik, Wortschatz und Struktur übergeordnetes Kriterium betrachtet werden, wobei es zwischen den Kategorien einer Ebene und zwischen den Kategorien unterschiedlicher Ebenen zu fließenden Übergängen und Überschneidungsbereichen kommt.

können und wie sie zu gewichten sind. Man findet in der Literatur[178] Ansätze, die alle Bewertungskategorien gleich gewichten; andere, die einigen Merkmalen mehr Gewicht verleihen als anderen, wobei dies oft auf Basis von Intuition oder Experteneinschätzung entschieden wird; wieder andere, bei denen die Gewichtung auf statistischen Analysen wie beispielsweise Faktorenanalysen beruht. Es ist ratsam, solche Entscheidungen einerseits auf Basis von Theorien und Experteneinschätzungen, eingebettet in das jeweilige Testkonstrukt, zu fällen, doch sollte komplementär dazu auch die empirische Datenlage mit entscheiden, was im jeweiligen Kontext machbar und angemessen ist.

Die oben getroffenen Aussagen zur Konstruktion von Skalenabstufungen haben auch für *rating scales* ihre Gültigkeit. Dabei muss die Handhabbarkeit über die Anzahl der Niveaus entscheiden; in der Regel werden nicht mehr als sechs Niveaus angesetzt. Gerade bei Bewertungsskalen sollten die Verbalisierungen der zur Abstufung genutzten Merkmale näher betrachtet werden, da sie die Entscheidungsgrundlage der Bewertung darstellen: Es scheint, wie beispielsweise Pollitt & Murray (1996: 86ff) bemerken, dass bestimmte Performanzmerkmale eher am oberen, andere eher am unteren Skalenspektrum auftauchen und sich scheinbar komplementär zueinander verhalten.[179] Deshalb spricht prinzipiell nichts dagegen, nicht alle Merkmale auf allen Niveaus abgestuft zu beschreiben, sondern relevante Merkmale auf den Niveaus, für die sie auch typisch sind. Die Entscheidung darüber muss je nach Situation und Bedarf gefällt werden, genau wie die Entscheidung darüber, ob man ausschließlich positive Merkmale beschreibt oder auch komplementär dazu einige typische Charakteristika auf negativem Weg beschreibt. Denn bei der Performanzanalyse und Beschreibung zeichnen sich, wie oben schon angedeutet, gerade die unteren Niveaus oft durch typische Mängel oder Fehlleistungen aus, die in einer *rating scale* durchaus aufgenommen werden können, um es den *raters* zu erleichtern, die Deskriptoren mit den Performanzen in Verbindung zu bringen. Auch unterstützen Hypothesen des Spracherwerbs solch ein Herangehen, denn die Interimsprachenhypothese etwa postuliert unter anderem, dass Fehlleistungen gerade beim Beginn des Sprachenlernen ein typisches Merkmal der Interimsprache sind und nicht als Defizit betrachtet werden müssen, sondern vielmehr als Lernanreiz und Quelle der Weiterentwicklung. In diesem Sinne könnten auch Negativformulierungen wertvolle Dienste leisten.

[178] Vgl. hierzu etwa Hughes 1986, Kroll 1998, Lehmann 1990, Lumley 2001, Milanovic, Saville & Shuhong 1996, Pollitt et al. 1996, Shohamy et al. 1992.

[179] In der Untersuchung von Pollitt & Murray hat sich gezeigt, dass die *raters* bei guten Probanden andere Kriterien ("What does he say?") angewandt haben als bei schlechteren ("How does he say it?"). Dieses Herangehen scheint so nicht reliabel, denn eine faire Bewertung verlangt, dass alle Probanden an denselben Kriterien gemessen werden. Es müssen nicht alle Merkmale auf allen Stufen erscheinen, doch die den Merkmalen zugrunde liegenden Kriterien müssen für alle Probanden dieselben sein, um zu einer reliablen und validen Bewertung zu kommen. Pollitt & Murray haben versucht, aus der Analyse von *Rating*-Prozessen relevante Merkmale und deren Abstufungen abzuleiten, die sie dann als Skala der Bewertung zugrunde legten – eine Art der intuitiven Skalenentwicklung, die erst validiert werden müsste.

3.3.1.1 Typen von *Rating Scales*

Es gibt je nach Art des Beschreibungsgegenstands und Zweck der *rating scale* unterschiedliche Typen von Skalen zur Bewertung: Soll mit einer Skala eine globale Einstufung vorgenommen werden, so wird die Skala nur eine holistische Bewertungskategorie beschreiben, während am anderen Ende des Spektrums analytische Skalen denkbar sind, die eine Vielzahl einzelner, abgegrenzter Bewertungskategorien beschreiben, die sich wiederum aus einem oder mehreren Bewertungskriterien zusammen setzen können. Entsprechend werden solche Skalen auch *holistische* respektive *analytische* Skalen genannt. Zwischen den beiden Enden lassen sich noch weitere Ansätze ausmachen, deren bedeutungsvollste der *primary trait approach* und *der multiple trait approach* sind. Im Folgenden sollen Eigenschaften und Einsatzmöglichkeiten dieser vier Skalenarten mit ihren jeweiligen Vor- und Nachteilen beleuchtet werden.[180]

Holistische Skalen geben den Gesamteindruck einer Leistung wieder, im Sinne dessen, was die Probanden „insgesamt“ können; dabei ergibt sich dieser Eindruck aus der Betrachtung des „Ganzen“ und wird nicht auf wenige, eventuell gar über- oder unterbeachtete Kriterien gestützt. Diese Art des Bewertens mag für Einstufungstests oder *proficiency tests* angemessen sein, bei denen es um die Feststellung geht, ob ein bestimmtes Niveau erreicht ist respektive auf welchem Niveau die Lerner sich generell befinden. Auch ermöglichen holistische Skalen einen Gesamteindruck, der sich nicht als Summenscore aus einzelnen *ratings* ergibt, sondern aus der jeweiligen Performanz als Ganzes. Allerdings kann ein Globalurteil dadurch nur unzureichend die Stärken und Schwächen einer Performanz wiedergeben: “But sometimes, a text is so internally complex (e.g., highly developed but fraught with grammatical errors) that it requires more than a single number to capture its strengths and weaknesses." (Hamp-Lyons 1996b: 760). Holistische Skalen ermöglichen also keine Rückmeldung diagnostischer Informationen und sind nur schwer zu interpretieren. Es ist nicht sichergestellt, dass sich alle *raters* auf dieselben Merkmale konzentrieren, noch dass sie diese gleich gewichten, um zu ihrem Urteil zu kommen. Die Bewertung kann durch *Halo*-Effekte verzerrt werden. *Halo*-Effekte liegen vor, wenn ein besonders auffälliges Merkmal alle anderen Merkmale überlagert und somit die Bewertung verzerrt. Darüber hinaus ist manchmal nicht eindeutig, ob sich die Skala auf eine bestimmte Fertigkeit wie die des Schreibens, oder auf das Sprachkönnen, die *proficiency* allgemein bezieht. Zudem werden holistische Skalen den oft unterschiedlich ausgeprägten Teilfertigkeiten nicht gerecht, denn sie implizieren, dass sich alle bei dieser Leistung eingesetzten Teilfertigkeiten in etwa parallel entwickeln – diese Annahme ist jedoch so nicht korrekt; wie in Kapitel 1.3.1 gezeigt entwickeln unterschiedliche Lerner sich auf sehr unterschiedlichen Wegen:

> The descriptors [of a holistic scale; Anm. d. V.] imply a pattern of development common to all language learners. They assume that a particular level of grammatical ability will always be associated with a particular level of lexical ability. This is, to say the least, highly questionable. The potential lack

[180] Die Ausführungen hier basieren auf Cohen 1994, Cumming 1990, Hamp-Lyons 1996a+b, Hamp-Lyons & Kroll 1996, Hughes 1986, Kroll 1998 und Shohamy et al. 1992.

> of fit in individuals between performances in the various subskills leads naturally to a consideration of analytic methods of scoring. (Hughes 1986: 91).

Analytische Skalen beschreiben – oft als Set von Einzelskalen – voneinander abgrenzbare, für den der Bewertung zugrunde liegenden *task* relevante Kriterien. Dieses Herangehen kann den Facetten einer komplexen Leistung eher gerecht werden, die damit möglichst unabhängig voneinander bewertet werden können. Allerdings ist das Ganze etwas anderes als die Summe seiner Teile, so dass dabei der Blick „aufs Ganze" leicht verloren geht. Auch ist nicht immer gewährleistet, dass die Kriterien wirklich unabhängig voneinander bewertet werden. Dennoch können analytische Skalen diagnostische Informationen liefern. Zudem wird nach Hughes (1986) über die Abgabe mehrerer Einzelbewertungen die Reliabilität der Bewertung insgesamt erhöht. Cohen (1994) jedoch führt an, dass die analytische Bewertung insofern unreliabel sein könnte, als dass solche Performanzen bevorzugt werden könnten, in denen sich die Merkmale der Skalen leicht wiederfinden lassen.

Ein weiterer Weg der Bewertung, der aufgrund der Nachteile des holistischen Bewertens entwickelt wurde, findet sich im ***primary trait scoring***: Dabei wird durch den Testinstrumentkonstrukteur das für diesen *task* und diese Bewertung entscheidende Merkmal, der *primary trait*, festgelegt. Der *task* muss natürlich auf die Erfassung dieses Merkmals ausgelegt sein und für eine valide Bewertung muss es für jedes wesentliche Merkmal einen oder mehrere solcher *tasks* geben, so dass sich die Bewertung im Idealfall aus genügend *samples* zu den wichtigsten Merkmalen in einem gegebenen Test zusammensetzt. Solch ein Vorgehen scheint aufwändig, jedoch kann es im Idealfall zu einem detaillierten Profil führen. Der Vorteil liegt darin, dass man sich auf einen Aspekt zu einer Zeit konzentrieren kann, doch ob dies von den *raters* auch so umgesetzt wird, ist nicht sichergestellt. Auch stellt Cohen (1994) in Frage, ob man solch einen *primary trait* validieren kann und ob dieser Ansatz nicht zu eng ist, indem er nur ganz bestimmte Merkmale betrachtet und andere, eventuell ebenso wichtige, außer Acht lässt. Diesem Ansatz fehlt ebenso wie bei dem analytischen der Blick aufs Ganze.

Hamp-Lyons (1996b) schlägt vor, statt der oben skizzierten Typen ein ***multiple trait scoring***[181] zu benutzen, um die erwähnten Nachteile auszugleichen:

> Holistic scoring obscures a pattern of consistent overemphasis or underemphasis on basic language control, but a multiple- trait instrument, in which language control is a trait to be judged together with traits found salient in the context, and in which the reader [rater, Anm. d. V.] is free to attend to the multidimensionality of ESL writing, is likelier to facilitate a balanced response to the strengths and weaknesses of the writer's text. (Hamp-Lyons 1996b: 234).

Diese Herangehensweise ist ganzheitlich ausgelegt, wobei *task* und elizitierte Performanz in die Bewertung einfließen: Auf der Basis von Analysen relevanter Schreibsituationen und *tasks* werden mehrere charakteristische Merkmale oder *multiple traits*, die elizitiert und bewertet werden sollen, in je einer Skala beschrieben. Zur Bewertung selbst müssen wiederum mehrere *tasks* entwickelt werden, die genau diese *traits* auch elizitieren. Auf diese Weise sollte man zu einem

[181] Zur Entwicklung dieser Herangehensweise vgl. Hamp-Lyons 1991.

balancierten und validen Gesamtergebnis kommen, unter Beachtung aller für diese Bewertung relevanten Merkmale. Cohen (1994) jedoch hält es für fraglich, ob solche *traits* tatsächlich empirisch belegt und validiert werden können. Sind die Merkmale bewertet, so sollen die Einzelbewertungen nach Hamp-Lyons (1996b) auch als Profil rückgemeldet werden und nicht zu einem Summenscore verrechnet werden, so dass auch dabei der Blick aufs Ganze verloren gehen dürfte.

Alle vier hier skizzierten Ansätze sollten je nach Kontext eingesetzt werden, um die genannten Nachteile zu minimieren und die erwähnten Vorteile zu nutzen. Konstrukteure eines Bewertungsinstruments müssen demnach immer situations- und kontextbezogene Entscheidungen über Teildimensionen, Abstufungen, Beschreibungsgegenstände und Typen des *rating* treffen – in vielen Fällen dürfte eine Kombination der verschiedenen Wege, immer begründet im jeweiligen Testkontext, zum gewünschten Ziel führen.

3.3.1.2 Die Rolle der Deskriptoren

Sind die grundlegenden Entscheidungen über Beschreibungsgegenstand, Dimensionen, Abstufungen und Ausrichtung der *rating scale(s)* getroffen, so muss der Status der Deskriptoren betrachtet werden, die Rolle also, die die Deskriptoren im Entscheidungsprozess spielen sollen beziehungsweise tatsächlich spielen. Den Idealfall stellt eine Skala validen Inhalts dar, welche die in einem *task* elizitierte Performanz und die dabei zum Einsatz kommenden Fertigkeiten und Wissensbestände in validen Abstufungen beschreibt. Die in der Skala beschriebenen Merkmale werden dann an der zu bewertenden Performanz festgemacht, um diese auf das wahrscheinlichste Niveau der Skala einzustufen. Dabei sollen die Deskriptoren als Basis und Ausgangspunkt der Bewertung dienen und im Sinne eines Werkzeugs oder Modells helfen, die möglichen Reaktionen auf einen *task* hin vergleichbar zu bewerten. Wie erwähnt stellen sie als Modell eine Simplifizierung der Realität dar – dies darf nicht vergessen werden. Die Deskriptoren sollen nach Hughes (1986: 89) den *raters* die Bewertung erleichtern, indem sie gleiche Erwartungen schaffen und bedeutungsvolle Umschreibungen statt wenig aussagekräftiger Punktwerte geben. Zudem bilden *rating scales* nach Hamp-Lyons (1996b: 234) die Basis der Rückmeldung, indem sie nützliche Informationen kommunizieren, die geteilt werden können mit den Probanden, deren Lehrenden und anderen davon Betroffenen im Bildungssystem, in dem die Bewertung stattfand.

Die Rolle der Skalen kann jedoch aus verschiedenen Perspektiven betrachtet werden: Upshur & Turner (1985) beispielsweise schlagen vor, sich der Bewertung produktiver *tasks* mittels empirisch entwickelter binärer Fragen zu nähern, wobei sich diese binären Fragen auf die Grenzen zwischen den Niveaus beziehen, in der Skala also nicht die zu bewertende Performanz beschrieben wird. Vielmehr besteht diese Skala aus Entscheidungsfragen, die aus der Analyse

von acht Performanzbeispielen entwickelt wurden: Upshur & Turner haben konkrete charakteristische Merkmale analysiert, mithilfe derer sie behaupten, eine gegebene Performanz reliabel und valide aufgrund einiger aufeinander folgender Ja/Nein-Entscheidungen einstufen zu können: “These scales require the rater to make a series of binary choices about features of student performance, that define boundaries between score levels. They are … empirically derived, binary-choice, boundary-definition (EBB) scales.” (Upshur & Turner 1985: 6). Bei genauerer Betrachtung jedoch kann dieses Herangehen nicht überzeugen: Während sich die oben skizzierten *rating scales* auf das Gemeinsame und Prototypische eines Niveaus beziehen und darüber feststellen, ob ein Niveau mit einer bestimmten Wahrscheinlichkeit schon erreicht ist[182], fokussieren diese binären Fragen darauf, ob eine konstruierte Grenze überschritten ist, wobei dies an nur einem einzigen Merkmal festgemacht wird – ohne dass man Aussagen darüber treffen könnte, welche der anderen für das Niveau charakteristischen Merkmale außerdem vorhanden sind.[183]

Lumley (2002) sieht *rating scales* ebenfalls nicht als Beschreibung der Performanz, sondern eher als “…a set of negotiated principles that the raters use as a basis for reliable action, rather than a valid description of language performance“ (ebd.: 286). Allerdings kann es vorkommen, wie Lumley (2002) bei der Untersuchung der Prozesse, die beim *rating* ablaufen, festgestellt hat, dass die Skalen nicht als Basis der Entscheidungsfindung genutzt werden, sondern nach der intuitiv gefällten Entscheidung als Rechtfertigung herangezogen werden. Wie dies verhindert werden kann, wird unten bei den Ausführungen in Kapitel 3.3.3 *Rater-Training* gezeigt.

3.3.2 *Rating*-Prozesse

Es gibt eine Vielzahl von Untersuchungen, die sich mit den Prozessen beschäftigen, die dem *rating* zugrunde liegen – meist werden Techniken des „lauten Denkens“ (*thinking-aloud techniques)* und Datenanalysen eingesetzt, um herauszufinden, welche Prozesse im Kopf der *raters* ablaufen, welche Strategien sie einsetzen, wie sie die Deskriptoren interpretieren und benutzen, auf welche Merkmale sie besonders achten und wie sie einzelne Kriterien gewichten. Diese Erkenntnisse können genutzt werden, um valide Skalen zu konstruieren und um erfolgreiche Prozesse und Strategien in einem *Rater*-Training zu vermitteln. Im Folgenden werden Untersuchungsergebnisse von Connor-Linton 1996, Cumming 1990, Lumley 2002, Milanovic, Saville & Shuhong 1996, Pollitt & Murray 1996 und Shohamy 1992 zusammengefasst.

[182] Die Wahrscheinlichkeit zeigt sich am Vorliegen bestimmter prototypischer Merkmale: Je mehr Merkmale eines Niveaus auf eine Performanz zutreffen, desto wahrscheinlicher dürfte das Niveau erreicht sein.

[183] Konkret lautet die erste Frage dieser EBB-Skala: „Sind mehr als 10 Zeilen vorhanden?“ Hier wird über die Länge alle weitere Bewertung limitiert, als ob die Länge das entscheidende Kriterium darstelle – dieser Annahme fehlt jedoch die empirisch belegte Basis. Die zweite Frage nach *Klarheit des Textes*, falls die erste denn mit *Ja* beantwortet wurde, entscheidet dann über das Erreichen der nächsten Stufe. Hat man die erste Frage jedoch verneint, so entscheidet als zweite Frage die Frage nach *Themenbezug* über die endgültige Einstufung. Die beiden Fragen auf der zweiten Ebene sind also nicht mehr vergleichbar. Im Extremfall bedeutet dies, dass eine Performanz mit beispielsweise 100 Zeilen, die keinen Themenbezug hat, aber klar aufgebaut ist, mindestens auf Stufe 3 bewertet wird, ohne dass eine Aussage bezüglich der dabei verwendeten Sprache getroffen wird. Ob dies eine valide Form der Bewertung ist, sei dahingestellt. Zudem sind die einzelnen Kriterien, die sich in den Fragen niederschlagen, nicht mehr unabhängig voneinander bewertbar – kein Kennzeichen einer reliablen Bewertung.

3.3.2.1 Studien zu *Rating*-Prozessen:

Cumming (1990) untersuchte *Rating*-Prozesse, indem er untrainierte Laien und Experten bei der Abgabe eines Globalurteils beobachtete. Dabei wurden ihnen weder abgestufte Kriterien noch Hinweise zur Gewichtung der Kriterien vorgegeben. Sie sollten lediglich die Facetten *language use, rhetorical organisation* und *substantive content* bewerten und dabei über lautes Denken offenbaren, welche Prozesse in ihren Köpfen abliefen. Beide Gruppen unterschieden zwischen sprachlichen Fähigkeiten einerseits und der Schreibfertigkeit andererseits, wobei die Experten größere Konsistenz zeigten. Das aufgrund der *thinking-aloud* Protokolle entwickelte Kodierschema[184] zum Erfassen der *Rating*-Prozesse brachte folgendes Ergebnis: Es ließen sich 28 *decision making behaviours* differenzieren, die sich aus interpretativen und evaluativen Strategien zusammensetzen. Unter die interpretativen Strategien fallen etwa das Lesen der Aufgabenstellung und des Textes, das Vorstellen der Situation (des Lerners, der Aufgabenstellung, der Textproduktion, des gedachten Rezipienten), das Identifizieren von Wirkfaktoren oder das Interpretieren der Wirksamkeit einer Textpassage (beispielsweise durch Auflösen von Mehrdeutigkeiten). Zu den evaluativen Strategien zählen die persönliche Haltung gegenüber der Textqualität, das Lesen der Kriterien, um eben diese Qualität festzustellen, und das Vergleichen von Aufsätzen untereinander, um zu evaluieren, welche Bedeutung die interpretierten Faktoren haben. Die Variabilität der eingesetzten Strategien war jedoch von *rater* zu *rater* sehr hoch.

Auch Lumley (2002) hat *Rating*-Prozesse über *thinking-aloud* Protokolle analysiert, um Abläufe, Interpretationsverhalten und Schwierigkeiten beim *rating* zu untersuchen. Er arbeitete mit vier erfahrenen *raters*, die Aufsätze anhand einer Skala mit vier Kategorien (*task fulfillment and appropriacy, cohesion and organisation, conventions of presentation, grammatical control*) bewerten sollten. Er fand 147 Kategorien von Verhalten, die er in drei grobe Typen einteilte: Management-Verhalten (wie beispielsweise Selbstkontrolle oder erste Kommentare), Leseverhalten (als Teil der Interpretation) und *Rating*-Verhalten (Zuordnung der Skalenniveaus). Er fand heraus, dass alle *raters* ähnlichen Prozessen in drei Schritten folgten: erstes Lesen, Einstufen der Kriterien, Überdenken der Entscheidung. Die Beziehung zwischen den Skaleninhalten und der Textqualität blieb jedoch im Dunkeln. Die Aufgabe der *raters*, namentlich das Zusammenführen von Texteindruck, Textmerkmalen und Skalendeskriptoren, ließ dabei „Freiräume“ in Interpretation und Zusammenführung von Text und Skalen, welche die *raters* mit verschiedenen Strategien ausfüllten: Die Spannung zwischen den *Rating*-Vorgaben und dem intuitivem Texteindruck wird von jedem anders aufgelöst; darin sieht Lumley eine der Ursachen für Inkonsistenzen beim *rating*. Diese könnten aber durch adäquates Training, begleitende Schulung und durch Bewertungsrichtlinien abgefangen werden. Hier eine Übersicht über die von Lumley identifizierten Schritte beim *rating*:

[184] Bei der Kodierung wurden die drei folgenden Kriterien angesetzt: Es musste sich erstens um relevante, logische, unterscheidbare kognitive Verhaltensweisen handeln, die zweitens mit genügender Häufigkeit auftraten und drittens mit genügender Genauigkeit kodierbar waren.

Stage	Rater's focus	Observable behaviours
1. First reading (pre-scoring)	• Overall impression of text: global and local features	• Identify script • Read text • Comment on salient features
2. Rate all four scoring categories in turn	• Scale and text	• Articulate and justify scores • Refer to scale descriptors • Reread text
3. Consider scores given	• Scale and text	• Confirm or revise existing scores

Abb. 9 (Lumley 2002: 255): Model of the stages in the rating sequence

Milanovic, Saville & Shuhong (1996) haben anhand der Bewertung offener direkter Schreibaufgaben aus FCE und CPE[185] das Entscheidungsverhalten beim *rating* mit einer rein verbal abgestuften holistischen Skala untersucht. Damit sollten Messfehler der Bewertung minimiert werden, die in den *raters* oder den *Rating*-Prozeduren ihre Ursache haben können; daneben sollten Training und *rating scales* verbessert werden. Methodisch wurden retrospektive geschriebene Berichte, introspektive verbale Berichte und Gruppeninterviews eingesetzt, um die vier Gruppen von *raters* (Erfahrung mit FCE, Erfahrung mit CPE, EFL-Lehrer ohne *Rating*-Erfahrung, erfahrene Muttersprach-Englischlehrer) zu untersuchen. Man fand vier Ansätze der Bewertung: a) Prinzipiell zweimaliges Lesen der Texte; b) pragmatisch gehandhabtes zweites Lesen nur bei Unklarheiten; c) einmaliges Lesen mit anschließendem Einstufen auf Basis des ersten Eindrucks; d) einmaliges Lesen, wobei es zu erster vorläufiger Bewertung schon beim Lesen kommt. Folgendes Modell veranschaulicht das in dieser Studie gefundene typische *Rater*-Verhalten[186]:

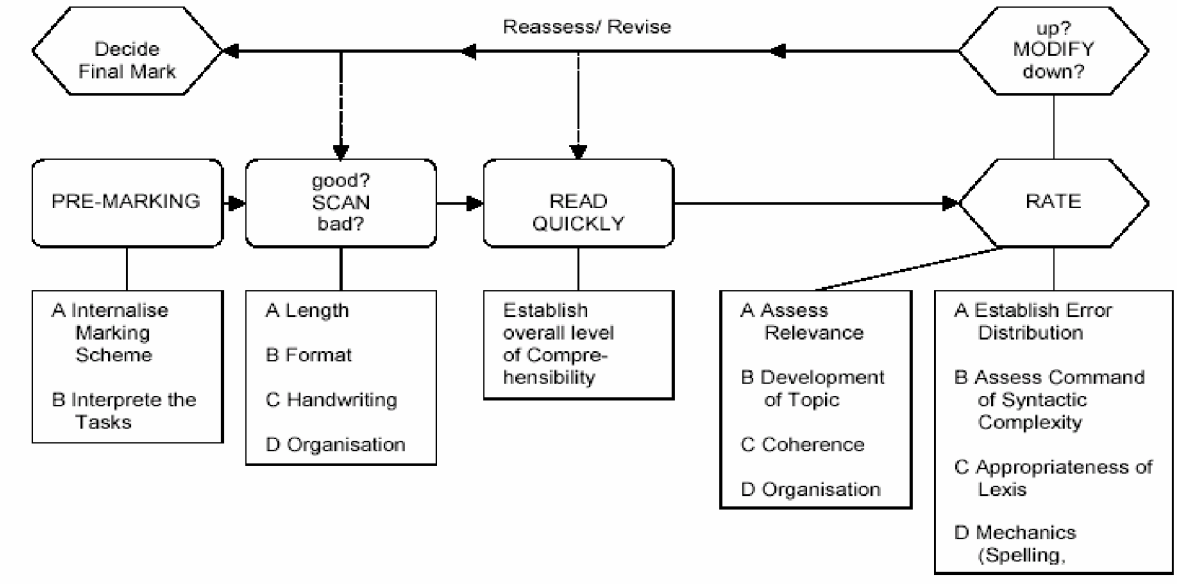


Abb. 10 (Milanovic, Saville & Shuhong 1996: 95): A model of the decision-making process in composition marking

[185] *Cambridge TESOL Exams: First Certificate of English, Certificate of Proficiency in English*.
[186] Zum Einfluss des Hintergrunds der *raters* vgl. unten Kapitel 3.3.2.2 *Reliabilitätsaspekte*.

Pollitt & Murray (1996) untersuchten, wie und was *raters* beim *rating* wahrnehmen und erfahren. Dazu nutzten sie zum einen eine Methode aus der Konstruktpsychologie nach Kelly[187], die *Repetory Grid Procedure*. Diese versucht über die Elizitierung von Wahrnehmungsprozessen und mittels Vergleich von Ähnlichkeiten und Unterschieden in der Wahrnehmung, die Konstrukte und Denkprozesse herauszufinden, die bei der Wahrnehmung im Kopf existieren oder ablaufen. Die zweite Methode, die Pollitt & Murray einsetzten, war die der *Paired Comparisons*[188], welche Objekte auf Basis von direkten Vergleichen auf ein Kontinuum einordnet, "… thereby providing a means of measuring their [i. e. the objects'] relative values according to the degree to which subjects perceive them as encapsulating whatever attribute the researcher is concerned with" (ebd.: 78). Eine Reihe verschiedener Performanzen wurden jeweils paarweise verglichen, um herauszufinden, welche der beiden höher einzustufen sei. Beide hier eingesetzten Methoden nutzten den Kontext des Vergleichens. Die erstere sollte helfen, Ähnlichkeiten und Unterschiede bei den *Rating*-Prozessen herauszufinden, um darauf aufbauend die Konstrukte der Performanz, die die *raters* einsetzten, zu analysieren, und um eventuelle Korrelationen zwischen diesen Konstrukten und den jeweils zugewiesenen *proficiency levels* zu bestimmen. Die Methode der *Paired Comparisons* sollte helfen, Konsistenzen der *raters* zu überwachen. Pollitt fand u. a. heraus, dass die Präzision des Urteils durch die Konzentration auf bestimmte Attribute geschärft wird; dies spricht für eine Aufteilung holistischer Urteile oder zumindest für eine nähere Beschreibung der Aspekte, die dem Globalurteil zugrunde liegen sollen. Des Weiteren zeigten die *raters* in dieser Studie zwei Zugänge zum *rating*: Einige ließen sich vom ersten holistischen Eindruck leiten, den sie mit schon bewerteten Performanzen eines bestimmten Niveaus verglichen. Dabei zeigte sich, dass der erste Eindruck alles Weitere überlagerte, da alle weiteren Charakteristika ebenfalls diesem intuitiv zuerst gefundenen Niveau zugeschrieben wurden. Andere *raters* limitierten ihre Bewertung auf tatsächlich beobachtbares Verhalten und kamen ebenfalls zu einem Gesamturteil, doch nicht über den alles überlagernden ersten Eindruck, sondern über das Identifizieren beobachtbarer Phänomene.

Solche Erkenntnisse können in *Rater*-Schulungen und bei der Skalenkonstruktion genutzt werden, um gewünschtes Verhalten zu fördern und somit die Reliabilität und Validität der Bewertung zu erhöhen.

3.3.2.2 Reliabilitätsaspekte

Wie kann die Reliabilität von *Rating*-Verfahren gewährleistet werden? Wie müssen *raters* sich verhalten, um solch eine Bewertung reliabel zu machen? Sind Laien und Experten gleichermaßen zum *raten* geeignet? Hilft Erfahrung oder Training, reliabler und valider zu *raten*?

[187] Vgl. Kelly 1955.
[188] Diese Methode stellt die Operationalisierung von Thurstones *Law of Comparative Judgement* dar, vgl. hierzu Thurstone 1959.

Alderson (1991a) stellt fest, dass sich absolute Reliabilität in der Bewertung nur über das „Klonen“ von *raters* erzielen ließe. Da dies natürlich nicht durchführbar ist, wird im Allgemeinen doppelt-blind bewertet, das heißt jede Performanz wird unabhängig von zwei *raters* bewertet. Alderson (ebd.) meint, dass die Reliabilität der *raters* abhängt von validen Aufgabenstellungen und einem validen Bewertungsschema. Messtheoretisch schlägt er vor, sich der *Rater*-Reliabilität über die Intra-*Rater*-Reliabilität zu nähern, die sich über Korrelationen zwischen zwei Bewertungen derselben Performanzen durch denselben *rater* in gewissem Abstand errechnen lässt. Nach Alderson (1991b) lässt sich Reliabilität in der Bewertung über Konsensfindung der *raters* bezüglich der Bewertungskriterien und deren Abstufungen erreichen. Dazu benötigt man leicht zu interpretierende und handhabbare Deskriptoren, die prototypisches, empirisch beobachtbares Verhalten beschreiben und die in ihren Abstufungen aufeinander aufbauen. Alderson kommt zu dem Schluss, dass Training und Bewertungsrichtlinien unabdingbar sind; zudem sollte die Reliabilität nach dem Training an standardisierten Skripten überprüft werden.

Cumming (1990, vgl. oben) verglich das Verhalten der Laien und Experten, die an seiner o. g. Untersuchung zu *Rating*-Prozessen teilnahmen. Beide Gruppen erhielten wie gesagt keine Schulung oder nähere *Rating*-Anweisungen. Auffällig war, dass die Experten mehr Selbstreflexion (im Sinne von Kontrollstrategien) zeigten; mehr auf Schlüsselmerkmale im Text achteten ebenso wie auf Abgrenzung der Kriterien; der Qualität des Inhalts mehr Aufmerksamkeit schenkten; bei der Beurteilung der Sprache mehr auf den Gesamteindruck achteten; beim Kriterium der rhetorischen Organisation mehr auf *key features* schon beim Lesen achteten und *linking language* in die Bewertung der Kohärenz einbezogen; und schließlich bei Fehlern die Klassifikation in beispielsweise *slips* und *errors* als eine Möglichkeit sahen, die Qualität der Sprache zu bewerten. Diese Beobachtungen können als Hinweis interpretiert werden, erfahrene oder geschulte *raters* einzusetzen.

Shohamy et al. (1992) untersuchten, ob Training oder der Hintergrund der *raters* größeren Effekt auf die *Rater*-Reliabilität haben. Dazu entwickelten sie auf Basis empirischer Analysen Skalen, die von insgesamt 20 *raters* angewandt werden sollten. Die *raters* waren je zur Hälfte Laien beziehungsweise Lehrer. Die Hälfte von ihnen, darunter Lehrer und Laien, wurde im Umgang mit den entwickelten Skalen geschult, indem die Skalen erläutert und gemeinsam durchgesprochen wurden, die Skalenbenutzung demonstriert wurde und von den Trainees selbst ausprobiert werden konnte, und die Demo-*Ratings* schließlich diskutiert wurden, bis ein hinreichender Konsens gefunden war. Die Vergleichsgruppe, ebenfalls aus Lehrern und Laien zusammengesetzt, erhielt nur die Skalen mit Terminologieerläuterungen als Grundlage der Bewertung. Dann bewertete jede dieser vier Gruppen dieselben 50 Arbeiten. Auf Basis dieser Daten wurden Inter- wie Intra-*Rater*-Reliabilitäten errechnet, mit folgendem Ergebnis: Bei den Inter-*Rater*-Reliabilitäten gab es keinen signifikanten Unterschied zwischen Laien und Experten, doch wurde der Effekt des Hintergrunds leider nicht über längere Zeiträume kontrolliert. Die

Intra-*Rater*-Reliabilitäten wurden lediglich für die Gruppe der trainierten Experten kontrolliert und ergaben über einen Zeitraum von drei Wochen stabile *ratings*.

Milanovic, Saville & Shuhong (1996) kommen zu dem Ergebnis, dass sowohl Hintergrund der *raters* als auch Training positiven Effekt auf die Stabilität der *ratings* haben. Diese wird aber beeinflusst durch die erwähnten *Halo*-Effekte und das *sequencing* der Texte[189], so dass auch diese Aspekte in einer Schulung und beim eigentlichen *rating* kontrolliert werden müssen, um zu reliablen *ratings* zu kommen.

Shale (1996) meint, dass man das Problem der Subjektivität beim *rating* weder dadurch lösen kann, dass man sich diesen Aspekten gar nicht mehr zuwendet und nur mehr objektivierbare Kriterien bewertet, die dann wiederum wesentliche qualitative Aspekte der Performanz außer Acht lassen, noch dadurch, dass man *raters* so auswählt und drillt, bis alle zum selben Ergebnis kommen. Er schlägt vielmehr vor, das Problem des *rater disagreement* durch entsprechendes Training zu minimieren und mittels genauer Spezifizierung von *tasks*, von Bewertungskriterien, von *rating scales* und deren Verwendung das „Universum" genau zu beschreiben, das die *raters* als „interpretative Gemeinschaft" bewerten sollen. Die *ratings* dieser Gemeinschaft seien dann auf dieses spezifizierte Universum hin verallgemeinerbar. Auf dieser Basis könne man zu stabilen und generalisierbaren *ratings* kommen, die auch den Anforderungen an Inter- wie Intra-*Rater*-Reliabilitäten genügten (Shale 1996: 93):

> There is no logical or philosophical basis to support such a proposition [i.e. the concept of identical marker behaviour, Anm. d. V.]. Does it not make more sense to accept that markers *naturally* vary in their judgments of texts and to settle on a measurement theory that allows us to accommodate this reality? Generalizability theory provides this structure, permitting us to specify markers as a facet in a study and to estimate the variation that is due to them – and to remove the effect of this variation from our considerations of other factors that may be of more direct interest.

Der oben erwähnte Schritt des Lesens zu bewertender Texte verdient noch zusätzliche Beachtung: Jeder Mensch hat andere Erwartungen an Texte, deren Aufbau, Inhalt, sprachliche Darstellung, Ton und Stil. Je eher ein gegebener Text mit unseren Erwartungen zusammentrifft, desto „besser" werden wir ihn finden. Diese individuell geprägte Einstellung und Erwartungshaltung an Texte beeinflusst unsere Wahrnehmung von Texten vom ersten Lesen an. Diese Haltung wirkt sich auch auf die Bewertung von Texten aus. Um Texte vergleichbar und reliabel zu bewerten, muss unter den *raters* ein gemeinsames Vorgehen bei der Textrezeption und Bedeutungserschließung etabliert werden, um zu gemeinsamen Erwartungshaltungen bezüglich dessen zu kommen, was einen „guten" Text im jeweiligen Bewertungskontext ausmacht.

Die Basis einer reliablen Bewertung ist in der Konstruktion valider *tasks* und Bewertungskriterien sowie in der genauen Spezifizierung des zu bewertenden „Universums" zu finden. Dieses muss den *raters* in einer Schulung vermittelt werden, zusammen mit den grundlegenden Prozessen

[189] *Sequencing*-Effekte beziehen sich auf die Reihenfolge, in der Texte bewertet werden – folgt beispielsweise ein guter Text auf viele schlechte, so kann es passieren, dass der gute Text besser eingestuft wird als er eigentlich ist.

und Strategien des *rating*, welche sich wie folgt charakterisieren lassen: Nach der ersten Lektüre von Aufgabenstellung und Text werden zu bewertende Merkmale im Text identifiziert und interpretiert (beispielsweise kommunikationsbelastende Fehler, Wortschatzbreite, Auftreten bestimmter Strukturen…). Darauf aufbauend wird der Gebrauch und der Einsatz dieser Merkmale bewertet nach zuvor zu definierenden Kriterien der Verständlichkeit, Korrektheit, Angemessenheit, Relevanz, Breite, etc.: Beispielsweise muss *linking language* zuerst identifiziert werden, ehe deren Verwendung interpretiert werden kann (Welche sprachlichen Mittel treten wo auf? Sind sie angemessen eingesetzt? Fehlt etwas?). Zum Schluss werden Angemessenheit und Effizienz der interpretierten Merkmale in Abgleichung mit *rating scales*, *benchmarks* und schon bewerteten Performanzbeispielen beurteilt. Daneben spielen Strategien der Selbstkontrolle und Reflexion eine Rolle. Durch entsprechendes Training kann eine interpretative Gemeinschaft gebildet werden, denn wie die o. g. Untersuchungen gezeigt haben, ist eine gemeinsame Vorstellung dessen, was in welcher Weise bewertet werden soll und ein gemeinsames Verständnis der Bewertungskategorien und der Niveaus der *rating scales* unabdingbar für eine reliable und valide Bewertung. Zudem sollten *raters* über einen gewissen gemeinsamen Hintergrund verfügen, auf dessen Basis die erwünschten Strategien erst geschult werden können.

3.3.3 *Rater*-Training

An dieser Stelle können nur wesentliche Charakteristika eines solchen Trainings umrissen werden, denn die Inhalte müssen im jeweiligen Kontext konkret gefüllt werden. Die Ausführungen stellen wiederum die Bewertung der Schreibfertigkeit in den Mittelpunkt, doch sie können in gewissem Rahmen auch auf die Bewertung der Sprechfertigkeit übertragen werden. Ein konkretes Trainingsprogramm im Rahmen des DESI-Projekts wird in Anhang 29 dokumentiert.

Generell sollten der Ansatz der Positivbewertung, die Bedeutung von *Rating*-Skalen und deren Charakteristika, und die Prozesse und Strategien des *rating* die Basis des Trainings bilden. Dazu können die Forschungsgrundlagen, je nach Hintergrund der *raters*, durchaus transparent dargestellt werden, so dass man sich je nach Interesse weiter in eine Themenstellung einarbeiten kann. Die Ausführungen oben umreißen diese Forschungslage und können entsprechend in eine Schulung aufgenommen werden.

Ergänzend sollte zu Beginn der Schulung die Notwendigkeit des Trainings selbst thematisiert werden, ebenso wie die Ziele, die damit erreicht werden sollen. Im Folgenden wird eine kleine Auswahl möglicher Quellen genannt, die dazu herangezogen werden können:

Cumming (1990) betont, dass gerade für Laien ohne Bewertungserfahrungen Training wichtig ist, um zu gemeinsamen Kriterien und Abstufungen zu kommen, und um gemeinsame, vergleichbare Strategien einzusetzen, die das Ergebnis konsistent und reliabel machen. Denn Laien hätten zwar Potential, müssen aber erst Erfahrungen im Umgang mit Texten sammeln, um

weg von der Oberfläche in die Tiefe zu gelangen: Abstraktionsprozesse bei der Textrezeption und Bedeutungserschließung müssen erfahren und eingeübt werden, um relevante Textmerkmale zu identifizieren und in ein Gesamtbild des Textes zu integrieren.

Shohamy et al. (1992) und Weigle (1994) fanden heraus, dass *Rater*-Training eine wichtige Rolle spielt, besonders wenn dabei Bewertungskriterien und deren Abstufungen geklärt werden und dadurch transparent und handhabbar für die *raters* werden; wenn die verwendete Terminologie definiert wird, so dass Konsens herrscht über die Bedeutung und Abgrenzung der Kriterien, die Interpretation der Kriterien und der Texte; und wenn Konsens erzielt wird über das eigentliche Vorgehen (die *rating steps*). Dies erhöht die Reliabilität der Bewertungen.

Lumley (2002: 267) betrachtet *Rater*-Training ebenfalls als Voraussetzung für reliables *rating*:

> Rating is certainly possible without training, but in order to obtain reliable ratings, both training and reorientation are essential in order to allow raters to learn or (re)develop a sense of what the institutionally sanctioned interpretations are of task requirements and scale features, and how others relate personal impressions of text quality to the rating scale provided.

Er gibt aber gleichzeitig zu bedenken, dass Training und Skalen manchmal lediglich als Rechtfertigung genutzt werden und nicht als Basis der Urteilsfindung. Solche Grenzen der Schulbarkeit müssen ebenfalls im Training thematisiert werden.

Die Ziele eines Rater-Trainings fasst etwa Cohen (1994: 336) übersichtlich zusammen:

> 1. Make sure that the raters gain the ability to give each assessment category the designated focus, whether or not it be equal focus.
> 2. Make sure that the raters use the **same** criteria for rating and that they all have the **same** understanding of what these criteria mean.
> 3. Strive to have novice raters approximate expert raters in terms of their rating behaviour.
> 4. If possible and if appropriate, provide for all raters training that will help them be sensitive to the rhetorical strategies of writers from other language and cultural backgrounds.

Am Ende sollten die *raters* eine interpretative Gemeinschaft bilden, die sich auszeichnet durch ein gemeinsames Verständnis von und ein gemeinsames Herantreten an: Aufgabenstellung und deren Anforderungen; zu bewertende Texte und deren Rezeption; Bewertungskategorien beziehungsweise Kriterien und deren Abgrenzung und Gewichtung; Abstufungen und Textzuordnungen; Anwendung und Einsatz von *Rating*-Strategien. Nicht zuletzt sollten die zukünftigen *raters* im Umgang mit möglichen Grenzen und Problemen des *rating* vertraut worden sein, wie unten erläutert wird.

Selbstverständlich müssen das Testkonzept, das der Beurteilung zugrunde liegt, das konkrete Testkonstrukt und der Task vorgestellt werden, ebenso wie das Bewertungsinstrument im Detail besprochen werden muss. Diesem sollte vermehrte Aufmerksamkeit geschenkt werden, müssen doch die *raters* das Instrumentarium vergleichbar interpretieren und einsetzen. Das Verständnis der Kategorien, der innerhalb der Kategorien angesetzten Kriterien, der Abgrenzung der Kriterien und der Abstufungen der Skalen bildet den Dreh- und Angelpunkt der Schulung. Sollten die Bewertungsskalen an ein Referenzsystem angelehnt sein, so muss auch dieses in der Schulung thematisiert werden. Da der Umgang mit Skalen und deren Bedeutung

erläutert und erfahren werden muss, bietet es sich an, Deskriptoren von den *raters* skalieren zu lassen oder selbst einige entwerfen zu lassen. So können etwa Grenzen der Versprachlichung erfahren werden oder Inkonsistenzen in den Deskriptoren überarbeitet werden, wenn möglich und machbar gemeinsam mit den *raters*. Dieses Vorgehen kann das Verständnis der Skalenkonstruktion und der damit einhergehenden Problemstellungen vertiefen und so zu einer angemessenen Verwendung der Skalen beitragen.

Um zu vergleichbarem Umgang mit den zu bewertenden Texten zu kommen, können Modelle der Textrezeption und Bedeutungskonstruktion helfen, sei es nun in Bezug auf geschriebene oder gesprochene Texte. Die Ausführungen etwa von Kintsch & vanDijke (1978) können dabei wertvolle Dienste leisten: Sie beschreiben Konstruktionsprozesse bei der Textrezeption, die vom Ziel der Rezeption abhängig sind. Dazu identifizieren sie semantische Oberflächen- und Tiefenstrukturen, die bei der Bedeutungskonstruktion eine Rolle spielen und die sich an Propositionen auf Mikro- respektive auf Makroebene eines Textes manifestieren. Sie entwickeln darauf aufbauend ein Prozessmodell des Textverstehens (ebd.: 368): Die Verstehensleistung ist gekennzeichnet durch das in der Regel automatisierte Erstellen eines Netzwerks kohärenter Propositionen auf Mikro- wie auf Makroebene des fraglichen Textes. Die jeweiligen Leseabsichten wirken dabei wie ein Filter, der bestimmt, was als relevant oder bedeutsam wahrgenommen wird.

Solche Erkenntnisse können in einer *Rater*-Schulung genutzt werden: Beispielsweise kann man gemeinsam untersuchen, wie man an die zu bewertenden Texte herantritt und unter welchen Gesichtspunkten man deren Bedeutung interpretieren soll. Denn um bei einer Performanzbewertung zu einer vergleichbaren Rezeption der Texte mit vergleichbaren Bewertungszielen zu kommen, müssen *individuelle* Erwartungen an einen „guten“ Text in den Hintergrund treten zugunsten einer *gemeinsamen* Erwartungshaltung, die von allen *raters* geteilt wird.

Neben den theoretischen Grundlagen und praxisorientierten Erfahrungen, die in einer Schulung vermittelt werden müssen, darf man emotionale und behaviorale Aspekte nicht vergessen. Man sollte beispielsweise den Effekt einer möglichen Verunsicherung der *raters* nicht unterschätzen, die sich ergeben könnte, weil bisherige Erfahrungen und Erwartungen aufgebrochen und erweitert werden müssen. Schon der Positivansatz ist im Allgemeinen schwer umzusetzen, wenn die *raters* aus einem Bildungssystem kommen, in dem traditionell nur Negativkorrekturen erfahren wurden. Solch eine beispielsweise schulisch bedingte Sozialisation muss vorsichtig erweitert werden, um nicht Abwehrreaktionen hervorzurufen. Dasselbe gilt für das Herausbilden eines Erwartungshorizonts an gute Texte im Beurteilungskontext: Dabei müssen intuitive, meist nicht reflektierte Erwartungen aufgegeben werden und neue, nicht aus der eigenen Erfahrung stammende Haltungen eingenommen werden; dies könnte als Bedrohung angesehen werden und muss in der Schulung aufgefangen werden.

Um Verunsicherungen möglichst früh auffangen zu können, muss die Schulung neben kognitiven Verfahrensweisen auch erfahrende, praktische Momente enthalten, so dass die *raters* im

Lauf des Trainings genügend eigene Erfahrungen sammeln können im Umgang mit den neuen Strategien, im Umgang mit den Texten und nicht zuletzt im Umgang mit sich selbst. Früh sollten authentische Textbeispiele die theoretischen Grundlagen konkretisieren und Übungsmomente eingebaut werden, so dass die *raters* in „sicherer" Umgebung erste Erfahrungen machen können. Beispielsweise kann das oben Gesagte zur Bedeutungskonstruktion an ausgesuchten Texten nachvollzogen werden, indem deren Mikro- wie Makrostrukturen gemeinsam analysiert und konstruiert werden. Die *raters* sollten früh in der Schulung die Möglichkeit bekommen, an ausgesuchten und schon eingestuften *Benchmark*-Texten erste Erfahrungen im Umgang mit *Rating*-Strategien zu machen. Wenn man dazu schon eingestufte Texte nutzt, so reduziert man die Komplexität des *ratings* und lässt den *raters* genügend Kapazitäten, sich zunächst auf eine Sache auf einmal konzentrieren zu können, wie beispielsweise das isolierte Identifizieren der Merkmale des Niveaus, auf dem der Text schon eingestuft wurde. Im nächsten Schritt dann kann die Interpretation der identifizierten Merkmale geübt werden, und daran anschließend die eigentliche Einstufung und Niveauzuordnung. Im Verlauf des Trainings nimmt die Komplexität der Aufgaben und Übungen, die die *raters* erledigen müssen, sukzessive zu, bis die Prozesse so weit automatisiert sind, dass hinreichender Konsens beim Vorgehen geschaffen wurde. Dieser kann beispielsweise am Ende der Schulung überprüft werden durch die Ermittlung von Intra- wie Inter-*Rater*-Reliabilitäten in Bezug auf während und nach der Schulung bewertete Texte.

Die *raters* sollten auch für mögliche Probleme im Verlauf des *rating* sensibilisiert werden, sei deren Ursache nun in den *raters*, den *Rating*-Prozeduren oder in den zu bewertenden Performanzen zu finden:

Probleme, die sich etwa wegen Müdigkeit oder Konzentrationsmangel ergeben, sind am leichtesten zu kontrollieren: Das Arbeiten an einem hellen, ruhigen Ort mit genügend Pausen über nicht zu lange Zeiträume hinweg sorgt hier für Abhilfe. Schwieriger schon wird es, Konsistenz in der Beurteilung über längere Zeiträume zu zeigen, denn man hat nicht jeden Tag dieselbe Form: Dabei hilft, sich auf jede *Rating*-Sitzung neu einzustimmen, indem man die Aufgabenstellung und die *Benchmark*-Texte liest und sich erneut mit den *rating guidelines* beschäftigt. Der fortlaufende Vergleich einzustufender Texte mit *Benchmark*-Texten und mit schon bewerteten Texten trägt ebenfalls zur Konsistenz bei. Zusätzlich sollten Strategen der Selbstkontrolle eingesetzt werden, beispielsweise nach dem Motto: "Am I doing what I am supposed to do?"

Probleme, die sich aus den eigentlichen *Rating*-Prozeduren ergeben, sind beispielsweise die unter Kapitel 3.3.1.1 respektive 3.3.2.2 erwähnten *Halo*-Effekte und *Sequencing*-Effekte. Daneben könnte sich das Verständnis der Dimensionen oder Abstufungen im Lauf des *rating* verschieben. Um diese Probleme abzufangen, müssen sich die *raters* an das geschulte Vorgehen und die *guidelines* halten, Vergleiche mit *Benchmark*-Texten und schon bewerteten Texten nutzen, und regelmäßig prüfen, ob sie noch das Gewünschte betrachten. Solche „Selbst-Kalibrierungen" können und sollen durch begleitende Schulungen und Feedback auf Basis

statistischer Analysen ergänzt werden. Der *Halo*-Effekt lässt sich durch klar definierte und (so weit möglich) abgegrenzte Kriterien minimieren; der *Sequencing*-Effekt kann durch die Zusammenstellung gut gemischter *Rating*-Pakete, Vergleiche mit *benchmarks* und schon bewerteten Texten und durch *ranking* der neu zu bewertenden Texte kontrolliert werden.

Eine weitere Problemquelle könnte in der künstlichen Situation der Textsorte *writing in and for a test* (nach Shale 1996) zu finden sein: Der Zweck des Schreibens liegt im Test, das (kommunikative) Ziel ist nicht in den Probanden oder deren Lebenswelten zu finden. Es handelt sich bei den dabei entstehenden Texten also um nicht-authentische Produkte, die unter nicht-authentischen Bedingungen erstellt werden. Dieses Kunstprodukt muss dann von den *raters* bewertet werden, die sich ja ebenfalls in einer Kunstsituation befinden, müssen sie sich doch in die fiktiven Adressaten versetzen, um beispielsweise die kommunikative Wirksamkeit der Texte zu beurteilen. Allerdings sind auch diesem Hineinversetzten in die Adressaten Grenzen gesetzt, wie Elbow (1996: 128) bemerkt: "…we mustn't be *too* pure about taking 'the real reactions of real readers' as our only standard for judgement". Deshalb sollten die *raters* Sensibilität entwickelt haben für solche Schreibsituationen und deren besondere Bedingungen, beispielsweise über den Besuch von (akademischen) Schreibkursen, in deren Verlauf sie selbst solche Kunsttexte verfassen müssen. Ein letzter Problempunkt sei noch erwähnt: der Umgang mit Verweigerern, die in der Testsituation zwar einen validen Text erstellen, dennoch beispielsweise über unangemessene Inhalte sehr deutlich machen, dass sie eigentlich „keine Lust" dazu haben. Die Gründe hierfür sind vielfältig, doch können sie in dieser Arbeit nicht näher betrachtet werden. Entscheidend für die *raters* ist, dass sie solche Verweigerer nicht „bestrafen" dürfen, sondern dass sie lernen, auch an extreme Texte neutral heranzutreten und die Stärken und Schwächen auch solcher Arbeiten zu identifizieren. Im Einzelfall kann es sinnvoll sein, zusätzliche Bewertungskriterien anzusetzen, wie beispielsweise *Handschrift* oder *swear words*, um der „Bestrafung" vorzubeugen, indem solche Aspekte vor oder nach der eigentlichen Beurteilung eigens kodiert werden.

Abschließend kann man mit Shohamy et al. (1992) sagen, dass mehr *raters* trainiert werden sollten als benötigt, da sich nicht alle Menschen in gleicher Weise für diese Tätigkeit eignen. Während der Schulung stehen, wie Lumley (2002) bemerkt, die *raters* im Zentrum der Aufmerksamkeit, und nicht die Skalen, Stimuli oder Performanzen. Die *raters* entscheiden, auf welche Merkmale sie Ihre Aufmerksamkeit lenken, wie sie die Deskriptoren interpretieren und wie ihr erster Eindruck einer Performanz im Hinblick auf Vorgaben und Anforderungen aus Skalen und Training zu rechtfertigen ist. Lumley (ebd.) fand Hinweise, dass *raters* auch nach einem erfolgreich abgelegten Training ihren Stil beibehalten ebenso wie die Komplexität ihres Denkens und Urteilens, und dass nicht alle *raters* nach dem Training gleich milde oder streng werden. Dennoch kann solch eine Schulung helfen, ihnen eine gemeinsame Richtung vorzugeben und ihr Urteil mit den jeweiligen Vorgaben abzugleichen.

3.4 Der Skalenansatz des GER

Nachdem die Bedeutung von Skalen in der Beurteilung erörtert und in diesem Bereich die Anwendung von Skalen bei der Bewertung produktiver *tasks* näher betrachtet wurde, wird nun der GER auf seinen Skalenansatz und dessen Verwendungsmöglichkeiten hin analysiert.

Dem Skalensystem und den Referenzniveaus des GER ist ein eigener Abschnitt 3 im GER gewidmet. Dieser zeichnet sich weitgehend durch Verständlichkeit und Transparenz aus. Das zeigt sich beispielsweise daran, dass die GER-Niveaus an die existierenden Niveaus des Europarats angebunden werden und anhand dreier Beispielskalen illustriert werden, und daran, dass klare Aussagen zur Benutzung der GER-Skalen getroffen werden. Zur Transparenz dieses Abschnitts trägt ebenfalls bei, dass Fragen der Beschreibung und der Messverfahren im Skalenansatz angesprochen werden und in den Anhängen A und B des GER allgemeine Aspekte der Skalenkonstruktion sowie Aspekte der konkreten Entwicklung der Skalen im GER erhellt werden.

Das Skalensystem des GER erhebt den Anspruch, einen gemeinsamen Referenzrahmen zu bilden, der alle relevanten Aspekte des Sprachvermögens, der *proficiency* also, in einer für alle Beteiligten bedeutungsvollen und verständlichen Weise beschreibt. Dieser Rahmen will die Verknüpfung von Curriculumplanung, Lernzielbeschreibung, Lehr-/Lernmaterialentwicklung, Leistungsbeurteilung und Zertifizierungen ermöglichen, ebenso wie er zu transparenten und kohärenten Vergleichen verschiedener Bildungssysteme, Prüfungen oder Populationen führen will.[190]

Um zu beurteilen, ob die Skalen zu all diesen Funktionen eingesetzt werden können, muss der Status der Skalen analysiert werden, denn wie oben erläutert bestimmen Skalenkonstruktionsprozess und Beschreibungsgegenstand die Einsatzmöglichkeiten der betreffenden Skalen. Deshalb wird im Folgenden die Entwicklung der Skalen des GER im Hinblick auf die Auswahl der horizontalen Dimensionen, den Ursprung der Deskriptoren, die Konstruktion der Abstufungen und die Validierung der Skalen dargestellt. Anschließend wird das Selbstverständnis des Skalenansatzes, das dem GER zugrunde liegt, herausgearbeitet. Darauf aufbauend werden Beschreibungsgegenstand und Aspekte der Versprachlichung an exemplarischen Skalen analysiert, um letztlich den Status der Deskriptoren und damit auch die im GER angegebenen Verwendungsmöglichkeiten dieser Skalen zu beurteilen. Abschließend werden die GER-Aussagen bezüglich der Funktionen, die seine Skalen im Kontext der Beurteilung übernehmen können, eingeschätzt.

3.4.1 Konstruktion der GER-Skalen

Die Entwicklung der Skalen des GER war Teil des seit 1971 laufenden Großprojektes des Europarats, einen gemeinsamen Referenzrahmen für das Sprachenlernen in Europa zu entwickeln. Vorarbeiten zu den Niveaus des GER-Systems sind mit den Niveaubeschreibungen *Waystage*,

[190] Vgl. North & Schneider (1998: 219), GER (2001: 8ff).

Threshold Level und *Vantage Level* des Europarats[191] geleistet worden, auf die sich die Niveaus des GER-Systems beziehen.[192] Die konkrete Skalenkonstruktion erfolgte im Rahmen eines Projekts[193] des *Schweizer Nationalfonds zur Förderung der wissenschaftlichen Forschung*, das sich in zwei Abschnitten über drei Jahre (von 1993 bis 1996) zog. In der ersten Projektphase wurden mündliche Interaktion und mündliche Produktion in Englisch untersucht, in der zweiten Phase wurde die Studie erweitert um Lesen und Hörverstehen in Englisch, Französisch und Deutsch.

3.4.1.1 Dimensionsauswahl

Da in den Skalen des GER das Sprachvermögen beschrieben werden soll, müsste die horizontale Einteilung des Sprachvermögens in bedeutungsvolle Kategorien im Idealfall auf einer empirisch validierten Beschreibung der *proficiency* beruhen. Dies ist jedoch nicht möglich, da es wie erwähnt noch keine umfassende Theorie auf diesem Gebiet gibt. Dennoch ist es sinnvoller, sich in den Bereichen, die sich (noch) nicht durch entsprechende Theorien beschreiben lassen, auf den gesunden Menschenverstand zu stützen und Entscheidungen pragmatisch zu begründen, statt ganz auf den Versuch zu verzichten, sich der Beschreibung dieser Bereiche (hier: der *proficiency*) zu nähern (vgl. dazu u. a. North 2000: 32). Diese Entscheidungen jedoch müssen transparent dokumentiert werden.

Wie in North (ebd.: 41-54), nicht jedoch im GER selbst, umfassend dargestellt, speist sich der Begriff der *proficiency*, der der GER-Skalenkonstruktion zugrunde gelegt wurde, aus verschiedenen Modellen der kommunikativen Kompetenz: Es werden Modelle der Kompetenz, Performanz und des Sprachgebrauchs von Canale & Swain 1980, Canale 1983, Chomsky 1975 & 1980, Hymes 1971 & 1972b und Gumpertz 1982 & 1984 analysiert. Mit diesen Modellen werden Modelle der kommunikativen Kompetenz und des Sprachvermögens von Bachmann 1990 & 1991, Canale 1983 und van Ek & Trim 1990 kontrastiert, um relevante Kategorien des Sprachvermögens abzuleiten (vgl. North 2000: 62-65). Diese Kategorien werden jedoch nicht nur in den gerade erwähnten Modellen verankert; vielmehr werden darüber hinaus schon existierende Skalen und deren Einteilungen analysiert, ebenso wie Forschungsarbeiten zur Definition, Unterscheidung und Abgrenzung solcher Kategorien herangezogen werden (vgl. dazu die Ausführungen in North 2000: 74-122). Abschließend werden die in der Studie benutzten Kategorien, die sich in den GER-Abschnitten 4 und 5 wiederfinden, vorgestellt (ebd.: 123-129).[194] Grob lassen sich die Bereiche *kommunikative Sprachaktivitäten*, *Strategien* und *qualitative Aspekte des Sprachvermögens* unterscheiden:

[191] Vgl. van Ek & Trim 1990, 1991 und 1997.
[192] Vgl. GER (2001: 42ff).
[193] Dieses „Schweizer Forschungsprojekt" (GER: 210) scheint keinen eigenen Namen zu besitzen. Für detaillierte Projektdarstellungen vgl. North & Schneider 1998, North 2000, und Schneider & North 2000 (zur zweiten Phase des Projekts). Für einen Überblick vgl. GER (2001: 3 und 7 sowie Anhang B).
[194] Wie in Kapitel 1.2.5.3 dieser Arbeit bereits dargestellt, sind in GER-Abschnitt 4 Unterkategorien zu *kommunikativen Aktivitäten* und *Strategien* beschrieben, während GER-Abschnitt 5 sich den denjenigen der *kommunikativen Sprachkompetenzen* zuwendet.

Communicative Activities	**Strategies**	**Qualitative Aspects of Proficiency**
Reception Interaction Production	Reception Interaction Production	Pragmatic Linguistic socio-linguistic
(correspond to Bachmann's *real life approach* and Alderson's *constructor-oriented classification*)		(correspond to Bachmann's *interactive-ability approach* and Alderson's *assessor-oriented classification*)
	Socio-cultural Knowledge	

Abb. 11: Kategorien der *proficiency*, nach North & Schneider (1998: 227).

Nicht zu allen im GER aufgeführten (Sub-)Kategorien sind Skalen konstruiert worden.[195] Die folgende Übersicht (GER 2001: 214f) zeigt diejenigen Kategorien, zu denen je eine Beispielskala entwickelt wurde:

Dokument B1 – Beispielskalen in Kapitel 4: Kommunikative Aktivitäten

REZEPTION	**Mündlich**	- Hörverstehen allgemein	Seite 71
		- Gespräche zwischen Muttersprachlern verstehen	Seite 72
		- Als Zuschauer/Zuhörer im Publikum verstehen	Seite 72
		- Ankündigungen, Durchsagen und Anweisungen verstehen	Seite 73
		- Radiosendungen und Tonaufnahmen verstehen	Seite 73
	Audiovisuell	- Fernsehsendungen und Filme verstehen	Seite 77
	Schriftlich	- Leseverstehen allgemein	Seite 74
		- Korrespondenz lesen und verstehen	Seite 75
		- Zur Orientierung lesen	Seite 75
		- Information und Argumentation verstehen	Seite 76
		- Schriftliche Anweisungen verstehen	Seite 76
INTERAKTION	**Mündlich**	- Mündliche Interaktion allgemein	Seite 79
		- Muttersprachliche Gesprächspartner verstehen	Seite 80
		- Konversation	Seite 80
		- Informelle Diskussion	Seite 81
		- Formelle Diskussion und Besprechungen	Seite 82
		- Zielorientierte Kooperation	Seite 83
		- Transaktionen: Dienstleistungsgespräche	Seite 83
		- Informationsaustausch	Seite 84
		- Interviewgespräche	Seite 85
	Schriftlich	- Schriftliche Interaktion allgemein	Seite 86
		- Korrespondenz	Seite 86
		- Notizen, Mitteilungen und Formulare	Seite 87
PRODUKTION	**Mündlich**	- Mündliche Produktion allgemein	Seite 64
		- Zusammenhängendes monologisches Sprechen: Erfahrungen beschreiben	Seite 64
		- Zusammenhängendes monologisches Sprechen: Argumentieren (z. B. in einer Diskussion)	Seite 65
		- Öffentliche Ankündigungen/Durchsagen machen	Seite 65
		- Vor Publikum sprechen	Seite 66
	Schriftlich	- Schriftliche Produktion allgemein	Seite 67
		- Kreatives Schreiben	Seite 67
		- Berichte und Aufsätze schreiben	Seite 68

[195] Zu Gründen hierfür vgl. North (2000: 125), North & Schneider (1998: 235) und Schneider & North (2000: 25-37). Beispielsweise erwiesen sich einige Kategorien als nicht skalierbar, andere Dimensionen waren nicht Gegenstand der Untersuchung; vgl. auch unten die Ausführungen zu *Skalierung der Deskriptoren*.

Dokument B2 – Beispielskalen in Kapitel 4: Kommunikative Strategien

REZEPTION		
	- Hinweise identifizieren und erschließen	Seite 78
INTERAKTION		
	- Sprecherwechsel	Seite 88
	- Kooperieren	Seite 89
	- Um Klärung bitten	Seite 89
PRODUKTION		
	- Planen	Seite 70
	- Kompensieren	Seite 70
	- Kontrolle und Reparaturen	Seite 70

Dokument B3 – Beispielskalen in Kapitel 4: Textarbeit

TEXT		
	- Notizen machen (in Vorträgen, Seminaren etc.)	Seite 98
	- Texte verarbeiten	Seite 98

Dokument B4 – Beispielskalen in Kapitel 5: Kommunikative Sprachkompetenz

LINGUISTISCH		
Spektrum:	- Spektrum sprachlicher Mittel (allgemein)	Seite 110
	- Wortschatzspektrum	Seite 112
Beherrschung:	- Wortschatzbeherrschung	Seite 113
	- Grammatische Korrektheit	Seite 114
	- Beherrschung der Aussprache und Intonation	Seite 117
	- Beherrschung der Orthographie	Seite 118
SOZIOLINGUISTISCH		
	- Soziolinguistische Angemessenheit	Seite 121
PRAGMATISCH		
	- Flexibilität	Seite 124
	- Sprecherwechsel	Seite 124
	- Themenentwicklung	Seite 125
	- Kohärenz und Kohäsion	Seite 125
	- Flüssigkeit (mündlich)	Seite 129
	- Genauigkeit	Seite 129

Abb. 12: Beispielskalen im GER (vgl. GER 2001: 214f).

Erst bei Betrachtung der Hintergrundliteratur zu Vorarbeiten und zum Skalierungsprojekt wird der umfassende Begriff der *proficiency*, der dem Kategoriensystem des GER in seinem Beurteilungsansatz zugrunde liegt, deutlich. Hätten die Autoren des GER die Quellen und Vorarbeiten, die hierzu stattfanden, in einem eigenen Abschnitt etwa zum *Begriff des Sprachvermögens in der Beurteilung* offen gelegt, hätte dies viel zur postulierten Transparenz des Dokuments beigetragen.

3.4.1.2 Ursprung der Deskriptoren

Die Deskriptoren, die die oben genannten Kategorien versprachlichen, sind wie erwähnt nicht alle neu entwickelt worden. Vielmehr griff man auf 41 schon existierende Skalen der *proficiency* zurück (North & Schneider 1998: 224), die auf ihre Kategorien und ihren Beschreibungsgegenstand

hin analysiert und in ihre Deskriptoren zerlegt wurden.[196] Dokument B6 (GER 2001: 217) gibt eine Übersicht über die Skalen, die als Quellen benutzt wurden:

Holistische Skalen der Mündlichen Kompetenz allgemein
- Hofmann: Levels of Competence in Oral Communication 1974.
- University of London School Examination Board: Certificate of Attainment – Graded Tests 1987.
- Ontario ESL Oral Interaction Assessment Bands 1990.
- Finnish Nine Level Scale of Language Proficiency 1993.
- European Certificate of Attainment in Modern Languages 1993.

Skalen für verschiedene kommunikative Aktivitäten
- Trim: Possible Scale for a Unit/Credit Scheme: Social Skills 1978.
- North: European Language Portfolio Mock-up: Interaction Scales 1991.
- Eurocentres/ELTDU Scale of Business English 1991.
- Association of Language Testers in Europe, Bulletin 3, 1994.

Skalen für die vier Fertigkeiten
- Foreign Service Institute Absolute Proficiency Ratings 1975.
- Wilkins: Proposals for Level Definitions for a Unit/Credit Scheme: Speaking 1978.
- Australian Second Language Proficiency Ratings 1982.
- American Council on the Teaching of Foreign Languages Proficiency Guidelines 1986.
- Elviri et al.: Oral Expression 1986 (in Van Ek 1986).
- Interagency Language Roundtable Language Skill Level Descriptors 1991.
- English Speaking Union (ESU) Framework Project: 1989.
- Australian Migrant Education Program Scale (Listening only).

Bewertungsskalen für mündliche Prüfungen
- Dade County ESL Functional Levels 1978.
- Hebrew Oral Proficiency Rating Grid 1981.
- Carroll B.J. and Hall P.J Interview Scale 1985.
- Carroll B.J. Oral Interaction Assessment Scale 1980.
- International English Testing System (IELTS): Band Descriptors for the Speaking & Writing 1990.
- Goteborgs Univeritet: Oral Assessment Criteria.
- Fulcher: The Fluency Rating Scale 1993.

Rahmenpläne für curriculare Inhalte und Beurteilungskriterien für pädagogische Lernerfolgsstufen bzw. Abschlussprüfungen
- University of Cambridge/Royal Society of Arts Certificates in Communicative Skills in English 1990.
- Royal Society of Arts Modern Languages Examinations: French 1989.
- English National Curriculum: Modern Languages 1991.
- Netherlands New Examinations Programme 1992.
- Eurocentres Scale of Language Proficiency 1993.
- British Languages Lead Body: National Language Standards 1993.

Abb. 13: Quellskalen des GER-Systems (vgl. GER 2001: 217).

Die Deskriptoren wurden demnach aus ganz unterschiedlichen Quellen gewonnen: Wie North (2000:13ff) darstellt, sind die meisten existenten *proficiency scales* eigentlich ***rating scales***, die in der Regel beobachtbares (oder tatsächlich beobachtetes) Verhalten in einem bestimmten Test beschreiben. Es wurden auch Skalen verwendet, die **Prüfungsniveaus** spezifizieren sollen. Dieser Typus Skalen beschreibt entweder die Erwartungen an Inhalt und Performanz für jedes Niveau einer gegebenen Prüfung, teils basierend auf realen Erfahrungen. Oder die Prüfungsniveaus werden durch eine Abfolge verschiedener abgestufter Prüfungen als Skala

[196] Das Vorgehen im Einzelnen wird in Schneider & North (2000: 39-48) beschrieben.

präsentiert, wie es beispielsweise bei den *Cambridge-Exams* oder den Niveaus der ALTE geschieht (vgl. North 2000: 15f). Ein dritter Typus der dem GER zugrunde liegenden Skalen beschreibt nach North (ebd.: 16f) *stages of attainment*, **Fertigkeitsniveaus** also. Dazu werden Fertigkeiten, Charakteristika der Sprachproduktion und der sprachlichen Handlungen auf verschiedenen Stufen beschrieben, um in einem gegebenen Bildungssystem einen Rahmen zu schaffen, in dem Ziele, Beurteilung und Zertifizierung verortet werden können. Solche Skalen sind performanzbezogen, können sowohl Prozesse als auch Produkte beschreiben und geben teils detaillierte inhaltliche Spezifizierungen.

Wichtig zu bedenken ist dabei, dass nicht alle Skalen, die diesen drei Typen zugeordnet werden können, zugleich auch *scales of proficiency* darstellen müssen: Um diesen Status jedoch bestimmen zu können, müsste nachvollziehbar sein, wie man ursprünglich zu den Deskriptoren der Quellskalen gekommen ist, ob dort beobachtetes oder beobachtbares Verhalten beschrieben wurde, Performanzen analysiert wurden, Erwartungen oder Annahmen formuliert wurden und in welchen Theorien die Deskriptoren verankert wurden. Dies ist allerdings nicht der Fall. Eine Analyse der Grundlagen der Beschreibung der 41 Quellskalen würde den Rahmen der vorliegenden Arbeit sprengen, doch letztlich müssen die Skalen des GER auch ohne Kenntnis ihres Ursprungs tragfähig und einsetzbar sein – zu Zwecken, die dem Beschreibungsgegenstand und den Annahmen, die hinter den Niveaus stehen, auch gerecht werden. Deshalb soll unter Kapitel 3.4.3 dieser Arbeit am Beispiel ausgewählter GER-Skalen aufgezeigt werden, welche Folgen der unklare Status der GER-Deskriptoren hat.

3.4.1.3 Skalierung der Deskriptoren[197]

Wie ist man zu den Niveaus des GER gekommen? Welche Methoden[198] wurden dabei eingesetzt?

Die Skalen des GER wurden auf Basis einer Kombination aller drei in Kapitel 3.2.2 dieser Arbeit vorgestellten Möglichkeiten konstruiert. (GER 2001: 202) Wie gerade erläutert, dienten dabei schon existente Deskriptoren als Ausgangspunkt. Die Möglichkeit, Stichproben von Leistungen als Basis zu nehmen, wurde laut GER nicht gewählt, da diese „…nur benutzt werden kann, um Deskriptoren zur Bewertung von Leistungen zu entwickeln“ (ebd.) – und das war nicht Anliegen der GER-Skalenkonstruktion.

Zunächst wurden die genannten Quellskalen analysiert, deren Deskriptoren in einzelne Sätze zerlegt und dekontextualisiert. Die relevantesten und tragfähigsten Deskriptoren wurden

[197] Für eine detaillierte Darstellung vgl. North & Schneider 1998, North 2000 und GER (2001: Anhang B). Hier wird lediglich ein knapper Aufriss gegeben, soweit es für die vorliegende Arbeit relevant erscheint.

[198] In Kapitel 3.2.2 dieser Arbeit wurden die drei grundsätzlichen Methoden der Skalenkonstruktion – intuitive, qualitative und quantitative – beschrieben. Es darf an dieser Stelle noch einmal auf den allgemeinen Überblick über insgesamt zwölf Methoden der Skalenentwicklung verwiesen werden, der im GER im Anhang A (2001: 202-205) gegeben wird, wobei lediglich bei zwei Methoden (Nr. 9 und 12) der Hinweis zu finden ist, dass sie bei der Entwicklung des Gemeinsamen Referenzsystems zum Einsatz kamen. Erst im darauf folgenden Anhang B des GER werden die bei der GER-Skalenkonstruktion eingesetzten Methoden genauer beschrieben. Detailliertere Darstellungen der Entwicklungsmethoden sind wiederum in North & Schneider 1998, North 2000 und Schneider & North 2000 (dort insbesondere in den Kapiteln 5 und 7) gegeben.

ausgewählt, ergänzt um einige neu verfasste (vor allem im Hinblick auf Strategien und interaktives Hörverstehen), und vorläufig den oben dargestellten Kategorien zugeordnet. Die dabei auftretenden Schwierigkeiten betrafen insbesondere die Zuweisung von Performanzdeskriptoren zu Kompetenzkategorien. Deshalb musste nach North & Schneider (1998: 227) "…a broad view of pragmatic competence (…) and a broad view of fluency…" gewählt werden. Wiederholungen, negative Formulierungen und normorientierte Deskriptoren wurden fallengelassen. Diese Vorgehensweisen sind den intuitiven Methoden zuzurechnen (vgl. GER: 210).

Der so entstandene *Pool* von etwa 1000 Deskriptorenentwürfen wurde in Workshops von Lehrkräften nach Kategorien und Niveaus sortiert, die Entwürfe wurden kommentiert, ergänzt oder verworfen, und die klar formulierten, nützlichen und relevanten Deskriptoren wurden identifiziert. Damit sollte zum einen überprüft werden, ob die Kategorien adäquat versprachlicht wurden respektive ob die Deskriptoren alle relevanten Kategorien reflektierten, und zum anderen sollten die Deskriptoren den vorläufigen Kompetenzniveaus zugeteilt werden. Dieses Vorgehen ist einerseits den erwähnten intuitiven Methoden zuzuordnen, als hierbei die Erfahrungen und die impliziten Annahmen – also auch die Intuition – der beteiligten Lehrkräfte herangezogen wurden. Die Methode „Nr. 3 *Auf Erfahrung beruhend*" (GER 2001: 203) dürfte hier die zutreffendste sein, obwohl dies im GER nicht explizit ausgesagt wird. Andererseits ist diese Überarbeitung den qualitativen Methoden zuzurechnen: Der GER erwähnt explizit, dass Methode „Nr. 9 *Sortieraufgaben*" bei der Konstruktion zum Einsatz kam (vgl. ebd.: 204 resp. 211).

Zur Überprüfung des Kategoriensystems wurde die Metasprache der Lehrenden mit einbezogen: Dazu wurde die in North & Schneider (1998: 228) dargestellte Technik benutzt, Lehrkräfte Videoaufzeichnungen von Gesprächen unter Lernenden kommentieren und diskutieren zu lassen im Hinblick auf qualitative Aspekte des Sprachgebrauchs. Die dabei elizitierte und aufgezeichnete Metasprache sollte helfen, die Kategorien des Systems und die Sprache der Deskriptoren qualitativ zu bestätigen. Insofern ist die Analyse der Aufzeichnungen der Lehrerdiskussionen den qualitativen Methoden zuzuordnen (vgl. auch GER 2001: 211).

Die auf diese Weise kategorisierten und reformulierten Deskriptoren wurden vertikal abgestuft, indem sie in Fragebogenform von etwa 300 Lehrenden zur Einschätzung ihrer circa 2800 Lernenden genutzt wurden. Durch diese auf Erfahrung beruhende – und damit den intuitiven Methoden zuzurechnende – Einschätzung erhielt man Schwierigkeitswerte der Deskriptoren, die mithilfe einer quantitativen Methode psychometrisch skaliert werden konnten. Diese Methode wird im GER als Methode „Nr. 12 *Item-Response-Theorie (IRT) oder ‚Latent Trait'-Analyse*" (GER 2001: 204) bezeichnet: Dabei kommt das so genannte Rasch-Modell zum Einsatz. Da die im GER (ebd.: 204f) hierzu gegebene Erklärung leicht verständlich ist, sei sie an dieser Stelle zitiert:

> IRT stellt eine Gruppe von Mess- oder Skalierungsmodellen zur Verfügung. Das direkteste und stabilste ist das *Rasch*-Modell, benannt nach dem dänischen Mathematiker Georg Rasch. Die IRT ist eine Weiterentwicklung, basierend auf der Probabilitätstheorie, und wird vor allem dazu benutzt, den Schwierigkeitsgrad einzelner Testaufgaben in einer Itembank zu bestimmten. Fortgeschrittene Lernende haben hohe Chancen, eine elementare Frage richtig zu beantworten, Anfänger haben sehr

geringe Chancen, eine anspruchsvolle Aufgabe zu lösen. Diese einfache Tatsache ist beim Rasch-Modell zu einer Skalierungsmethode entwickelt worden, die man benutzen kann, um Items auf der gleichen Skala zu kalibrieren. Eine Weiterentwicklung dieses Ansatzes kann sowohl zur Skalierung von Deskriptoren der Kommunikationsfähigkeit als auch zur Skalierung von Testitems benutzt werden. Bei einer Rasch-Analyse können verschiedenen Tests oder Fragebögen zu einer überlappenden Kette zusammengefügt werden, indem man 'Ankeritems' benutzt, die den aneinander grenzenden Elementen gemeinsam sind. Im folgenden Diagramm sind die Ankeritems grau schattiert. Auf diese Weise können die Test- oder Fragebögen auf bestimmte Gruppen abgestimmt werden, sie bleiben aber mit eine gemeinsamen Skala verknüpft. Allerdings muss man bei diesem Prozess sehr sorgfältig vorgehen, weil das Rasch-Modell die jeweils besten und niedrigsten Ergebnisse bei jedem Test verzerrt.

Der Vorteil einer Rasch-Analyse ist, dass sie ein stichproben- und skalenunabhängiges Maß liefern kann, d. h. eine Skalierung, die unabhängig ist von den Stichproben und den Tests/Fragebögen, die bei der Analyse benutzt wurden. Sie liefert Skalenwerte, die bei zukünftigen Gruppen konstant bleiben, vorausgesetzt, die zukünftigen Probanden können als neue Gruppen innerhalb der gleichen statistischen Population gelten. Systematische Veränderungen in den Werten im Verlauf der Zeit (z. B. aufgrund curricularer Veränderungen oder von Prüfertraining) können quantifiziert und in Anpassungen berücksichtigt werden. Ebenso kann systematische Variation zwischen Lernertypen bzw. Typen von Beurteilenden quantifiziert und ausgeglichen werden (Wright & Masters 1982; Linacre 1989).

Deskriptoren, die technische Probleme bei der Skalierung zeigten, wurden fallengelassen.[199] Dies betraf nach North & Schneider (1998: 28-231) vor allem solche Deskriptoren, die Strategien, sozio-kulturelle Kompetenz, die Berufswelt und negative Konzepte beschrieben. Deskriptoren, welche die Schreibfertigkeit zum Gegenstand haben, wurden nicht empirisch kalibriert, sondern laut GER (2001: Anmerkung S.67) „ … durch eine Kombination von Elementen aus anderen Skalen erstellt." Auch die Skala zu Orthographie wurde nicht empirisch skaliert (vgl. GER 2001: 118). Lediglich auf S. 212 des GER findet sich der Hinweis, dass sich Deskriptoren zu „Aktivitäten außerhalb des Klassenraums", wie etwa Briefe oder Aufsätze schreiben, als nicht skalierbar erwiesen. Aus welchen Gründen sie nicht skalierbar waren oder weshalb diese Aktivitäten nicht auch im Klasszimmer stattfinden sollten, bleibt unklar. Auf derselben Seite steht bezüglich der Schreibfertigkeit lediglich Folgendes: „Schreiben stand nicht im Mittelpunkt der Untersuchung. Die Deskriptoren für das Schreiben, die sich in Kapitel 4 finden, wurden vor allem aus denen für die mündliche Produktion entwickelt." Im GER selbst wird keine Begründung oder Erklärung für dieses Vorgehen gegeben – wiederum kein Beitrag zur Transparenz.

Die auf Basis der Rasch-Skalierung erhaltene Skala, auf der die Deskriptoren entsprechend ihrer durch die Skalierung zugewiesenen Schwierigkeitswerte liegen, wurde zunächst, wie in North & Schneider (1998: 231ff) dargestellt, aufgrund messtheoretischer Überlegungen in zehn gleich breite Kompetenzniveaus eingeteilt. Diese Niveaus wurden dann auf Logik der Versprachlichung und Progression hin überprüft, und daraufhin, ob sich qualitative Unterschiede zwischen den einzelnen Niveaus zeigten. Zuletzt wurden die zehn Niveaus auf die sechs Niveaus des

[199] Für eine Erörterung dieser technischen Probleme vgl. North 1995.

Europarats hin adaptiert[200], die auch dem GER zugrunde gelegt wurden: *Breakthrough*, *Waystage*, *Threshold*, *Vantage*, *Effective Operational Proficiency* und *Mastery* (vgl. GER 2001: 33f und 42ff). Im GER wird folgendes Verzweigungsmodell vorgestellt:

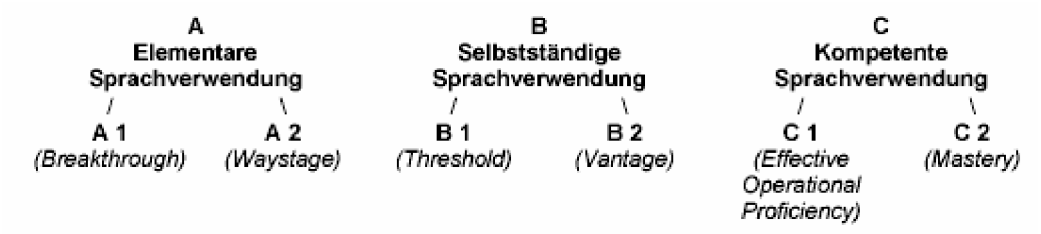


Abb. 14: Die Referenzniveaus des GER (GER 2001: 34)

Auch bei der vertikalen Kalibrierung der Deskriptoren muss bedacht werden, dass implizite Annahmen der Lehrkräfte über Lernfortschritte und Fertigkeitsniveaus ihrer Lernenden, und damit auch implizite Annahmen über Progression und Spracherwerb, mit in die Skalierung einflossen. Gerade diese impliziten Annahmen die horizontale wie vertikale Einteilung betreffend bleiben jedoch im Dunkeln und werden im GER nirgends thematisiert. Es lässt sich auch kein theoretisches Modell – sei es nun beispielsweise hinsichtlich des Spracherwerbs oder der Lernfortschritte generell, oder bezogen etwa auf die Entwicklung bestimmter Teilfertigkeiten – ausmachen, auf dem die vertikale Kalibrierung fußt.

3.4.1.4 Validierung des Skalenkonstrukts[201]

Aufgrund der gerade ausgeführten Unklarheiten stellt sich die berechtigte Frage, inwieweit die Skalen des GER empirisch validiert sind. Wo im Konstruktionsprozess ist eine Anbindung an empirisch beobachtetes Verhalten, an empirisch beschriebene Performanz oder empirisch belegte Charakteristika des Sprachvermögens auszumachen? Welchen Beitrag zur Validierung der Skalen haben die eingesetzten Konstruktionsmethoden geleistet?

Die intuitiven und qualitativen Methoden beziehen sich zunächst darauf, das Kategoriensystem, die horizontalen Dimensionen des Konstrukts *proficiency* zu validieren. Wie oben erläutert, sind die Dimensionen konstruiert worden auf Basis von theoretischen Modellen, von Forschungen im Bereich der *proficiency* und von existierenden Einteilungen der Quellskalen. Wo immer es keine Basis gab, hat man sich auf pragmatische Entscheidungen verlassen. Das Konstrukt der Dimensionen des Skalensystems im GER, das demnach nicht auf empirischer Basis steht, wird dann von Lehrkräften, erfahrenen Praktikern also, intuitiv wie qualitativ validiert: Die Kategorien, die das Konstrukt der *proficiency* operationalisieren, werden auf Nützlichkeit und Sinn überprüft, basierend auf den Erfahrungen und impliziten Vorstellungen der Lehrkräfte – also wiederum nicht auf expliziter empirischer Basis. Die Sprache der Deskriptoren wird an der

[200] Für eine detaillierte Darstellung vgl. North 2000 (insb. Kapitel 6 und 7) und Schneider & North 2000 (insb. Kapitel 8.4).
[201] Vgl. hierzu auch die Ausführungen in North & Schneider (1998: 228ff) und North (2000: 65-74).

Metasprache der Lehrer validiert. Die Basis des Beschreibungsgegenstands der einzelnen Deskriptoren wird nicht eigens validiert: Lediglich über die Zuordnungen der Deskriptoren zu den betreffenden Kategorien wird überprüft, ob die Deskriptoren die Kategorien reflektieren – was diesem Gegenstand jedoch zugrunde liegt, was die Deskriptoren tatsächlich beschreiben (seien es nun Performanzbeispiele oder Erwartungen an Performanzen, Verhaltensbeschreibungen aufgrund von Beobachtungen, von Stichproben von Leistungen oder aufgrund von Erwartungen, oder aber Generalisierungen auf zugrunde liegende Kompetenzen), bleibt im Dunkeln.[202] Die vertikalen Abstufungen, die ja schon in den aus existenten Skalen stammenden Deskriptoren angedeutet sind, werden durch die Anwendung der Deskriptoren in der realen Einschätzung der den Lehrkräften bekannten Lernenden überprüft. Auch hierbei spielen wie gesagt implizite Vorstellungen eine Rolle, doch zumindest kommt die Realität in den Validierungsprozess mit herein, welcher zusätzlich durch die quantitative Skalierung mittels des Rasch-Modells gestützt wird.

Kategorien und Abstufungen des GER-Skalensystems können demnach nicht uneingeschränkt als empirisch validiert betrachtet werden. So findet sich bei H. Christ (2003: 62) die Forderung nach Validierung des Skalensystems in der Praxis, denn auch eine Skalierung durch Experten sei kein Validitätsbeweis. Königs (2003) hält es zumindest für fraglich, ob eine empirische Skalierung der auf vielen „Ebenen angesiedelten Deskriptoren … für die wissenschaftliche Absicherung“ des GER genügt, denn die „Grundannahmen des Referenzrahmens“ seien damit keineswegs abgesichert, „sondern allenfalls das methodologische Instrumentarium, dessen sich der GER zur Durch- bzw. Umsetzung dieser Grundlagen bedient.“ (ebd.: 118). Darüber hinaus beschreibt Quetz (2003) das Phänomen des „auseinanderfallenden Mittelbereichs“ bei Sortieraufgaben: Die Zuordnungen von entsprechenden Deskriptoren auf die untersten respektive obersten Niveaus einer Skala erfolgt in der Regel mit größerer Übereinstimmung als die Zuordnungen im Mittelbereich der Skala. Die dort entstehenden Diskrepanzen lassen sich nicht durch eine psychometrische Skalierung auflösen (ebd.: 153). Eine Möglichkeit ist dann natürlich, die entsprechenden Deskriptoren einfach fallen zu lassen, doch ob dies zur Validität der betreffenden Skala beitragen kann, sei dahingestellt. Quetz stellt ferner fest (ebd.: 153f), dass Zuordnungen von sprachlichen Mitteln auf bestimmte Niveaus, wie es etwa von Curriculumplanern oder Testentwicklern benötigt wird, im momentanen GER-System nicht valide erfolgen können.

North (2000) bezeichnet denn auch die Reichweite des *common framework* wie folgt: “(…) common, that is, to the raters who produced it, and to raters like them, in relation to the learners who were assessed, and to learners like them.” (ebd.: 73) Also bezieht sich das Gemeinsame des *Gemeinsamen Referenzrahmens* auf die Erfahrungen, Vorstellungen und Annahmen einer Lehrerschaft vergleichbar der in den oben genannten Workshops, und auf eine Lernerschaft vergleichbar derer, welche die Lehrenden der Workshops unterrichteten. Deshalb wird an dieser

[202] Deshalb werden wie gesagt unter Kapitel 3.4.3 dieser Arbeit einige ausgesuchte Skalen des GER untersucht, um aufzuzeigen, welche Auswirkungen der unklare Status der Deskriptoren auf die Verwendbarkeit der Skalen hat.

Stelle eine Kurzübersicht (vgl. GER 2001: 210) gegeben, aus welchen Ländern und Bildungsbereichen diese Bezugsgruppen stammen:

> Es waren Lehrende aus den deutsch-, französisch-, italienisch- und romanisch-sprachigen Regionen der Schweiz involviert, wobei allerdings die Zahlen aus den italienisch- und romanisch-sprachigen Regionen sehr begrenzt waren. In beiden Untersuchungsjahren unterrichteten ungefähr ein Viertel der beteiligten Lehrenden ihre Muttersprache. […]
>
> Die Lernenden waren wie folgt auf die Sekundarstufen I und II, die Berufsbildung und die Erwachsenenbildung verteilt:

	Sekundarstufe I	Sekundarstufe II	Berufsbildung	Erwachsenenbildung
1994	35%	19%	15%	31%
1995	24%	31%	17%	28%

Tabelle 4: Verteilung der Lernenden (GER 2001: 210)

Wenn also Experten aufgrund ihrer Erfahrungen und impliziten Annahmen „im Erziehungssektor der Schweiz“ (GER 2001: 211) Kategorien und Abstufungen überprüfen – stellt dies dann eine qualitative und quantitative empirische Validierung des Konstrukts *Scales of Proficiency* dar? Es könnte sich vielmehr um eine „interne“ Validierung – intern bezogen auf das, was North “common“ nennt, also bezogen auf eine Lehrer- und Schülerschaft, die der bei der Validierung beteiligten vergleichbar ist – handeln, bei der die reale Welt gar nicht oder nur implizit (etwa über die Erfahrungen der Lehrenden) mit hereinspielt: Wissenschaftliche Experten entwickeln Modelle, die der Skalenkonstruktion zugrunde gelegt werden; existente Skalen werden wieder in Einzeldeskriptoren zerlegt; Experten aus der Praxis diskutieren und kommentieren die Kategorien und ordnen die Deskriptoren bestimmten Niveaus zu – wo wird dieses System an der Realität überprüft? Es könnte sich denn auch um einen Zirkelschluss handeln, wenn der Konstruktionsprozess der Skalen selbst zugleich als Validierungsmethode betrachtet wird, werden doch dabei lediglich Annahmen unter Experten gegenseitig validiert – insofern muss sich das Skalensystem des GER erst in der Praxis bewähren und durch seine Verwendung validiert werden. Ein Beispiel für solch einen empirischen Validierungsversuch wird, wie gesagt, in Kapitel 4 dieser Arbeit gegeben: die Entwicklung von *rating scales* im DESI-Projekt, welche ihren Ausgangspunkt in der abgestuften Beschreibung von Charakteristika realer Schüleraufsätze haben und erst nach Entwicklung tragfähiger Deskriptoren mit relevanten Skalen des GER abgeglichen wurden, um zu sehen inwieweit sich die empirisch gefundenen Merkmale in Kategorien und Niveaus des GER wiederfinden lassen.

3.4.2 Selbstverständnis des GER bezüglich seines Skalenansatzes

Wie oben ausgeführt ist die Grundlagenvalidierung der Deskriptoren der GER-Skalen nicht erfolgt, da die Deskriptoren von bereits bestehenden Skalen abgeleitet wurden, die ihrerseits

verschiedene Grundlagen beschreiben. Es ist somit nicht möglich, die Basis der Beschreibungen der GER-Skalen – und damit ihre Verwendungszwecke – genau zu definieren. Deshalb lohnt ein näherer Blick auf das „Selbstverständnis" des Skalenansatzes im GER, auf den Status, den die Konstrukteure den Skalen zuschreiben und auf die Verwendungszwecke, die im GER angegeben sind. Diese werden abschließend nach der erwähnten exemplarischen Skalenanalyse in Kapitel 3.4.4 dieser Arbeit beurteilt.

Das Referenzsystem des GER will, wie gesagt, alle relevanten Aspekte des kommunikativen Sprachvermögens, der *proficiency*, abdecken. Wie oben in den Kapiteln 1.2.5.3, 2.5.2 und 3.4.1 dieser Arbeit erläutert, fallen darunter verschiedenste Aspekte, die je andere Grundlagen haben und je anders beschrieben werden müssen. Man kann sich dem Sprachvermögen auf verschiedenen Ebenen nähern und je nach Ebene verschiedene Facetten beleuchten; diese Betrachtungsweise findet sich beispielsweise auch in Bachmanns Modell der kommunikativen Kompetenz. Für diese systemische Ansicht bietet der GER ein mehrschichtiges „Pyramiden"-Modell an (vgl. GER 2001: 48): An der Spitze steht eine sehr einfach gehaltene Sichtweise auf das generelle Sprachvermögen, im GER dargestellt durch die eher grobkörnige Globalskala auf S.35 des GER. Dieser holistische Blick wird in immer mehr Einzeldimensionen aufgefächert, deren Beschreibung zunehmend komplexer und detaillierter wird, je weiter man sich im System nach unten bewegt. Beispielskalen in unterschiedlicher Detailliertheit zu verschiedenen Aspekten auf den einzelnen Ebenen finden sich wie gesagt in den GER-Abschnitten 4 und 5.

Folgende Abbildung (nach de Jong 2004: 21f) illustriert das Modell der Dimensionen im GER:

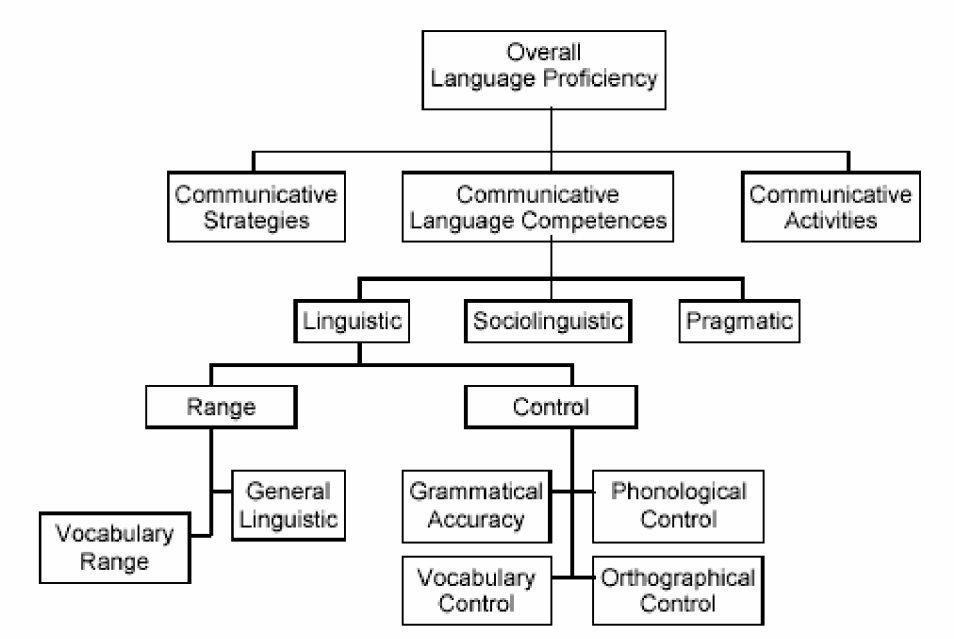


Abb. 15: Modell der Dimensionen im GER

Zur Betrachtung der Niveaus des Skalensystems schlägt der GER (ebd.: 28f) ein konisches Modell der Zusammenhänge vor: Je weiter nach oben man sich in einer Skala bewegt, desto

breiter werden die Niveaus, da das Sprachvermögen nicht nur vertikal-qualitativ im Sinne des Vertiefens eines Bereichs zunimmt, sondern auch horizontal-quantitativ i. S. des Verbreiterns der Könnensbereiche anwächst. Die Niveaus der Skalen des GER wollen deshalb nicht als „lineare Mess-Skala“ (ebd.: 29) verstanden werden: Sprachlernen ist kein linearer Prozess, sondern verläuft hoch individuell. Daher unterliegt jeder Versuch, Kompetenzen, Wissen oder Fertigkeiten bestimmten Niveaus zuzuweisen, einer gewissen Willkürlichkeit (vgl. ebd.: 28). Diese Aussage scheint in gewissem Widerspruch zu den Behauptungen auf S. 39 des GER zu stehen, namentlich dass die Deskriptoren auf Erfahrungen vieler im Bereich der Definition von Kompetenzniveaus bewanderter Institutionen beruhten und „objektiv kalibriert“ seien. Die GER-Skalen reflektieren ein gemeinsames Verständnis der an der Konstruktion Beteiligten und beschreiben die Realität – wie alle Modelle – in simplifizierender Weise. Man muss sich der Grenzen dieser Modelle bewusst sein, um sie angemessen einzusetzen.

Das Skalensystem des GER hat Rahmencharakter: Das Niveausystem wird wie gesagt als „flexibles Verzweigungssystem“ vorgestellt, das ebenso wie die Dimensionen je nach Bedarf verfeinert und differenziert werden kann (vgl. ebd.: 40ff). Die Referenzpunkte sind über den „Wortlaut der Deskriptoren“ (ebd.: 34) gegeben: Sie wollen helfen, externe Systeme auf die Niveaus oder die Kategorien des Referenzsystems zu beziehen. Solch einem Referenzrahmen könnte man einen *Werkbank*-Charakter zuschreiben, denn er kann und soll nach Meinung seiner Autoren um die Erfahrung der Institutionen erweitert und verfeinert werden, in denen er benutzt wird, jeweils flexibel auf die jeweiligen Verwendungskontexte hin adaptiert (vgl. ebd.: 34 und 39). Der Referenzrahmen stellt „kriterienbezogene Aussagen zum Kontinuum der fremdsprachlichen Kompetenz“ zur Verfügung; er will dabei jedoch „holistisch bleiben, um einen Überblick zu ermöglichen“ (beide Zitate: ebd.: 39).

Wie nun stellt der GER die Verwendungsmöglichkeiten seiner Skalen dar? Im GER lassen sich die Hinweise auf Verwendungsmöglichkeiten grob in drei Kategorien einteilen: (a) Allgemeine Hinweise, die sich an verschiedenen Stellen des GER finden lassen; diese Hinweise werden in der vorliegenden Arbeit gleich im Anschluss diskutiert. (b) Hinweise, konkret auf die jeweiligen Skalen der Abschnitte 3, 4 und 5 bezogen und auch jeweils dort zu finden; diese Hinweise werden in Kapitel 3.4.3 der vorliegenden Arbeit bei den konkreten Skalenanalysen betrachtet. (c) Hinweise auf Verwendungsmöglichkeiten der Skalen bei der Beurteilung von Sprachvermögen, in Abschnitt 9.2 des GER[203] zu finden. Diese beurteilungsbezogenen Aussagen des GER-Abschnitts 9.2 wurden in dieser Arbeit bereits in Kapitel 2.5.4 vorgestellt; hier nun werden sie in Kapitel 3.4.4 wieder aufgenommen und im Anschluss an die erwähnte Analyse ausgewählter GER-Skalen untersucht, um ihre Einschätzung auf solide Basis stellen zu können.

[203] GER-Abschnitt 3.8 verweist auf GER-Abschnitt 9, welcher „beschreibt, wie man die Skala mit Gemeinsamen Referenzniveaus als Hilfsmittel bei der Beurteilung von Sprachkompetenz benutzen kann“ (ebd.: 46) – wiederum ein Indiz für die Beurteilungslastigkeit des Dokuments.

An dieser Stelle der vorliegenden Arbeit sollen diejenigen Aussagen untersucht werden, die sich allgemein auf das Selbstverständnis des GER bezüglich seiner Skalen und damit auch auf die Nutzungsmöglichkeiten der Skalen beziehen: In GER-Abschnitt 3.8, der sich konkret mit der Verwendung der GER-Skalen beschäftigt, wird der folgende, nicht zu unterschätzende Hinweis gegeben: „Sehr wichtige Fragen bei der Erörterung von Skalen der Sprachkompetenz sind jedoch (a) die genaue Identifikation des Zwecks, dem die Skala dienen soll, sowie (b) eine diesem Zweck angemessene Formulierung der Deskriptoren." (ebd.: 46). Anschließend werden im GER die drei Skalentypen *Skalen für Benutzer, für Beurteilende und für Testautoren* vorgestellt (vgl. zu den Zwecken, die diese Typen erfüllen können, auch die Ausführungen unter Kapitel 3.1 dieser Arbeit) und mit Beispielen aus externen Skalen belegt, die diesen Typen und den mit ihnen einhergehenden Funktionen und Zwecken zuzuordnen sind. Es finden sich im GER jedoch keinerlei Hinweise, welchem oder welchen dieser drei Typen welche GER-Skalen zuzurechnen sind. Somit wird – entgegen der gerade zitierten Forderung des GER – der Zweck, dem die Skalen „dienen" sollen, nicht „genau" identifiziert. Statt dessen wird im GER darauf hingewiesen, dass sich Probleme ergeben können, „wenn eine Skala, die für einen bestimmten Zweck konstruiert wurde, für einen anderen Zweck eingesetzt wird – es sei denn, die Formulierung ist nachweislich auch dafür angemessen" (ebd.: 46). Es finden sich jedoch keine Aussagen dazu, wie solch ein Nachweis auszusehen hätte. Im Anschluss an die Charakterisierung der drei Skalentypen wird auf S. 48 diese Übersicht gegeben:

Zusammenfassend kann man Skalen der Sprachkompetenz eine oder mehrere der folgenden Orientierungen zuweisen:

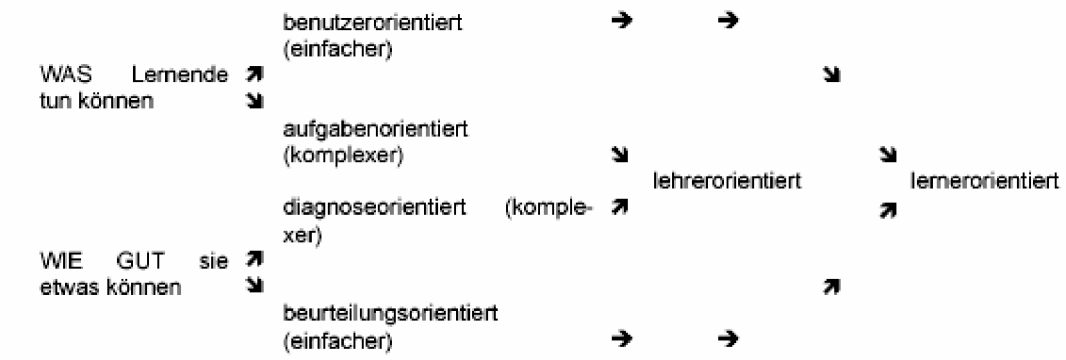


Abbildung 6

Alle diese Orientierungen sind für den *Gemeinsamen Referenzrahmen* relevant.

Abb. 16: Skalenorientierungen (GER 2001: 48)

Die Unterschrift zu dieser Übersicht gibt den einzigen Hinweis auf das Verhältnis der Referenzskalen zu diesen Typen.[204] Wenn jedoch alle diese Orientierungen für den GER „relevant" sind, bedeutet das dann, dass seine Skalen all die mit diesen Typen verbundenen Zwecke erfüllen

[204] Es darf auf das englische Original verwiesen werden, das sich an dieser Stelle wie folgt liest: "All those orientations can be considered relevant to a common framework." (CEF: 39). Im Original bezieht sich diese Aussage demnach nicht auf die Skalen des CEF – in der deutschen Übersetzung jedoch schon! Es darf nicht verwundern, wenn der Europäische Referenzrahmen in den verschiedenen Ländern aufgrund unterschiedlicher Übersetzungen zu verschiedenen Zwecken eingesetzt wird. Eine Überarbeitung zumindest der deutschen Version scheint dringend geboten.

(können)? Bei dieser Auslegung ist Vorsicht geboten: Beispielsweise ist es nicht möglich, dass Deskriptoren, die in generalisierender Form Kompetenzen beschreiben (seien sie nun daraufhin ausgerichtet, *was* Lernende tun können oder *wie gut* sie etwas können), zugleich beurteilungsorientiert sein können – denn eine beurteilungsorientierte Skala im Sinne einer *rating scale* beschreibt, wie unter Kapitel 3.1.2 dieser Arbeit ausgeführt, konkrete, empirisch identifizierte Merkmale, die im jeweiligen Testkonzept verankert sind, in dessen Rahmen die Skala eingesetzt werden soll. Man kann im Idealfall die *rating scale* in eine *reporting scale* überführen, die dann zur Vermittlung etwa von Testergebnissen eingesetzt wird[205], doch dürfen diese beiden Skalentypen und ihre Verwendungszwecke nicht miteinander verwechselt werden. Des Weiteren fällt an der obigen Übersicht auf, dass sich dort eine beurteilungsorientierte Skala nur auf den Aspekt *Wie gut etwas gekonnt wird* bezieht – obwohl die Beurteilung von Sprachvermögen die beiden Aspekte *Was getan werden kann* und *Wie gut etwas gekonnt wird* umfasst (dies wird im GER auch auf S.48 oben anerkannt). Ebenfalls auffällig an der obigen Übersicht ist die Unterscheidung hinsichtlich *Einfachheit* (bei den benutzer- und beurteilungsorientierten Skalen) versus *Komplexität* (bei den aufgaben- und diagnoseorientierten Skalen): Diese Unterscheidung greift zu kurz, denn wie im GER auf S. 46 festgestellt wird, muss neben dem Komplexitätsgrad einer Skala auch die Formulierung der Deskriptoren nachweislich für die jeweiligen Zwecke angemessen sein – diese Überlegung scheint jedoch in Abbildung 6 des GER keine Entsprechung zu finden.

Weitere allgemeine Aussagen zu Verwendungszwecken sind beispielsweise bei der Beschreibung der *Eigenschaften* der Deskriptoren der Abschnitte 4 und 5 (GER 2001: 39) zu finden:

> - Die Deskriptoren sind relevant für die Beschreibung … tatsächliche(r) Lernerfolge (…); sie können deshalb auch realistische Lernziele darstellen.
>
> - Sie stellen eine Sammlung von genau beschriebenen kriterienbezogenen Aussagen zum Kontinuum der fremdsprachlichen Kompetenz dar; man kann sie flexibel benutzen, um kriteriumsorientierte Beurteilungen zu entwickeln. Sie können auf jedes lokale System bezogen werden; man kann sie auch auf der Basis lokaler Erfahrungen erweitern und/oder dazu benutzen, neue Systeme von Lernzielen zu entwickeln.

Für welche konkreten Zwecke wurden die Skalen des GER aber konstruiert? Sind die Formulierungen diesen Zwecken angemessen? Und wo im GER wird diese Angemessenheit nachgewiesen? Oder bedeutet die obige „Relevanz"-Aussage, dass die Skalen des GER auf all diese Verwendungszwecke hin adaptiert werden können? Dazu müsste jedoch für jede Adaption nachgewiesen werden, dass die adaptierten Formulierungen dem Zweck auch angemessen sind.

An dieser Stelle können die konkreten Fragen nach den Verwendungsmöglichkeiten der Skalen noch nicht beantwortet werden, da dazu zunächst – die Hintergründe hierfür wurden oben in den Kapiteln 3.1 und 3.2 dieser Arbeit erläutert – der Status und die im GER angegebenen Einsatzmöglichkeiten der verschiedenen Beispielskalen des GER untersucht werden müssen.

[205] In Kapitel 4 dieser Arbeit wird ein Beispiel dafür gegeben.

3.4.3 Skalenanalyse: Beschreibungsgegenstand, Sprache, Verwendbarkeit

In diesem Kapitel werden Analysen ausgesuchter Skalen aus den Abschnitten 3, 4 und 5 des GER vorgestellt, um Probleme im Zusammenhang mit dem Beschreibungsgegenstand, der Basis der Deskriptoren und ihrer Versprachlichung aufzuzeigen. Denn diese Probleme führen dazu, dass die Skalen nur bedingt zu den Zwecken einsetzbar sind, die im GER jeweils konkret dafür angegeben werden. Drei „Arten" von Skalen lassen sich im GER ausmachen: Zunächst die drei Beispielskalen, die in GER-Abschnitt 3 die Referenzniveaus illustrieren sollen; dann die Skalen die *kommunikativen Aktivitäten* des GER-Abschnitts 4 betreffend und schließlich die Skalen zu den *kommunikativen Sprachkompetenzen* des GER-Abschnitts 5.[206] Um diesen drei Ausrichtungen der Skalen im GER gerecht zu werden, wird jede in je einem eigenen Unterkapitel dieser Arbeit analysiert. Diese Form der Darstellung mag zwar zu Redundanzen führen, doch nur eine detaillierte, auf die spezifischen Skalen hin ausgelegte Analyse lässt transparente Schlussfolgerungen in Bezug auf die konkrete Verwendbarkeit[207] der jeweiligen Skalen zu.

Die GER-Skalen werden im Folgenden nach einem Schema analysiert, welches einerseits an die GER-Kategorien *Situationen* (GER 2001: 53ff), *Themen* (ebd.: 58f) und *Bedingungen/Einschränkungen* (ebd.: 55f) und andererseits an das Analyseschema des so genannten *Dutch Grid*[208] angelehnt ist. Die in dieser Arbeit angesetzten **Analysekategorien** gehen aus den Tabellen der Anhänge 14 mit 18 hervor: Dabei bezieht sich die dort erwähnte Kategorie *Operation* auf die sprachliche Handlung; die Kategorie *Was/Wie* bezieht sich darauf, was man mit der Operation auf welche Weise ausführen kann; die Kategorie *Einschränkungen/Bedingungen* umfasst Charakteristika, die eine Operation beeinträchtigen oder beschränken; die Kategorien *Situationen* und *Themen* sind selbsterklärend. Jede hier analysierte Skala wird auf allen sechs Niveaus untersucht: Dazu werden zunächst die Deskriptoren in ihre „Bestandteile" zerlegt und in Form der gerade erwähnten Tabellen den einzelnen Analysekategorien zugewiesen. Dabei werden die Niveaus von unten (A1) nach oben (C2) ansteigend präsentiert und analysiert, da sich die Annahmen hinsichtlich der Zunahme an Sprachvermögen auf diese Weise anschaulicher darstellen lassen. Jedes Niveau wird durch eine eigene Tabelle repräsentiert, jeder Deskriptor in einer neuen Zeile dargestellt. Ein Eintrag in der entsprechenden Zelle einer Tabelle erfolgt natürlich nur, wenn sich dazu eine Aussage im Deskriptor findet. In den Tabellen sind Inkonsistenzen innerhalb eines Niveaus oder über die Niveaus hinweg *kursiv* gesetzt. Abweichungen von diesem Kategoriensystem werden jeweils in dem Kontext begründet, in dem sie

[206] Alle bei der Skalenanalyse angesprochenen Skalen finden sich in ihrer Originalform in den Anhängen 1 mit 13 dieser Arbeit.
[207] Vgl. die Ausführungen oben unter Kapitel 3.4.2 dieser Arbeit zu den drei Kategorien von Verwendungshinweisen im GER: Hier nun wird Kategorie (b) untersucht.
[208] Der *Dutch Grid*, entwickelt im *Dutch CEF Construct Project 2004*, gibt Spezifikationen für die Entwicklung und Klassifikation von Testitems innerhalb des GER-Systems, bezogen auf die Fertigkeiten des Lesens und Hörverstehens (vgl. Alderson 2004). Auf ihn wird in Kapitel 3.5 dieser Arbeit näher eingegangen. Er ist unter www.ling.lancs.ac.uk/cefgrid einsehbar (Zugriff am 08.11.2004). Inzwischen wurde von der ALTE ein *grid* bezogen auf die Schreibfertigkeit entwickelt (vgl. http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual/, Zugriff am 22.08.2005).

notwendig werden. Die **Analyseergebnisse** werden (in den Kapiteln 3.4.3.1 mit 3.4.3.3 dieser Arbeit) unter nachstehenden Gesichtspunkten dargestellt:

- *Strukturiertheit*: Dabei wird beurteilt, ob die Kategorisierung der Merkmale stringent erfolgt, ob die Merkmale über die Skala hinweg kohärent beschrieben sind und ob die einzelnen Niveaus innere Konsistenz aufweisen.

- *Lücken*: Es werden Lücken identifiziert, die sich ergeben, weil beispielsweise nicht alle relevanten Bereiche einer Kategorie abgedeckt werden, nicht alle für einen Bereich relevanten Operationen erwähnt werden, nicht alle Merkmale auf den für sie relevanten Niveaus beschrieben sind, oder relevante Aspekte wie etwa *Themen* oder *Situationen* fehlen.

- *Sprache*: Dabei werden Terminologieprobleme erörtert, die in fehlenden Definitionen ihren Ursprung haben, oder die sich ergeben, weil unklar ist, ob verschiedene Termini synonym oder in unterschiedlicher Bedeutung benutzt werden. Auch Probleme der verbalen Abstufung fallen unter diesen Analyseaspekt.

- *Verwendbarkeit*: Zunächst wird auf Basis der obigen Analysen der Beschreibungsgegenstand, sein Generalisierungsgrad und die Einfachheit/Komplexität der Darstellung beurteilt. Darauf aufbauend wird die Ebene im GER-System identifiziert, auf der die Skala angesiedelt ist. Dabei helfen die Kriterien der Einfachheit respektive Komplexität der Beschreibung und des Abstraktionsgrads der Darstellung. Daneben tritt die Betrachtung der Basis der Deskriptoren hinsichtlich der Frage, was dort auf welcher Grundlage auf welche Weise beschrieben wird. Unter Beachtung der Aussagen, die sich im GER konkret bezüglich der Zwecke und Einsatzbereiche der jeweils zu analysierenden Skala finden lassen, wird schließlich die Verwendbarkeit der Skala beurteilt.

Die Ergebnisse der exemplarischen Analysen der GER-Skalen zur Schreibfertigkeit decken sich weitgehend mit denen, die im Zusammenhang mit der Erstellung des erwähnten *Dutch Grid* identifiziert wurden: Dort hat man aufgrund einer umfassenden Analyse der Skalen zu Lesen und Hörverstehen folgende Probleme festgestellt: Terminologieprobleme, Lücken, Inkonsistenzen in den Niveaubeschreibungen und fehlende Definitionen der verwendeten Termini (vgl. Alderson et al. 2004: 7ff). Es ist im Rahmen dieser Arbeit natürlich nicht möglich, alle Skalen des GER solch einer Analyse zu unterziehen, weshalb neben den folgenden Analysen auf die umfangreichen Arbeiten im *Dutch Grid*-Projekt zu den dort analysierten Skalen verwiesen werden darf. In den folgenden drei Unterkapiteln dieser Arbeit wird gezeigt, dass sich die im Projekt identifizierten Probleme noch in weiteren Skalen des GER finden. Die Bedeutung dieser Probleme in Bezug auf den Status der Deskriptoren und damit bezüglich der Verwendbarkeit einer Skala muss für jede Skala eigens untersucht werden, da jede Skala auf einer anderen Ebene im unter Kapitel 3.4.2 dieser Arbeit erläuterten GER-System angesiedelt ist, einen anderen Beschreibungsbereich abdeckt und sich die genannten Probleme je anders auswirken dürften.

3.4.3.1 Die Beispielskalen des GER-Abschnitts 3

Die Referenzniveaus werden in Abschnitt 3 des GER anhand dreier Skalen vorgestellt: einer *Globalskala* (GER: 2002: 35), eines *Rasters zur Selbstbeurteilung* (ebd.: 36) und eines *Beurteilungsrasters zur mündlichen Kommunikation* (ebd.: 37f). Die Deskriptoren dieser drei Skalen wurden aus „einer Datenbank“[209] (ebd.: 38) zusammengefasst – bei der Datenbank dürfte es sich um den oben erwähnten *Pool* handeln, in dem alle Deskriptoren der GER-Skalen ihren Ursprung haben. Die Konstruktion dieser drei Skalen ist den Angaben auf S. 205 des GER zufolge ebenfalls im oben erwähnten Projekt der GER-Skalenkonstruktion erfolgt.

Wenden wir uns zunächst der Betrachtung der erwähnten **Globalskala** (GER 2001: 35, hier in Anhang 1 zu finden) zu, die das generelle Sprachvermögen hinsichtlich der Bereiche der Rezeption, Produktion und Interaktion beschreiben will. Folgende Probleme ergeben sich bei näherer Analyse dieser Skala:

- Die Struktur innerhalb der einzelnen Niveaus ist inkonsistent, ebenso wie die Struktur über die Niveaus hinweg: Man würde erwarten, dass die Bereiche Rezeption, Interaktion und Produktion kohärent auf jedem Niveau beschrieben werden, doch nicht auf jedem Niveau wird jeder Bereich durch einen entsprechenden Deskriptor dargestellt: Beispielsweise findet sich auf A1 ein Deskriptor zur Rezeption und Produktion, während die Interaktion durch zwei Deskriptoren beschrieben wird. Auf C2 hingegen findet sich kein Deskriptor zur Interaktion. Die Struktur des Niveaus C1 ist in sich inkohärent: Zuerst wird die Rezeption beschrieben, dann folgt ein Deskriptor zur Produktion, einer zur Interaktion und schließlich wieder einer zur Produktion.

Eine weitere Inkonsistenz zeigt sich an der nicht konsequenten Unterscheidung zwischen mündlichem und schriftlichem Sprachgebrauch: Auf A1 etwa wird in einem Deskriptor auf „Gesprächspartner“ Bezug genommen, während in einem anderen Deskriptor lediglich vom Verstehen von Ausdrücken und Sätzen die Rede ist – ob sich dieses Verstehen auf mündlichen oder schriftlichen Sprachgebrauch bezieht, bleibt unklar. Diese Undifferenziertheit zieht sich bei der vorliegenden Skala durch alle Niveaus.

Die hier gezeigte Unstrukturiertheit trägt nicht zur Handhabbarkeit der Skala bei, denn um sich, andere, eine Leistung oder etwa ein Lernziel einem Niveau zuordnen zu können, bedarf es eines vergleichbaren Aufbaus der Niveaus und eines transparenten Beschreibungsgegenstands.

[209] Es findet sich nirgends im GER eine Aussage dahingehend, ob die hier zitierte Datenbank, der vermutliche Ursprungs-*Pool* also, der Öffentlichkeit zugänglich ist – im Internet ist sie jedenfalls nicht zu finden.
Die Datenbank, in welcher die GER-Deskriptoren zu Hörverstehen, Interaktion, mündlicher Produktion und Lesen nach Schwierigkeit und nach Kategorien sortiert sind, ist jedoch zugänglich unter: http://www.unifr.ch/ids/Portfolio/descriptors.htm, Zugriff am 25.01.2005.

- Es finden sich Lücken in der Skala: Wie aus den Tabellen in Anhang 14 dieser Arbeit ersichtlich, kommt es zu verschiedenen Leerstellen: Die *Einschränkungen* beispielsweise nehmen von unten nach oben ab; das ist nachvollziehbar und liegt in der Natur des Zuwachses an Sprachvermögen. Warum aber häufig Spezifikationen zu *Situationen* und *Themen* fehlen, ist nicht nachvollziehbar. Die auf den jeweiligen Niveaus beschriebenen Operationen sind ebenfalls lückenhaft; die Zuordnung von Operationen auf bestimmte Niveaus erweckt einen selektiven und arbiträren Eindruck: Beispielsweise erscheint das *Bewältigen von Situationen auf Reisen im Sprachgebiet* nur auf B1 (und dort wird nichts weiter zur Interaktionsfähigkeit ausgesagt), während das *Zusammenfassen verschiedener Informationsquellen*[210] nur auf C2 beschrieben wird. Unverständlich bleibt auch, warum die *Fähigkeit zur Anwendung verschiedener Mittel der Textverknüpfung* überhaupt in dieser Globalskala und dann nur auf C1 erwähnt wird. Es wurde zwar in Kapitel 3.2.3 dieser Arbeit dargestellt, dass nicht alle Merkmale einer Beschreibungskategorie auf allen Niveaus einer Skala versprachlicht werden müssen, doch gerade bei holistischen Skalen, die mehrere Kategorien (hier: Rezeption, Produktion und Interaktion) auf einem Niveau beschreiben, sollten doch wenigstens, wie bei der Kritik zur mangelnden Strukturiertheit oben schon eingefordert, die Kategorien auf jedem Niveau konsistent in den jeweils charakteristischen Merkmalen beschrieben werden. Wie die Benutzer mit den Lücken dieser Skala umgehen sollen, bleibt ihnen überlassen.

- Probleme der Versprachlichung lassen sich sowohl im Hinblick auf die Terminologie als auch auf die Abstufungen feststellen: Die Verben, die die *Operationen* versprachlichen, variieren von Niveau zu Niveau, wobei nicht immer ersichtlich ist, welche Bedeutung sie tragen und ob sie synonym verwendet werden. Beispielsweise findet sich im Bereich der Rezeption das Verb *verstehen* auf allen Niveaus, wohingegen sich kein Verb findet, das ebenso konsequent für den Aspekt der Sprachproduktion eingesetzt würde. In diesem Bereich finden sich vielmehr Verben wie *verwenden*, *sich vorstellen* (A1); *beschreiben* (A2); *berichten, beschreiben* (*Begründungen, Erklärungen*) *geben* (B1); *ausdrücken, erläutern, angeben* (B2); *ausdrücken*, (*Sprache*) *gebrauchen, sich äußern* (C1); und *wiedergeben, ausdrücken* (C2). Es böte sich an, ein Verb wie *ausdrücken* oder *beschreiben* stringent zu nutzen. Der Bereich der Interaktion schließlich wird gerade auf den beiden obersten Niveaus undurchsichtig versprachlicht: Während sich *verständigen* auf A1, A2 und B2 findet, und auf B1 (*Situationen auf Reisen*) *bewältigen* benutzt wird, findet sich auf C1 und C2 lediglich *ausdrücken*, das sich aber offenbar eher auf die Produktion bezieht. Auf C1 ist zusätzlich noch die Rede von (*Sprache*) *gebrauchen*.

Die Verbalisierung der Abstufungen ist wie bereits angedeutet nicht immer problemlos: Es lassen sich in der obigen Skala eine Vielzahl von Adjektiven ausmachen, die der verbalen Abstufung dienen, wie etwa: *vertraut, alltäglich, ganz einfach, konkret* (A1); *häufig gebraucht,*

[210] Warum in obiger Globalskala die Textverarbeitungsfähigkeit nur auf einem Niveau beschrieben wird, bleibt unbegründet. An dieser Stelle darf auf die Skala *Texte verarbeiten* auf S.98 des GER (in dieser Arbeit in Anhang 4 zu finden) verwiesen werden und die dortige Abstufung, die das *Zusammenfassen von Texten* in verschiedenen Ausprägungen auf den Niveaus B1 mit C2 beschreibt.

einfach, routinemäßig, direkt (A2); *klar, vertraut, einfach und zusammenhängend* (B1); *komplex, konkret und abstrakt, spontan, fließend, klar und detailliert* (B2); *anspruchsvoll, spontan und fließend, wirksam und flexibel, klar, strukturiert und ausführlich, komplex, angemessen* (C1); *praktisch (alles), mühelos, zusammenhängend, spontan, sehr flüssig und genau* (C2). Auch hier wäre eine konsistente Terminologie wünschenswert, finden sich doch Adjektive wie *klar* auf B1, B2 und C1, und *spontan* auf B2, C1 und C2. Daneben werden teils rein verbale Abstufungen genutzt, wie etwa *ganz einfach – einfach – einfach und zusammenhängend* (A1 mit B2). Zudem haben diese Adjektive, ebenso wie die oben genannten Verben, einen Bedeutungsbereich, der interpretiert werden muss.

Es findet sich nirgends eine Definition der verwendeten Termini; teils handelt es sich um Adjektive und Verben mit breitem Bedeutungsbereich, teils bezeichnen die Verben (dann aber erweitert um eine Nominalphrase) sehr konkrete Handlungen. Wie unter Kapitel 3.2.2 und 3.2.3 dieser Arbeit bereits erläutert, kann man diese Probleme der Versprachlichung nicht immer umgehen. Dennoch könnten sie abgeschwächt werden, indem etwa eine Operation durch immer dasselbe Verb dargestellt wird. Auch wenn dies manchen Nutzer stilistisch nicht befriedigen mag, so führt diese Redundanz wenigstens zu mehr Transparenz und erleichtert das Verständnis.

- Die Verwendbarkeit der Globalskala wird wie folgt eingeschätzt: Sie deckt auf oberster Ebene des GER-Systems die Bereiche Rezeption, Produktion und Interaktion ab; diese werden teils durch eher abstrakte kommunikative Handlungen (beispielsweise *Kann sich spontan und fließend ausdrücken*), teils durch sehr konkrete Handlungen (wie etwa *Kann anderen Leuten Fragen zu ihrer Person stellen*) dargestellt; die Deskriptoren sind in Form von generalisierenden KANN-Aussagen verfasst. Die Basis der Aussagen bleibt im Dunkeln, ebenso wie die der Abstufungen. Da die Benutzer nicht wissen, ob es sich beispielsweise um tatsächlich beobachtetes Verhalten, um intuitive Annahmen oder um erfahrungsbasierte Erwartungen an die Niveaus handelt, ist es schwer, einen konkreten Verwendungsbereich für diese Skala anzugeben. Sie ist für Aufgabenkonstruktion, diagnostische Zwecke oder Performanzbewertung zu grobkörnig und zu generalisierend gehalten; sie könnte eventuell für eine holistische Einschätzung von Lernenden genutzt werden, die der einschätzenden Person jedoch sehr gut in allen Aspekten der Sprachverwendung bekannt sein müssten; insofern könnte sie auch für eine grobe Selbsteinschätzung genutzt werden. Der Grad der Generalisierung der meisten Deskriptoren dieser Skala rückt sie in den Bereich der *reporting scales*: Eventuell könnte sie zum Berichten der verallgemeinerten Ergebnisse eines Sprachvermögenstests genutzt werden, der die dort beschriebenen Teilbereiche und Handlungen jedoch abdecken müsste. Doch die Grobkörnigkeit, die Lücken und Inkonsistenzen sowie die Versprachlichungsprobleme weisen darauf hin, dass sie auch für diese Zwecke überarbeitet werden müsste. Nun lohnt der Blick darauf, zu welchem Zweck diese Skala im GER konkret angeführt wird. Dazu findet sich auf S.34 des GER folgende Aussage:

> Es ist auch wünschenswert, dass die gemeinsamen Referenzpunkte für unterschiedliche Zwecke auf unterschiedliche Weise präsentiert werden. Für einige Zwecke wird es genügen, das System der Gemeinsamen Referenzniveaus in einfachen, holistischen Abschnitten zusammenzufassen wie in Tabelle 1. Eine solche einfache "globale" Darstellung macht es leichter, das System Nichtfachleuten zu vermitteln, und es kann zugleich Lehrenden und Curriculumplanern Orientierungspunkte geben.

Es ist gut, wenn eine auf den jeweiligen Zweck hin adaptierte Darstellung anerkannt wird; dennoch bleiben Zweifel, ob diese Skala helfen kann, „Nichtfachleuten" das Niveausystem näher zu bringen – wie gerade gezeigt, charakterisiert sie die Niveaus nicht in konsistenter Weise. Die genannten „Orientierungspunkte" dürften auch nur sehr grober Natur sein – ob sie wirklich helfen können, Lehrenden oder Curriculumplanern Orientierung zu geben, muss die Praxis zeigen. Solange jedoch nicht offen gelegt wird, ob die Abstufungen beispielsweise auf angenommener oder tatsächlicher Progression (beispielsweise in den Bereichen des Schweizer Bildungssystems, in denen die Lehrenden der genannten Workshops tätig sind) oder etwa auf impliziten Erwartungen an den Spracherwerb generell oder auf Beobachtungen konkreter Lernfortschritte beruhen, ist solch eine Skala nur mit größter Vorsicht verwendbar. Als Entscheidungsbasis gerade im Bereich der Curriculumplanung kann sie mangels empirisch abgesicherter Grundlagen sicher nicht dienen.

Die beiden auf S.36 respektive S.37 des GER folgenden Skalen sind insofern erwähnenswert, als zwei Beurteilungsskalen an den Anfang des GER gestellt werden; diese prominente Platzierung lässt wieder einmal auf die Beurteilungslastigkeit des GER schließen. Eine detaillierte Analyse dieser Skalen bringt an dieser Stelle keine neuen Einsichten, dennoch sind einige Punkte erwähnenswert: Beide Skalen, sowohl das *Raster zur Selbstbeurteilung* (GER 2001: 36) als auch das *Beurteilungsraster zur mündlichen Kommunikation* (ebd.: 37f) sind – ebenso wie die gerade vorgestellte Globalskala – aus Deskriptoren zusammengesetzt worden, die dem oben erwähnten *Pool* des Schweizer Konstruktionsprojekts entstammen (vgl. GER 2001: 38). Die Raster wurden demnach analog zu den GER-Skalen der GER-Abschnitte 4 und 5 (und damit unabhängig von diesen Skalen) aus der genannten Deskriptorensammlung entwickelt. Dieses Vorgehen widerspricht den Aussagen im GER (2001: 175f), dass man die Skalen der GER-Abschnitte 4 und 5 nutzen könne, um Beurteilungsraster abzuleiten. Es könnte jedoch insofern richtungweisend sein, als es eine von den GER-Skalen unabhängige Skalenkonstruktion erlaubt, sich jedoch auf denselben Deskriptoren-*Pool* bezieht, der auch den GER-Skalen zugrunde liegt. Deshalb wird es unter Kapitel 3.4.4 dieser Arbeit wieder aufgenommen, wenn die Verwendbarkeit der GER-Skalen abschließend beurteilt wird.

Im **Selbstbeurteilungsraster** (GER 2001: 36, hier in Anhang 2 wiedergegeben) werden kommunikative Aktivitäten in den Bereichen *Verstehen* (Hören und Lesen), *Sprechen* (an Gesprächen teilnehmen und zusammenhängendes Sprechen) sowie *Schreiben* (nicht differenziert in produktives und interaktives Schreiben) dargestellt – eine Einteilung, die in dieser Art nicht zwingend notwendig ist und eigentlich einer Begründung, wenigstens jedoch einer Erläuterung

bedarf. Die Abdeckung dieser Bereiche lässt darauf schließen, dass diese Skala eine Ebene unterhalb der Globalskala angesiedelt ist, jedoch immer noch einen globalen Blick auf das Sprachvermögen wirft. Jede der Kategorien wird auf sechs Niveaus durch Deskriptoren beschrieben, die ähnlich derer in der oben analysierten Skala teils abstrakt, teils konkret gehalten sind und das Sprachkönnen auf generalisierende Weise beschreiben. Dies ist für eine erste grobe Profilbildung seitens der Lernenden durchaus angemessen, denn sie können, wo immer sie es für nötig oder wünschenswert halten, in detailliertere Skalen des GER – die wiederum eine Ebene tiefer im System angesiedelt sind – einsteigen, um den jeweiligen Gegenstandsbereich auf dem grob identifizierten Niveau differenzierter zu betrachten. Dieser Zweck wird im GER auch auf S. 37 angegeben. Allerdings scheint es ratsam, Lernende zuerst in den Umgang mit Selbstbeurteilung, Skalen und Checklisten einzuführen und ihnen die Referenzniveaus des GER zuerst zu erläutern, beispielsweise anhand der Charakterisierungen der Niveaus, wie sie etwa in Abschnitt 3.6 des GER gegeben werden. Aufbauend darauf können die Lernenden im Umgang mit Selbstbeurteilung und einem Sprachenportfolio vertraut gemacht werden. An dieser Stelle darf auf die umfangreichen Arbeiten zur Entwicklung von Sprachenportfolios in verschiedenen Projekten verwiesen werden.[211]

Das **Beurteilungsraster zur mündlichen Kommunikation** (GER 2001: 37f, hier in Anhang 3 abgebildet) soll eine weitere Orientierung der Skalen im GER illustrieren, namentlich eine Skala zur Bewertung der „generelle[n] Aspekte der Kompetenz, die immer in der mündlichen Performanz sichtbar werden“ (GER 2001: 187). Man kann diese Skala als fokussierenden Ausschnitt des gerade erwähnten Selbstbeurteilungsrasters betrachten: Es werden „qualitative Aspekte des Sprachgebrauchs“ (ebd.: 37) in den Kategorien *Spektrum, Korrektheit, Flüssigkeit, Interaktion* und *Kohärenz* beschrieben. Die Darstellung bleibt jedoch immer noch generalisierend, ohne Performanzmerkmale wiederzugeben, die sich konkret auf eine bestimmte Prüfung beziehen. Problematisch dabei ist, dass mit einem Raster, das keine Performanzmerkmale zur Bewertung bietet, mündliche Performanz auf Kompetenzniveaus eingestuft werden soll (vgl. die Aussage auf S. 187 GER) – ein Widerspruch zur Aussage auf S.174f des GER, dass sich die „Beurteilung der Sprachkompetenz nicht ... auf eine spezielle einzelne Leistung beziehen … sollte“. Der Rückschluss von einem Performanzbeispiel auf das zugrunde liegende Können ist nur denkbar in Kontexten, in denen die Bewerter die Probanden kennen und damit der Beurteilung des Könnens eben doch mehr als nur ein einziges Performanzbeispiel zugrunde legen.

Hier zeigt sich wiederum die konzeptionelle Inkonsistenz des Skalenansatzes: Da die Herkunft der Beschreibungen in den Deskriptoren nicht bekannt ist, werden lediglich die Kategorien (durch die Zuordnung zu den verschiedenen Ebenen des oben beschriebenen Pyramiden-Modells) und die Niveaus jeweils in ihrer Feinheit oder Grobheit adaptiert, nicht jedoch die Basis der Beschreibungen mit bedacht. Wie oben schon erwähnt, deckt die Unterscheidung Grobheit

[211] Vgl. etwa http://www.coe.int/portfolio, Zugriff am 03.02.2005.

– Feinheit (oder anders ausgedrückt, die Unterscheidung Einfachheit – Komplexität) jedoch nur eine Facette der Aspekte ab, die letztlich den Verwendungsbereich einer Skala bestimmen. Sie ist somit für die Entscheidung, welche Deskriptoren für welche Zwecke eingesetzt werden können, nicht hinreichend.

In den nächsten beiden Unterkapiteln dieser Arbeit werden exemplarisch Skalen die kommunikativen Aktivitäten betreffend, und Skalen, die sich auf die sprachlichen Kompetenzen beziehen, analysiert, da dies die Bereiche sind, für die der GER Beispielskalen in seinem Abschnitt 4 respektive Abschnitt 5 bereitstellt. Dabei bietet es sich an, solche Skalen auszuwählen, die im DESI-Projekt bei der Konstruktion der *rating scales* im Rahmen des Bewertungsschemas der semikreativen Schreibaufgabe herangezogen wurden, welches in Kapitel 4 dieser Arbeit dokumentiert wird. Gegenstand der Analyse ist zunächst der Bereich, der in den betreffenden Skalen beschrieben wird; es soll betrachtet werden, ob er sich deckt mit dem Bereich, der in den Abschnitten 4 und 5 des GER den Beispielskalen jeweils vorangestellt definiert oder zumindest umrissen ist. Im Anschluss werden wiederum die Strukturierung der Niveaus, Lücken in den respektiven Skalen, Aspekte der Versprachlichung und Aussagen im GER bezüglich der Verwendbarkeit der jeweiligen Skalen beurteilt. Das dabei verwendete Analyseschema ist zu Beginn des Kapitels 3.4.3 dieser Arbeit vorgestellt worden – Abweichungen davon werden im Kontext begründet.

3.4.3.2 GER-Skalen zu kommunikativen Aktivitäten

Im Folgenden konzentriert sich die vorliegende Arbeit im Bereich der kommunikativen Aktivitäten auf das produktive Schreiben: Es wird die vergleichende Analyse der GER-Skalen *Schriftliche Produktion* (GER 2001: 67), *Kreatives Schreiben* (ebd.), *Berichte und Aufsätze schreiben* (ebd.: 68) und *Schreiben* (aus dem Selbstevaluationsraster, ebd.: 36) vorgenommen. Bei dieser Analyse wird die oben erwähnte Kategorie *Situationen* in der tabellarischen Darstellung (vgl. Anhang 15) fallen gelassen, da sich dazu keine Aussagen in den vier Skalen finden. Ab dem Niveau B2 sind auch keine *Einschränkungen* mehr in den Skalen aufgeführt, dafür finden sich detailliertere Aussagen zur Qualität der Texte, weshalb ab B2 die Analysekategorie *Einschränkungen* wegfällt zugunsten der Aufteilung der Kategorie *Was/Wie* in die beiden Unterkategorien *Was* und *Wie*. Die Deskriptoren sind wiederum zeilenweise in den Tabellen in Anhang 15 dieser Arbeit dargestellt. Dort sind Inkonsistenzen und Auffälligkeiten *kursiv* gesetzt; sie werden in der folgenden Analyse besprochen:

- Gegenstandsbereich: Zur schriftlichen Textproduktion[212] finden sich Aussagen an verschiedenen Stellen des GER: Die Definition in GER-Abschnitt 4.4.1.2, in dem sich auch die Skalen *Schriftliche Produktion*, *Kreatives Schreiben* und *Berichte und Aufsätze schreiben* finden, ist sehr knapp und wenig aufschlussreich: „Bei *produktiven schriftlichen Aktivitäten* (beim *Schreiben*) produzieren die Sprachverwendenden als Autoren einen geschriebenen Text, der von einem oder mehreren Lesern rezipiert wird." (GER 2001: 66). Auch die Auflistung einiger Beispiele für Schreibaktivitäten ist nicht sehr hilfreich. Den Skalen nachgeschaltet sind Ausführungen zu *Produktionsstrategien*, die den Aspekten der Planung, Ausführung, Kontrolle und Reparatur Rechnung tragen. Sie werden in drei eigenen Beispielskalen dargestellt, welche jedoch nicht gemeinsam mit der Schreibfertigkeit skaliert wurden. Unter GER-Abschnitt 4.5 *Kommunikative Sprachprozesse* findet sich der schriftliche Produktionsprozess beschrieben: Er umfasst „kognitive und sprachliche" sowie „manuelle Fertigkeiten", wobei zur Formulierung „lexikalische, grammatische und …orthographische Fertigkeiten" beitragen (vgl. ebd.: 93). In GER-Abschnitt 4.6 werden *Texte* im Sinne „aller sprachlichen Produkte, die Sprachverwendende/Lernende empfangen, produzieren oder austauschen" (ebd.: 95) charakterisiert. Allerdings ist die Auflistung der Textsorten nicht nachvollziehbar, werden dort doch beispielsweise „Bücher, Literatur und Sachbücher einschließlich literarischer Zeitschriften" als Textsorten bezeichnet (ebd.: 97).

Schreibfertigkeit[213] wird in den in dieser Arbeit analysierten Skalen nach obiger Kurzdefinition unter je verschiedenen Gesichtspunkten (vgl. dazu auch die Analyse zum Aspekt der Versprachlichung unten) und in unterschiedlichem Detailliertheitsgrad beschrieben, ohne dass auf dabei involvierte Prozesse, Strategien oder sprachliche Kompetenzen eingegangen wird. In der Skala *Schriftliche Produktion* finden sich globale Aussagen zur Schreibfertigkeit, ohne dass Textsorten thematisiert werden; die Skalen *Kreatives Schreiben* und *Berichte und Aufsätze schreiben* beschreiben hingegen konkretere Arten des Schreibens (vgl. unten die Ausführungen zur Verwendbarkeit): In der Skala *Berichte und Aufsätze schreiben* werden selbstverständlich Bericht und Aufsatz beschrieben, die Skalen *Kreatives Schreiben* und *Schreiben* zeigen jedoch bezüglich der Textsorten einige Inkonsistenzen respektive Intransparenzen (vgl. unten).

- Strukturierung: Die Niveaus, auf denen die Schreibfertigkeit abgestuft wird, sind über die Skalen hinweg relativ konsistent voneinander abgegrenzt und in vergleichbarer Weise charakterisiert: Während sich etwa A1 durch *isolierte Wendungen* auszeichnet, finden sich auf B1 *zusammenhängende, unkomplizierte Texte zu vertrauten Themen*, und auf C2 schließlich kommen zu *flüssigen und anspruchsvollen Texten* Merkmale wie *angemessener Stil, effektive Struktur* und *Leserbezogenheit* mit herein. In den Skalen *Kreatives Schreiben* und *Berichte und Aufsätze*

[212] Die schriftliche Interaktion wird in GER-Abschnitt 4 grundsätzlich von der schriftlichen Produktion unterschieden; im hier analysierten Selbstbeurteilungsraster (GER 2001: 36) wird Schreiben jedoch nicht in *produktiv* und *interaktiv* differenziert, weshalb sich in dessen Rahmen auch Aspekte des interaktiven Schreibens finden.

[213] Zusätzlich zu den Ausführungen hier darf auf Kapitel 4.2.5 dieser Arbeit verwiesen werden: Dort wird ausgehend vom Testkonstrukt der Schreibfertigkeit im DESI-Projekt das Verständnis der Schreibfertigkeit im GER im Detail analysiert. Auch wenn dieses Vorgehen zu Redundanzen führen mag, so ist doch das Augenmerk hier auf den Gegenstandsbereich der Skalen gerichtet, während es in Kapitel 4.2.5 auf die Konzeption der Schreibfertigkeit gerichtet ist.

schreiben scheint die Zunahme an Textsorten, die verfasst werden können, ebenso logisch wie die thematische Progression von *Vertrautem* über *persönliche Interessengebiete* hin zu *komplexen Themen*, welche in allen vier Skalen zu finden ist. Auch die Abstufung der zunehmend umfangreicheren Operationen ist nicht kontra-intuitiv.

Dennoch gibt es einige Aspekte, die nicht einleuchten. Beispielsweise sind die Bereiche dessen, *was* auf welchem Niveau verschriftlicht werden kann, nicht immer konsistent: Nur auf A2 finden sich in der Skala *Kreatives Schreiben* die Genres *Biographie* und *Gedicht*, lediglich auf B1 werden *Sachinformationen* erwähnt; in der Skala *Schriftliche Produktion* findet sich nur auf A2 die Angabe konkreter *Konnektoren* – wieso diese Aspekte so selektiv an diesen Orten auftauchen, ist nicht nachvollziehbar. Auch die Operationen sind nicht konsistent auf allen Niveaus „dekliniert". Finden sich Aussagen zur Texterstellung auf allen Niveaus, so finden sich solche zur Verarbeitung fremder Texte lediglich bei den Skalen *Schriftliche Produktion* (auf dem Niveau B2), *Berichte und Aufsätze schreiben* (auf den Niveaus B1, B2- und C2), und *Schreiben* (auf C2). Diese Aussagen erwecken einen eher inkonsistenten Eindruck; in diesem Zusammenhang muss auch die Frage erlaubt sein, ob der Aspekt *Texte verarbeiten* (wozu es ja eine eigene Skala auf S. 98 des GER gibt, die jedoch wiederum von den oben dargestellten Merkmalen abweicht) denn überhaupt unter der Kategorie *Schriftliche Produktion allgemein* und insbesondere unter der Kategorie *Berichte und Aufsätze schreiben* dargestellt werden sollte.

Beim Vergleich der Deskriptoren der vier Skalen innerhalb der Niveaus fällt auf, dass es auch hier Inkonsistenzen gibt: Auf A1 etwa sind bei der Skala *Schriftliche Produktion* keine Angaben zu den Themen gemacht, wohingegen die Skala *Kreatives Schreiben* konkrete Informationen darüber gibt. Auf A2 beispielsweise sind die unterschiedlichen Gegenstände dessen, was verschriftlicht werden kann, sehr detailliert in der Skala *Kreatives Schreiben* aufgeführt, während sich bei der Skala *Schriftliche Produktion* nichts dazu findet, dafür aber die erwähnte Angabe konkreter Konnektoren. Auf B1 fällt die Skala *Berichte und Aufsätze schreiben* aus dem Rahmen: Insgesamt ist dieses Niveau durch *unkomplizierte, einfache* und teils *detaillierte Texte* gekennzeichnet, nur in der Skala *Berichte und Aufsätze schreiben* werden *Berichte* als *sehr kurz* und Aufsätze als *kurz* charakterisiert – eine nicht verständliche Angabe, zumal in besagten Berichten *Gründe für Handlungen* oder *Sachinformationen gegeben werden können* und *kurz* als Einschränkung auf A1 und A2 genutzt wird. Eventuell ist der Grund für diese Inkonsistenz der Skala *Berichte und Aufsätze schreiben* in den fehlenden Deskriptoren zu den untersten beiden Stufen zu suchen. Das Niveau B2 ist über die vier Skalen hinweg konsistent beschrieben, jedoch ist nicht einleuchtend, warum die Verschriftlichung der *Bedeutung von Ereignissen und Erfahrungen* lediglich in der Selbstbewertungsskala thematisiert wird und warum *erörtern* und *abwägen* nur bei der Skala *Berichte und Aufsätze schreiben* eine Rolle zu spielen scheint. Auf den Niveaus C1 und C2 ist lediglich der Gebrauch des Adjektivs *klar* inkonsistent, denn es dient schon auf B2 zur Charakterisierung, doch zu sprachlichen Aspekten gleich mehr.

- Lücken: Bezüglich der Textsorten oder Genres, die verschriftlicht werden können, gibt es sowohl im Vergleich der vier Skalen innerhalb eines Niveaus als auch über die Niveaus hinweg beträchtliche Unterschiede: Während in der Skala *Schriftliche Produktion* nichts zu Genres gesagt wird, ist in der Skala *Kreatives Schreiben* die Zuordnung der Textsorten *Biographie, Gedicht, Bericht, Beschreibung* und *Geschichte* weder abschließend noch zwingend; dies gilt auch für die der Skala *Berichte und Aufsätze schreiben* (neben *Bericht* und *Aufsatz*) zugeordneten Genres „kritische Würdigung eines literarischen Werks" oder „Zusammenfassung einer größeren Menge an Sachinformationen". Die Zuordnung der verschiedenen Textsorten auf bestimmte Niveaus ist auch in der Skala *Schreiben* nicht transparent. Im Bereich der schriftlichen Textproduktion wäre eine Systematisierung und empirische Validierung der Genres, die den einzelnen Niveaus und Skalen zugeordnet werden, ratsam.

Die im Zusammenhang mit der Schreibfertigkeit relevanten Aspekte der Situationen und Kontexte, in denen geschrieben wird, fehlen ganz. Kommunikative Schreibanlässe werden genauso wenig thematisiert wie die kommunikative Wirkung, die die Texte erzielen. Lediglich ab C1 kommt die Leserperspektive hinsichtlich des Stils und des Aufbaus herein. In Bezug auf den Stil eines Textes wird auch nicht unterschieden zwischen informellem und formalem Stil, einer nicht unwichtigen Differenzierung der Schreibfertigkeit; lediglich eine Aussage wie „persönlicher Brief" impliziert informellen Stil. Ab welchen Niveau beispielsweise formale Geschäftsbriefe verfasst werden können, geht aus diesen Skalen nicht hervor.

- Wenden wir uns nun der Sprache der Deskriptoren zu, insbesondere den damit einhergehenden Problemen:

Terminologie: Auch in diesen Skalen lassen sich die bei der vorangegangenen Skalenanalyse bereits dargestellten Probleme beobachten. Man kann beispielsweise die die *Operationen* beschreibenden Verben mehreren Aspekten der schriftlichen Produktion zuordnen: a) „neutrale" Textproduktion, dargestellt durch Verben wie *schreiben, verfassen, beschreiben, sich ausdrücken, angeben,* etc; b) „argumentierende" Textproduktion, beschrieben durch Verben wie *erörtern, entwickeln, erläutern*, (*Argumente*) *geben*, etc.; c) „wertende" Textproduktion, charakterisiert durch Verben, die eine persönliche Wertung implizieren, wie (*Stellung*) *nehmen, abwägen, würdigen*, etc.; d) „Verarbeitung von Texten", gekennzeichnet durch Verben wie *zusammenfassen* oder *zusammenführen*; e) „Konventionen der Textgestaltung", dargestellt durch Verben wie *strukturieren*, (*Schluss*) *abrunden*, (*Stil*) *wählen* oder (*Aufbau*) *geben*; und schließlich f) „Fokus im Text", beschrieben durch Verben wie (*deutlich*) *machen*, *hervorheben*, oder (*durch Beispiele*) *stützen*. Oft ist jedoch innerhalb dieser Aspekte nicht klar, ob die Verben synonym oder in unterschiedlichen Bedeutungen gebraucht werden. Auch in diesem Zusammenhang muss, wie oben schon gefordert, die Terminologie definiert, vereinheitlicht und konsistent genutzt werden.

Verbale Abstufungen: Meist sind die Abstufungen stringent durch Adjektive charakterisiert, die jedoch gelegentlich, wie oben schon angedeutet, auf anderen als den für sie charakteristischen

Niveaus gebraucht werden. Beispiele hierfür sind *klar* auf B2, C1 und C2, oder *detailliert* auf B1 mit C1. Zudem werden Verstärker wie *sehr* oder *ganz* (*kurz*) genutzt und zu interpretierende Adjektive wie *angemessen*.

- Wie also sind diese Skalen verwendbar? Betrachten wir wieder ausgehend vom Grad der Detailliertheit respektive der Abstraktion die Ebenen im Beschreibungssystem des GER, auf denen die vier Skalen jeweils angesiedelt werden können. Darauf aufbauend soll die Basis der Deskriptoren helfen, den Verwendungsbereich festzustellen.

Ein Vergleich der vier Skalen zeigt, dass die Skala *Schriftliche Produktion* die einfachste und abstrakteste ist, die grobe Beschreibungen bietet, ohne auf konkrete Details wie etwa Textsorten einzugehen. Die ebenfalls global gehaltene Skala *Schreiben* zeigt mehr Details, gerade im Hinblick auf Themen und Textsorten; in ihr werden auch konkrete Schreibaktivitäten genannt. Die Skala *Kreatives Schreiben* ist die detaillierteste; sie geht konkret auf Themen, Stil und Aspekte des Verschriftlichens von eigenen Wertungen ein. In der im Vergleich zu *Kreatives Schreiben* weniger detaillierten Skala *Berichte und Aufsätze schreiben* werden teils abstrakte Aussagen gemacht (wie beispielsweise *Kann klare, gut strukturierte Ausführungen zu komplexen Themen machen* auf C1), teils finden sich jedoch auch konkretere Beschreibungen (wie etwa *Kann in einem Aufsatz oder Bericht etwas erörtern, dabei Gründe für oder gegen einen bestimmten Standpunkt angeben und die Vor- und Nachteile verschiedener Optionen erläutern* auf B2-).

Auf welchen Ebenen des GER-Beschreibungssystems sind diese vier Skalen angesiedelt? Die Skala *Schriftliche Produktion* ist sicherlich auf einer Ebene unterhalb der obersten, globalen Ebene angesiedelt, da sie die Schreibfertigkeit als einen Teilaspekt des globalen Sprachvermögens holistisch beschreibt. Dies tut die Skala *Schreiben* ebenfalls, jedoch bleibt dabei unklar, ob sie als Verfeinerung zu *Schriftliche Produktion* gedacht ist oder ob die größere Detailliertheit darauf beruht, dass sie aus anderen Quellen entwickelt wurde (vgl. oben). Die Skalen *Kreatives Schreiben* und *Berichte und Aufsätze schreiben* dürften als verfeinerte Skalen zu *Schriftliche Produktion* gedacht sein, die Informationen aus der Skala *Schriftliche Produktion* detaillierter wiedergeben und darüber hinaus auch neue Aspekte beschreiben, denen die Globalschreibskala nicht gerecht wird: Beispielsweise kann die Detailliertheit bezüglich der Themen in der Skala *Kreatives Schreiben* als Verfeinerung der globalen Aussagen dazu in der Skala *Schriftliche Produktion* betrachtet werden; die in der Skala *Kreatives Schreiben* beachteten Aspekte der Verschriftlichung von Wertungen und Gefühlen hingegen können als neue Aspekte betrachtet werden, die in der Skala *Schriftliche Produktion* nicht thematisiert sind; dies gilt auch für die Aussagen in Bezug auf Textsorten in der Skala *Kreatives Schreiben*. Das Verhältnis der Skalen *Kreatives Schreiben* und *Berichte und Aufsätze schreiben* zueinander ist mit der Taxonomie Einfachheit - Detailliertheit nicht zu beschreiben. Denn die Skala *Kreatives Schreiben* ist wesentlich detaillierter als die Skala *Berichte und Aufsätze schreiben*, beide Skalen beschreiben jedoch ganz bestimmte Arten des Schreibens, so dass es logisch erscheint, beide derselben Ebene im zuzuordnen, einer Ebene, auf der konkrete Schreibaktivitäten angesiedelt sind.

Da die Skalen zur Schreibfertigkeit wie gesagt nicht empirisch kalibriert worden sind, sondern aus „anderen Skalen“ (GER 2001: 67) kombiniert worden sind, wäre es hilfreich, den Ursprung der Deskriptoren und eine Begründung für die Kategorisierung der Kriterien und Merkmale zu erfahren. Auffallend ist in diesem Zusammenhang, dass die Skala *Berichte und Aufsätze schreiben* für A1 und A2 keine Deskriptoren zur Verfügung stellt, wohingegen in den drei anderen (ja ebenfalls „kombinierten“) Skalen durchaus verwendbare Deskriptoren vorhanden sind. Am interessantesten für die Beurteilung der Verwendbarkeit ist jedoch, wie oben erläutert, die theoretische und/oder empirische Basis der Deskriptoren und der Abstufungen dieser vier Skalen. Momentan ist es, wie bei den oben analysierten Skalen auch, nicht möglich, die Basis der Beschreibungen in diesen vier Skalen festzustellen. Deshalb kann für diese vier Skalen, in Analogie zu den Ausführungen oben bei der Analyse der Globalskala, kein konkreter Zweck abgeleitet werden, zu dem sie eingesetzt werden könnten.

Welchen der vier genannten Orientierungen (vgl. GER 2001: 48) können diese vier Skalen zugeordnet werden? Die Skalen beschreiben jeweils, *was* jemand *wie gut* kann, und das wie gesagt in unterschiedlichen Detailliertheits- und Abstraktionsgraden. Insofern können sie dem Orientierungssystem im GER nicht eindeutig zugewiesen werden. Wenden wir uns deshalb den konkreten Aussagen zur Verwendbarkeit der Skalen der kommunikativen Aktivitäten zu, die in den GER-Abschnitten 3 und 4 zu finden sind:

Auf S. 39 des GER wird ausgesagt, dass das Deskriptorensystem zur Entwicklung kriterienorientierter Beurteilung genutzt werden kann und dazu auf die jeweiligen Kontexte hin erweitert und adaptiert werden kann. Trifft dies auf obige Skalen zu? Die Skalen *Schriftliche Produktion* und *Schreiben* beschreiben in holistischer Weise abgestufte Kriterien der Schreibfertigkeit – insofern sind sie zur groben Einschätzung dieser Fertigkeit auch geeignet; die Skala *Schreiben* bietet zur Selbsteinschätzung darüber hinaus wertvolle Konkretisierungen. Die Skalen *Kreatives Schreiben* und *Berichte und Aufsätze schreiben* sind teils konkret genug, um existente Bewertungskriterien mit den kriterienbezogenen Aussagen der Skalen abzugleichen – dennoch scheint eine Ableitung von Bewertungskriterien allein auf der Grundlage der obigen Skalen schon aufgrund der fehlenden empirischen Validierung der Basis der Beschreibungsgegenstände sehr gewagt. Dazu tritt das Problem, dass die Einordnung der Charakteristika auf den respektiven Niveaus auf den jeweiligen Bewertungskontext zutreffen muss, im Idealfall also jeweils eigens empirisch validiert werden müsste. Diese Aussagen werden in Kapitel 4 dieser Arbeit konkretisiert, weshalb an dieser Stelle auf Beispiele verzichtet werden kann.

Im GER wird behauptet, dass die Deskriptoren „relevant für die Beschreibung … tatsächlicher Lernerfolge“ seien und deshalb „auch realistische Lernziele darstellen“ können (ebd.: 39), ohne dass allerdings diese Aussagen näher erläutert oder begründet würden. Nur wenn die Deskriptoren auf der Beschreibung tatsächlicher Lernerfolge beruhen, können sie eine relevante Beschreibung derselben darstellen; das gilt auch für die Beschreibung der Lernziele. Das Referenzsystem ist auch deswegen als alleinige Grundlage zur Ableitung von Lernzielen nicht direkt

geeignet, da es keinerlei Aussagen bezüglich der Ansiedlung von Lernzielen in den unterschiedlichen Ebenen oder Fächern der jeweiligen Bildungsinstitution respektive bezüglich ihres Konkretisierungsgrads treffen kann.[214] Nur wenn in einem gegebenen Bildungsbereich analysiert oder belegt ist, warum welches Lernziel auf welchem Niveau angesetzt oder erwartet wird, kann diese konkrete Lernzielverortung mit dem Referenzsystem des GER verglichen werden. Wie soll beispielsweise das Merkmal des Niveaus B2- *Kann Rezensionen zu Filmen verfassen* als Lernziel in einem bestimmten Bildungsgang für ein bestimmtes Kursniveau, beispielsweise im Gymnasium in der Sekundarstufe II, angesetzt werden? Stellt es dabei ein fächerübergreifendes oder fachlegitimierendes Ziel, ein Grob- oder Feinziel dar? Um diese Entscheidungen auf valide Basis zu stellen, müssen Bedarfsanalysen und je nach Institution Curriculum- und Lernstandsanalysen vorgeschaltet werden, und diese Ergebnisse müssen mit den Niveaus des GER und idealerweise mit einem weiteren Außenkriterium wie etwa mit den oben erwähnten Lernzielkatalogen des Europarats abgeglichen werden.

In GER-Abschnitt 4 wird ausgesagt, dass aus dem „System von Parametern und Kategorien“ Erwartungen formuliert werden können bezüglich dessen, was Lernende „tun können“ oder „wissen sollten, um handlungsfähig zu sein“ (GER 2001: 51). Da jedoch nicht geklärt ist, ob in den Deskriptoren beobachtetes oder beobachtbares Verhalten oder real vorhandene Wissensbestände beschrieben wurden, stellt sich dasselbe Validierungsproblem wie bei den anderen Verwendungszwecken auch: Skalen können nur in dem Bereich eingesetzt werden, den sie auch valide beschreiben – und dazu muss wie gesagt nachgewiesen werden, dass die Formulierungen dem Zweck auch angemessen sind. Wie oben erwähnt unterliegt jedoch die Validierung von Erwartungen an das Können oder Wissen einem nicht validierbaren Zirkelschluss, wenn sie wiederum an Erwartungen ausgerichtet ist.

Ebenfalls auf S. 51 des GER findet sich die Feststellung, dass die „Gesamtstruktur von Abschnitt 4 als eine Art Checkliste“ dienen kann für Kursplaner, Mitgestaltende an Lehrwerken, Lehrende und Prüfende: Sie können ihre Entscheidungen bezüglich etwa „Textinhalte(n), Übungen, Aktivitäten, Tests, etc.“ anhand dieser Checkliste reflektieren. Auf S. 52 des GER wird deutlich gemacht, dass der GER den Verantwortlichen diese Entscheidungen aufgrund der Komplexität solcher Situationen nicht abnehmen kann oder will – das heißt, die Autoren des GER sind sich des Rahmencharakters des Referenzsystems bewusst: Er dient vorwiegend der Reflexion und der Standortbestimmung der Entscheidungsträger in einem System, das im Idealfall von allen in vergleichbarer Weise verstanden wird – doch von diesem Idealfall sind wir (noch) weit entfernt. Dennoch liegt das Potenzial des Abschnitts 4 zu kommunikativen Aktivitäten genau in diesem Bereich: Es kann bei vielen Fragen zum Thema, welche Aspekte in welchen Kategorien von Sprachhandlungen wie relevant sind, nützliche Denkanstöße geben. Die Beispielskalen dieses Abschnitts können, wenn auch nicht auf solider empirischer Basis,

[214] Vgl. hierzu auch die Ausführungen in Kapitel 1.3.3 dieser Arbeit.

mögliche Abstufungen illustrieren beziehungsweise können die dort erfolgten Niveauzuweisungen als Diskussionsanlass dienen.

3.4.3.3 GER-Skalen zu kommunikativen Sprachkompetenzen

Im Bereich der kommunikativen Sprachkompetenzen (vgl. GER-Abschnitt 5) werden exemplarisch die Skala zu *Orthographie* (ebd.: 118), die beiden Skalen zu *Wortschatzspektrum* und *Wortschatzbeherrschung* im Vergleich (ebd.: 112f), und schließlich die Skala *Kohärenz und Kohäsion* (ebd.: 125) analysiert. Um die Verwendbarkeit einzuschätzen, werden wiederum die Aussagen aus GER-Abschnitt 3.8 herangezogen, die sich konkret auf die Skalen der GER-Abschnitte 4 und 5 beziehen. Auffallend ist, dass sich in Abschnitt 5 – anders als bei Abschnitt 4 – keine konkreten Aussagen zur Verwendbarkeit der dortigen Beispielskalen finden. Lediglich auf S. 110 des GER wird der vermutlich auf alle Skalen der sprachlichen Kompetenzen abzielende Hinweis gegeben: „Fortschritte, die Lernende bei der Nutzung sprachlicher Mittel machen, lassen sich in Skalen fassen; sie werden im Folgenden in dieser Form dargestellt: [dann folgt die erste Beispielskala des Abschnitts 5]“. Demnach beschreiben die Skalen des GER-Abschnitts 5 im nicht nur die kommunikativen Sprachkompetenzen, sondern auch Lernfortschritte.[215] Die folgende Analyse soll unter anderem zeigen, ob diese Behauptung zutrifft.

Bei der Analyse der **Skala *Orthographie*** werden die oben vorgestellten Analysekategorien angesetzt, mit der folgenden Abweichung: Zur Kategorie *Situationen* gibt es keine Aussagen in der Skala, weshalb sie nicht genutzt wird; stattdessen finden sich teils *konkrete Beispiele*, so dass diese in einer eigenen Kategorie erfasst werden. Die zugehörigen Tabellen sind in Anhang 16 dieser Arbeit dargestellt.

- Gegenstandsbereich: Auf S. 117 des GER wird *Orthographie* knapp definiert: Sie umfasst die „Kenntnis der Symbole, aus denen geschriebene Texte bestehen sowie die Fertigkeit, sie wahrzunehmen und zu produzieren.“ Das in diesem Bereich relevante rezeptive wie produktive Wissen umfasst Buchstaben, die korrekte Schreibweise von Wörtern, Zeichensetzung und typographische Konventionen. Dies ist eine brauchbare Beschreibung des Gegenstandsbereichs, die jedoch nicht stringent in der Skala umgesetzt wird, wie folgende Befunde zeigen:

- Strukturierung: An der Skala *Orthographie* fällt auf, dass der Bereich der Orthographie nicht sauber kategorisiert wurde: Die unter orthographischen Gesichtspunkten beschriebenen Operationen des *Abschreibens*, des *Buchstabierens* sowie der *Beachtung* oder *Einhaltung von Konventionen der Rechtschreibung und Zeichensetzung* fallen sicherlich unter den Bereich der orthographischen Fertigkeiten. Doch das *Schreiben von zusammenhängenden, durchgängig* respektive *klar verständlichen Texten* (B1 resp. B2) wäre zutreffender dem Bereich der Schreibfertigkeit allgemein oder konkreter der Fertigkeit, kohärente Texte zu erstellen, zugeordnet

[215] Zur Diskussion der beiden Perspektiven „Momentaufnahme unterschiedlich weit entwickelter Kompetenzen“ versus „Lernfortschritte über bestimmte Zeiträume hinweg“ darf auf Kapitel 3.2.4 dieser Arbeit verwiesen werden.

worden, genau wie die textuellen Aspekte der *Gestaltung* und der *Gliederung in Absätze* (B2 und C1). Zudem ist die Struktur innerhalb der einzelnen Niveaus nicht kohärent: Es leuchtet zwar ein, dass der Bereich der Zeichensetzung in den unteren Niveaus noch nicht vorhanden ist; letztere werden jedoch wenigstens durch konkrete Beispiele dessen, was wiedergegeben werden kann, illustriert. Wieso aber auf den Niveaus B1 mit C1 die Bereiche *Rechtschreibung, Zeichensetzung, Textgliederung und Absatzeinteilung* in je anderen Kombinationen in einem oder zwei Deskriptoren beschrieben werden und nicht mehr durch Beispiele illustriert werden, ist nicht nachvollziehbar. Auf C2 wird nur mehr die *orthographische Fehlerfreiheit von schriftlichen Texten* beschrieben. Die Interpretation der für die jeweiligen Niveaus typischen Fehler im Bereich der Orthographie bleibt den Benutzern ebenso überlassen wie die Interpretation dessen, was generell unter *Orthographie* verstanden wird.

- Lücken: Es fehlen nähere Angaben zum Wortschatz, der korrekt geschrieben werden kann, seien diese nun thematischer Art, oder bezogen auf die Frequenz oder Vertrautheit der Ausdrücke, oder bezogen auf die inhärente Schwierigkeit der Wörter (ein Wort wie *has* oder *name* dürfte aufgrund der Tatsache, dass es so geschrieben wird wie es ausgesprochen wird, leichter zu schreiben sein wie beispielsweise *gorgeous* oder *coughing*, deren Aussprache sich für deutsche Lernende signifikant von der Schreibung unterscheidet). Lediglich auf A1 und A2 finden sich Qualifizierungen wie *kurze, vertraute Wörter* beziehungsweise *kurze Sätze*. Ab dem Niveau B1 werden, wie gesagt, keine Beispiele mehr gegeben, obwohl diese die Bedeutung der Niveaus erhellen könnten. Eine weitere Lücke zeigt sich auf dem Niveau C2: Dort finden sich neben der *orthographischen Fehlerfreiheit* keine Aussagen bezüglich dessen, was auf diesem Niveau beherrscht wird.

- Sprache: Es finden sich Termini wie *nicht notwendigerweise übliche Rechtschreibung* oder *orthographische Fehler*, die im Licht dessen, was alles in der Skala unter dem Bereich Orthographie beschrieben wird, definiert werden müssten, um vergleichbar interpretiert werden zu können. Die verbalen Abstufungen sind nicht immer einleuchtend: Die unteren Niveaus werden durch Adjektive wie *kurz* und *vertraut* charakterisiert; sie werden zusätzlich durch Beispiele konkretisiert. Auf den Niveaus A2 mit C2 wird der Aspekt der Korrektheit durch die Ausdrücke *einigermaßen akkurat – exakt genug (so dass Texte meistens verständlich sind) – hinreichend korrekt (aber Einflüsse der Muttersprache können sich zeigen) – richtig (abgesehen von gelegentlichem Verschreiben) – frei von orthographischen Fehlern* abgestuft. Durch die Zusätze (hier in Klammern dargestellt) ist die verbale Abstufung verständlich und scheint intuitiv logisch. Allerdings fällt auch unter dem Aspekt der Versprachlichung auf, dass der Deskriptor, der C2 repräsentiert, aus dem Rahmen fällt: Er definiert das Können über die Fehlerfreiheit von Texten, ohne relevante Bereiche des Könnens positiv zu formulieren. Überhaupt finden sich in dieser Skala nicht durchgängig positive Aspekte, so etwa auf A2: *benutzt dabei nicht notwendigerweise die übliche Rechtschreibung*; auch das Prinzip der Kann-Beschreibung ist auf den Niveaus B1 mit C2 nicht stringent eingehalten.

- Verwendbarkeit: In diesem Zusammenhang interessiert die Frage, was in der Skala beschrieben wird, wenn es nicht nur das Können ist. Teils lesen sich die Deskriptoren eher wie eine Beschreibung textueller Merkmale, die jedoch, wie gerade dargestellt, nicht alle dem Bereich der Orthographie zuzuordnen sind. Der Beschreibungsbereich ist demnach nicht sauber kategorisiert und die Basis der Beschreibung ist nicht nachvollziehbar. An dieser Stelle sei noch einmal darauf verwiesen, dass die Skala nicht im oben erwähnten Projekt konstruiert wurde, ihre Kategorisierung nicht in den Workshops überprüft wurde und ihre Deskriptoren nicht mithilfe des Rasch-Modells skaliert wurde. Diese Skala kann man als Beispiel dafür anführen, was passieren kann, wenn eine Skala ohne empirische Grundlagen entwickelt wird. Aufgrund dieser Mängel wird von einer Beurteilung ihrer Verwendbarkeit abgesehen, denn es ist offensichtlich, dass diese Skala nicht in den Einsatzbereichen, wie sie in Abschnitt 3.8 des GER angegeben werden, verwendet werden kann: Ausgehend von der Skala *Orthographie* können weder orthographische Kompetenzen oder Lernfortschritte auf valider Basis beschrieben werden, noch Bewertungskriterien entwickelt oder Erwartungen an das Können auf einem bestimmten Niveau formuliert werden; auch Lernziele sollten von dieser Skala mangels valider Beschreibungen nicht abgeleitet werden. Zuerst müsste der Bereich der Orthographie definiert werden, die Niveaus aus empirischen Beschreibungen von Lernertexten, Lernfortschritten und/oder Theorien bezüglich der Entwicklung der orthographischen Fertigkeiten abgeleitet werden und im Idealfall die Deskriptoren empirisch kalibriert werden.

Wenden wir uns nun der **vergleichenden Analyse der Skalen *Wortschatzspektrum* und *Wortschatzbeherrschung*** zu. Dabei werden die oben unter Kapitel 3.4.3 vorgestellten Analysekategorien angesetzt; nur auf dem obersten Niveau werden die *Einschränkungen* fallen gelassen zugunsten der Aufteilung der Kategorie *Was/ Wie* (vgl. die Tabellen in Anhang 17 dieser Arbeit). Folgende Befunde ergeben sich:

- Beschreibungsgegenstand: Im GER wird die lexikalische Kompetenz definiert als „die Kenntnis des Vokabulars einer Sprache, das aus lexikalischen und aus grammatischen Elementen besteht, sowie die Fähigkeit, es zu verwenden“ (GER 2001: 111). Dort werden diese lexikalischen und grammatischen Elemente mit konkreten Beispielen erläutert, so dass die Terminologie meist transparent und verständlich wird. Wenn es dennoch zu Terminologieproblemen kommt, sind diese in der folgenden Analyse dargestellt. Der Bereich der lexikalischen Kompetenz wird in den beiden Skalen unter den Gesichtspunkten des Umfangs respektive der Korrektheit in der Verwendung in folgender Weise abgedeckt:

- Strukturierung: In der Skala *Wortschatzspektrum* wird der Umfang des zur Verfügung stehenden Wortschatzes stringent auf allen Niveaus beschrieben, wobei die unteren Niveaus A1 mit B1 durch die Angabe von Situationen und Themen konkretisiert werden und der Umfang des Wortschatzes durch Einschränkungen bestimmt wird; auf den oberen Niveaus nimmt der

Umfang zu und wird charakterisiert durch *Umschreibungs*- und *Variationsmöglichkeiten*, die ab B1 respektive B2 auftreten, sowie ab C1 durch Kenntnisse *idiomatischer Ausdrücke* und *umgangssprachlicher Wendungen*. Einzig die Operation des *Beherrschens* eines bestimmten Umfangs auf C1 und C2 gehört nicht in die Skala *Wortschatzspektrum*, denn dazu gibt es ja die Skala *Wortschatzbeherrschung*. Ansonsten sind die Abstufungen in der Skala *Wortschatzspektrum* intuitiv nachvollziehbar und logisch, dennoch müssten auch sie durch empirische Beobachtungen und Beschreibungen bestätigt werden. Die Strukturierung der Skala *Wortschatzbeherrschung* hingegen ist nicht ganz so stringent: Auf A1 gibt es keinen Deskriptor, da sich auf diesem Niveau wohl noch keine Aussagen zur Wortschatzbeherrschung machen lassen. Auf A2 mit B2 ist die *Beherrschung* des Wortschatzes beschrieben, abgestuft durch Einschränkungen wie *begrenzter Wortschatz* (A2) *Grundwortschatz* (B1) oder *großer Wortschatz* (B2) – insofern kommt auch bei der Skala *Wortschatzbeherrschung* der Umfang mit herein, der beherrscht wird. Des Weiteren sind die Niveaus A2 und B1 charakterisiert durch Situationen und Themen, die Niveaus B1 mit C1 jedoch auch durch Fehlleistungen – gerade der Deskriptor auf C1 leistet keinen Beitrag zur Positivformulierung. C2 schließlich ist gekennzeichnet durch *korrekte und angemessene Verwendung* – ob jedoch die Angemessenheit erst ab C2 gegeben ist, müsste ebenfalls durch empirische Untersuchungen belegt werden.

Betrachten wir nun die Kohärenz innerhalb der Niveaus über die beiden Skalen hinweg: Auf A2 *verfügt* man *über ausreichend/genügend Wortschatz*, um sich zu *Vertrautem, Elementarem* äußern zu können – ein *begrenzter Wortschatz* im Rahmen *konkreter Alltagsbedürfnisse* wird *beherrscht*; abgesehen davon, dass die Adjektive einen breiten Interpretationsrahmen lassen, impliziert der Vergleich der beiden Skalen, dass der Umfang des bekannten Wortschatzes größer sein dürfte als der Umfang, der produktiv beherrscht wird – eine intuitiv nachvollziehbare Aussage, die auch der Erfahrung entspricht (vgl. beispielsweise Bleyhl & Timm 1998: 269), dennoch müsste auch sie empirisch belegt werden. In ihren Aussagen zu *Themen* zeigen sich die beiden Skalen auf B1 kohärent: Wortschatz ist vorhanden in Bezug auf *die meisten Themen des eigenen Alltagslebens* (Skala *Wortschatzspektrum*), doch bei *wenig vertrauten Themen* können Fehler auftreten (Skala *Wortschatzbeherrschung*). Auf dem Niveau B2 haben die Deskriptoren der beiden Skalen keine offensichtlichen Bezugspunkte, können jedoch als komplementär betrachtet werden: der Umfang ist *groß*, *Formulierungen können variiert werden*, und die Beherrschung wird über die *Verwendung mit großer Genauigkeit* definiert, ohne dass Aussagen über den Umfang des Beherrschten (wie es etwa auf A2 und B1 geschieht) getroffen werden. Auch die Deskriptoren auf C1 und C2 zeigen keine Berührungspunkte. Auf diesen Stufen wird der Aspekt der Beherrschung in der Skala *Wortschatzspektrum* beschrieben, so dass es scheint, man habe in der Skala *Wortschatzbeherrschung* andere Aspekte betrachtet, namentlich *Fehlleistungen im Gebrauch* auf C1 und die *angemessene und korrekte Verwendung* auf C2. Gerade die beiden obersten Niveaus sollten sauber kategorisiert werden, um die beiden Perspektiven des Umfangs und der Beherrschung kohärent darzustellen.

- Lücken: Nicht immer wird in der Skala *Wortschatzbeherrschung* auf den Umfang des beherrschten Wortschatzes eingegangen, doch wenn die beiden Skalen als komplementär betrachtet werden, so stellt dies kein Problem dar. Problematisch ist jedoch die Lücke auf C1 in der Skala *Wortschatzbeherrschung*, denn dort wird nichts über Beherrschung ausgesagt. Hilfreich wären auch konkrete Beispiele (für die verschiedenen Sprachen, in denen der GER vorliegt), um den Wortschatz eines Niveaus zu illustrieren – doch dies dürfte ein Unterfangen sein, das sich nur im Lauf der Zeit durch Verwendung und Überprüfung der Skalen in der Praxis realisieren lassen wird.

- Sprache der Deskriptoren: Die oben schon angedeuteten Interpretationsprobleme ergeben sich trotz der vorgeschalteten Definition. Auch bei diesen Skalen tritt die bekannte Synonym-Problematik auf: Sind beispielsweise Ausdrücke wie *gute Beherrschung* und *Verwendung mit großer Genauigkeit* gleich zu verstehen? Haben die „Wortpaare" *Verwendung* und *Wortgebrauch* oder *Genauigkeit* und *Angemessenheit* einen semantischen Deckungsbereich? Daneben sind auch die verbalen Abstufungen nicht immer selbsterklärend (man denke etwa an *ausreichend – ausreichend groß – groß*), werden aber durch Angaben von Einschränkungen, Bedingungen, Situationen und Themen hinreichend verdeutlicht.

Nicht alle Deskriptoren genügen dem Kriterium der Positivformulierung: So werden beispielsweise in der Skala *Wortschatzbeherrschung* auf B1 und C1 negative Konzepte beschrieben, um den Grad an Beherrschung näher zu bestimmen – dies scheint aber (nicht zuletzt aufgrund der Charakteristika der Interimsprache) ein tragfähiges Vorgehen, da nicht nur Negativformulierungen genutzt werden.

- Verwendbarkeit: Der Bereich der lexikalischen Kompetenz wird aus den Perspektiven des Wissens und der Anwendbarkeit, des Könnens beschrieben – ob diesen Beschreibungen jedoch empirische Beobachtungen zugrunde liegen oder ob sie auf Erfahrung, Intuition oder Erwartungen beruhen, bleibt unklar: Die Aussagen bezüglich des Umfangs des Wortschatzes (*Grundwortschatz, großer Wortschatz*, etc.) könnten aus der Erfahrung oder aus Performanzanalysen gewonnen worden sein; die Charakteristika der Wortschatzproduktion könnten aus der Beschreibung von Lernerproduktionen stammen (wie etwa das Auftreten von *Umschreibungen* oder *Variationen*); die beschriebenen Fehlleistungen könnten ebenfalls aus Performanz- respektive Fehleranalysen oder aber aus Interimsprachanalysen gewonnen worden sein.

Wie lassen sich diese Skalen nun im System des GER (vgl. GER-Abschnitt 3.8) einordnen? Beide Skalen beschreiben einen begrenzten Teilbereich, ein Detail innerhalb des Bereichs der sprachlichen Kompetenzen, so dass sie schon deswegen auf einer relativ tiefen Ebene im System des GER angesiedelt sein müssen. Die Skala *Wortschatzspektrum* geht dabei noch mehr ins Detail als die Skala *Wortschatzbeherrschung*; dies dürfte aber in der Natur des Gegenstands liegen: Wenn bezüglich des Wortschatzes erst der Umfang detailliert beschrieben ist, so lassen sich Aussagen zur Beherrschung darauf beziehen. In beiden Skalen finden sich sowohl abstrakte

Darstellungen als auch Konkretisierungen, letztere gerade im Hinblick auf Themen (vgl. etwa *Wortschatzspektrum*: B1). Die Aussagen sind jedoch immer generalisierend: Nirgends findet sich eine konkrete Niveauzuweisung von sprachlichen Mitteln, und es dürfte schwer sein, eine konkrete Performanz auf ein bestimmtes Niveaus einzustufen. Insofern stellen diese Skalen einen eng umrissenen Bereich des Sprachvermögens aus zwei sich ergänzenden Perspektiven (was wird gekonnt – wie gut wird es gekonnt) dar und können auf derselben Ebene angesiedelt werden. Aufgrund ihres Grads an Generalisierung können sie in dieser Form eine holistische Beschreibung der lexikalischen Kompetenz (die Aspekte Wissen und Anwendbarkeit betreffend) geben, doch auch für sie gilt, dass die Skalen nicht für Zwecke verwendet werden sollten, für die sie nicht geschaffen sind. Da auch bei diesen Skalen die Basis der Beschreibung – wie gerade dargestellt – nicht transparent ist, können sie zwar für eine grobe (Selbst- wie Fremd-) Einschätzung der lexikalischen Kompetenzen genutzt werden, doch wiederum nur, wenn die Lernenden den Einschätzenden hinreichend bekannt sind. Die Angaben beispielsweise zu (Umschreibungs-)Strategien oder Fehlleistungen können zwar diagnostische Hinweise geben, doch für eine detaillierte Diagnose sind die Skalen weder komplex noch konkret genug. Dasselbe gilt für den Zweck der Aufgabenerstellung mithilfe dieser Skalen. Lediglich an Aussagen zu Themen oder Situationen könnten Aufgaben ausgerichtet werden – eine m. E. nicht ausreichende Basis.

Um die Zwecke abzudecken, die in GER-Abschnitt 3.4 vorgeschlagen werden (vgl. oben, Entwicklung von kriterienorientierter Beurteilung, Beschreibung tatsächlicher Lernerfolge, oder Darstellung realistischer Lernziele), müssten diese Bereiche zuvor empirisch untersucht und die Ergebnisse den Skalenniveaus valide zugeordnet werden, seien es nun Erwartungen an Leistungen oder die über einen bestimmten Zeitraum beobachteten Lernfortschritte oder die Charakteristika von Momentaufnahmen tatsächlicher Leistungen. Im derzeitigen Zustand sind die Deskriptoren zu generell und abstrakt, um aus ihnen Skalen für konkrete Zwecke abzuleiten. Sie können jedoch als Grundlage für einen Abgleich von „neuen“ Skalen dienen, die auf bestimmte Zwecke hin konstruiert werden: Diese zu konstruierenden Skalen müssen selbstverständlich auf valider Basis stehen und ihren Beschreibungsgegenstand offen legen; dann können durch einen Abgleich beispielsweise Merkmalsausprägungen identifiziert werden, die sich auch in GER-Deskriptoren finden. So können Niveauzuweisungen von Merkmalen in den „neuen“ Skalen und den betreffenden GER-Skalen gegenseitig validiert werden.

Abschließend wird aus dem Bereich der pragmatischen Kompetenzen die Analyse der **Skala *Kohärenz und Kohäsion*** vorgestellt. Dabei werden die drei oben erläuterten Analysekategorien *Operation, Was* und *Wie* angesetzt. Daneben treten zwei weitere Kategorien: die Kategorie *konkrete Beispiele*, da zumindest auf den unteren Niveaus solche gegeben werden, und die Kategorie *Zweck*, da meist beschrieben wird, zu welchem Zweck eine bestimmte Operation ausgeführt werden kann (vgl. die Tabellen in Anhang 18). Die Analyse bringt folgende Ergebnisse:

- Beschreibungsgegenstand: Auf S. 123 des GER wird die Diskurskompetenz als ein Unteraspekt der pragmatischen Kompetenzen definiert, im Sinne der Fähigkeit, kohärente Textpassagen erstellen zu können. Dazu werden Wissen und Fähigkeiten hinsichtlich beispielsweise der Textstrukturierung, der thematischen Entwicklung, der Reihenfolge oder der „Kohärenz und Kohäsion“ (ebd.) gerechnet – wenn nun aber Kohärenz unter anderem mit sich selbst und Kohäsion gar nicht definiert wird, so lässt sich der Beschreibungsgegenstand der Skala *Kohärenz und Kohäsion* nur aus der Skala selbst ableiten: Dort finden sich die Aspekte der Verknüpfung/Verbindung von Worten, Wortgruppen und Sätzen durch Konnektoren (Kohäsion), des inhaltlichen Zusammenhangs eines Textes, der Gliederung und Strukturierung von Texten und der Anwendung sprachlicher wie inhaltlicher Verknüpfungsmittel – insgesamt ist der Bereich der Kohärenz damit grob abgedeckt. Betrachten wir nun aber die Kategorisierung und die Darstellung der einzelnen Niveaus, um die Strukturiertheit der Skala beurteilen zu können:

- Strukturierung: Die Aspekte der Kohärenz und Kohäsion sind nicht über die gesamte Skala hinweg stringent beschrieben: Auf A1 und A2- wird lediglich die Verwendung von *Konnektoren zur Verknüpfung von Wörtern oder Wortgruppen*, also der Einsatz von Kohäsionsmitteln auf Satzebene, beschrieben und nützlicherweise durch die Angabe von konkreten Beispielen illustriert. Es scheint logisch, dass auf diesen Niveaus Kohärenz auf Textebene noch nicht erzielt werden kann. Auf den Niveaus A2+ mit B2- wird der Aspekt der *Satzverknüpfung* thematisiert und mittels der respektiven Charakterisierungen *einfache Aufzählung – lineare, zusammenhängende Äußerung – klarer, zusammenhängender Text* abgestuft. Auf A2+ und B2- werden Kohäsionsmittel erwähnt, auf B1 jedoch unerklärlicherweise nicht. B2+ hingegen wendet sich wieder den *Verknüpfungswörtern zur Verdeutlichung inhaltlicher Beziehungen* zu, sagt jedoch nichts aus über sonstige Aspekte der Kohärenz. C1 fällt insofern aus dem Rahmen als dort auf *klares, sehr fließendes und gut strukturiertes Sprechen* eingegangen wird – die anderen Niveaus unterscheiden nicht zwischen mündlicher und schriftlicher Produktion. Des Weiteren wird erst auf C1 die *Beherrschung der Mittel der Gliederung und Verknüpfung* thematisiert – auf den unteren Niveaus werden lediglich Umfang der Mittel und deren Verwendungszweck beschrieben. C2 schließlich schließt ab mit zwei Deskriptoren, die einer Kurzdefinition von Kohärenz und Kohäsion gleichkommen: Es wird die *Erstellung eines gut gegliederten und zusammenhängenden Texts unter angemessenem Einsatz einer Vielfalt von Gliederungs- und Verknüpfungsmitteln* thematisiert.

Der Blickwinkel der Niveaus ist demnach je ein anderer: Im unteren Bereich geht es um Konnektoren auf Satzebene, im mittleren um Satzverknüpfungen auf Textebene und im oberen Bereich um angemessene Verwendung von Mitteln und Beherrschung der kohärenten Texterstellung – wenn sich diese Progression in der Realität belegen lässt, so ist dagegen nichts einzuwenden. Die Abstufung erscheint insgesamt nicht kontra-intuitiv; die Niveaus sind nicht nur sprachlich abgestuft, sondern sie werden durch Angaben zum Umfang der Mittel und zu Textcharakteristika die Kohärenz betreffend zusätzlich qualitativ abgestuft. Doch auch für diese Skala gilt,

dass ihre Deskriptoren und deren Abstufungen einer empirisch fundierten Beschreibungsgrundlage bedürfen.

- Lücken: Die Skala zeigt, wie oben unter Strukturierung schon angedeutet, einige Lücken: Mag die Abwesenheit der Kohärenzbeschreibung auf den unteren Niveaus noch nachvollziehbar sein, so tragen das Fehlen der Konnektoren auf B1, die Nicht-Thematisierung relevanter Kohärenzaspekte wie Gliederung oder Struktur des Textes auf B2+ und die fehlende Charakterisierung des Umfangs an Kohäsionsmitteln auf C1 nicht zur Handhabbarkeit der Skala bei.

- Sprache der Deskriptoren: Die Termini zur Bezeichnung der Kohäsionsmittel (*Konnektoren, Verknüpfungsmittel, Verknüpfungswörter, Mittel der Gliederung*) könnten in der der Skala vorangestellten Kurzdefinition des Beschreibungsgegenstands erhellt und durch Beispiele (ähnlich wie beim Wortschatz, vgl. oben) illustriert werden.

In Bezug auf den Umfang der Mittel werden im untersten Bereich der Skala die abstufenden Adjektive *sehr einfache* (*Konnektoren*) und *einfache* (*Konnektoren*) durch Beispiele gestützt; im weiteren Verlauf der Skala jedoch wird dieses Prinzip nicht stringent durchgehalten: Es finden sich die Abstufungen *die häufigsten* (A2+), *begrenzte Anzahl* (B2-), *verschiedene* (B2+) und schließlich *Vielfalt* (C2), doch ab A2+ werden keine konkreten Beispiele mehr gegeben.

Die Abgrenzung der beiden obersten Niveaus wird durch die Verbalisierung nicht unbedingt erleichtert: Wie unterscheidet sich die *Beherrschung der Mittel der Gliederung sowie inhaltlichen und sprachlichen Verknüpfung* vom *angemessenen Einsatz einer Vielfalt von Mitteln zur Gliederung und Verknüpfung*? Doch insgesamt betrachtet werden die Abstufungen meist durch die Angabe von textuellen Merkmalen transparent.

- Verwendbarkeit: Die Skala beschreibt in Form generalisierender positiver Kann-Aussagen die Fertigkeiten im Bereich der Verwendung von Kohäsions- und Kohärenzmitteln. Diese Beschreibungen könnten, in Analogie zu den bisher analysierten Skalen, entweder auf Intuition, auf Erfahrungen, auf Beobachtungen des Lernfortschritts oder auf der empirischen Beschreibung von schriftlichen wie mündlichen Lernerperformanzen beruhen. Aufgrund des Gegenstands, der den produktiven Fertigkeiten zuzuordnen ist, und des Grades an Generalisiertheit kann die Skala auf einer mittleren Ebene der produktiven Fertigkeiten im GER-System angesiedelt werden, unterhalb der globalen Ebene, denn die Diskurskompetenzen wirken auf alle produktiven Sprachprozesse ein. Insofern die Kohäsionsmittel in ihrem Umfang beschrieben sind, könnten diese auf einer noch tieferen Ebene angesetzt werden, vergleichbar der, auf der die sprachlichen Kompetenzen angesiedelt sind.

Aufgrund des unklaren Status der Deskriptoren und des Grades an Generalisiertheit gilt wie bei den anderen Skalen, die kommunikative Sprachkompetenzen beschreiben, dass ein konkreter Verwendungszweck nicht angegeben werden kann. Zur Selbsteinschätzung und zur globalen Beurteilung, wenn die Lernenden nicht nur aufgrund einer einzigen Leistung beurteilt

werden, kann die Skala hilfreich sein, doch für alle weiteren Zwecke müsste sie erst an der Realität überprüft werden.

3.4.4 Fazit: Der Skalen-Ansatz des GER und seine Verwendungsmöglichkeiten

Im Folgenden werden die Ansprüche, die im GER an die Deskriptoren gestellt werden, zusammengestellt und mit den obigen Analyseergebnissen kontrastiert, um den Status der Deskriptoren einschätzen zu können. Auf dieser Basis werden dann die generellen Aussagen im GER zur Verwendung seines Skalensystems beurteilt, ehe abschließend die in GER-Abschnitt 9.2 genannten, konkret auf Sprachbeurteilungskontexte bezogenen Einsatzmöglichkeiten der GER-Skalen untersucht werden.

3.4.4.1 GER-Deskriptoren: Ansprüche und Realität

Im GER finden sich an verschiedenen Stellen Ansprüche an die Sprache der Deskriptoren, die effektive Kompetenzbeschreibungen darstellen wollen (vgl. GER 2001: 39), Ansprüche an das Referenzsystem, das alle relevanten Aspekte des Sprachvermögens abbilden will, und Aussagen bezüglich der Verwendungsmöglichkeiten der Skalen und des Referenzsystems. Letztere basieren auf den Ansprüchen an die Sprache der Deskriptoren und an das System, in welchem sie angesiedelt sind. Wenden wir uns deshalb zunächst zusammenfassend den Aspekten der Sprache der Deskriptoren zu, ehe die Ansprüche an das Referenzsystem und dessen Funktionen beleuchtet werden. Auf S.39 des GER wird behauptet:

> Die Deskriptoren erfüllen die in Anhang A umrissenen Kriterien für effektive Kompetenzbeschreibungen: Sie sind kurz, klar und transparent sowie positiv formuliert; sie beschreiben etwas Bestimmtes und sind unabhängig, d. h. sie können für sich alleine stehen und ohne Bezug auf die Formulierungen in anderen Deskriptoren verstanden und interpretiert werden.

Diese Kriterien werden nun auf dem Hintergrund der obigen Skalenanalyse geprüft:

- *Kürze der Formulierungen*: Das Kriterium der Kürze wird im GER auf S. 201 als nicht „länger als 25 Wörter“ definiert. Die Wortanzahl kann jedoch nicht alleine über die Kürze entscheiden – auch die Anzahl der Merkmale, die in einem Deskriptor beschrieben werden, spielt hier mit herein. Es werden holistische Skalen kritisiert, die das Typische eines Niveaus in einem umfangreichen Deskriptor erfassen wollen (ebd.: 201). Solche Deskriptoren lassen sich jedoch auch im GER ausmachen: Beispielsweise sind in der oben besprochenen Globalskala teils mehrere Merkmale in einem Deskriptor zusammengefasst (vgl. etwa die Globalskala, dort jeweils die letzten Deskriptoren der Niveaus B1 und B2). Bei solchen Deskriptoren ist keine eindeutige Ja/Nein-Entscheidung, wie sie auf S.201 des GER gefordert wird, möglich: Eventuell kann man *über Erfahrungen berichten*, aber (noch) keine *Begründungen* abgeben. Es gibt jedoch eine ganze Anzahl Deskriptoren, die den Kriterien der Kürze (und damit auch der Ja/Nein-

Entscheidungsmöglichkeit) genügen (vgl. beispielsweise die Deskriptoren der Skala *Wortschatzspektrum* auf A1 und A2). Die Benutzer der Skalen müssen sich die Deskriptoren, mit denen sie arbeiten wollen, jeweils einzeln betrachten und entscheiden, ob die dort gemeinsam beschriebenen Merkmale auch in dieser Kombination beobachtbar respektive feststellbar sind.

- *Klarheit und Transparenz der Beschreibungen*: Auf S. 201 des GER wird *Klarheit* über *Verständlichkeit und Transparenz*, und letztere über das *Fehlen von Jargon* definiert. Die exemplarisch aufgezeigten Terminologieprobleme und die Inkonsistenzen bei den verbalen Abstufungen widersprechen teils der Forderung nach Transparenz und Verständlichkeit. Fachjargon wird sich nicht ganz vermeiden lassen, könnte aber über eine Definition der verwendeten Termini verständlich gemacht werden. Auch eine „klare logische Struktur", wie sie auf S. 201 des GER gefordert wird, ist wie gezeigt, nicht immer gegeben. Um diese zu erreichen, müssten die Niveaus in sich und über die Skala hinweg sauber strukturiert werden. Damit und mit einer Terminidefinition dürfte sich die Forderung nach Klarheit erfüllen lassen.

- *Transparenz der Deskriptoren* im Sinne von *Bestimmtheit, Eindeutigkeit*: Im GER findet sich auf S. 201 die Forderung, vage Formulierungen und rein verbale Abstufungen zu vermeiden. Wie gezeigt sind viele Deskriptoren jedoch so generell und auf solch abstraktem Niveau, dass sie vage erscheinen und einer Interpretation bedürfen: Wie sind beispielsweise die Verben in Formulierungen wie *kann sich verständigen – kann sich ausdrücken* zu verstehen? Sind sie synonym verwendet oder haben sie einen unterschiedlichen Bedeutungsbereich? Was hat man unter den *meisten Situationen auf Reisen* zu verstehen? Wie ist das Adjektiv *angemessen* zu interpretieren? Die Probleme im Zusammenhang mit den verbalen Abstufungen sind ausführlich in der obigen Skalenanalyse dargestellt – jedoch werden sie sich nie ganz vermeiden lassen, wie unter Kapitel 3.2.3 dieser Arbeit schon erläutert, weshalb wiederum für eine Terminologiedefinition plädiert wird. Dem Kriterium der Bestimmtheit/Eindeutigkeit können demnach nicht alle Deskriptoren gerecht werden. Die zum Teil generellen oder vereinfachenden Aussagen, die bei einer globalen Beschreibung der kommunikativen Kompetenzen oder des allgemeinen Sprachstands teils unumgänglich sind, sind diesem Beschreibungsgegenstand jedoch auch angemessen. Denn je nach Ausrichtung und Zweck einer Skala muss generalisiert werden, oder es muss konkret spezifiziert werden – ein und dieselbe Skala kann nicht allen Anforderungen an alle Ausrichtungen gerecht werden. Es scheint, die GER-Skalen sollen zu viele Zwecke abdecken, denen die momentanen Formulierungen nicht gerecht werden können. Es ist unmöglich, alle Kriterien der Versprachlichung in allen Skalenarten erfüllt zu wollen, da sich die angemessenste Versprachlichung oft aus Kompromissen ergibt (vgl. die Ausführungen in Kapitel 3.2.3 dieser Arbeit) Deshalb wäre es sinnvoll, notwendige Kompromisse auch offen zu legen, statt generell zu behaupten, dass alle prototypischen Anforderungskriterien an die Versprachlichung auch erfüllt seien.

- *Positive Formulierungen*: Diesem Kriterium werden diejenigen Deskriptoren gerecht, die als positive KANN-Aussagen formuliert sind. Wie oben gezeigt, gibt es jedoch Deskriptoren, die von diesem Schema abweichen, wie etwa die Formulierung der *Orthographie*-Skala auf C2 oder die

der Skala *Wortschatzbeherrschung* auf C1. Auch für dieses Kriterium gilt, dass es nicht zwingend für alle Skalenarten angesetzt werden muss: Wie in Kapitel 3.2.3 dieser Arbeit ausgeführt, gibt es Kontexte, in denen es Bewertern erleichtert wird, konkrete Merkmale in Performanzen zu identifizieren, wenn sie negativ formuliert sind; in einer *reporting scale* jedoch, die das Können in generalisierter Form beschreibt, können diese Merkmale in positive Formulierungen überführt werden. Insofern die Skalen des GER *proficiency scales* darstellen, ist die Positivformulierung angemessen, doch wenn in den GER-Skalen davon abgewichen wird, so sollten diese Abweichungen auch begründet werden. Sonst könnten die diesbezüglichen Aussagen im GER an Kraft verlieren und den Eindruck von leicht widerlegbaren Tatsachenbehauptungen erwecken.

- *Unabhängigkeit der Deskriptoren*: Diese Forderung widerspricht im Grunde der Forderung nach Kohärenz eines Niveaus und Kohärenz einer Skala über alle Niveaus hinweg. Die Deskriptoren im GER können durchaus für sich alleine stehen, doch auch für sie gilt, dass erst alle Deskriptoren eines Niveaus zusammen dessen Bedeutung illustrieren: Die Niveaus selbst sollen zwar voneinander unabhängig sein, dennoch erhalten sie ihre Bedeutung auch im Vergleich zu den anderen Niveaus einer Skala. Die Ableitung auf S. 201 des GER jedoch, dass Deskriptoren allein aufgrund der Erfüllung des Kriteriums der Unabhängigkeit als „eigenständiges Lernziel dienen“ könnten, scheint nicht zwingend: Ein Deskriptor kann natürlich als Lernziel-Check genutzt werden, wenn er ein solches Ziel auch thematisiert und konkret beschreibt. Man kann aus einem Deskriptor wie *Kann die Hauptpunkte verstehen, wenn klare Standardsprache verwendet wird und wenn es um vertraute Dinge usw. geht* das Lernziel „Hauptpunkte werden verstanden, wenn es um vertraute Dinge aus Arbeit, Schule oder Freizeit geht“ ableiten – allerdings muss noch thematisiert werden, in welchen Kontexten (rezeptiv oder interaktiv) und in welcher Art von geschriebenen oder gesprochenen Texten. Wie jedoch soll beispielsweise ein so vager Deskriptor wie *Kann sich spontan und fließend äußern* als Lernziel umgesetzt werden, wenn nicht Situationen und Themen näher bestimmt werden? Lernziele sind, wie in Kapitel 1.3.3 dieser Arbeit erläutert, zu unterscheiden nach ihrer fachunabhängigen, fächerübergreifenden respektive fachlegitimierenden Ausrichtung. Auf dieser Basis können Lernziele wie gesagt in Grob-, Richt- und Feinziele differenziert werden. Die GER-Deskriptoren müssten zunächst unterschieden werden hinsichtlich der Lernzielebene(n), die sie abdecken. Sodann muss eine solide Auswahl getroffen werden, welche Deskriptoren als Lernziele in welchem Rahmen und in welcher Konkretisierungsstufe dienen können und welche nicht, etwa weil sie nicht operationalisierbar sind.

3.4.4.2 Der Status der GER-Deskriptoren

Die Deskriptoren werden im GER (2001: 32) vorgestellt als Referenzniveaus, die einen gemeinsamen Rahmen bilden, der bei der Beschreibung von Kompetenzniveaus und beim Vergleich verschiedener Qualifikationssysteme helfen soll. Es werden vier Kriterien angeführt, die Skalen eines solchen Referenzrahmens erfüllen sollten: (a) *Kontextfreiheit*, um Raum für Generalisierungen zu

lassen; dennoch so viel *Kontextrelevanz*, um für alle Funktionen geeignet zu sein, die die Skalen übernehmen sollen. Auf diese scheinbare Widersprüchlichkeit wird gleich im Anschluss noch eingegangen. (b) *Verankerung der Beschreibungen in Theorien der Sprachkompetenz*: Wie unter Kapitel 3.4.1 dieser Arbeit gezeigt wurde, basieren die Kategorien des GER auf solchen Theorien, so dass dieses Kriterium erfüllt ist. Dennoch bleibt die Forderung nach empirischer Basis der Beschreibungen bestehen. (c) Die *vertikale Kalibrierung* der Niveaus soll auf *objektiven Messverfahren* beruhen – auch dieses Kriterium ist erfüllt, doch hier gilt ebenso wie bei der Basis der Beschreibungen, dass die vertikale Zuordnung der Deskriptoren zusätzlich empirisch belegt werden muss. (d) Die *Anzahl der Niveaus* soll geeignet sein, Fortschritte zu beschreiben und hinreichende Differenzierung zu ermöglichen. Dieses Kriterium dürfte durch das oben vorgestellte „flexible Verzweigungssystem" des GER erfüllt sein (vgl. ebd: 33f und 40ff).

Welche Aussagen bezüglich des Status der Skalen aufgrund ihres Beschreibungsgegenstands werden daneben getroffen? Auf S. 39 des GER findet sich die Behauptung, dass die Deskriptoren eine „Sammlung von genau beschriebenen kriterienbezogenen Aussagen zum Kontinuum der fremdsprachlichen Kompetenz" darstellen. Wie jedoch an den oben analysierten Skalen gezeigt, kann von *genauer* Beschreibung nicht immer die Rede sein; oft gibt es inkonsistente Beschreibungen und auf die Terminologieprobleme darf noch einmal verwiesen werden.

Die Aussagen der Deskriptoren sind auf *Kriterien* bezogen: Je nach Gegenstand einer Skala, also je nach der dort beschriebenen horizontalen (Teil-)Dimension, sind ein oder mehrere für diesen Bereich relevante Kriterien in der Skala dargestellt. Die Kriterien basieren, wie gesagt, auf Modellen der kommunikativen Kompetenz und des Sprachvermögens und wurden von Experten überprüft – insofern sind sie theoretisch fundiert und an Erfahrungen aus der Praxis validiert. Ein Problem im Zusammenhang mit den Kriterien in den Skalen könnte sich nur ergeben, wenn die Kriterien nicht auf allen Niveaus „durchdekliniert" werden: Dies kann zu inkonsistenten Beschreibungen eines Kriteriums führen, wie oben bei den Skalenanalysen gezeigt.

Da die Skalen Aussagen zum *Kontinuum der fremdsprachlichen Kompetenz* (oder zutreffender zur *proficiency*) wiedergeben sollen, legt dies ihre Verwendungsmöglichkeiten fest: Es handelt sich wie schon gesagt um generalisierende, holistische Aussagen, die nicht für eine direkte Anwendung in konkreten Kontexten geeignet sind. Insofern zeichnen sie sich durch eine gewisse Kontextfreiheit aus. Für jeden Anwendungsfall muss eine angemessene Adaption auf den jeweiligen Kontext hin erfolgen. Dabei muss die Realität etwa auf Basis empirischer Beschreibungen, Beobachtungen oder Analysen mit hereingebracht werden, um die Skalen in den jeweiligen Anwendungskontexten zu validieren. Da die GER-Skalen und damit das Referenzsystem insgesamt – bedingt durch die Konstruktionsmethoden – das Verständnis und die Annahmen der an der Konstruktion Beteiligten widerspiegeln und die Basis der Deskriptoren nicht empirisch validiert wurde, haben die Deskriptoren einen eher hypothetischen Charakter: Sie stellen Annahmen dar, wie das Sprachvermögen, die *proficiency*, als Momentaufnahme über das gesamte Spektrum des Könnens hinweg beschrieben werden kann; die Deskriptoren

beruhen jedoch nicht auf der empirischen Beobachtung von Lernfortschritten oder Kompetenzzuwachs über längere Zeiträume hinweg.

Diese beiden Perspektiven, die Beschreibung des Sprachvermögens als Momentaufnahme und die Beschreibung von Fortschritten in der Entwicklung des Sprachvermögens, sind, wie oben bereits erläutert, nicht direkt ineinander überführbar. Auch wenn jede skalare Darstellung eines Könnensbereichs ein Anwachsen, eine Zunahme des in der Skala beschriebenen Könnens impliziert, kann dennoch nicht von einer Momentaufnahme des Spektrums an zunehmendem Können auf Lernfortschritte Einzelner rückgeschlossen werden, denn es ist nicht belegt (und auch zumindest in den oben analysierten Skalen nicht beschrieben), wie individuelle Lernende ihr Sprachvermögen jeweils entwickeln.

Wenn nun die GER-Skalen das Sprachvermögen in generalisierender Form aus möglichst vielen relevanten Perspektiven auf verschiedenen Niveaus beschreiben wollen, um einen Referenzrahmen mit Orientierungspunkten zu schaffen (vgl. GER 2001: 32 und 34), so hat diese abstrakte und generalisierende Beschreibungsperspektive, die ihre Basis zudem in existenten Skalen und nicht in der Empirie hat, ihren Preis: Die GER-Skalen sind so weit dekontextualisiert worden, dass sie m. E. nicht wieder rückführbar sind in Skalen für konkrete Zwecke in konkreten Verwendungskontexten, wie es im GER behauptet wird (vgl. ebd.: 32): Diese Kompetenzskalen (zutreffender: *scales of proficiency*) können nicht mehr „auf alle nur denkbaren relevanten Kontexte bezogen“; sie sind, wie in der obigen Skalenanalyse gezeigt, nicht generell „…für die Funktionen geeignet (…), die sie dort [in allen denkbaren relevanten Kontexten] übernehmen sollen“ (ebd.: 32). Zudem muss man sich bewusst machen, was solch eine Verwendung in allen relevanten Funktionen hinsichtlich des Nachweises der angemessenen Formulierung (vgl. ebd.: 46) bedeutet: Jede Formulierung müsste für sich auf ihre Angemessenheit im jeweiligen Verwendungskontext überprüft werden, sowohl hinsichtlich der Gegenstände oder der Merkmalsausprägungen, die sie auf einem bestimmten Niveau beschreibt, als auch hinsichtlich der dabei verwendeten Sprache. Erst auf dieser Basis könnten Deskriptoren adaptiert werden auf konkrete Verwendungskontexte.

3.4.4.3 Verwendungsmöglichkeiten der GER-Skalen generell

Aufbauend auf die obigen Überlegungen folgt die zusammenfassende Beurteilung der Aussagen im GER zu den Verwendungsmöglichkeiten seines Skalensystems: Im Folgenden werden überblicksartig alle generellen Aussagen im GER zusammengestellt und kommentiert, die sich auf die Verwendbarkeit seines Skalensystems und dessen Deskriptoren beziehen. Im Anschluss daran werden die konkreten Aussagen des GER-Abschnitts 9.2 zur Verwendbarkeit der Skalen in der Beurteilung herausgegriffen und auf Basis des bisher Gesagten diskutiert.

- *Formulierung von Erwartungen an die Niveaus/Hilfe bei Lernzielbeschreibung*: Die GER-Skalen sollen helfen, Erwartungen an die Niveaus (GER 2001: 27) und an das, was Lernende mit der Fremdsprache tun können oder wissen sollten (ebd.: 51), zu formulieren. Auf Basis dieser Erwartungen sollen die Deskriptoren „realistische Lernziele darstellen" (ebd.: 39) beziehungsweise „zur Entwicklung von transparenten und realistischen Beschreibungen von globalen Lernzielen beitragen" (ebd.: 27). Hierbei ist wie gesagt Vorsicht geboten, denn es könnte sich um einen Zirkelschluss handeln, wenn die Niveaus aufgrund von Erwartungen und/oder Erfahrungen im Zusammenhang mit Lernzielen definiert wurden (man denke an diejenigen Quellskalen, die Erwartungen an Prüfungsniveaus formulieren, oder an die Lehrenden, die ihre Erwartungen und Erfahrungen ebenfalls mit in die Skalierung haben einfließen lassen) und nun wiederum zur Formulierung von Erwartungen und/oder Lernzielen herangezogen werden sollen. Wie oben bei den Analysen gezeigt, sind viele Deskriptoren aufgrund des Grades an Abstraktion und aufgrund der mangelnden empirischen Basis nicht geeignet, Lernziele darzustellen, da sie diese nicht beschreiben. Inwieweit man mit ihrer Hilfe realistische Lernziele ableiten kann, muss die Praxis zeigen. Es scheint nach dem momentanen Stand der Dinge ratsam, Lernziele unabhängig vom Referenzsystem des GER zu entwickeln und sie posthoc mit dem Referenzrahmen abzugleichen, um sie beispielsweise in einem gegeben Bildungssystem transparent und vergleichbar darzustellen.

- *Beschreibung von Lernfortschritten und Leistungen*: Die Deskriptoren werden im GER (ebd.: 39) als „relevant für die Beschreibung … tatsächliche(r) Lernerfolge" charakterisiert. Des Weiteren wird ausgesagt, dass die Skalen der Abschnitte 4 und 5 Lernfortschritte dadurch beschreiben würden, dass die Fertigkeiten der Lernenden „auf verschiedenen, aufeinander folgenden Niveaus" (ebd.: 131) dargestellt werden. Dieselbe Aussage findet sich in GER-Abschnitt 5, direkt vor der ersten dortigen Skala (ebd.: 110; vgl. oben bei den Analysen in Kapitel 3.4.3.3 dieser Arbeit). Es trifft zwar zu, dass die Skalen Fertigkeiten beschreiben, doch implizieren diese abgestuften Fertigkeitsbeschreibungen auch empirisch beobachtete Lernfortschritte? Eine auf Erwartungen oder Erfahrungen beruhende Beschreibung verschiedener abgestufter Fertigkeitsmerkmale muss nicht zwingend auf Lernfortschritte deuten. Auch wenn die Deskriptoren u. a. über die Einschätzung realer Lernender skaliert wurden, so wurden sie doch nicht benutzt, um Lernfortschritte über längere Zeiträume einzuschätzen. Es mag sein, dass die Referenzniveaus nützlich sein können bei der Bestimmung von Lernfortschritten, doch die Skalen beschreiben diese nicht direkt und können mangels angemessener empirischer Fundierung der Beschreibungen und mangels Nachweis, dass die Formulierungen dieser Funktion angemessen wären, nicht direkt für die Beschreibung von Lernfortschritten genutzt werden. Zur Beschreibung des generellen Leistungstands können die Skalen aufgrund ihres Grades an Generalisiertheit als benutzerorientierte *reporting scales* hilfreiche Dienste leisten.

- *Beurteilung von Leistungen und/oder Lernfortschritten*: Im GER findet sich die Aussage, dass die Sprachkompetenzbeschreibungen des GER hilfreich sein können bei der Beurteilung

von Lernleistungen und Fortschritten in der Sprachkompetenz (ebd.: 28), ebenso wie sie helfen können, Erwartungen an Kompetenzniveaus in Tests und Prüfungen zu formulieren (ebd.: 32). Die Deskriptoren können laut Ausführung des GER (ebd.: 39) genutzt werden zur Entwicklung kriterienorientierter Beurteilungen. Auf S. 46 des GER dann wird noch einmal auf die Verwendung in der Beurteilung hingewiesen, welche in GER-Abschnitt 9.2 eigens behandelt wird. Schließlich wird festgestellt, dass es wichtig sei, „bei Skalen zwischen (a) der Definition von Stufen der Sprachkompetenz (wie in den Gemeinsamen Referenzniveaus) und (b) der Bewertung von erreichten Leistungen in Bezug auf Ziele auf einem bestimmten Niveau zu unterscheiden“ (ebd.: 49). Die Autoren des GER stellen deutlich klar, dass die GER-Skalen nicht (direkt) zur Bewertung genutzt werden können – wie sie im größeren Kontext der Beurteilung eingesetzt werden können, wird gleich im Anschluss dargestellt. An dieser Stelle der Arbeit sei nur darauf verwiesen, dass Lernleistungen ein Indikator für die Sprachkompetenz sind und zur Beurteilung der beiden Aspekte *Leistung/Performanz* und *Kompetenz* je anders formulierte Beschreibungen benötigt werden.

- *Hinweise auf Sequenzierung*: Folgende Aussage des GER weist darauf hin, dass die Skalen Ordnungsprinzipien implizieren können (ebd.: 148):

> Der *Referenzrahmen* ersetzt keine Grammatikbücher und bietet keine strenge Reihenfolge an (obwohl das Skalieren eine Auswahl und somit einige globale Sequenzierungen beinhalten kann); er stellt jedoch einen Rahmen für die Entscheidungen der Praktiker dar, die sie anderen mitteilen wollen.

Letztere Funktion kann der Referenzrahmen sehr wohl erfüllen, doch da die Skalen nicht auf Spracherwerbstheorien oder empirischen Beobachtungen von Erwerbssequenzen beziehungsweise auf Analysen unterrichtlicher Progression und deren Auswirkungen auf das Lernen beruhen, sollte man mit Aussagen zur Sequenzierung, die man aus diesem Referenzsystem ableiten könnte, vorsichtig sein. An dieser Stelle darf nochmals auf die Ausführungen unter Kapitel 3.4.1.4 dieser Arbeit und das Problem des dort erörterten Zirkelschlusses verwiesen werden: Wenn implizite Vorstellungen der Lehrenden zu Progression und Sequenzierung mit in die Skalierung eingeflossen sind, so ist die nun implizit in den Abstufungen vorhandene Sequenzierung nicht als empirisch belegt zu verstehen – vielmehr müssen auch in diesem Bereich die impliziten Annahmen, die hinter den Abstufungen stecken, zuerst an der Realität überprüft werden.

- *Abstimmung von Lernzielen und Lernmaterialien auf Lernfortschritte und Progression*: Das Referenzsystem kann gemäß GER hilfreich sein bei der Berücksichtigung von Progression und Lernfortschritten sowie bei der Abstimmung von Lernzielen und Lernmaterialien auf Progression und Fortschritte (ebd.: 28). Das Kategoriensystem des Abschnitts 4 wird als „Checkliste“ für Kursplaner, Lehrwerksentwickler, Lehrende und Prüfende vorgestellt (ebd.: 51). Dies wäre eine brauchbare Anwendungssituation, wenn mithilfe eines Systems alle relevanten Komponenten des Lehr- und Lernprozesses aufeinander bezogen werden könnten. Doch dürfte sich dies mit dem derzeitigen System noch nicht erreichen lassen, denn wie erläutert beschreiben die Skalen weder Lernfortschritte noch Progression auf valider empirischer Basis; auch Lernziele lassen

sich nur an solchen Deskriptoren festmachen, die konkret genug sind und solche auch beschreiben; dazu tritt das Problem, dass es schwer sein dürfte, beispielsweise konkrete sprachliche Mittel oder Texte unterschiedlicher Schwierigkeitsstufen einem bestimmten Niveau zuzuweisen.

- *Hilfe bei der Erstellung von Curricula/Abschlussprüfungen*: Der GER verweist darauf, dass eine „Skala von klar definierten Niveaustufen" für „praktische Zwecke, wie z. B. die Erstellung sprachlicher Curricula oder für Abschlussprüfungen" (ebd.: 28) nützlich sein kann. Nähere Ausführungen dazu finden sich in Abschnitt 8 des GER. Dort wird das Taxonomiesystem der Abschnitte 4 und 5 des GER als Rahmen vorgeschlagen, innerhalb dessen Curriculumplaner ihre Ziele und Entscheidungen ansiedeln könnten. Als Orientierungspunkte können die Kategorien und Deskriptoren vielleicht hilfreiche Dienste leisten, doch sollten Curricula und Abschlussprüfungen tunlichst in ihren Kontexten entwickelt und begründet werden, ehe sie auf den Referenzrahmen bezogen werden.

- *Vergleichbarkeit*: Das Referenzsystem soll „den Vergleich von Lernzielen, Niveaustufen, Materialien, Tests und von Lernerfolgen in unterschiedlichen Systemen und Situationen" erleichtern (ebd.: 28), ebenso wie „den Vergleich zwischen verschiedenen Qualifikationssystemen" (ebd.: 32). Auch dabei gilt, dass nur Beschreibungen desselben Gegenstands verglichen werden können bzw. dass Materialien oder Tests zuerst valide in ihren jeweiligen Kontexten entwickelt und verankert werden müssen, ehe sie in einem übergeordneten Referenzsystem eingeordnet und auf dessen Basis verglichen werden können. Derzeit dürften diese Vergleiche noch nicht auf valider Basis durchführbar sein. An dieser Stelle darf auf die unten folgenden Ausführungen zur Testanbindung an die Niveaus des GER verwiesen werden.

- *Formulierung von Standards*: Nicht zuletzt kann das Referenzsystem helfen, im Rahmen der Schulaufsicht Standards zu formulieren oder die Zusammenarbeit verschiedener Bildungssektoren zu erleichtern (GER 2001: 32). Doch auch für diese Zwecke muss das Referenzsystem auf valider Basis stehen und die Standards müssen außenvalidiert werden, das heißt sich an Kriterien der realen (Bildungs-)Welt überprüfen lassen.

3.4.4.4 Verwendung der GER-Skalen bei der Beurteilung des Sprachvermögens

Die folgenden Ausführungen konzentrieren sich auf die Verwendung der GER-Skalen in der Beurteilung, da diese Funktion im GER besonders betont wird.[216] Auffallend ist, dass im GER grundsätzlich die Rede von Verwendbarkeit im generellen Kontext der *Beurteilung* ist – erst in Abschnitt 9.2 wird von *Beurteilung und Bewertung* gesprochen, wobei *Bewertung* sich auf konkrete Auswertungen konkreter Tests bezieht – wie oben analysiert, sind die Skalen jedoch für

[216] Auf die Beurteilungslastigkeit des Dokuments wurde an verschiedenen Stellen dieser Arbeit hingewiesen; an dieser Stelle darf verwiesen werden auf die beiden Beurteilungsraster zu Beginn des die Skalen einführenden Abschnitts 3 im GER; auf den Hinweis zur Verwendung der Skalen in der Beurteilung gleich zu Beginn des GER-Abschnitts 3.8, der sich mit der Nutzung der Skalen beschäftigt; und nicht zuletzt auf den GER-Abschnitt 9.2, der sich explizit der Verwendung der Skalen in der Beurteilung widmet.

konkrete Verwendungszwecke nicht unbedingt geeignet. Betrachten wir die drei Möglichkeiten aus Abschnitt 9.2 des GER, die bereits in Kapitel 2.5.4 dieser Arbeit vorgestellt wurden, näher:

(a) *Die inhaltliche Beschreibung von Tests und Prüfungen*: Sie soll abgeleitet werden aus GER-Abschnitt 4 bezüglich der kommunikativen Aktivitäten, der Kontexte der Sprachverwendung und der Texte, die in einem Test verwendet werden. Daneben soll GER-Abschnitt 7 zur Rolle und Schwierigkeit kommunikativer Aufgaben helfen, die angemessenen Aufgaben für den jeweiligen Test zu bestimmen. GER-Abschnitt 5.2 soll die inhaltliche Basis für das Erstellen von Testaufgaben stellen, ergänzt um konkrete Lernzielbeschreibungen auf den verschiedenen Niveaus des Europarats, wie sie etwa in der Bibliographie des GER zu seinem Abschnitt 2 aufgeführt sind. Die Begründung dieser Aussage, dass die Skalen des Abschnitts 4.4 zur Testspezifikation, die Skalen des Abschnitts 5.2 hingegen zur Testerstellung geeignet sein sollen, bleiben die Autoren des GER jedoch schuldig.

Wie oben in Kapitel 2 dieser Arbeit ausgeführt, sollte eine Analyse der Erwartungen, Testinhalte und Aufgabenstellungen im Zusammenhang mit Prüfungen immer im jeweiligen Prüfungskontext erfolgen. Innerhalb diesem müssen die Inhalte der zu entwickelnden Tests und Prüfungen verankert und validiert werden. Sind Anforderungen an Tests hinsichtlich der abzudeckenden Inhalte, Aufgaben, Formate und Schwierigkeiten charakterisiert, so können diese Spezifikationen mit dem Referenzsystem abgeglichen werden. Wenn gemäß dieser Spezifikationen Testitems entwickelt worden sind, so kann das Referenzsystem des GER helfen, diese Items zu einzustufen. Doch es kann aufgrund seiner generellen Natur eben nur ein Referenzpunkt, eine Orientierungshilfe sein und nicht die Ausgangsbasis konkreter inhaltlicher Testbeschreibung oder Testentwicklung darstellen. Bezogen auf die Skalen des GER würde diese Funktion sie als aufgabenorientierte Skalen darstellen – wie jedoch in der Skalenanalyse gezeigt wurde, sind die meisten Skalen dafür nicht konkret genug, ganz abgesehen davon, dass sie nicht Aufgaben in deren charakteristischen Merkmalen beschreiben, sondern Kompetenzen, Strategien und Aktivitäten.

Im erwähnten *Dutch CEF Construct Project*, aus dem der *Dutch Grid* resultiert, wurde ebenfalls festgestellt, dass der GER in seinem momentanen Zustand zur inhaltlichen Spezifizierung von Tests nicht geeignet ist: “[I]t is far from clear that the still relatively abstract Can-Do descriptors in the CEF can be turned into items that illustrate or exemplify the different CEF levels.“ (Alderson et al. 2004: 3) Es folgt der Hinweis, dass im DIALANG-Projekt (vgl. hierzu auch Anhang C im GER) dieselbe Erfahrung gemacht wurde: “The experience of the DIALANG project was that is was necessary to develop additional specifications before the CEF could be used as the basis for test development“ (ebd.). Im *Dutch Project* wurden wie gesagt die relevanten Skalen zu Lesen und Hörverstehen analysiert, Probleme und Lücken identifiziert und letztere durch die Ergänzung relevanter Aspekte in Bezug auf die Charakterisierung von Testitems geschlossen. Der so entstandene *Grid* wurde angewandt zur Analyse einer Reihe von Testspezifikationen

und Testitems, um seine Tragfähigkeit zu untersuchen. Er hat sich erwiesen als "…useful instrument for the description of test items and tasks in terms of the CEF." (ebd.: 17). Er wird vorgestellt als "…framework, based on the … CEF, for analysing language test items, test tasks and test specifications, in order to help test developers relate their examinations to the CEF" (ebd.: 20). Der *Dutch Grid* ist web-basiert und Testitems können mit seiner Hilfe unter www.ling.lancs.ac.uk/cefgrid oder unter http://www.ealta.eu.org/dutch/grid.htm im System des GER qualitativ eingestuft werden, indem das Niveau abgeschätzt wird, auf welchem die Testitems am wahrscheinlichsten anzusiedeln sind.

Die an der Entwicklung des *Dutch Grid* Beteiligten mahnen jedoch zur Vorsicht bei der Entwicklung von Testitems auf einem bestimmten Niveau des GER: "The empirical research that we have been able to conduct (…) suggests that the CEF does not provide sufficient guidance to enable item writers to develop tests at specific levels of the CEF." (Alderson et al. 2004: 21). Es wird Bezug genommen auf das erwähnte *Manual*, das bei der Anbindung von Tests an das Niveausystem des GER helfen soll: Der *Dutch Grid* kann als Ergänzung zum *Manual* betrachtet werden, das Tests ebenfalls unter anderem über inhaltliche Spezifikationen an den GER anbindet. Das *Manual* wird gleich im Anschluss unter Kapitel 3.5 dieser Arbeit besprochen.

(b) *Kriterien für das Erreichen eines Lernziels*: Die zweite Verwendungsmöglichkeit der GER-Skalen im Kontext der Bewertung wird auf S. 174 des GER vorgestellt: Die Skalen werden als „Quelle bei der Entwicklung von Bewertungsskalen" angeboten, wobei es um die Bewertung des Erreichens von Lernzielen geht, welche wiederum sehr breit ausgelegt werden (es kann sich bei solch einem Lernziel um ein Referenzniveau insgesamt handeln oder aber um ein sehr spezifisches Lernziel, vgl. GER 2001: 174). Bis dahin war im GER die Rede davon, dass die Skalen hilfreich sein können bei der Identifizierung von Kriterien in der Beurteilung – dass aus ihnen jedoch konkrete Bewertungsskalen abgeleitet werden sollen, ist ein „neuer" Aspekt, der erst in Abschnitt 9.2.2 des GER angesprochen wird. Dort wird nun unterschieden zwischen den Skalen der Abschnitte 4 (*Aktivitäten*) und 5 (*Kompetenzen*) des GER.

Wie oben bereits dargestellt, wird im GER vorgeschlagen, die Skalen des Abschnitts 4 zur „*Beurteilung* durch Lehrende und zur *Selbstbeurteilung* im Hinblick auf realitätsbezogene Aufgaben" (GER 2001: 174) zu nutzen – wie bei den Skalenanalysen gezeigt, ist dies eine Funktion, die diese Skalen übernehmen können, wenn die zu Beurteilenden den Beurteilenden hinreichend bekannt sind. Im Kontext der Selbstbeurteilung darf auf die Bedeutsamkeit der Skalen bei der Entwicklung von Sprachenportfolios verwiesen werden – ein zukunftsweisender Ansatz in der Selbstbeurteilung, dessen Analyse und Bewertung jedoch den Rahmen der vorliegenden Arbeit sprengen würde.

Die des Weiteren angeführte Funktion der *Aufgabenerstellung* auf Basis der Deskriptoren für kommunikative Aktivitäten (die schon in GER-Abschnitt 9.2.1 abgehandelt wird und unerklärlicher-

weise in Abschnitt 9.2.2 wieder aufgegriffen wird, vgl. ebd.: 175) ist gerade schon als nicht erfüllbar zurückgewiesen worden. Daneben tritt die Funktion der *Rückmeldung* von Ergebnissen (ebd.) – insoweit in den gerade angesprochenen Beurteilungskontexten nicht einzelne Leistungen bewertet werden, sondern das Sprachvermögen generell beschrieben werden soll, kann die Beschreibung dieses generalisierten Könnens auch zur Rückmeldung genutzt werden.

In Bezug auf die Skalen des GER-Abschnitts 5, welche die linguistischen Kompetenzen beschreiben, finden sich auf S. 175f des GER ebenfalls die Verwendungsfunktionen der *Selbstbeurteilung* und der *Beurteilung durch Lehrende* in Form von Checklisten, wofür positive und unabhängige Deskriptoren nötig sind – wie oben bei den entsprechenden Skalenanalysen gezeigt, erfüllen nicht alle Deskriptoren der Beispielskalen diese Anforderungen, weshalb jeweils im Einzelfall betrachtet werden muss, welche Deskriptoren in welche Beurteilungsskala aufgenommen werden können. Des Weiteren wird im GER angegeben, dass die Kompetenzskalen des GER-Abschnitts 5 des in der „Beurteilung von *Performanzen*“ als „Ausgangspunkte für die Entwicklung von Beurteilungskriterien“ genutzt werden können und dass die Deskriptoren für diesen Zweck entweder als „Sprachkompetenzskala“ oder als „Bewertungsskala für Prüfungen“ zusammenstellen und präsentieren kann (ebd.: 176): Erstgenannte „Sprachkompetenzskala“ wird charakterisiert als eine Skala, die zu bestimmten Kategorien ausgewählte Niveaus präsentiert, auf denen dann eine Leistung eingestuft wird. Wie man jedoch eine einzelne Performanz auf Kompetenzniveaus einordnen will, bleibt fraglich, zumal in den Skalenanalysen gezeigt wurde, dass die Kategorien nicht immer stringent in den Skalenniveaus beschrieben sind, die Basis der Beschreibungen nicht transparent ist und es sich dabei um generalisierte Kompetenzbeschreibungen handelt. Diese können, wie in Kapitel 3.1 und 3.4.3.3 dieser Arbeit gezeigt, nicht genutzt werden, um konkrete Performanzen zu bewerten; dazu müssten die Skalen schon konkrete Performanzmerkmale beschreiben.[217] Letztgenannte „Bewertungsskala für Prüfungen“ erweist sich im Zug der Beschreibung auf S. 176 des GER als „standard-orientierte“ Skala: Zu jeder relevanten Kategorie in einer gegebenen Prüfung soll ein Deskriptor aus den Skalen des GER-Abschnitts 5 ausgewählt oder selbst definiert werden, der den Standard des Bestehens der Prüfung widerspiegelt. Neben diesem Deskriptor können weitere Deskriptoren etwa aus angrenzenden Niveaus der jeweiligen Ursprungsskala (soweit denn eine Skala aus GER-Abschnitt 5 zugrunde gelegt wurde) zur Beschreibung von Leistungen unterhalb oder auch überhalb des Standards herangezogen werden. Unverständlich an diesem Vorschlag ist, dass an dieser Stelle des GER die Sprachkompetenzskalen als geeignet zur Bewertung dargestellt werden, wo doch an anderer Stelle im GER (ebd.: 49) auf die Bedeutsamkeit der Unterscheidung dieser beiden Skalentypen hingewiesen wird. Abgesehen von dieser Widersprüchlichkeit finden sich in den Kompetenzskalen des GER-Abschnitts 5 keine konkreten Deskriptoren, die auf einer empirischen Beschreibung von Performanzmerkmalen beruhen und somit zur direkten Bewertung

[217] Unerklärlich ist in diesem Zusammenhang, wieso nur in Bezug auf die Skalen zu kommunikativen Aktivitäten der Hinweis erfolgt, dass diese nicht zur Bewertung einer einzelnen Leistung genutzt werden könnten (GER 2001: 174f), wo dies doch wie gerade gezeigt auch auf die Skalen der Sprachkompetenzen zutrifft.

einer solchen Performanz herangezogen werden könnten. Ehe GER-Deskriptoren in eine Bewertungsskala übernommen werden können, müsste ihre Angemessenheit hinsichtlich des Beschreibungsgegenstands, seiner Abstufungen und der verwendeten Sprache nachgewiesen werden. Es darf auf Kapitel 4.4 dieser Arbeit und die dort vorgestellten *rating scales* verwiesen werden.

(c) *Beschreibung der Kompetenzniveaus in Tests und Prüfungen als Hilfe bei Vergleichen*: Die GER-Skalen werden in ihrer dritten Verwendungsmöglichkeit als Metasystem eingesetzt: Dabei geht es darum, verschiedene Systeme vergleichbar zu machen, indem auf Basis des Metasystems des GER gemeinsame Standards diskutiert und im Idealfall gefunden und beschrieben werden können, die über verschiedene Prüfungen hinweg vergleichbar interpretiert werden. Im Rahmen einer solch standard-orientierten Beurteilung sind *benchmarks*, also „Beispiele von Arbeiten in Bezug auf standardisierte Definitionen“, wie es der GER auf S. 177 beschreibt, von großer Bedeutung bei der Herausbildung eines gemeinsamen Verständnisses. Auch die Entwicklung der *benchmarks*, ihre Diskussion und Zuordnung zu den Niveaus des Referenzsystems, muss in der Praxis erfolgen – es darf noch einmal auf Kapitel 4 dieser Arbeit verwiesen werden. Das Referenzsystem der Skalen und Deskriptoren will ein Begriffsraster darstellen, das helfen soll, „nationale und institutionelle Systeme vermittels des *Referenzrahmens* aufeinander zu beziehen“ und „die Ziele bestimmter Prüfungen und Kursmodule vermittels der skalierten Referenzniveaus abzubilden“ (GER 2001: 177). Doch wie im GER an dieser Stelle eingeräumt wird, ist die Entwicklung eines gemeinsamen Verständnisses solcher Standards langwierig und muss in der Praxis erfolgen durch Prozesse der Veranschaulichung und des Austausches von Meinungen (ebd.). Dort muss sich zeigen, wie praktikabel dieses Metasystem „als Mittel zum Herstellen von Bezügen“ (ebd.: 187) ist und wo seine Lücken und Schwächen sind.[218]

Zusammenfassend kann die Verwendbarkeit der GER-Skalen wie folgt beurteilt werden: Das Skalensystem in seinem derzeitigen Stand wirft einen generellen und damit eher kontextfreien Blick auf das, was Sprachvermögen ausmacht, auf die „gemeinsame Schnittmenge“ der vielen relevanten Bereiche und Perspektiven des Sprachvermögen – dieser Blick ist eines Referenzsystems durchaus angemessen. Deshalb dürfte sich der Verwendungsbereich der GER-Skalen allgemein eher mit *Referenz- und Bezugsrahmen für extern entwickelte Skalen oder Systeme* angeben lassen als mit *Ausgangsbasis zur Entwicklung von Skalen für konkrete Verwendungszwecke*.

Sucht man jedoch nach einer Möglichkeit, Skalen für konkrete Zwecke aus den Deskriptoren des GER zu entwickeln, so darf vorsichtig auf die oben bei der Analyse der beiden Beurteilungsraster des GER-Abschnitts 3 angedeutete Möglichkeit verwiesen werden, Skalen aus dem Deskriptoren-Pool des Schweizer Konstruktionsprojekts (in welchem die GER-Skalen konstruiert wurden) zu „kombinieren“: Die erwähnte Deskriptorensammlung könnte bei der Entwicklung

[218] Vgl. in diesem Kontext auch Alderson 2002 und das erwähnte *Manual* (Council of Europe 2003a), das gleich im Anschluss vorgestellt wird.

spezifischer Skalen wertvolle Hilfe leisten: Wenn in dieser Datenbank kategorisierte und kalibrierte Deskriptoren zusammengestellt werden, die zusätzlich bezüglich der Basis ihres Beschreibungsgegenstands gekennzeichnet sind, so könnten daraus je nach Beschreibungsbasis Skalen für konkrete Zwecke abgeleitet respektive zusammengestellt werden. Neben die generalisierenden Deskriptoren könnten auch solche treten, welche beispielsweise konkrete Textmerkmale oder spezifische Aspekte von Performanzen auf empirischer Basis beschreiben und dann etwa zur Bewertung eben solcher Performanzen oder Texte herangezogen werden könnten. Ob dies machbar ist, müsste wiederum die Praxis zeigen. In solch einem Ansatz könnten (neu zu entwickelnde) Beispielskalen illustrieren, wie eine Skala auf einen bestimmten Ausschnitt hin für spezifische Zwecke aus solch einer Datenbank konstruiert werden könnte. Dann hätten auch die vier Orientierungen des GER-Abschnitts 3.8 ihre Berechtigung, wenn es zu jeder der Orientierungen auch Deskriptoren mit nachweislich angemessenen Formulierungen gäbe. So aber werden die existierenden Beispielskalen des GER als Quelle zur Entwicklung und Ableitung spezifischer Skalen vorgestellt – eine Funktion, die sie nicht übernehmen können.

Der GER-Skalenansatz ist aufgrund seiner Kontextfreiheit notgedrungenerweise für eine konkrete Anwendung in der Praxis zu generell und grobkörnig. Er kann jedoch als Reflexionsanstoß, als Mittel zur Überprüfung eigener Systeme, zum Abgleichen einmal entwickelter Skalen oder Beurteilungssysteme und als Metasystem in der Beschreibung und Beurteilung dessen, was Sprachvermögen ausmacht, nützlich sein.

3.5 Testanbindung an die Niveaus des GER: Das *Manual*

Um Tests, Prüfungen oder Qualifikationssysteme an das System des GER anzubinden, bedarf es detaillierterer Anleitung als sie in GER-Abschnitt 9 gegeben wird. Dieser Aspekt wird jedoch durch das erwähnte *Manual*[219] abgedeckt, das sich explizit mit technischen Fragen der Testanbindung an die Niveaus des Referenzrahmens beschäftigt. Beim Dokument des *Manual* handelt es sich um einen vorläufigen Entwurf, der derzeit in Projekten pilotiert und erprobt wird, welche sich mit der Anbindung konkreter Tests oder Prüfungen an das System des GER beschäftigen. Die praktischen Erfahrungen aus der Pilotierung sollen zur Verbesserung und Weiterentwicklung des *Manual* beitragen; eine revidierte Fassung des *Manual* wird voraussichtlich Ende 2005/Anfang 2006 veröffentlicht (vgl. Council of Europe 2003b: 12). Die Beurteilung des *Manual* in der vorliegenden Arbeit erfolgt aus theoretisch-didaktischer Perspektive. An dieser Stelle darf noch einmal darauf hingewiesen werden, dass psychometrische Messmodelle nicht Gegenstand dieser Arbeit sind und somit die Angemessenheit der technischen Verfahren der Anbindung im *Manual* nicht beurteilt werden kann.

[219] Vgl. Council of Europe 2003a.

Testanbindung ist grundsätzlich nur möglich, wenn dem anzubindenden Test und dem Referenzrahmen dasselbe Konstrukt zugrunde liegt, genau wie verschiedene Tests nur dann verglichen oder zur Validierung genutzt werden können, wenn sie dasselbe Konstrukt messen.[220] Das Referenzsystem des GER ist im momentanen Zustand nicht in Testitems operationalisierbar, weil die Beschreibungen des Sprachvermögens wie oben gezeigt kontextfrei und generell gehalten sind, keine konkreten Inhalte abbilden, nicht auf eine spezifische Sprache bezogen sind, und weil es sich wie oben gezeigt bei den Skalen eher um benutzerorientierte *reporting scales* als um aufgabenorientierte Skalen zur Testkonstruktion handelt. Daher kann das Referenzsystem zwar als Instrument zum Kommunizieren von generalisierten Testergebnissen in Form bedeutungsvoller Deskriptoren genutzt werden, ist jedoch zur Anbindung von Tests nicht hinreichend. Dazu bedarf es komplexer Prozesse, die transparent und zuverlässig dokumentiert werden müssen: "The existence of such a relation is not a simple observable fact, but is an assertion for which the test constructor has to provide sufficient evidence, theoretical as well as empirical." (*Manual* 2003: 99)

Zu folgenden vier Aspekten der Anbindung gibt das *Manual* Anleitung: Da vor jeder Nutzung des Referenzsystems eine gewisse Vertrautheit mit seiner Struktur, seinen Kategorien und Niveaus erreicht werden muss, wird die so genannte *familiarisation* mit dem GER (im Folgenden mit *Familiarisierung* wiedergegeben) als eigener Aspekt behandelt. Darüber hinaus finden sich Ausführungen zur *test specification* (im Folgenden *Spezifizierung* genannt), der Beschreibung des Testkonstrukts, der Testinhalte und Gegenstandsbereiche in Relation zu den Kategorien und Niveaus der Abschnitte 4 und 5 des GER; der oben erwähnte *Dutch Grid* wie auch der erwähnte *ALTE Grid for Writing Tasks* können in diesem Zusammenhang als Ergänzung und Erweiterung der Ausführungen zur Spezifizierung von Testinhalten betrachtet werden. Der dritte im *Manual* behandelte Aspekt ist der der so genannten *standardisation* (in dieser Arbeit mit *Standardisierung* wiedergegeben); unter *Standardisierung* versteht das *Manual* die Anbindung gegebener Tests und ihrer Ergebnisse an die Niveaus des GER durch Experteneinschätzung der Testschwierigkeiten respektive der Niveaus der Performanzen; auch hierbei kann der *Dutch Grid* in den Bereichen Lese- und Hörverstehen seinen Beitrag leisten. Schließlich wird die *empirical validation* im *Manual* dargestellt, die empirischen Prozeduren und Methoden betreffend, die bei der Validierung der Testanbindung zum Einsatz kommen (sollen).

Zur eigentlichen Testkonstruktion und Validierung der entwickelten Tests kann und will das *Manual* nicht beitragen – der Nachweis der Reliabilität und Validität der anzubindenden Tests muss natürlich vor der eigentlichen Anbindungsprozedur erfolgen. Das im *Manual* beschriebene Vorgehen ist als Reflexionsanstoß und als Rahmen gedacht, welcher auf die jeweiligen Kontexte und Bedürfnisse hin abgestimmt werden muss; die vier gerade vorgestellten Aspekte oder Phasen sollten jedoch in jedem Fall abgedeckt werden. Das *Manual* gibt auf S. 4 eine Übersicht (vgl. Anhang 19 dieser Arbeit), die die Anbindungsprozeduren veranschaulicht.

[220] Vgl. dazu die Ausführungen oben zu den Testgütekriterien, insbesondere zur Validität, in Kapitel 2.3 dieser Arbeit.

Im Folgenden werden diese vier Phasen vorgestellt und kritisch betrachtet.

3.5.1 Phase der Familiarisierung

Das *Manual* stellt in seinem Abschnitt 2 den GER in Kurzform vor, aufbauend auf den Ausführungen der GER-Abschnitte 2 und 3. Es wird auf die Funktionen eingegangen, die der GER (gemäß den Aussagen seines Abschnitts 9.2.3) übernehmen will. Nicht zuletzt wird davor gewarnt, schon kalibrierte Tests oder Performanzbeispiele zu frühzeitig als *benchmarks* oder als Anker zu nutzen – die betreffenden Tests müssen ihre Anbindung nicht nur behaupten, sondern auch belegen können. Abschnitt 3 des *Manual* wendet sich dem eigentlichen Familiarisierungsprozess zu: Dort finden sich Hinweise, wie die Niveaus und Kategorien des GER vorgestellt und „erfahren" werden können, beispielsweise über die Diskussion der Niveaubeschreibungen des GER-Abschnitts 3.6 oder über Selbsteinschätzung des Könnens in einer Fremdsprache mittels des Rasters aus GER-Abschnitt 3. Aufbauend darauf können qualitative Analysen der GER-Skalen erfolgen, etwa indem einzelne Deskriptoren einer zerlegten Skala ihren geschätzten Niveaus zugeordnet werden oder indem das in seine Deskriptoren zerlegte Selbstbewertungsraster rekonstruiert wird. Die im *Manual* gegebenen Tipps sind hilfreich, es werden mögliche Aktivitäten, ihr Aufwand und das dazu nötige Material beschrieben – lediglich die Zeitangaben dürften zu knapp bemessen sein, denn die dort vorgestellten einführenden Aktivitäten dürften mehr als 45 Minuten in Anspruch nehmen, ebenso wie die qualitative Analyse der Skalen und die Diskussion der Analyseergebnisse vermutlich nicht in 60 Minuten erfolgen kann.

3.5.2 Phase der Spezifizierung

In Abschnitt 4 des *Manual* wird die Phase der Spezifizierung eines gegebenen Tests, der an das GER-System angebunden werden soll, dargestellt: Dabei trifft das *Manual* zunächst Aussagen zur internen Validität des betreffenden Tests; diese Aussagen werden dann in Relation zum Kategorien- und Niveausystem des GER gesetzt, um die externe Validität des gegebenen Tests, also seine Validität in Bezug auf ein Außenkriterium zu beschreiben. Wie bereits gesagt genügen die Angaben im GER selbst jedoch nicht, um das Testkonstrukt, die Aufgabenschwierigkeiten, die Testinhalte, die Aktivitäten, die mit dem Test erfasst werden sollen und die dabei zu testenden Aspekte des kommunikativen Sprachvermögens hinreichend zu konkretisieren und zu spezifizieren. Die Behauptung des *Manual*, der GER sei bei der Spezifizierung der Hauptbezugspunkt (*Manual* 2003: 29), ist so nicht haltbar. Zwar wird im *Manual* die interne Aufgabenbeschreibung zunächst im jeweiligen Testkontext vorgenommen und erst im Zug der externen Validitätsbetrachtung auf den GER bezogen, doch Testspezifikationen müssen sprach- und

kontextspezifisch erfolgen – diese Spezifik findet sich in den kontextfreien und nicht auf eine bestimmte Sprache bezogenen GER-Skalen jedoch nicht.

In diesem Zusammenhang kann zumindest für die Fertigkeiten des Lese- und Hörverstehens auf den *Dutch Grid* rückgegriffen werden, der wie gesagt die Lücken, Inkonsistenzen und Terminologieprobleme des GER in diesen Bereichen analysiert und überwunden hat. Er kann helfen, die interne Testspezifizierung in transparenter Weise an die externen Kategorien und Niveaus des GER anzubinden. In allen anderen Bereichen muss die Analyse und Überwindung der Problematiken, wie sie auch bei den obigen Skalenanalysen im Bereich des Schreibens[221] zutage traten, erst noch angegangen werden. Die im *Manual* zu den verschiedenen Bereichen angebotenen Checklisten und GER-Skalen können einen möglichen Ausgangspunkt der Analyse bieten, doch wie gesagt genügen sie für eine transparente und kohärente Anbindung nicht.

Es darf an dieser Stelle darauf hingewiesen werden, dass die externe Validierung auch in umgekehrter Richtung interpretiert werden kann: Testspezifizierungen, die unabhängig vom GER-System erstellt werden, können als externe Validierung der GER-Kategorien und Niveaus betrachtet werden, wenn wie gesagt das fragliche Testkonstrukt eine Vergleichsbasis im GER hat. Gerade auf dem Gebiet des Schreibens ist dieser Aspekt interessant, denn dort sind die GER-Skalen nicht empirisch kalibriert worden, weshalb der GER in diesem Bereich kein empirisch fundiertes externes Validitätskriterium darstellen kann. Hier muss die Spezifizierung fraglicher Textproduktionstests auf „eigenen Füßen“ stehen, ehe man die jeweiligen Spezifizierungen mit den betreffenden Kategorien und Skalen des GER abgleichen kann – im Sinne einer „gegenseitigen“ externen qualitativen Validierung.

Die Ausführungen im *Manual* zu den Zwecken dieser Spezifizierung (vgl. ebd.: 29) sind kritisch zu betrachten: Die Schaffung von Bewusstsein dahingehend, dass Inhaltsanalysen von Prüfungen im Rahmen ihrer internen Validierung eine gewichtige Bedeutung zukommt und dass die Anbindung von Tests an das Referenzsystem zum Zweck der Vergleichbarkeit notwendig ist, erscheint ein sinnvolles Anliegen. Doch dass dort von Möglichkeiten, dieses Referenzsystem bei der *Planung* von Sprachprüfungen einzusetzen, gesprochen wird, ist m. E. bedenklich – aufgrund der oben aufgezeigten Probleme im Zusammenhang mit dem Skalensystem kann der GER nicht als Ausgangsbasis der Testplanung oder Erstellung benutzt werden. Neben den genannten bewusstseinsbildenden Aspekt tritt das Ziel, mithilfe der Spezifizierungen Minimalstandards zu setzen hinsichtlich der inhaltlichen Beschreibung von Prüfungen oder Tests und hinsichtlich der Anbindungsprozesse selbst. Dazu bietet das *Manual* eine solide Basis: Es werden detaillierte Fragebögen zur Testanalyse (*Manual*: 32-40, Formulare A1 mit A7) und zur

[221] Im Bereich des Schreibens kann zukünftig eventuell auf den erwähnten *ALTE Grid for Writing Tasks* zurückgegriffen werden; doch da er bei Erstellung dieser Arbeit noch nicht vorlag, konnte er in seiner Konzeption und Bedeutsamkeit nicht beurteilt werden. Aus dem Dokument unter http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual/GridINput.pdf?L=E geht auch nicht hervor, ob diesem *Grid* eine Analyse der entsprechenden Kategorien und Skalen des GER zugrunde liegt; das Dokument erweckt vielmehr den Eindruck, es sei noch in Bearbeitung (Zugriff am 23.08.2005).

Beschreibung der Tests in Bezug auf die Kategorien und Skalen des GER (*Manual*: 41-63, Formulare A8 mit A23) angeboten, die ein standardisiertes Vorgehen ermöglichen.

3.5.3 Phase der Standardisierung

Die in der Phase der Spezifizierung aufgestellten Behauptungen bezüglich der Verortung anzubindender Tests im GER-System müssen in der sich anschließenden Phase der Standardisierung gestützt werden. Hierfür werden Experteneinschätzungen genutzt, die die spezifizierten Tests ihren GER-Niveaus zuweisen. Dazu werden in Abschnitt 5 des *Manual* zwei Herangehensweisen vorgestellt: Produktive Aufgabenstellungen werden zunächst in einer Probandengruppe getestet und die dabei elizitierten Performanzen werden von geschulten Experten auf GER-Niveaus eingestuft. Testitems hingegen, die rezeptive Fertigkeiten erfassen, werden in ihren Schwierigkeiten von geschulten Experten eingeschätzt und dadurch einem Niveau im GER-System zugeordnet. Selbstverständlich wird auch dabei ein gezieltes Training (vgl. unten) vorgeschaltet, um das Referenzsystem in vergleichbarer Weise bei der Einschätzung von Performanzen als *benchmarks* für die Niveaus des GER respektive bei der Einschätzung von Itemschwierigkeiten in Bezug auf die GER-Niveaus anzuwenden. Damit sollen die Grenzen zwischen den einzelnen Niveaus, die so genannten *cut-off points*, belegt werden und somit der jeweilige Minimalstandard für das Erreichen eines Niveaus festgelegt werden. Dieser Prozess wird “standard-setting“ genannt (*Manual* 2003: 65); auf ihn wird unten näher eingegangen.

Die Anleitung zum erwähnten Training (ebd.: 71-86) gibt wiederum hilfreiche Tipps, stellt die erforderlichen Materialien und den prototypischen Ablauf von Schulungssitzungen vor, und gibt Beispiele, wie *Rating*-Ergebnisse der Schulungsphase dokumentiert werden können. Allerdings sind auch hierbei die Zeitangaben (vgl. ebd.: 81) mit Vorsicht zu genießen – gerade die Diskussion der *Benchmark*-Texte dürfte mehr als 60 Minuten in Anspruch nehmen. Problematisch an den Ausführungen zum Training ist, dass immer wieder Bezug genommen wird auf den Einsatz schon kalibrierter Performanzbeispiele – dem derzeitigen Stand der Dinge spiegelt dieser Vorschlag jedoch nicht wider, denn es gibt beispielsweise im Bereich des Schreibens keine kalibrierten Performanzen, die als Schulungs- oder *Rating-Benchmarks* herangezogen werden könnten.[222]

Die Ausführungen zum “benchmarking“ von Performanzen (*Manual* 2003: 87ff) geben Informationen über den Zeitbedarf und die Zusammensetzung der erforderlichen Beispielperformanzen und sie gehen auf einige grundsätzliche Probleme des *rating* ein, wie sie oben in Kapitel 3.3 dieser Arbeit schon angesprochen wurden. Prekär in diesem Zusammenhang ist, dass das *Manual* rät, dann andere Performanzbeispiele auszuwählen, wenn sich die gegebenen zu sehr vom GER-Format unterscheiden (ebd.: 88) – das würde aber bedeuten, dass der gesamte Test, der die fraglichen Beispiele elizitiert hat, in Frage gestellt wird und somit nicht an

[222] Es darf nochmals auf Kapitel 4 dieser Arbeit verwiesen werden, in welchem *benchmarks* aus dem DESI-Projekt zur Diskussion gestellt werden.

den GER angebunden werden kann. Ebenfalls problematisch ist der Vorschlag im *Manual*, relevante GER-Deskriptoren als *rating scales* zu nutzen (ebd.), sollte sich beispielsweise das Bewertungsraster (Tabelle 3 aus Abschnitt 3 des GER) als unangemessen herausstellen – es darf auf die obigen Analysen und Schlussfolgerungen im Zusammenhang mit der Nicht-Verwendbarkeit der GER-Skalen in der Funktion von Bewertungsskalen verwiesen werden (vgl. Kapitel 3.4.3 und 3.4.4 dieser Arbeit).

Im darauf folgenden Abschnitt (ebd.: 89-97) stellt das *Manual* Prozeduren des "standard-setting" vor. Es finden sich viele hilfreiche Tipps und eine detaillierte Anleitung, wie die einschätzenden Experten auf hinreichenden Konsens geschult werden können. Doch auch dort wird auf schon kalibrierte Testitems[223] verwiesen – wobei fraglich ist, ob es für alle anzubindenden Testformate entsprechend kalibrierte *items* gibt. Ebenfalls bedenklich ist, dass Testitems in ihrer Schwierigkeit beurteilt werden sollen anhand von GER-Deskriptoren, die wie oben analysiert solche Testcharakteristika gar nicht beschreiben. In diesem Zusammenhang darf wiederum ergänzend zu den Ausführungen im *Manual* auf den *Dutch Grid* verwiesen werden, der auf Basis von Testspezifizierungen das GER-Niveau identifiziert, auf dem das betreffende Testitem am wahrscheinlichsten eingestuft werden kann.

Zur Bestimmung der Niveaugrenzen, dem eigentlichen *standard-setting* nach dem Training, werden zwei Methoden im *Manual* exemplarisch vorgestellt – die jeweils angemessenste soll je nach Kontext auch unter Hinzuziehung von weiterer Fachliteratur gefunden werden (vgl. *Manual* 2003: 90ff). Die erste ist nach Angoff, dem Entwickler dieser Methode, benannt: Experten sollen imaginäre Lerner, die am untersten Ende eines bestimmten Niveaus stehen, hinsichtlich der Wahrscheinlichkeit einschätzen, mit der sie die jeweiligen Testitems lösen würden. All diese Wahrscheinlichkeiten aufsummiert ergeben dann den erwarteten Test-*Score* dieser imaginären Person – nun müssen diese erwarteten Test-*Scores* noch über alle Experten hinweg gemittelt werden und man hat den *cut-off score*, den Wert, der in diesem imaginären Test den Beginn des betreffenden GER-Niveaus bezeichnet. Fragwürdig an dieser Methode ist, dass die Realität an keiner Stelle hereinkommt – wie zuverlässig Wahrscheinlichkeitsabschätzungen in Bezug auf imaginäre Lernende sind, welche noch dazu auf einer imaginären Niveaugrenze gedacht werden müssen, sei dahingestellt – doch empirisch fundiert ist diese Methode nicht. Ebenfalls fragwürdig ist, wie diese geschätzten Wahrscheinlichkeitswerte sich in Bezug auf eine empirische Skalierung der betreffenden Testitems und deren Lösungen durch reale Testprobanden verhalten würden.

Die zweite Methode wurde im DIALANG-Projekt eingesetzt: Dabei wurden keine imaginären Probanden eingeschätzt, sondern es musste die folgende Frage für jeden Test oder jedes Testitem einer Prüfung beantwortet werden: "At what CEF level can a test taker already answer the following item correctly?" (*Manual* 2003: 91).[224] Auf diese Weise können Datenbanken kalibrierter

[223] Die *Item*-Bank, die im DIALANG-Projekt erstellt wird, könnte hierbei hilfreich sein. Doch über das Internet ist sie nicht zugänglich.
[224] Details dieser Methode können nachgelesen werden in Kaftandjieva et al. (1999).
Technische Aspekte sind beschrieben im *Reference Supplement, Section B* zum *Manual* 2003, im Internet erhältlich unter http://culture2.coe.int/portfolio//documents/CEF%20reference%20supplement%20version%202.pdf, Zugriff am 02.02.2005.

Testitems geschaffen werden. Auch wenn bei diesem Vorgehen keine Wahrscheinlichkeiten eingeschätzt werden müssen, ist doch die Interpretation des Testitems als Operationalisierung eines entsprechenden "Can Do"-Deskriptors (vgl. ebd.: 91) höchst fragwürdig – denn wie oben festgestellt beschreiben die Deskriptoren keine operationalisierbaren Konzepte, sondern generalisiertes Sprachvermögen. Da auch diese Methode, wie die Angoff-Methode, keinen Bezugspunkt zur Realität hat, kann sie ebenfalls nicht zur empirischen Anbindung von Tests oder Testitems beitragen, sie ist aber ein Beitrag zur qualitativen Validierung der Anbindungsprozedur.

Die Ausführungen zur Standardisierung im *Manual* schließen mit den Hinweisen auf die transparente Dokumentation dieser Einschätzungen, einschließlich der quantitativen Analysen der Inter-*Rater*- und Intra-*Rater*-Reliabilitäten der Einschätzenden, um den Validierungsprozess der Anbindung von Tests an das Niveausystem des GER belegen zu können. Die Tabelle auf S. 98 des *Manual* gibt einen hilfreichen Überblick über die Prozesse der Standardisierung.

3.5.4 Phase der empirischen Validierung

Die Prozeduren der Spezifizierung und Standardisierung müssen durch empirische Validierungsmethoden bestätigt und gestützt werden. Diese letzte Phase der Anbindung wird in Abschnitt 6 des *Manual* dargestellt. Von diesem Abschnitt würde man erwarten, dass jene vier Anbindungsmöglichkeiten[225], die in der Einleitung des *Manual* (ebd.: 9f) kurz vorgestellt werden, charakterisiert und konkretisiert werden, doch dieser Abschnitt setzt ganz andere Schwerpunkte: Dort werden zunächst interne und externe Validierungsaspekte (s. u.) beschrieben, ehe psychometrische Methoden der jeweiligen Validierungsarten vorgestellt werden. Die eigentlichen Anbindungsmöglichkeiten werden jedoch nicht wieder aufgenommen, sondern lediglich im Rahmen der Ausführungen zur externen Validierung an Beispielen (ebd.: 113 mit 122) illustriert. Deshalb ist es mühsam, nähere Informationen zu diesen Anbindungsmöglichkeiten aus Abschnitt 6 des *Manual* zu ziehen. Auch fehlt teils der Bezug zu den vorangegangenen Phasen; während beispielsweise bei den Ausführungen zur Standardisierungsphase zwischen produktiven und rezeptiven Aufgabenstellungen unterschieden wird, wird diese Differenzierung in Abschnitt 6 nicht fortgeführt, so dass in Bezug auf die externe Validierung produktiver Aufgabenstellungen keine konkreten Aussagen zu finden sind.

Im vorliegenden Unterkapitel dieser Arbeit werden aufgrund dieser Problematik zunächst die grundlegenden Validierungsaspekte, die im *Manual* erwähnt werden, betrachtet, um den Rahmen der externen Validierung zu stecken. Nachfolgend werden die erwähnten vier Anbindungsmöglichkeiten aus der Einleitung des *Manual* auf die oben beschriebenen Standardisierungsprozesse bezogen und kritisch beurteilt. Erst im Anschluss daran werden jene beiden psychometrischen

[225] Auf diese vier Möglichkeiten wird unten näher eingegangen.

Methoden herausgegriffen, die das *Manual* im Rahmen der externen Validierung vorstellt, um deren empirische Reichweite in Bezug auf die vier Anbindungsmöglichkeiten einschätzen zu können.

Bei der empirischen Validierung wird die Realität in Form von Datenerhebungen und Datenanalysen mit in den Validierungsprozess der Testanbindung einbezogen. Dabei unterscheidet das *Manual* zu Beginn seines Abschnitts 6 zwischen Methoden der internen und der externen Validierung: Erstere sollen helfen bei der Konstrukt- und Inhaltsvalidierung, die die Test- und Itemcharakteristika und die psychometrische Qualität des Tests belegen muss – eigentlich nicht die Aufgabe des *Manual*; letztere sollen helfen, die Behauptungen in Bezug auf die Anbindung an das Außenkriterium GER, die während der beiden vorangegangenen Phasen der Spezifizierung und Standardisierung aufgestellt wurden, zu stützen und zu belegen.

Im Rahmen der internen Validierung darf auf die Ausführungen des Kapitels 2.3.1 dieser Arbeit verwiesen werden, in welchem grundlegende Validitätskonzepte erläutert werden. Die interne Validierung muss bei jeder Testkonstruktion erfolgen, ganz unabhängig davon, ob der fragliche Test an ein Referenzsystem angebunden werden soll oder nicht. Das *Manual* gibt auch in diesem Bereich eine hilfreiche Übersicht, denn "internal validation is a prerequisite for acceptable linking to the CEF" (*Manual* 2003: 100).[226] Es wird eine Checkliste (ebd.: 100) angeboten, mittels derer Aspekte aus der Spezifizierungsphase anhand von Daten aus Prätests oder Pilotierungen überprüft werden können: Homogenität und Trennschärfe der Testitems, angemessene Inhalte, Formate und Schwierigkeiten, Reliabilität von *ratings* und Korrelationen mit anderen Testmodulen innerhalb einer Prüfung mögen hier als Stichworte genügen. Im Anschluss daran werden im *Manual* die dabei üblicherweise eingesetzten Methoden der klassischen Testtheorie, der qualitativen Analysen, der Generalisierbarkeitstheorie, der Faktorenanalyse und der *Item-Response*-Theorie kurz charakterisiert – für eine genauere Darstellung darf auf die *reference supplements*[227] verwiesen werden (vgl. *Manual* 2003: 102-108).

Bei der Testanbindung selbst spielt die externe Validierung die entscheidende Rolle. Dabei werden Daten aus der Administrierung der anzubindenden Tests auf ein unabhängiges Validierungskriterium bezogen. Diese Bezüge werden mittels quantitativer Methoden erfasst und analysiert, wobei die sich ergebenden Zusammenhänge ergründet werden müssen. Dazu benötigt wird statistisches und psychometrisches Know-how, um mögliche Fehlerquellen zu erkennen und die Daten angemessen zu interpretieren. Dadurch sollen die qualitativen Behauptungen und Einschätzungen aus der Spezifizierungs- und Standardisierungsphase hinsichtlich der Bezüge des jeweiligen Tests zum Referenzsystem und hinsichtlich des *standard-setting* empirisch

[226] Allerdings steht diese Aussage in gewissem Widerspruch zu der Aussage auf S.1 des *Manual*: Dort distanzieren sich die Autoren von Aspekten der internen Validierung. Nichtsdestotrotz sind die in Abschnitt 6 des *Manual* zu findenden Ausführungen zur internen Validierung hilfreich.

[227] Hier erhalten interessierte Benutzer, sollten sie keine Psychometriker sein, die Möglichkeit, sich so weit zu informieren, dass sie in gewissem Rahmen Verständnis für die theoretischen und statistischen Grundlagen entwickeln können. Vgl. Council of Europe 2004, im Netz erhältlich unter http://culture2.coe.int/portfolio//documents/CEF%20reference%20supplement%20version%202.pdf, Zugriff am 02.02.2005.

belegt werden. Dazu schlägt das *Manual* (vgl. ebd.: 9f und 108-113) grundsätzlich zwei Wege vor: Entweder werden schon kalibrierte Tests als Vergleichskriterium eingesetzt, oder es werden Lehrereinschätzungen von Performanzen respektive Lernenden mithilfe bereits kalibrierter Deskriptoren (gemeint sind vermutlich die GER-Deskriptoren) genutzt, wobei die Lehrenden natürlich mit dem GER vertraut sein müssen. Folgende Abbildung illustriert den Prozess:

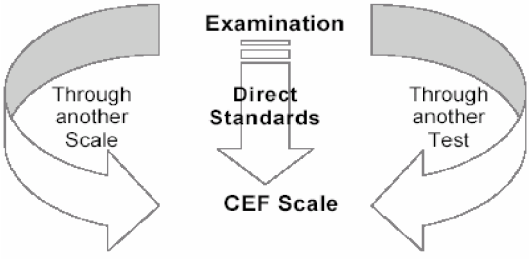


Abb. 17 (nach Takala 2003): Link to CEF

Das *Manual* stellt (in seiner Einleitung) zu beiden Wegen je zwei Varianten vor, eine basierend auf klassischer Testtheorie und eine auf probabilistischer Testtheorie, so dass man letztlich vier Validierungsmöglichkeiten erhält (vgl. ebd.: 9f, Übers. und Herv. d. V.):

(a) Der neu anzubindende Test und ein bereits *in valider und reliabler Form* an den GER angebundener Test (als Anker) werden in derselben Zielgruppe administriert, um Vergleichsdaten zu erhalten. Testergebnisse des anzubindenden Tests werden korreliert mit den Ergebnissen des bereits an den GER angebundenen Tests. Mittels Regressionsanalysen können die neuen Testergebnisse auf die Niveaus des GER bezogen werden.

(b) Beide Tests werden in derselben Zielgruppe administriert und gemeinsam skaliert, so dass der anzubindende Test auf die Skala des Ankertests und somit auf die GER-Niveaus kalibriert werden kann.

(c) Lehrende nutzen *Deskriptoren* [*vermutlich*, um die Probanden, die den anzubindenden Test abgelegt haben, *einzuschätzen*, Anm. d. V.]; die Deskriptoren wurden zuvor auf valide und reliable Weise auf das Niveausystem des GER kalibriert [gemeint sind daher vermutlich GER-Deskriptoren, Anm. d. V.]. Diese Einschätzungen werden mit den Test-*Scores* des anzubindenden Tests korreliert. Nun können die Testergebnisse des anzubindenden Tests mittels Regressionsanalysen auf die GER-Niveaus bezogen werden.

(d) Wie (c), nur werden nun die Skalenwerte der schon kalibrierten Deskriptoren [gemeint sind vermutlich die Werte aus dem Schweizer GER-Skalenkonstruktionsprojekt, Anm. d. Verf.] als *Anker* benutzt, um den anzubindenden Test direkt auf die Skala zu kalibrieren, die hinter den GER-Niveaus liegt.

Folgende Aspekte, die oben zur Verdeutlichung kursiv gesetzt sind, verdienen im Rahmen der Beurteilung der Bedeutsamkeit der erwähnten vier Anbindungsmöglichkeiten nähere Betrachtung:

Zunächst interessiert bei den Möglichkeiten (a) und (b), wie die *valide und reliable Weise* aussieht, mit der diese Tests auf die GER-Niveaus kalibriert wurden – denn solch eine Anbindung über Korrelation und Regression oder über eine gemeinsame Skalierung stellt an und für sich eine empirische Validierungsmethode dar. Doch dieser Verweis auf schon kalibrierte Tests erscheint wie das „Henne und Ei"-Problem: Hat man erst einen empirisch kalibrierten Test, so kann dieser, bei vergleichbarem Konstrukt, als Validierungskriterium genutzt werden – doch hat man diesen nicht, was tut man dann? Was verbirgt sich hinter der zitierten *valid and reliable fashion* konkret? Die Antwort darauf findet sich im *Manual* in Abschnitt 6 bei einem Fallbeispiel, das die Nutzung der Lehrereinschätzungen illustrieren soll (vgl. *Manual* 2003: 117ff): Dort wird ausgesagt, dass in den Fällen, in denen es noch kein kalibriertes Testitem gibt, Lehrereinschätzungen der entsprechenden Testpopulation ein geeignetes Validierungsmaß darstellen könnten. Allerdings sollte m. E. dabei bedacht werden, dass sich die Lehrereinschätzung auf dieselbe Probandengruppe und auf mit dem anzubindenden Test vergleichbare Konstrukte beziehen muss und zunächst erst einmal selbst validiert werden müsste.

Lehrereinschätzungen werden auch bei den oben vorgestellten Anbindungsmöglichkeiten (c) und (d) eingesetzt. In diesem Zusammenhang stellen sich zwei Fragen: 1. Auf was genau beziehen sich die dort genannten Lehrereinschätzungen? 2. Wie können Test-*Scores* auf Deskriptoren-Skalenwerte kalibriert werden?

Die Antwort auf Frage 1 dürfte wiederum in oben zitiertem Fallbeispiel zu finden sein: Es muss sich nach den Ausführungen des *Manual* (ebd.: 117f) um Einschätzungen der Testpopulation, in der der anzubindende Test administriert wird, oder zumindest eines *samplings* dieser Population handeln. Es darf in diesem Zusammenhang auf die obigen Skalenanalysen und die Aussagen zur Verwendbarkeit der Skalen verwiesen werden – wie festgestellt, sind die GER-Deskriptoren für eine kriterienorientierte Bewertung von konkreten Performanzbeispielen nicht angemessen formuliert, sie sind keine *rating scales*. Lehrende können die Deskriptoren zwar zur globalen Einschätzung der ihnen bekannten Lernenden nutzen, doch um die Einstufung eines konkreten Tests über Lehrereinschätzungen zu validieren, müssen sich diese wie gesagt auf ein dem anzubindenden Test vergleichbares Konstrukt beziehen; diese Entsprechung des Testkonstrukts mit dem GER-System und dessen Beispielskalen ist jedoch nicht immer gegeben.

Zu Frage 2 findet sich, ebenfalls bei dem erwähnten Fallbeispiel (vgl. *Manual* 2003.: 117ff), ein Hinweis auf die vermutliche Antwort: Die Deskriptoren des GER wurden im oben beschrieben Schweizer Projekt kalibriert und durch die Rasch-Skalierung wurde ihnen ein Skalenwert zugewiesen. Wenn nun bereits kalibrierte Deskriptoren als "rating scale test items" (ebd.: 119) betrachtet werden und mit ihnen Testprobanden eingeschätzt werden, so kann diese Einschätzung gemäß *Manual* als Test-*Score* betrachtet werden, so dass man Deskriptoren-Skalenwerte und die Test-*Scores* des anzubindenden Tests aufeinander beziehen kann. Dennoch bleibt die Frage unbeantwortet, ob es psychometrisch korrekt ist, die als Test-*Scores* interpretierten

Skalenwerte, die ja durch die Einschätzung innerhalb einer bestimmten Probandengruppe[228] erzielt wurden, als Anker zu nutzen, wenn der anzubindende Test in einer anderen Probandengruppe administriert wurde. Die Antwort darauf müssen Psychometriker geben.

Neben den genannten offenen Fragen finden sich im Zusammenhang mit den Anbindungsmöglichkeiten (c) und (d) Widersprüchlichkeiten, wenn man sich die obige Abbildung 17 und die Übersicht im *Manual* (ebd.: 129, vgl. Anhang 30 dieser Arbeit) betrachtet: In Abbildung 17 findet sich die Aussage, dass Tests "through another scale" angebunden werden können – was genau ist mit der Anbindung durch eine andere Skala gemeint? Aus dem *Manual* geht m. E. nicht hervor, dass eine andere, vom GER unabhängige Skala zum Einsatz käme. In der besagten Übersicht (vgl. Anhang 30 dieser Arbeit) finden sich in der rechten Spalte die vier Anbindungsmöglichkeiten wieder, nur dass dabei Möglichkeit (c) durch die Aussage "Correlating ratings with CEF descriptors" dargestellt wird. Dabei stellt sich die Frage, wieso das *Manual* in dieser Übersicht rät, *ratings* mit den Skalenwerten der GER-Deskriptoren zu korrelieren. Denn bisher war die Rede davon, die GER-Deskriptoren zur Probandeneinschätzung zu nutzen; dies würde dann zu den besagten *ratings* führen, welche jedoch mit den Test-*Scores* des anzubindenden Tests korreliert werden sollten, und nicht mit den GER-Deskriptoren. Wenn es sich bei den genannten *ratings* jedoch nicht um die Probandeneinschätzung mithilfe von GER-Deskriptoren handelt, stellt sich die Frage, welche *ratings* man dann mit den GER-Deskriptoren korrelieren will; denn wie gesagt ist nirgends im *Manual* von anderen als den GER-Skalen die Rede, die zum *rating* verwendet werden können. Nun wäre noch denkbar, dass sich die „andere Skala" und die *ratings* auf offene Aufgabenstellungen beziehen, die mittels einer anderen Skala bewertet wurden (daher die *ratings*, vgl. auch die Anmerkungen in Fußnote 229 unten) – doch davon ist im *Manual* bei der Vorstellung der vier Möglichkeiten nicht die Rede (vgl. *Manual* 2003: 10). Wie die Nutzer des Manuals mit diesen Widersprüchlichkeiten umgehen sollen, bleibt offen.

Bei den erwähnten Anbindungsmöglichkeiten (c) und (d) bleibt, ganz abgesehen von den gerade erörterten Widersprüchlichkeiten, fraglich, ob es sich um ein empirisch quantifizierbares Außenkriterium handeln kann, wenn Lehrereinschätzungen der Testpopulation als Validierungskriterium genutzt werden, um zu überprüfen, ob Test-*Scores* in ein bestimmtes GER-Niveau fallen, das den Testitems zuvor durch Experteneinschätzung zugewiesen wurde. Zumindest sollte deutlich gemacht werden, dass immer dann, wenn es noch keine kalibrierten Tests gibt, Einschätzungen aus der Standardisierungsphase durch Lehrereinschätzungen qualitativ validiert werden; die Empirie kommt insofern herein, als sich die Lehrereinschätzungen auf die Population beziehen, in der der anzubindende Test auch administriert wurde.

Aufgrund der hier angedeuteten Verständnisschwierigkeiten und Widersprüchlichkeiten wäre es wie gesagt hilfreich, würde Abschnitt 6 des *Manual* explizit an die in der Einleitung vorgestellten vier Möglichkeiten (a) mit (d) anknüpfen, sie in ihren Anwendungskontexten beschreiben

[228] Vgl. Kapitel 3.4.1.4, Tabelle 4 dieser Arbeit, in der die Schweizer Probandengruppe vorgestellt wird, auf die sich die Skalierung der GER-Deskriptoreneinschätzungen bezog.

und konkret auf die Prozesse der Standardisierungsphase beziehen. Auch wäre es wünschenswert, wenn der Aspekt, wie produktive Testaufgaben an den GER angebunden werden können, transparent in den Ausführungen zur empirischen Validierung thematisiert würde.[229]

Im Rahmen der allgemeinen Prinzipien der externen Validierung stellt Abschnitt 6 des *Manual* zwei quantitative Prozeduren vor, namentlich Korrelationen und das Zuweisen quantitativer Test-*Scores* zu den qualitativen GER-Niveaus (vgl. Manual 2003: 108-113): Einmal kann man mittels Korrelationskoeffizienten zwischen dem fraglichen Test und dem Ankertest in derselben Population den Grad des Zusammenhangs zwischen den beiden Tests darstellen – eine übliche Methode der Konstruktvalidierung und somit beim Vergleich zweier Tests eine Grundvoraussetzung, die Zusammenhänge zwischen ihnen überhaupt sinnvoll interpretieren zu können. Im *Manual* werden so genannte *Scattergrams* vorgestellt, die verschiedene Korrelationskoeffizienten anschaulich graphisch darstellen:

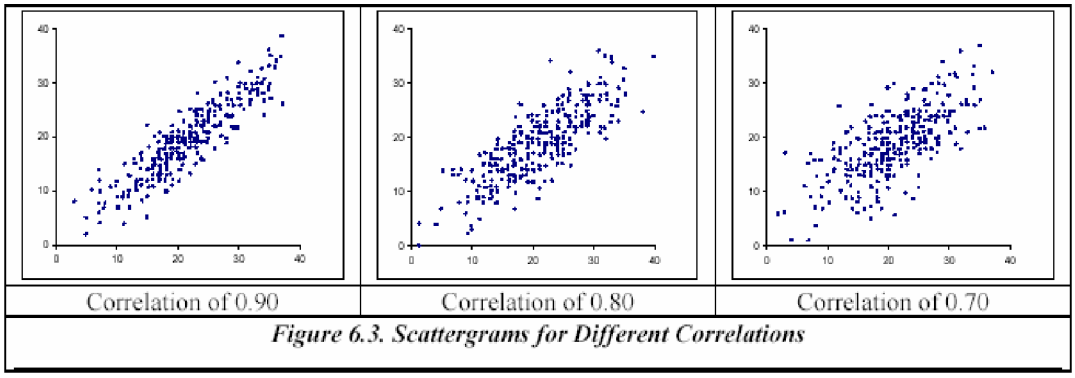


Abb. 18: *Scattergrams* verschiedener Korrelationskoeffizienten; jeder Proband ist durch einen Punkt dargestellt; horizontal ist die Leistung eines Probanden im zu validierenden Test repräsentiert, vertikal die Leistung im Ankertest (vgl. *Manual* 2003: 111).

Wenn nun der Zusammenhang zwischen dem individuellen Probandenverhalten in beiden Tests hoch genug ist (eine Interpretation, die in Zusammenarbeit mit Psychometrikern gefunden werden sollte), ist dies ein Hinweis, dass beide Tests dasselbe Konstrukt in ähnlicher Weise erfassen. Wenn beide Tests gemeinsam skaliert werden können, so können auch Niveaugrenzen vergleichbar gezogen werden.

Die zweite Prozedur wird im *Manual* als *classification match* bezeichnet: Diese Prozedur soll die Zuweisung quantitativer Test-*Scores* zu qualitativen Kategorien verifizieren, wie es beim *standard-setting* erfolgt ist. Gemäß *Manual* geht es dabei um die Beantwortung der Frage “To what extent can one be sure that a test taker classified by the criterion test as B1 will also be classified in the same category by the test under study?“ (ebd.: 109). Dabei zeigt sich, dass die

[229] Eine solche Möglichkeit könnte wie folgt aussehen (vgl. auch Kapitel 4.7.1 dieser Arbeit): Wenn man davon ausgeht, dass in der Standardisierungsphase in Bezug auf produktive Tasks *benchmarks* bestimmt wurden, so müsste diese Niveauzuweisung nun empirisch validiert werden. Vermutlich kann man auch dabei Lehrereinschätzungen nutzen: Man administriert die Aufgabe und lässt die Performanzen nach dem mit der Aufgabe entwickelten Bewertungsschema *raten* (wobei die *benchmarks* helfen, die Performanzen zu bewerten). Dann lässt man die Testpopulation im Hinblick auf die getesteten Fertigkeiten durch Experten, denen die Testprobanden sehr gut bekannt sein müssen, auf die Niveaus des GER einstufen, allerdings nicht auf Grundlage konkreter Performanzen, sondern generell hinsichtlich des „Könnens“ in der getesteten (Teil-)Fertigkeit. Die GER-Einschätzungen können dann mit den *ratings* der konkreten Performanzen korreliert werden.

Zuweisung quantitativer Werte zu qualitativen Kategorien nicht durch empirisch belegbare, „harte“ Fakten erfolgen kann, sondern dass es sich dabei immer um die Einschätzung von Wahrscheinlichkeiten handelt. Wie zutreffend diese Einschätzungen (hier in Bezug auf das *standard-setting*) waren, kann wie folgt überprüft werden (vgl. ebd.: 109 und 111-113): Zunächst muss der Korrelationskoeffizient einen Zusammenhang zwischen zu verankerndem Test und Ankerkriterium nahe legen. Dann erst macht es Sinn, die Zuweisungen der Testitems oder Probandenfähigkeiten zu den GER-Niveaus zu betrachten. Dabei kommen so genannte Kreuztabellen zum Einsatz:

Decision Table: Classification according to Test and to Criterion

Criterion	Test: A2	Test: B1	Test: B2	Total
A2	68	7	0	75
B1	20	90	20	130
B2	2	3	90	95
Total	90	100	110	300

Abb. 19: Kreuztabelle (vgl. *Manual* 2003: 111)

Die Tabelle ist wie folgt zu lesen: Die Probandengruppe von 300 Lernenden ist im als Validierungskriterium dienenden Test und im anzubindenden Test dieselbe. In diesem Fall werden keine quantitativen Testscores oder Skalenwerte aus einer psychometrischen Skalierung genutzt, sondern es geht um die Einstufung auf die qualitativen Niveaus des GER: Horizontal finden sich zeilenweise die Zuweisungen der Lernenden auf die entsprechenden Niveaus des GER (hier A2, B1 und B2), die ja beim Validierungskriterium „auf valide und reliable Weise“ erfolgten (Kritik daran siehe oben). In den ersten drei Spalten wird die Zuweisung der Lernenden zu den entsprechenden GER-Niveaus aufgrund der Leistungen im anzubindenden Test dargestellt. Die *cut-off scores*, jene Testrohpunkte also, die den Beginn der jeweiligen GER-Niveaus bezeichnen, wurden ja beim *standard-setting* durch Einschätzung festgelegt. Nun kann man die unterste Zeile „Total“ (die sich auf die Einstufung im zu validierenden Test bezieht) mit der rechten Spalte „Total“ (die die Anzahl der Lernenden auf den entsprechenden GER-Niveaus im Validierungskriterium darstellt) vergleichen, um festzustellen, inwieweit die Niveaueinteilungen des anzubindenden Tests übereinstimmen mit denen des schon auf die GER-Niveaus kalibrierten Tests (ausgedrückt durch die Anzahl der jeweils in beiden Tests auf dasselbe Niveau eingestuften Lernenden – vergleiche die Diagonale von links oben nach rechts unten in der Tabelle). Diese Tabelle soll nun als „Entscheidungstabelle“ dienen (vgl. *Manual* 2003: 111f), um zu beurteilen, ob der fragliche Test hinreichend viele Probanden auf demselben Niveau einstuft, wie sie durch den Validierunsgtest eingestuft wurden. Sollte dies nicht der Fall sein, so müssten die Niveaugrenzen erneut gezogen werden (vgl. die Phase des *standard-setting*) und anschließend diese Tabelle neu berechnet werden, um zu sehen, ob die neuen Niveaugrenzen nun validere Ergebnisse liefern. Im Extremfall würde es also genügen, willkürliche Niveaugrenzen zu ziehen, solch eine Entscheidungstabelle zu berechnen und die Grenzen so lange zu adjustieren, bis ein

befriedigendes Ergebnis erzielt werden kann (vgl. ebd.: 113). Es fällt schwer, dieses *trial-and-error*-Verfahren als empirisch-quantitativen Validierungsprozess zu betrachten. Auch wenn im *Overview* zum *Manual* (Council of Europe 2003b: 4) ausgesagt wird, dass in der Phase der empirischen Validierung quantitative Verfahren zum Einsatz kommen, so lassen sich doch auch qualitative Methoden ausmachen, die für all die Aspekte, die nicht quantifiziert werden können, durchaus angemessen sein mögen – dies sollte jedoch auch transparent dargestellt werden.

3.5.5 Resümee zum Validierungsansatz des *Manual*

Zusammenfassend wird festgestellt, dass sich das *Manual* an Experten auf dem Gebiet der Testkonstruktion und Psychometrie wendet; es ist aufgrund seiner spezifischen Thematik nicht leicht zugänglich, doch können Literaturhinweise und die erwähnten *reference supplements* helfen. Die Einleitung gibt einen guten Überblick, die Abschnitte 2 mit 5 sind verständlich aufgebaut und geben nützliche Tipps und Hilfestellungen hinsichtlich der Phasen der Familiarisierung, Spezifizierung und Standardisierung. Das *Manual* stellt ein fundiertes Vorgehen bei der Testanbindung vor, welches richtungweisend ist, da es Standards setzt, an denen sich jede Testanbindung messen lassen kann. Allerdings sollte dabei beachtet werden, dass der GER keinesfalls als Ausgangspunkt der Beschreibung oder Testentwicklung genutzt werden sollte, jedoch als Referenzmittel in diesen Phasen hilfreiche Dienste leisten kann.

Dennoch gibt es im *Manual* Probleme, die sich insbesondere in Abschnitt 6 bei der Darstellung der Phase der empirischen Validierung zeigen: Dort fehlt teils die explizite Anknüpfung an die Prozeduren der vorangestellten Abschnitte des *Manuals*: Es scheint über weite Strecken, dass allgemein gültige Sachverhalte dargestellt werden, ohne den in Abschnitt 6 beschriebenen empirischen Validierungsprozess konkret auf die beiden vorangegangenen Phasen der Spezifizierung und Standardisierung zu beziehen. Dadurch bleibt dem „psychometrischen Laien“ manches Mal unverständlich, welcher Schritt der beiden vorangestellten Phasen durch welche Maßnahme in welchem Grad empirisch validiert werden kann. Wie oben schon erläutert, fehlt beispielsweise die explizite Thematisierung der Validierung der Anbindung produktiver Aufgabenstellungen. Zudem kommt es zu den o. g. Widersprüchlichkeiten zwischen der Kurzvorstellung der empirischen Validierung in der Einleitung des *Manual* und der Übersicht auf S. 129 des *Manual*.

In Bezug auf die empirisch-quantitative Validierungsphase ist festzustellen, dass nicht alle dort vorgestellten Prozeduren quantitativer Natur sind. Die erwähnte Übersicht auf S. 129 des *Manual* (vgl. Anhang 20 dieser Arbeit) fasst die verschiedenen Schritte der Testanbindungsprozedur zusammen, so dass an ihr die qualitativen Aspekte der empirisch-quantitativen Validierung illustriert werden können: Die vorgeschalteten qualitativen Phasen der Spezifizierung und Standardisierung haben ihre Berechtigung innerhalb der oben dargestellten Grenzen. In diesen Phasen werden qualitative Einschätzungen abgegeben (vgl. bei den Spezifizierungen die

unteren Zellen in der Übersicht, welche sich mit Aspekten der externen Validität beschäftigen respektive welche die Aspekte des *benchmarking* und *standard-setting* in der Standardisierungsphase darstellen). Diese qualitativen Einschätzungen sollen in der Phase der empirischen Validierung wiederum zum Teil anhand von Lehrereinschätzungen (vgl. die unteren Zellen der rechten Spalte) validiert werden, wobei der Zusammenhang beider Einschätzungen durch quantitativ-psychometrische Verfahren bestimmt wird. Auch die Erwähnung der Verwendung von bereits auf die GER-Niveaus kalibrierten Ankertests in der Validierung weist auf die dieser ursprünglichen Kalibrierung zugrunde liegende qualitative Validierung hin.

In diesem Zusammenhang darf der das *Manual* ergänzende Vorschlag aus dem Bericht zur Erstellung des *Dutch Grid* (Alderson et al. 2004: 22) zitiert werden, wie Tests an den GER angebunden werden könnten:

> • Describe the text and items using the dimensions of a classification system (The Frames and Grids).
> • Make a guess at the level of an item (guided by the classification system and the CEF scales), leading to an estimated CEF-level.
> • Pretest the items thus labelled, describing in detail the characteristics of the pilot sample.
> • Calibrate the items.
> • Do standard setting to set the boundaries of the levels on the scale coming from the calibration.
> • Assign a psychometric level to the items.
> • Assign a definitive level to the items. An item can only be assigned to a definitive level if the psychometric level falls within the band of the estimated level (in other words if the estimation based on the **analysed content** is comparable with the **psychometric** value).

Auch dort wird deutlich, dass man die Niveauzuweisung eines Items eher auf qualitativer Basis „erraten“ soll (“make a guess“). Und auch dort bleibt im Unklaren, wie ein psychometrisches Niveau mit dem abgeschätzten Niveau als Teil der *empirischen* Validierung in Verbindung gebracht werden kann, anders als durch die oben bei den Ausführungen zur empirischen Validierung vorgestellten *qualitativen* Methoden des *Einschätzens* durch Experten und/oder Lehrer (und die Nutzung von Paralleltests, so es denn adäquate gibt – doch auch diese wurden ursprünglich über qualitative Methoden den GER-Niveaus zugewiesen).

Nun ist gegen qualitative Validierungsmethoden, wenn sie angemessen eingesetzt werden, nichts einzuwenden, gerade in Bereichen, die sich nicht durch rein quantitative Verfahren erfassen lassen, doch man tut gut daran, die Grenzen der psychometrischen Quantifizierbarkeit offen zu legen und die qualitative Basis transparent darzustellen. Zweifelsohne verlangt gerade die empirische Validierung und auch die Beurteilung der Angemessenheit der dabei zum Einsatz kommenden Prozeduren Expertise auf dem Gebiet der Psychometrie, doch man sollte zumindest versuchen, die dabei auszuführenden Prozeduren so darzustellen, dass sie auch einem psychometrisch nicht bewanderten Publikum transparent und nachvollziehbar vermittelt werden können. Schließlich behaupten renommierte Testanbieter, ihre Prüfungen auf bestimmten Niveaus des GER ansiedeln zu können – und diese Behauptung muss auch gegenüber einem Laienpublikum belegt werden können.

Doch man muss berücksichtigen, dass diese Offenlegung der Validierungsprozesse individueller Anbindungen von Prüfungen nicht Aufgabe des *Manual* sein kann: Das *Manual*

(vgl. ebd.: 122f) will nicht „den einen“ Weg der Testanbindung an das Referenzsystem des GER vorschreiben; es will vielmehr das Bewusstsein für verschiedene Konzepte und Prozeduren der externen Validierung und die damit verbundenen Möglichkeiten und Grenzen fördern, so dass professionelle Testentwickler informierte und fundierte Entscheidung darüber treffen können, auf welche Weise sie neue Tests und Prüfungen an den GER anbinden wollen. Zudem befindet sich das *Manual* wie gesagt noch in der Pilotierungsphase. Deshalb finden sich in Abschnitt 7 des *Manual* Hinweise und Formulare zur Berichterstattung über die Erprobung der „vorläufigen Pilotversion“ des *Manual*. Diese Erfahrungen aus der Praxis werden zeigen, wo es noch Verbesserungsbedarf gibt.

3.5.6 Fallbeispiel für ein alternatives Vorgehen

An dieser Stelle soll kurz das Vorgehen im DESI-Projekt bei der Bestimmung der Niveaugrenzen am Beispiel des C-Tests beschrieben werden, um zu zeigen, wie der GER als reines Referenzmittel genutzt werden kann, wenn ein bestimmtes Testkonstrukt keine Entsprechung im GER-System hat. Auf Fragen des Testkonstrukts und der inhaltlichen Spezifizierung soll hier nicht konkret eingegangen werden, da der C-Test nicht Gegenstand dieser Arbeit ist.[230] Dennoch kann die Betrachtung des *standard-setting* in diesem DESI-Modul eine Alternative zum im *Manual* vorgeschlagenen Weg darstellen. Dabei wird die direkte Einschätzung von Items oder Tests auf GER-Niveaus umgangen, ebenso wie die empirische Validierung dieser Einschätzung auf Basis des Vergleichs der Probandenleistungen im fraglichen Test mit den Leistungen derselben Probanden in schon kalibrierten Tests. Auch die Adaption der Niveaugrenzen, die so lange wiederholt werden muss, bis die beiden verglichenen Tests stimmige Einstufungen ergeben, entfällt dabei.

Die im DESI-Projekt eingesetzten C-Tests wurden zunächst in ihren schwierigkeitsbestimmenden Merkmalen analysiert und unabhängig vom GER-System beschrieben. Jeder Text und jede darin befindliche Lücke (wobei hier die Lücken als Testitems verstanden werden) wurde hinsichtlich dieser abgestuft beschriebenen Aufgabenmerkmale eingeschätzt. Die Testitems wurden administriert und Rasch-skaliert. Die Rasch-Skala wurde unabhängig vom Referenzsystem des GER in Kompetenzniveaus eingeteilt. Die Niveaueinteilung gründet einerseits in den erwähnten schwierigkeitsbestimmenden Merkmalen. Mittels Korrelationsanalysen wurden die vorhersagestärksten Merkmale identifiziert, die in ein Regressionsmodell[231] eingeflossen sind, welches bestimmten Merkmalskombinationen bestimmte Schwierigkeiten vorhersagt. Diese wurden genutzt, um dort Schwellen zwischen den Niveaus zu ziehen, wo es eine Häufung von Items mit bestimmten Merkmalskombinationen, so genannte *Item-Cluster* gab. Andererseits gründen die Niveaueinteilungen in inhaltlichen Itemanalysen: Die Rasch-skalierten Testitems

[230] Für eine detailliertere Darstellung vgl. Harsch & Schröder 2005b.
[231] Vgl. Hartig 2005.

wurden – wiederum unabhängig vom GER-System – in ihren Anforderungen und Schwierigkeiten beschrieben, um sie in ihren Charakteristika auch von empirischer Seite zu erfassen. Diese Beschreibungen wurden auch für die endgültige Schwellenziehung herangezogen: Die Schwellen wurden dort gesetzt, wo es zu den besagten *Itemclustern* kam und die inhaltlichen Analysen eine solche Schwelle ebenfalls nahe legten.

Die so entstandenen Kompetenzniveaus werden durch Deskriptoren beschrieben, die ihre Basis einerseits in den schwierigkeitsbestimmenden Merkmalen haben und andererseits in den erwähnten Itembeschreibungen. Daraus ergibt sich die Beschreibung des konkreten Umgangs mit der Aufgabenstellung. Diese spezifischen Niveaubeschreibungen wurden mit relevanten Skalen aus dem GER[232] abgeglichen und dieser Abgleich wurde zum Rückschließen auf die zugrunde liegenden Kompetenzen genutzt, so dass auf den DESI-C-Test Kompetenzniveaus auch Generalisierungen in Bezug auf linguistisches Wissen, Lesestrategien und Fähigkeiten der Textrezeption beschrieben werden können. Dabei wurden die Wortlaute der jeweiligen Deskriptoren der DESI-Skala und der entsprechenden GER-Skalen abgeglichen (vgl. dazu GER 2001: 34 und 203): Es konnten Anknüpfungspunkte auf den Niveaus A1 mit B2+ identifiziert werden. Es wird jedoch nicht der Anspruch erhoben, die DESI-C-Test-Niveaus eindeutig den GER-Niveaus zuzuordnen, da das C-Test-Konstrukt eben keine Entsprechung im GER hat – dennoch können die GER-Deskriptoren, wo es inhaltliche Übereinstimmungen nahe legen, zur Beschreibung von Testergebnissen und Probandenfähigkeiten genutzt werden, eben in ihrer Funktion als benutzerorientierte Skalen zum Berichten von Testergebnissen mit Bezug auf ein gemeinsames Referenzsystem.

Zusätzlich schlägt beispielsweise North[233] vor, eine empirische Validierung der Anknüpfpunkte dadurch zu erzielen, dass die C-Tests in einer kleinen Probandengruppe administriert werden. Dieselbe Probandengruppe wird von einer Lehrkraft, die diese Gruppe sehr gut kennen muss, bezüglich ihres generellen Sprachstandes mithilfe von entsprechenden Deskriptoren aus dem Referenzsystem des GER eingeschätzt. Dann kann man, etwa über Korrelationskoeffizienten und Regressionsanalysen, die fraglichen C-Tests auf die Niveaus des GER kalibrieren.

[232] Relevante Skalen des *GER* (Seitenangabe) waren: Lesen (74f), Rezeptionsstrategien (78), Texte verarbeiten (98), Lexik (112f), Grammatik (114), Orthographie (118), Kohäsionsmittel (125), Spektrum allgemein (110).
[233] In einem persönlichen Gespräch auf der ALTE-Tagung in Berlin im Mai 2005 im Anschluss an den Vortrag von Harsch & Schröder zum Thema „C-Test-Kompetenzniveaus im DESI-Projekt und die Niveaus des GER".

4 Das Testmodul *Textproduktion Englisch* im DESI-Projekt

Das Untersuchungsdesign der in der Einleitung vorgestellten DESI-Studie sieht zum einen sprachlich-kommunikative Tests vor, um den Leistungsstand der Schülerinnen und Schüler und dessen Veränderungen zu erfassen. Zum anderen werden Fragebögen zur Erfassung von Hintergrundvariablen zur Situation der Schülerinnen und Schüler, ihrer häuslichen Bedingungen, des Unterrichts und der schulischen Bedingungen eingesetzt und eine videographische Studie des Englischunterrichts durchgeführt. Dadurch sollen Leistungsunterschiede über den Einfluss schulischer, unterrichtlicher und personaler Faktoren erklärt und Optimierungsansätze für den Unterricht aufgezeigt werden. Folgende Testmodule umfasst das DESI-Design: In Deutsch werden Kenntnisse und Fähigkeiten[234] in den Bereichen *Wortschatz*, *Rechtschreibung, Kommunikation und Argumentation*, *Leseverstehen*, *Textproduktion* und *Sprachbewusstheit* erfasst; im Englischen gibt es Module zu den rezeptiven Bereichen des *Lese-* und *Hörverstehens*, zu den produktiv-interaktiven Bereichen *freie Textproduktion* und *mündliche Sprachproduktion* und zu den reflexiven Bereichen *Sprachbewusstheit* und *interkulturelle Kompetenzen*; zusätzlich erfasst ein C-Test an der Schnittstelle zwischen Rezeption und Produktion den *globalen Sprachstand*.[235]

Deutschtests		**Englischtests**
Wortschatz	Leseverstehen	Hörverstehen
Rechtschreibung	Textproduktion	mündliche Produktion
Kommunikation/ Argumentation	Sprachbewusstheit	interkulturelle Kompetenzen
		globaler Sprachstand

Tabelle 5: Übersicht Testmodule im DESI-Projekt

Das Ziel der „Beschreibung der quantitativen Seite der interindividuellen Varianz der Leistungen“ (Klieme: 2004) rückt zwar quantitativ-psychometrische Verfahren in den Vordergrund. Dennoch wird die qualitative Seite der Leistungen nicht ganz außer Acht gelassen: In den einzelnen Testmodulen werden die Testanforderungen respektive die durch die Tests erzielten Leistungen qualitativ analysiert und beschrieben, so dass beide Perspektiven in gegenseitiger Ergänzung genutzt werden können.

[234] Im Konsortium des DESI-Projekts hat man sich auf folgende Terminologie geeinigt:
Der Begriff der *Kompetenz* wird synonym mit *Fähigkeit* verwendet und umfasst Wissensbestände und deren adäquate Anwendung. (Im Begriff der *Kompetenz* in DESI spiegelt sich der Begriff der *proficiency* wider.) *Fähigkeiten* in DESI umfassen die didaktischen Konzepte der *Fähigkeiten* (*abilities*) und der *Fertigkeiten* (*skills*). Zum Kompetenzbegriff in DESI vgl. auch Hartig & Klieme 2005.
In diesem Kapitel der vorliegenden Arbeit wird die didaktische Terminologie verwendet, wie sie in Kapitel 1 dieser Arbeit erläutert wurde. Soweit erforderlich, wird sie jeweils in Fußnoten kommentiert und in Bezug gesetzt zur Terminologie des DESI-Projekts.

[235] Die DESI-Module orientieren sich an den traditionellen vier Fertigkeiten (Hörverstehen, Lesen, Sprechen, Schreiben, ergänzt durch reflexive Fertigkeiten im Bereich der Bewusstheit bezüglich grammatischer, pragmatischer und interkultureller Phänomene). Die DESI-Module können (auf einer Ebene oberhalb der Fertigkeiten) den Bereichen der Rezeption, Produktion und Interaktion zugeordnet werden, die auch der GER (2001: 25f, wie in Kapitel 2 dieser Arbeit erläutert) zur Klassifizierung kommunikativer Aktivitäten vorschlägt: Dort werden die vier Grundtypen Rezeption, Produktion, Interaktion und Sprachmittlung differenziert, die jeweils noch in mündliche und schriftliche Aktivitäten unterschieden werden können. Insofern können die DESI-Module der Aktivitätstypologie des GER zugeordnet werden.

Aufgrund der Produktorientierung des DESI-Projekts können in der Studie keine Prozesse (seien es nun Spracherwerbsprozesse, Prozesse der Sprachproduktion oder auf spezifische Fertigkeiten bezogene Prozesse wie beispielsweise Prozesse der Textgenese) erfasst und untersucht werden. Dadurch können keine Aussagen darüber getroffen werden, welche Prozesse lernförderlich sind und wie sie effektiv vermittelt werden können. Eine Schulleistungsstudie wie DESI ist jedoch aus praktischen Gründen (begrenzte Ressourcen wie Zeit, finanzielle Mittel, Durchführbarkeit der Erhebungen, etc.) gezwungen, sich auf bestimmte Ziele zu beschränken und kann deshalb nicht alle Aspekte erfolgreichen Lernens und Lehrens erfassen. Um dennoch zu didaktisch verwertbaren Aussagen im Bereich von lernförderlichen Prozessen zu kommen, müssten Schulleistungsstudien ergänzt werden um beispielsweise Langzeitstudien zum Spracherwerb und zur Genese der einzelnen Teilfertigkeiten, ergänzt werden um Prozessanalysen der Sprachproduktion und Interaktion im mündlichen und schriftlichen Bereich, ergänzt werden um solche in den rezeptiven Bereichen und schließlich ergänzt werden um Untersuchungen der Lernkontexte und der Unterrichtsbedingungen in der Schulung der einzelnen Fertigkeitsbereiche.[236]

Die fehlende Prozessperspektive im DESI-Projekt wird allerdings zum Teil aufgefangen durch die Verankerung der Testkonstrukte in Theorien des Erwerbs und der Genese der jeweiligen Fertigkeitsbereiche. Beispielsweise ist das Konstrukt des Moduls *Textproduktion Englisch* unter anderem in der Forschung zur Schreibentwicklung und zu Schreibprozessen verankert (vgl. die Ausführungen unten in Kapitel 4.2). Zum Teil wird die fehlende Prozessperspektive auch aufgefangen durch die Videostudie des Englischunterrichts. Zwar können damit beispielsweise keine konkreten Schreibprozesse erfasst werden, dennoch lassen sich dadurch Aussagen zu lernförderndem Unterricht und damit verbundenen Lehr- und Lernprozessen treffen.

Die Studie umfasst unterschiedliche Phasen[237]: In der Planungsphase (2001/2002) wurden Konzepte und Konstrukte definiert, die die Grundlage für die Instrumentenentwicklung in der Designphase bildeten. Anschließend wurden die ersten Instrumente in der Phase der Präpilotierung (2002) informell an lokalen Schulen getestet, um ihre Güte und Durchführbarkeit zu bestimmen. Die auf Grundlagen der Präpilotierungen revidierten Instrumente kamen in der Pilotphase (September 2002 - März 2003) deutschlandweit in einem formellen Testlauf mit ungefähr 500 Probanden zum Einsatz, um die Instrumente zu validieren. Die Erkenntnisse aus Präpilotierung und Pilotierung wurden genutzt, um die Instrumente abschließend zu überarbeiten und sie adäquat auf die Hauptuntersuchung auszulegen. Die Hauptuntersuchung fand zu zwei Messzeitpunkten statt, im Herbst 2003 und im Sommer 2004, um die Veränderungen während eines Schuljahres zu erfassen. Die Tests wurden in einem Matrixdesign[238] eingesetzt; es kamen jedoch

[236] Vgl. hierzu etwa Cumming 1998.
[237] Es darf auf den Zeitplan unter http://www.dipf.de/desi/zeitplan.html verwiesen werden, Zugriff am 09.06.2005. Vgl. Anhang 21. Zu den verschiedenen Phasen in der Testentwicklung vgl. auch Kapitel 2.6 dieser Arbeit.
[238] Dabei werden nicht alle Tests von allen Probanden bearbeitet, da dies bei der Vielzahl der Tests in der gegebenen Zeit nicht machbar wäre. Vielmehr bearbeiten verschiedene Probandengruppen unterschiedliche Testteile, wobei diese so eingesetzt werden, dass es zu hinreichenden Überlappungen (so genannten Verankerungen) in der Bearbeitung der Testteile kommt. Dadurch können alle Probanden und alle Testvarianten auf einer gemeinsamen psychometrischen Skala abgebildet werden.

aus Zeitgründen nicht alle Testmodule zu beiden Messzeitpunkten zum Einsatz. So wurde beispielsweise das Testmodul *semikreatives Schreiben* nur zum zweiten Messzeitpunkt eingesetzt.

Die Testergebnisse, in die die oben erwähnten quantitativen und qualitativen Perspektiven einfließen, werden an die teilnehmenden Schulen rückgemeldet. Diese Rückmeldung wird begleitet durch eine Studie des Instituts für Schulentwicklungsforschung (IFS) an der Universität Dortmund. Damit soll untersucht werden, wie einzelne Schulen mit diesen Rückmeldungen umgehen. Die Ergebnisse der DESI-Studie insgesamt (Testleistungen und Zusammenhänge mit den oben erwähnten Hintergrundvariablen) werden der KMK berichtet, die ihrerseits die durch die Studie aufgedeckten Optimierungsansätze in den Schulen umsetzen muss.

Um die Ergebnisse der DESI-Studie in Bezug setzen zu können zu Vorarbeiten der KMK wie der Entwicklung von Curricula und Bildungsstandards, die sich am *Gemeinsamen europäischen Referenzrahmen* orientieren[239], wäre es hilfreich, wenn sich die DESI-Ergebnisse ebenfalls am GER orientieren könnten. Welche Bedeutung hat der GER aber im DESI-Projekt? Wie und bei welchen Aspekten wurde er verwendet? Wie hilfreich hat er sich dabei erwiesen? Die Kategorien und Skalen des GER sind, wie oben erläutert, auf fremdsprachliche Kontexte hin ausgelegt, so dass der GER nur in den Modulen die englische Sprache betreffend relevant ist. Der hier gegebene Kurzüberblick und die im Anschluss in den Kapiteln 4.1 mit 4.6 folgende Konkretisierung konzentrieren sich auf das Modul *Textproduktion Englisch*, da die Bedeutung und Verwendung des GER je nach Modul anders ausgeprägt ist: Bedingt durch die (in den vorangegangenen drei Kapiteln dieser Arbeit analysierten) Inhalte, Beschreibungsbereiche und Probleme des GER-Referenzsystems und nicht zuletzt bedingt durch die (im vorangegangenen Kapitel 3 analysierten) Verwendungsmöglichkeiten seiner Skalen wird der GER im Modul *Textproduktion Englisch* nicht als Ausgangspunkt der Testentwicklung gewählt, sondern seiner Ausrichtung gemäß als *Referenz*mittel genutzt. Das Testkonstrukt selbst steht auf „eigenen Füßen“, da es sich, wie unten gezeigt wird, nicht befriedigend im System des GER verorten lässt. Im Rahmen der Testauswertung basiert die Entwicklung des Bewertungsinstrumentariums, die unter Kapitel 4.4 dieser Arbeit dokumentiert wird, ebenfalls nicht auf dem GER, sondern auf theoretischen Überlegungen und empirischen Analysen von Schülerperformanzen. Der GER wird vielmehr als Vergleichsrahmen bei der Kriterienfindung und bei der Konstruktion der *rating scales* genutzt, denn wie oben in Kapitel 3 dieser Arbeit gezeigt, sind die Skalen des GER nicht zur direkten Bewertung von Performanzen geeignet. Bei der Rückmeldung der Testergebnisse mittels Kompetenzskalen[240], die aus den *rating scales* entwickelt werden, dient der GER als Referenzmittel, um die DESI-Deskriptoren zusätzlich zu validieren.[241]

[239] Vgl. Kultusministerkonferenz 2003 und Klieme, Avenarius et al. 2003.
[240] Wiederum sei der Hinweis gestattet, dass es sich im DESI-Projekt bei den so genannten Kompetenzskalen eher um Skalen des Sprachvermögens, in diesem Fall um Skalen das Schreibvermögen in der Fremdsprache betreffend handelt, wobei diese Kompetenzskalen sowohl konkrete aufgabenbezogene Fertigkeiten beschreiben als auch Generalisierungen in Bezug auf das Schreibvermögen insgesamt enthalten.
[241] Die Validierung geht wie erwähnt auch in umgekehrter Richtung: Wie unten gezeigt wird, können die DESI-Deskriptoren, die u. a. auf empirischen Aufsatzmerkmalen basieren, als zusätzliche inhaltliche Validierung der GER-Deskriptoren betrachtet werden.

In den folgenden Unterkapiteln wird die Konstruktion und Entwicklung des Testmoduls *Textproduktion Englisch* dokumentiert: Ausgehend vom Testkonzept wird die Verankerung des Testkonstrukts, die Entwicklung der Testinstrumente und des Bewertungsschemas und schließlich die Durchführung und Auswertung des Schreibtests dokumentiert, ehe auf Aspekte der Rückmeldung der Testergebnisse eingegangen wird. Am Ende eines jeden Unterkapitels wird die Bedeutung des GER diskutiert und gegebenenfalls seine Verwendung dokumentiert. Das Kapitel schließt ab mit einem Ausblick, wie Ergebnisse und Produkte solch groß angelegter Schulleistungsstudien in den täglichen Unterricht rückfließen könnten, um die systemische Validität[242] solcher Studien zu erhöhen.

4.1 Testkonzept

Im Modul *Textproduktion Englisch* soll erfassen, inwieweit Schülerinnen und Schüler der 9. Jahrgangsstufe eigenständige englische Texte verfassen können und inwieweit ihre Schreibfertigkeit entwickelt ist. Dabei ist zu bedenken, dass sich die Schreibfertigkeit, die Fertigkeit also, kommunikativ wirksame Texte zu verfassen, aus dem komplexen Zusammenspiel verschiedener Kompetenzen und Wissensbestände und aus deren Anwendbarkeit speist. Diese Leistungsdimensionen werden im Testkonstrukt definiert (siehe Kapitel 4.2), um eine Basis für die Operationalisierung der Aufgaben und für die Bestimmung der Bewertungskriterien (die zugleich die horizontale Einteilung der *rating scales* darstellen) zu erhalten. Darüber hinaus ist zu bedenken, dass sich *die* Schreibfertigkeit nicht global feststellen lässt, sondern immer bezogen ist auf bestimmte Schreibanlässe und Kommunikationssituationen, weshalb diese in der unter Kapitel 4.3.1 angeführten Aufgabenspezifikation näher bestimmt werden. Um für die Testpopulation relevante Anlässe und Situationen zu identifizieren und geeignete Testinstrumentarien zu entwickeln, ist es hilfreich, vorab die einschränkenden Bedingungen des Testmoduls festzuhalten. Diese dienen zusammen mit dem Testkonstrukt und der Aufgabenbeschreibung als Rahmen, innerhalb dessen von den gezeigten Performanzen auf zugrunde liegende Kompetenzen und Fertigkeiten generalisiert werden kann:

- Der Test als solches stellt eine künstliche Situation dar, in der die Probanden wissen, dass sie nicht zu realen Kommunikationszwecken schreiben, sondern für einen Bewerter, der die Rolle des Textrezipienten einnimmt. Der scheinbare Widerspruch zur didaktischen Forderung nach authentischen Schreibanlässen findet sich jedoch auch in der Unterrichtssituation. Nur selten wird zu realen Anlässen, etwa im Rahmen eines E-Mail Projektes, ein fremdsprachlicher Text verfasst. Es darf demnach davon ausgegangen werden, dass die Lernenden mit der erwähnten Pseudosituation vertraut sind und das Testergebnis dadurch nicht wesentlich verfremdet

[242] Vgl. dazu die Ausführungen in Kapitel 2.3.1 und 2.4 dieser Arbeit.

wird. Zudem stellt ein handlungsorientiertes kommunikatives Testformat eine Annäherung an die Realität dar, das gewisse Rückschlüsse auf die Handlungsfähigkeit im realen Leben zulässt.

- Der Horizont der Testpopulation bestimmt den Rahmen, innerhalb dessen valide Testinstrumente entwickelt werden können. Die Aufgabenstellung muss ausgelegt sein auf die Lebenswelt von Lernenden der 9. Jahrgangsstufe im Alter von etwa 14 bis 16 Jahren, auf deren Erfahrungen und allgemeinen Entwicklungsstand, und sie muss ausgelegt sein auf die Lernerfahrungen in der Fremdsprache Englisch. Deshalb basiert das Testkonstrukt unter anderem auf Curriculaanalysen (vgl. Kapitel 4.2.3). Dabei ist zu bedenken, dass eine Aufgabenstellung gewählt werden muss, die von Lernenden mit unterschiedlich ausgeprägten Kompetenzen gleichermaßen bearbeitet werden kann, da in der DESI-Studie alle Schulformen getestet werden. Freie Schreibformate können diese Bedingung erfüllen, wie im folgenden Kapitel 4.2 gezeigt wird. Die unterschiedlichen Entwicklungsstände müssen auch bei der Konstruktion des Bewertungsinstrumentariums bedacht werden: Dieses muss verschiedenen Wegen und Realisierungen der Aufgaben gerecht werden, weshalb ein Positivansatz gewählt wird, der in Kapitel 4.4 näher beschrieben wird.

- Wie bereits angedeutet unterliegen Schulleistungsstudien Beschränkungen, sei es nun hinsichtlich des zeitlichen Rahmens oder der Durchführbarkeit der einzelnen Tests. Aufgrund dieser Beschränkungen muss für das Modul *Textproduktion Englisch* in Kauf genommen werden, dass es nur zum zweiten Messzeitpunkt zum Einsatz kommen kann, weshalb sich innerhalb dieses Moduls keine Veränderungsmessung durchführen lässt. Man muss also bedenken, dass es sich bei dieser Testleistung um eine reine Momentaufnahme handelt, weshalb der empirische Blick auf die Perspektive der Schreibentwicklung innerhalb eines Schuljahres verschlossen bleibt. Für die Bearbeitung einer Schreibaufgabe stehen 20 Minuten zur Verfügung, weshalb die Aufgabenstellung so ausgelegt sein muss, dass sie in dieser Zeit bearbeitbar ist und einen Text elizitiert, der genügend lang ist, um bewertet werden zu können.

- Die erwähnte Produktorientierung des DESI-Projekts lässt es nicht zu, der Produktion zugrunde liegende Schreibprozesse zu beobachten oder zu erfassen. Bestimmte Merkmale der Produkte jedoch lassen auf Prozesse des Schreibens und des Erwerbs schließen: Beispielsweise kann von der Struktur eines Textes her auf Prozesse der Planung einerseits und auf Erwerbsprozesse andererseits geschlossen werden, allerdings nur auf Basis von theoretischen Überlegungen und Modellen, da sich die DESI-Studie den Prozessen nicht empirisch nähern kann. Entsprechende Ausführungen sind in Kapitel 4.2 zu finden.

Basierend auf diesen Rahmenbedingungen sind folgende grundlegende Entscheidungen getroffen worden, um valide Instrumente zum Erfassen der Schreibfähigkeit in der gegebenen Zielgruppe zu entwickeln:

- *Direktes Testen*: Die Schreibfertigkeit wird über ein direktes Format erfasst, da ein valider Test diejenigen Fertigkeiten, Prozesse und sprachlichen Handlungen erfassen muss, die auch in realen Situationen benötigt werden, um die entsprechende kommunikative Handlung auszuführen: "Since there is a strong general sense that good writing tests should involve students producing writing, indirect measurements of writing ability are not likely to remain viable options in the foreseeable future." (Grabe and Kaplan 1996: 399). Indirekte Formate lassen, wie seit Beginn der 70er Jahre kritisiert wird, keine validen und reliablen Rückschlüsse auf in realen Handlungssituationen zur Verfügung stehende Fertigkeiten zu. Bei einem authentischen, kommunikativen Format hingegen kommen alle am Schreiben beteiligten komplexen Wissensbestände und Fertigkeiten verschränkt zum Einsatz. Um die Schreibfertigkeit von anderen Fertigkeiten abzugrenzen[243], wird die Aufgabenstellung in der Muttersprache gegeben, so dass die Erfüllung der Aufgabe nicht etwa von der Lesefertigkeit abhängt. Kroll (1998) bemerkt in diesem Zusammenhang, dass sich Schreiben in Interaktion mit den anderen sprachlichen Teilfertigkeiten entwickelt, eine Parallelentwicklung jedoch nicht gegeben sein muss. Deshalb rät sie, die Teilfertigkeiten getrennt voneinander zu testen und in Ergänzung dazu den globalen Sprachstand zu erfassen. [244]

- *Freies Schreiben*: Aufgrund der Heterogenität der Probandengruppe wird ein Format gewählt, das allen Lernenden genügend Offenheit bietet, um die Aufgabe auf unterschiedlichen Wegen gemäß des jeweiligen Entwicklungsstands zu lösen und dabei individuelle Schwerpunkte setzen zu können. Um die Vergleichbarkeit der Produkte zu erzielen, wird hinreichend Lenkung gegeben, indem Adressat, Schreibanlass und Kommunikationssituation bestimmt werden. Die semikreative Aufgabenstellung, bekannt aus dem Bundeswettbewerb Fremdsprachen[245], erfüllt diese Anforderungen und besitzt einen hohen Grad an *face validity*[246]. Solche Formate, deren Realitätsbezug offensichtlich ist und deren Beantwortung die Probanden selbst mitsteuern können, wirken vermutlich motivierender auf die Probanden als geschlossene Formate, deren *face validity* nicht offensichtlich ist und die den Probanden keinen Spielraum in der Beantwortung lassen. [247]

- *Positivansatz, Kriteriumsorientierung, Rating-Verfahren*: Um eine reliable und valide Bewertung zu gewährleisten, die den unterschiedlichen Lösungsansätzen der Probanden gerecht wird, wird positiv an die Performanzen herangetreten: Nach Bleyhl (2003) wird man der Dynamik, Vielseitigkeit und Individualität des Lernprozesses dadurch gerecht, dass man das Positive betont und bewertet. Zudem kann die Positivbewertung Aufschlüsse geben über das Erreichen bestimmter Kriterien. Dazu werden solche Kriterien angesetzt, die relevant für die Produktion

[243] In der Fachliteratur findet sich etwa bei Börner 1989, Camp 1996, Hamp-Lyons 1996, Hughes 1986, Kroll 1998, Shohamy 1992 u. a. die Aussage, die einzelnen Fertigkeiten getrennt voneinander zu erfassen, um valide Generalisierungen aus den Testergebnissen ableiten zu können.
[244] Im DESI-Projekt ist die Vorgabe umzusetzen, die einzelnen Teilfertigkeiten möglichst unabhängig voneinander zu erfassen. Deshalb wird zusätzlich das integrative Format des C-Tests eingesetzt, um einen Blick auf das generelle Sprachvermögen zu erhalten.
[245] Vgl. hierzu etwa Finkenstaedt & Schröder 1989 oder Hertel 1994.
[246] Vgl. die Ausführungen zur *face validity* unter Kapitel 2.3.1 dieser Arbeit.
[247] Vgl. hierzu etwa Börner 1989, Shohamy et al. 1992.

sind und auf Theorien der Schreibproduktion und der kommunikativen Kompetenz beruhen (vgl. die Ausführungen unter Kapitel 4.2 und 4.4). Ziel der Untersuchung ist, den Leistungsstand der Probanden in Bezug auf diese Kriterien zu erfassen, und nicht etwa normorientiert festzustellen, wie gut die Probanden in Bezug auf die Probandengruppe sind. Um möglichst viele Kriterien adäquat erfassen zu können, die ihrerseits, wie in Kapitel 3.3.1 dieser Arbeit erläutert, die Reliabilität der Bewertung erhöhen, wird eine Kombination verschiedener *Rating*-Methoden angewandt. Die Validität der *rating scales* wird dadurch erzielt, dass sie auf der Beschreibung relevanter Aufsatzmerkmale basieren. Die Subjektivität der *ratings* wird durch *Rater*-Training, die Nutzung von *benchmarks* und Doppelt-Blind-Bewertung begrenzt. Die Objektivität wird zusätzlich erhöht durch die Skalierung der *ratings* auf Basis eines psychometrischen Modells, das den Strenge/Milde-Tendenzen der *raters* und den Aufgabenschwierigkeiten Rechung trägt.[248]

- *Generalisierbarkeit*: Da es sich bei Tests immer um Momentaufnahmen der Performanz handelt, die auch von der Tagesform der Probanden beeinflusst werden, ist die Generalisierbarkeit der Bewertung nur in begrenztem Rahmen gegeben. Um diese zu gewährleisten, werden insgesamt vier Aufgaben gestellt, denn je mehr Performanzen eines Probanden oder einer Probandengruppe vorliegen, desto valider kann auf zugrunde liegende Kompetenzen rückgeschlossen werden. Allerdings sind in der DESI-Studie aus den oben genannten Gründen Grenzen gesetzt, so dass jede Schülerin und jeder Schüler nur zwei Aufgaben bearbeiten kann; doch in jeder Klasse kommen alle vier Aufgaben zum Einsatz. Um die begrenzte Zeit möglichst effizient zu nutzen, werden keine Wahlmöglichkeiten gegeben, sondern die Zuteilung der zu bearbeitenden Aufgaben wird als *random effect* behandelt, dem in der Skalierung Rechnung getragen wird. Die Skalierung der vier Aufgaben ergab eine Gesamtskala Schreiben, in die alle vier Aufgaben und alle Kriterien einflossen, so dass auf Klassenebene auf Basis der erwähnten Aufgabenspezifikationen die Vergleichbarkeit der Produktionen als akzeptabel betrachtet wird.[249] In diesem Rahmen sind generalisierende Aussagen über den Leistungsstand einer Klasse im Bereich der Schreibfertigkeit möglich.

Bei diesen grundsätzlichen Überlegungen und Entscheidungen hat der kontextfrei und generell gehaltene GER eine untergeordnete Rolle gespielt, denn die Rahmenbedingungen einer Schulleistungsstudie wie DESI sind konkret zu füllen. Ein Vergleich des unabhängig vom GER entwickelten Testansatzes im DESI-Modul *Textproduktion Englisch* und des Testbegriffs des GER (vgl. Kapitel 2.5.3 dieser Arbeit) zeigt zwar, dass es Deckungsbereiche gibt hinsichtlich des modell-basierten Testansatzes und des kommunikativen, authentischen, handlungsorientierten Herantretens, hinsichtlich der Annahme der Verschränktheit der Wissensbestände und Kompetenzbereiche, die bei der Sprachproduktion zum Einsatz kommen, hinsichtlich des direkten

[248] Die Skalierungen finden an der Humboldt-Universität zu Berlin statt, da Professor Lehmann (Konsorte im DESI-Projekt) und seine Mitarbeiterin Frau Neumann Spezialisten auf dem Gebiet von *Rating*-Skalierungen sind.
[249] Vgl. hierzu auch die Ausführungen in Shale 1996 und in den Kapiteln 2.3, 3.1 und 3.3.2.2 dieser Arbeit.

Herantretens an produktive Fertigkeiten und hinsichtlich der Akzeptanz der Subjektivität bei der Bewertung offener Aufgaben sowie hinsichtlich der Möglichkeiten, diese zu objektivieren. Dennoch stellt etwa Abschnitt 9 des GER *Beurteilen und Bewerten* keine Erkenntnisse zur Verfügung, auf die eine anstehende Entscheidung gestützt werden könnte. Beispielsweise kann die Frage, ob man die Schreibfertigkeit indirekt oder direkt erfassen will, nicht auf Basis der dortigen Ausführungen (vgl. GER: 181f) entschieden werden, da diese Ausführungen die beiden Herangehensweisen lediglich definieren, sie jedoch nicht in ihren Bedingungen und Vor- und Nachteilen charakterisieren. Deshalb muss die Begründung für das eine oder andere Vorgehen stets aus den konkreten Testbedingungen und aus der Fachliteratur abgeleitet werden. Die Ausführungen auf S.182 des GER zu subjektiven Bewertungsverfahren und die sich anschließenden Vorschläge, wie solch eine Bewertung möglichst objektiv durchgeführt werden kann, sind hingegen aufschlussreicher und bieten zumindest eine hilfreiche Checkliste, ob alle Möglichkeiten der Objektivierung genutzt wurden. Der in Kapitel 2.6 dieser Arbeit diskutierte UGE, insbesondere die detaillierte Darstellung der einzelnen Phasen im Testentwicklungsprozess in seinem Abschnitt 2, ist eine zusätzliche Hilfe, um die Güte der Testkonstruktion fortlaufend zu kontrollieren. Doch letztlich müssen alle Entscheidungen in einem gegebenen Testentwicklungsprozess in ihren Kontexten begründet werden.

4.2 Konstrukt der Schreibfertigkeit im DESI-Projekt[250]

Die fremdsprachliche Schreibfertigkeit wird durch die Aktivität des Schreibens selbst entwickelt und verbessert, so dass das fremdsprachliche Schreibkonstrukt mit den Worten Camps definiert werden kann: "... Writing [is] a rich, multifaceted, meaning-making activity that occurs over time and in a social context, an activity that varies with purpose, situation, and audience and is improved by reflection on the written product and on the strategies used in creating it" (Camp 1996: 135).

Das Konstrukt der Schreibfertigkeit im DESI-Projekt ist dreifach verankert: Es basiert auf textlinguistischer Forschung, auf der didaktischen Schreibentwicklung- und Schreibprozessforschung und auf Curriculaanalysen der Lehrpläne für 8. und 9. Klassen aller Schulformen in allen Bundesländern.

[250] Die Schreibfertigkeit ist, wie in Kapitel 1.2.3 erläutert, Teil der handlungsbezogenen kommunikativen Kompetenz. Während man, wie gesagt, in der Terminologie der Fremdsprachendidaktik von Schreib*fertigkeit* spricht, wird diese im DESI-Projekts als Schreib*fähigkeit* bezeichnet.

4.2.1 Funktionale Linguistik

Im Testkonstrukt des DESI-Schreibmoduls werden Texte im Sinn der funktionalen Textlinguistik[251] verstanden als Versuch, einem gegebenen Adressaten eine bestimmte Sprechabsicht zu vermitteln: "Writing may be said to represent an attempt to communicate with the reader." (Grabe & Kaplan 1996: 41). Dabei ergibt sich die Wirksamkeit eines Textes aus der adressatenbezogenen Übermittlung der Sprechabsicht in der gegebenen Kommunikationssituation. Daher sind es nicht alleine sprachliche Kompetenzen, die die Wirksamkeit eines Textes konstituieren. Hinzu treten vielmehr Welt- und Sachwissen und pragmatisches Handlungswissen. Selbstverständlich wird die Wirksamkeit durch die Rezipienten konstituiert – in diesem Fall durch die *raters*, die sich dazu in die Rolle der fiktiven Adressaten versetzten müssen. Diesem Aspekt wird in der *Rater*-Schulung (siehe Kapitel 4.5.1) besondere Aufmerksamkeit geschenkt.

In der fremdsprachlichen Schreibsituation dient das muttersprachliche Handlungswissen als Folie aller Schreibprozesse. Während in der Muttersprache davon ausgegangen werden darf, dass gesellschaftlich bedingte textsortenspezifische Handlungsmuster vorhanden sind,[252] an denen sich die Beurteilung orientieren kann, ist dies so für die Fremdsprache nicht vorauszusetzen: Die Schülerinnen und Schüler haben nur begrenzte Erfahrungen im Umgang mit authentischen Texten und realen Schreibanlässen in der unterrichtlichen Situation; die wenigsten Lernenden dürften über entsprechende Erfahrungen in der fremdsprachlichen Zielgesellschaft verfügen. Deshalb müssen Curriculaanalysen zeigen, welches textsortenspezifische Handlungswissen vorhanden sein sollte und daher in den Aufgabenstellungen gefordert werden kann.

Die funktionale Textlinguistik kennt zahlreiche Kategorisierungsmöglichkeiten für Textsorten, doch gibt es bisher kein System, das allgemein anerkannt und in der Lage wäre, bestimmten Textsorten konstituierende Merkmale zuzuordnen.[253] Daher kann auch keine „Textnorm" angesetzt werden, an der die fremdsprachlichen Textprodukte gemessen werden könnten. Vielmehr gibt es unterschiedlichste Wege, einen Text zielgerichtet, adressatenbezogen und wirksam zu strukturieren und zu verfassen. Dieser Diversität der Produkte müssen das Bewertungsinstrumentarium und die Bewerter gerecht werden; auf diese Aspekte wird unter Kapitel 4.4 *Bewertungsschema* eingegangen.

4.2.2 Schreiberwerbs- und Schreibprozessforschung

Zur **Entwicklung der Schreibfertigkeit** findet sich bei Bereiter (1980) ein Modell, das aufbauend auf dem Modell der hierarchischen *Skill*-Integration von Schaeffer (1975) unterschiedliche Phasen der Entwicklung ansetzt, ohne jedoch zu behaupten, dass diese Phasen festgelegten

[251] Vgl. beispielsweise Grabe & Kaplan 1996 oder Halliday & Hasan 1989.
[252] Vgl. etwa Becker-Mrotzek 1997, Brinker 1988.
[253] Vgl. dazu beispielsweise De Beaugrande & Dressler 1981, Halliday & Hasan 1989 oder Schiffrin 1994.

Sequenzen einer natürlichen Erwerbsordnung entsprächen. Nach diesem Modell können Kinder (im Erstspracherwerb) und Jugendliche (beim Erlernen einer Fremdsprache) aufgrund mangelnder "information-processing capacities" (vgl. Bereiter 1980: 83) erst nach und nach die benötigten sprachlichen, intellektuellen, sozialen und kognitiven Fertigkeiten in ihr Wissenssystem *Schreiben* integrieren und automatisieren. Jede Phase ist demnach gekennzeichnet durch die Nutzung und Integration bestimmter Fertigkeiten; erst wenn diese automatisiert sind, werden wieder Kapazitäten frei, um weitere Fertigkeiten zu integrieren. Folgende Abbildung zeigt die Charakteristika der unterschiedlichen Phasen (vgl. Bereiter 1980: 84):

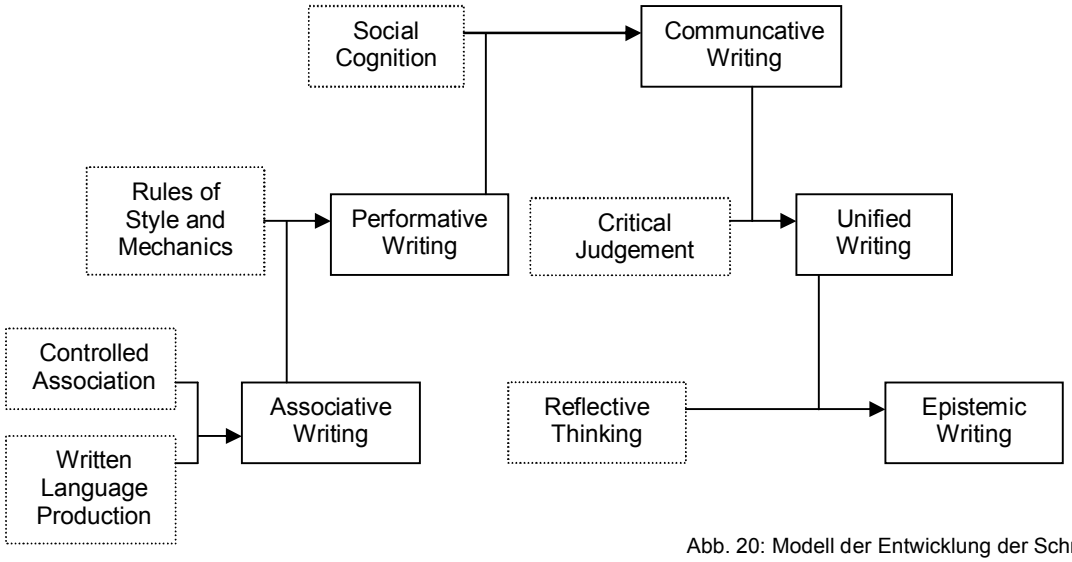


Abb. 20: Modell der Entwicklung der Schreibfähigkeit

Die erste Phase ist das *Associative Writing*, die einfachste Form des Schreibens, bei dem die Fertigkeiten *fluency of language* und *ideational fluency* integriert werden: Die Gedanken werden assoziativ niedergeschrieben, ohne Planung oder Beachtung formaler Regeln. Darauf aufbauend wird das System der Konventionen hinsichtlich Stil und sprachlicher Normen integriert und es kommt zur zweiten Phase: dem *Performative Writing*: Es werden Textsortenkonventionen, stilistische und orthographische Konventionen beachtet. Wenn nun *social cognition*, das Wissen um die Wirkungsweise von Texten, integriert wird, so treten die Lernenden in die dritte Phase des *Communicative Writing* ein. Hierbei wird die sprachliche Realisierung des Adressatenbezugs entwickelt; dies setzt natürlich voraus, dass sich die Schreibenden der Leserperspektive bewusst sind. Die darauf folgenden Phasen des *Unified* respektive *Epistemic Writing* sind bezogen auf den Fremdsprachenunterricht der 9. Jahrgangsstufe relativ irrelevant, da sie nach maximal 5 Jahren Unterricht nur in Ausnahmefällen erreicht werden können. In diesen Phasen geht es vorrangig um die Integration kritischer und evaluativer Lesefertigkeiten und um die Integration reflexiver Fertigkeiten, um die eigene Textproduktion kritisch zu reflektieren.

Bereiter stellt ausdrücklich fest, dass die verschiedenen Fertigkeiten und Schreibphasen durchaus in verschiedener Reihenfolge entwickelt respektive durchlaufen werden können, in Abhängigkeit von Lernervorwissen, Persönlichkeit und Schreibunterricht. So kann beispielsweise das assoziative Schreiben als eine Schreibtechnik in allen Phasen genutzt werden; oder es

kann sein, dass die Phase des *Performative Writing* niemals „gemeistert" wird, da nicht alle Subsysteme korrekt integriert werden; dies verhindert aber keineswegs eine Weiterentwicklung hin zur nächsten Phase, wenn mentale Kapazitäten zur Integration weiterer Systeme frei sind.

Dieses Schreibentwicklungsmodell dient bei der Konstruktion des Bewertungsinstrumentariums als eine theoretische Basis der Abstufungen. Hinzu tritt die empirische Verankerung des Bewertungsschemas in Aufsatzanalysen: Unter Kapitel 4.4 *Bewertungsschema* wird näher ausgeführt, wo es zu Berührungspunkten von Merkmalen der Schüleraufsätze aus der Präpilotierungsphase und Merkmalen aus Bereiters Modell kommt.

Auch wenn **Schreibprozesse** in DESI nicht erfasst werden können, helfen Modelle und Theorien der Schreibprozessforschung, die am Schreiben beteiligten Leistungsdimensionen zu identifizieren, um valide Aufgaben und Bewertungsinstrumente zu entwickeln. Bei Börner (1989) findet sich ein Prozessmodell, das eine Weiterentwicklung des Modells von Hayes & Flower (1980) darstellt. Börners Modell betrachtet den fremdsprachlichen Schreibprozess als einen dynamischen „Kreislauf von Interaktion zwischen Lehrer, Lerner und Texten" (Börner 1989: 354). Dieser Dynamik kann DESI nicht entsprechen, geht es doch dabei um eine Momentaufnahme. Dennoch können auch solche Momentaufnahmen wieder in den besagten Kreislauf zurückfließen, wie in Kapitel 4.7.2 dieser Arbeit gezeigt wird.

Börner unterscheidet zwei grundsätzliche Schreibtypen voneinander: den Typus *Aufsatz*, bei dem das Erstellen eines neuen Textes im Vordergrund steht, vom Typus der *Textarbeit*, bei der ein vorhandener Text umgearbeitet und fortgeschrieben wird. Damit erweitert Börner das Modell von Hayes & Flower um den Aspekt der Intertext-Funktion. Dieser Funktion kommt im realen Leben große Bedeutung zu, denn viele Schreibanlässe sind in der Reaktion auf einen gegebenen Text zu finden, so auch in DESI: Dort werden sowohl interaktive als auch produktive Formen des Schreibens erfasst, die sich zusätzlich auf Curriculaanalysen stützen.

Das Modell von Börner unterscheidet drei Ebenen: Die Ebene der *Schreibprozesse* (in der folgenden Graphik grau unterlegt) umfasst Strategien des Planens, Formulierens und Überarbeitens. Das DESI-Testmodul stellt den Aspekt des Formulierens in den Mittelpunkt, da aus zeitlichen Gründen Planungs- und Überarbeitungsstrategien vermutlich nur marginal zum Einsatz kommen können und deswegen auch nicht in die Bewertung einfließen. Börners zweite Ebene betrifft die *Schreibumgebung*, die Wissensbestände, die Aufgabenstellung und die Schreibhilfen in einer spezifischen Schreibsituation. In DESI gehen benötigte Wissensbestände in die Analyse der Leistungsdimensionen und damit auch in die Entwicklung der Bewertungskriterien ein; Aufgabenstellung und Schreibhilfen werden bei den Aufgabenmerkmalen definiert, um den Rahmen möglichst exakt zu bestimmen, innerhalb dessen auf die Schreibfertigkeit generalisiert werden kann. Die dritte Ebene in Börners Modell bezieht sich auf die *Lehrperspektive* und umfasst Aspekte wie beispielsweise Lehr- und Lernziele, Textauswahl, Methoden und

Übungsformen, Leistungsmessung und Bewertung. Diese Aspekte liegen der Beurteilung in DESI zugrunde und bestimmen deren Rahmen; sie werden durch curriculare Analysen erfasst. Auf den zyklischen Aspekt des Korrekturtextes und der Überarbeitung wird in Kapitel 4.7.2 *Ausblick* eingegangen, um Möglichkeiten aufzuzeigen, Textprodukte aus externen Beurteilungen wieder in den Unterricht rückfließen zu lassen.

Folgende Graphik zeigt das Schema von Börners Modell (vgl. Börner 1989: 355):

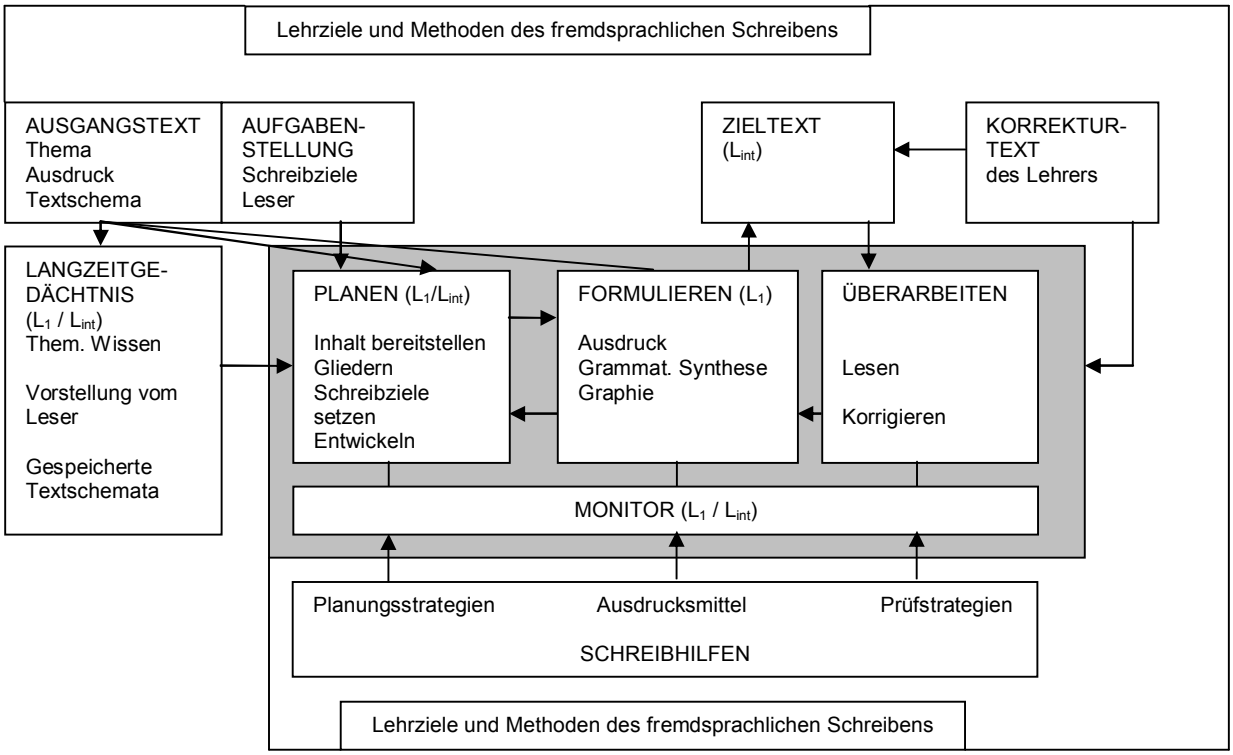


Abb. 21: Modell der fremdsprachlichen Schreibprozesse

In diesem Modell wird deutlich, dass sich muttersprachlicher Einfluss (L_1) bei den der Produktion zugrunde liegenden Wissensbeständen und Strategien (vgl. *Langzeitgedächtnis*), beim Planen und beim Formulieren bemerkbar macht, wie dies auch Krings (1986) feststellt: Die Muttersprache dient beim Schreiben in der Fremdsprache als Steuerungsmittel, so dass es zu einer Verschränkung muttersprachlicher und fremdsprachlicher Prozesse kommt, welche man jedoch nicht am textuellen Endprodukt direkt beobachten kann (ebd.: 423f). Beobachtbar am Endprodukt sind aber Einflüsse der Muttersprache auf die Interimsprache, so genannte Interferenzen, weshalb diese in DESI zur abgestuften Beschreibung der sprachlichen Kriterien genutzt werden. Die begrenzte Zeitvorgabe in DESI dürfte eine muttersprachliche Planung und anschließende Übersetzung in die Zielsprache erschweren, ebenso wie die Entscheidung, eine Wörterbuchnutzung nicht zu gestatten. Auf diese Weise werden automatisierte Schreibprozesse in der Fremdsprache stärker in den Vordergrund gerückt.

Das Schreibprozessmodell von Börner dient in DESI der Aufgabencharakterisierung, um möglichst alle für das DESI-Konstrukt relevanten Aspekte bei der Operationalisierung zu erfassen. Zusätzlich dient es bei der Entwicklung der Bewertungskriterien als Folie, so dass die

Bewertung zumindest die der Produktion zugrunde liegende Prozesse reflektiert, wenn letztere schon nicht eigens erfasst werden können.

Ergänzend zu den obigen Ausführungen tritt dieser Stelle der Blick auf **schreibdidaktische Ansätze**, um das Format des semikreativen Schreibens auch aus dieser Perspektive zu begründen. Die Vermittlung der Schreibfertigkeit in der Fremdsprache erfolgt im Allgemeinen „vom Wort zum Satz zum Text" (Kast 1999: 24f); die Fertigkeit, einen kohärenten Text zu erstellen, bedarf gewisser Vorerfahrungen und Übungen. Um diese zu vermitteln, können grob drei schreibdidaktische Ansätze nach Portmann (1991) herangezogen werden:

Direktive Ansätze üben das gelenkte Schreiben durch geschlossene Aufgabenformen mit engen Vorgaben. Dabei geht es noch nicht um freie Produktion, sondern um eine Vorstufe des produktiven Schreibens, um das „schriftliche Üben von (...) sprachlichen Elementen und Strukturen" (ebd.: 376). Der so genannte textlinguistische Ansatz dagegen stellt textkonstituierende Merkmale in dem Mittelpunkt; dementsprechend konzentriert sich der Unterricht auf die Erarbeitung dieser Merkmale und auf „Gegebenheiten, die fürs Herstellen von Texten relevant sind." (ebd.: 381). Dieses Herangehen ist aber nicht als Vorstufe des produktiven Schreibens zu sehen, sondern als Zyklus „des Kennenlernens, der Erarbeitung und Anwendung spezifischer Mittel und Verfahren" (ebd.). Als dritter Ansatz kann der prozessorientierte Ansatz genannt werden, der sich nicht mehr auf einzelne Teilfertigkeiten konzentriert, sondern Schreiben als Organisation verschiedener Arbeitsprozesse betrachtet, die die Erstellung angemessener Texte zum Ziel haben: „Das Schreiben eines Textes (freies Schreiben, produktives Schreiben) steht nicht mehr notwendig am Ende einer sorgfältig geplanten Folge von Übungen, in welchen einzelne Teilfertigkeiten oder Teilstrukturen isoliert und geübt werden. Vielmehr wird das Schreiben eines Textes zum Anlass und Zentrum des ganzen Bestrebens überhaupt" (ebd.: 385). Ziel des fremdsprachlichen Unterrichts ist die kommunikativ angemessene Formulierung sprachlich eigenständiger Texte.

Semikreatives Schreiben, wie es im DESI-Projekt eingesetzt wird, kann demnach auf jeder Lernstufe geübt werden; es zeigt sich, inwieweit die Lernenden die Fremdsprache frei zu kommunikativen Zwecken einsetzen können und in wieweit sie ihre Aufmerksamkeit auf Schreibprozesse, das Schreibprodukt selbst und die Adressaten lenken können.

4.2.3 Curriculare Analysen

In die Curriculaanalysen sind die Lehrpläne aller Bundesländer zum Englischunterricht der 8. und 9. Klassen eingegangen, die bis Ende 2002 vorlagen. Einige Bundesländer differenzieren nach Schulformen, etwa Bayern, Baden-Württemberg, Hamburg, Niedersachsen, Nordrhein-

Westfalen, Saarland, Sachsen-Anhalt und Thüringen, die anderen Bundesländer unterscheiden die Schulformen nicht voneinander. Ziel der Analyse ist nicht ein Vergleich der heterogenen Lehrpläne, sondern das Auffinden von Gemeinsamkeiten, um die semikreative Aufgabenstellung auch in den Curricula zu verankern. Die Analyse zur Aufgabenstellung in den 8. und 9. Klassenstufen ergibt folgende Befunde:

Mehrheitlich sind in den Englischcurricula freie, kommunikative Formen des Schreibens in allen Schulformen genannt, wobei die Lernenden von der 8. zur 9. Klassenstufe vom *guided writing* hin zu freieren Schreibformen geführt werden sollen. In einigen Bundesländern, beispielsweise in Hessen, Hamburg, Bremen und Niedersachsen, ist das (semi-)kreative Schreiben explizit als Aufgabenstellung und Übungsform erwähnt. Generell ist kreatives Schreiben eher an Realschulen und Gymnasien gefordert, doch wird freies Schreiben auch in einigen Hauptschul-Lehrplänen explizit genannt, so beispielsweise in Nordrhein-Westfalen, Niedersachsen und Bayern.

Im überwiegenden Teil der Lehrpläne werden die Textsorten persönlicher Brief, Bericht, Beschreibung von Bildern oder Personen und Erzählung (auch unter Einbezug der Phantasie) explizit genannt. Im Vordergrund stehen die kommunikative Absicht beim Verfassen von Texten und der freie Umgang mit der Fremdsprache; demgegenüber tritt das Verfassen formaler Texte deutlich in den Hintergrund.

In den meisten Lehrplänen tritt bei der Bewertung kommunikativer Formen des Schreibens die formale Korrektheit hinter die Verständlichkeit, hinter die Umsetzung der kommunikativen Absicht: So legen etwa Thüringen oder Hamburg bei der Bewertung offener Schreibaufgaben inhaltliche und kommunikative Kriterien neben formalen an, um die Subjektivität in der Bewertung zu minimieren. Sachsen-Anhalt oder Schleswig-Holstein beispielsweise fordern explizit, dass bei solchen Schreibanlässen die gelungene, verständliche Kommunikation vor die sprachliche Korrektheit tritt. Im Rahmenplan von Mecklenburg-Vorpommern (2001: 35) etwa ist folgende Aussage zur positiven Leistungsbewertung zu finden:

> Bei jeder Art der Bewertung muss davon ausgegangen werden, dass die Lernenden hauptsächlich und zuerst erfahren, was sie schon wissen und können und erst danach Leistungsdefizite aufgedeckt werden. [...] Eine wesentliche Funktion von Lernerfolgskontrollen ist die ***Lernförderung***. [...] Der Begriff „**Lernerfolgskontrolle**“ steht in engem Zusammenhang mit ***Positivbewertung***. Hierbei ist grundsätzlich das Erreichte Maßstab und Ausgangspunkt der Leistungseinschätzung. Fehler werden als etwas Normales und für die Sprachausbildung Wesentliches und Hilfreiches betrachtet.

Bezüglich der Themenbereiche divergieren die Lehrpläne hinsichtlich ihrer Kategorisierungen und Präzision. Es lassen sich übergreifend drei Themenbereiche ausmachen, die sich zum einen auf Familie, Freunde, Schule, Ausbildung beziehen, zum anderen auf die beiden Bereiche Alltagsleben in den USA respektive in Großbritannien. Des Weiteren finden sich Übereinstimmungen hinsichtlich der Bereiche Freizeitaktivitäten, Reisen, Probleme der Jugend, Beruf und Arbeitsleben, soziale Probleme sowie Kultur, Sport und Medien.

Die curricularen Analysen ergaben, dass die semikreative Aufgabenstellung in allen Bundesländern über alle Schulformen hinweg einsetzbar ist. Schreiben in sinnvollen kommunikativen

Zusammenhängen steht im Zentrum; dabei soll die kommunikative Absicht möglichst verständlich und angemessen umgesetzt werden. Es bieten sich die Textsorten persönlicher Brief bzw. Schülerzeitungsbericht zu den Themenbereichen Reise, Abenteuer und Erlebnisse, Beschreibung einer Person und ihres Lebensumfelds sowie Umgang mit Problemen Jugendlicher an.

4.2.4 Kompetenzmodell – Leistungsdimensionen – Bewertungskriterien

Die Leistung eines Probanden ergibt sich aus den Anforderungen einer Aufgabe und den zugrunde liegenden Kompetenzen des Probanden. Unter welchen Aspekten die Leistung beurteilt werden kann, hängt demnach vom angesetzten Kompetenzmodell und den Anforderungen der Aufgabenstellung ab. Das im DESI-Schreibmodul angesetzte Kompetenzmodell basiert auf Bachmanns Modell der kommunikativen Kompetenz (Bachmann 1991a), wie es in Kapitel 1.2.3 dieser Arbeit erläutert ist. Die im Testkonstrukt analysierten Bedingungen der Textproduktion und die daran beteiligten Teilkompetenzen werden als Leistungsdimensionen angesetzt, aus denen die Bewertungskriterien wie folgt abgeleitet werden:

Bachmanns “linguistic competences“ führen zu den sprachlichen Kriterien (jeweils bezogen auf Umfang, Korrektheit und Angemessenheit der sprachlichen Mittel): Die Kriterien *Orthografie, Lexik/Lexiko-Grammatik, Grammatik* lassen sich aus Bachmanns “grammatical competences“ ableiten, das Kriterium *linking language* aus Bachmanns Kategorie der “cohesion“ als Unterkategorie der “textual competence“. Das Kriterium *Textsorte/Aufbau* (bezogen auf die Makrostruktur eines Textes) ist neben dem Bezug auf die jeweilige Aufgabenstellung verankert in Bachmanns “rhetorical organization“ wiederum als Unteraspekt der textuellen Kompetenzen; zudem ist es verankert in den strategischen Kompetenzen, da sie zum Aufbau eines Textes beitragen. Das Kriterium der *Textlänge* kann insofern aus Bachmanns strategischen Kompetenzen abgeleitet werden, als die Länge u. a. von der Planung und Zeiteinteilung abhängt. Die Kriterien *Inhalt* und *kommunikative Wirksamkeit* lassen sich aus Bachmanns “pragmatic competences“ ableiten, allerdings unter Einbezug der jeweiligen Aufgabenstellung; auf diese Kriterien wirken neben den pragmatischen Kompetenzen auch Weltwissensbestände und strategische Kompetenzen ein.

Es muss bedacht werden, dass diese Kriterien nicht absolut unabhängig voneinander konstruiert werden können, da Teile der Kompetenzen miteinander in Wechselbeziehungen stehen, beziehungsweise da sich einige Kriterien aus mehreren Teilkompetenzen ergeben.[254] Man denke beispielsweise an das Kriterium der *Orthographie*, bei dem der Umfang des Wortschatzes bewertet wird unter dem Aspekt der korrekten Schreibung, wohingegen das Kriterium *Lexik* den Umfang des Wortschatzes auf angemessene Verwendung hin bewertet. Ein weiteres Beispiel für ein Kriterium, das mit anderen Kriterien in Wechselwirkung steht, ist das Kriterium *kommunikative*

[254] Vgl. in diesem Zusammenhang auch die Ausführungen in Kapitel 1.2.1 dieser Arbeit zum Prototypenmodell der innersprachlichen Organisation.

Wirksamkeit, das sich zusammensetzt aus der angemessenen und wirksamen Versprachlichung (vgl. die sprachlichen Kriterien) relevanter Ideen (vgl. das Inhaltskriterium) und aus dem Adressatenbezug. Dennoch wird darauf geachtet, den jeweiligen Fokus der einzelnen Kriterien möglichst unabhängig zu definieren (vgl. dazu die Ausführungen in Kapitel 4.4 dieser Arbeit). Um der Komplexität des Schreibens gerecht zu werden, wird vor der Beurteilung dieser analytischen Kriterien ein Globalurteil angesetzt, denn vermutlich ist das Ganze „etwas anderes" als die Summe seiner Teile.

Folgende Übersicht verdeutlicht die Zusammenhänge zwischen Kompetenzen, Leistungsdimensionen und Bewertungskriterien:

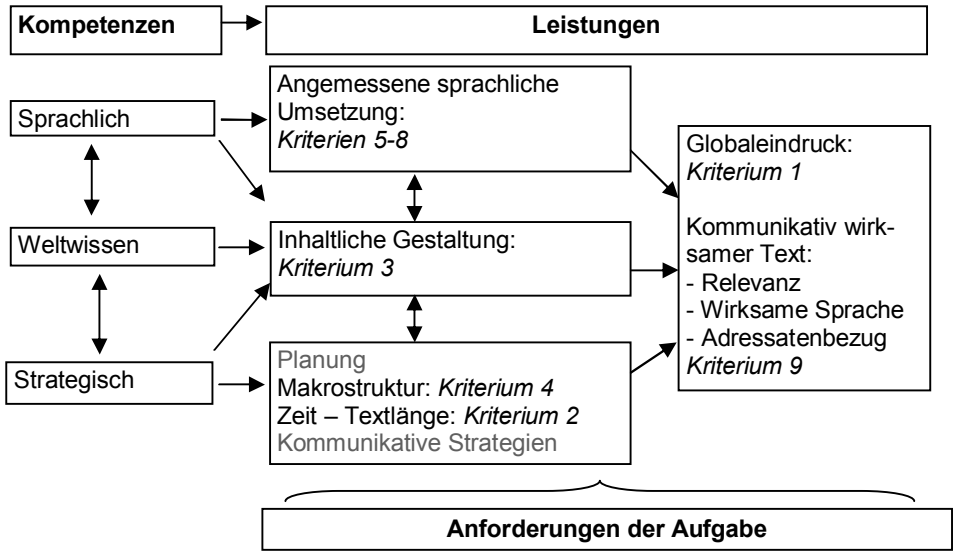


Abb. 22: Leistungsdimensionen

Die Anforderungen der Aufgabenstellung wirken dabei als exogene Variable auf die Leistungen, nicht aber auf die zugrunde liegenden Kompetenzen; sie werden im Anschluss in Kapitel 4.3.1 *Aufgabenbeschreibung* näher definiert, um die Aufgaben in ihren Schwierigkeiten und Anforderungen fassen zu können. Die prozessualen Aspekte des Schreibens können wie erwähnt in der DESI-Studie nicht erfasst werden, so dass folgende Bewertungskriterien aufgrund der analysierten Leistungsdimensionen abgeleitet wurden:

BEWERTUNGSKRITERIEN	
1. Globalurteil 2. Länge 3. Inhalt 4. Textsorte/ Aufbau (Makrostruktur)	5. Orthographie 6. Lexik u. Lexiko-Grammatik 7. Grammatik 8. Sprachliche Organisation 9. Kommunikative Wirkung

Abb. 23: Bewertungskriterien

An dieser Stelle wird lediglich der Zusammenhang zwischen Testkonstrukt und den daraus abgeleiteten Bewertungskriterien verdeutlicht. Zur inhaltlichen Definition der Kriterien darf auf das Handbuch in Anhang 27 verwiesen werden. Auf die vertikalen Abstufungen der Bewertungskriterien und die Verfahren der Bewertung wird unter Kapitel 4.4 *Bewertungsschema* näher eingegangen.

4.2.5 Die Bedeutung des GER bei der Entwicklung des Testkonstrukts

Wie kann nun der GER bei der Bestimmung des Testkonstrukts helfen? Welches Verständnis der Schreibfertigkeit lässt sich im GER ausmachen? In welchen Theorien ist dieses verankert? Welche Aussagen zur Schreibgenese und zu Schreibprozessen sind im GER zu finden? Welche Kategorien werden als relevant für die Bewertung von Schreibaktivitäten betrachtet?

Aussagen zur Aktivität des Schreibens werden im GER u. a. bei den *kommunikativen Aktivitäten* in GER-Abschnitt 4 behandelt. Es wird unterschieden zwischen produktiven und interaktiven schriftlichen Aktivitäten (GER 2001: 66f respektive 85f):

> Bei ***produktiven schriftlichen Aktivitäten*** (beim ***Schreiben***) produzieren die Sprachverwendenden als Autoren einen geschriebenen Text, der von einem oder mehreren Lesern rezipiert wird.
>
> Beispiele für schriftliche Aktivitäten sind:
> - Formulare und Fragebögen ausfüllen;
> - Artikel für Zeitungen, Zeitschriften, Rundschreiben usw. schreiben;
> - Plakate herstellen;
> - Berichte, Mitteilungen usw. schreiben;
> - Notizen zur späteren Verwendung anfertigen;
> - Mitteilungen nach Diktat schreiben;
> - Kreatives Schreiben;
> - persönliche Briefe oder Geschäftsbriefe usw. schreiben.
>
> Es stehen folgende Beispielskalen zur Verfügung:
> - Schriftliche Produktion allgemein;
> - Kreatives Schreiben;
> - Berichte und Aufsätze schreiben.
>
> Zur ***Interaktion im Medium der geschriebenen Sprache*** (Herv. d. V.) gehören sprachliche Aktivitäten wie:
> - Notizen, Memos usw. weitergeben und austauschen, wenn mündliche Interaktion nicht möglich oder nicht passend ist;
> - Korrespondenz durch Briefe, Fax, E-Mail usw.;
> - den Wortlaut schriftlicher Vereinbarungen, Verträge, Kommuniqués usw. durch Austausch von Entwürfen, Änderungen, Korrekturversionen usw. aushandeln;
> - an *online*- oder *offline*-Computerkonferenzen teilnehmen.
> (…)
> Beispielskalen stehen zur Verfügung für:
> - Schriftliche Interaktion allgemein;
> - Korrespondenz;
> - Notizen, Mitteilungen und Formulare.

Produktives Schreiben wird demnach definiert als schriftliche Textproduktion, die für einen Rezipienten bestimmt ist, wohingegen interaktives Schreiben nicht näher definiert wird. Die im GER gegebenen Beispiele spiegeln einen kleinen Ausschnitt der Schreibfertigkeit wider, sind jedoch keine hinreichende Beschreibung derselben. Die Beispielskalen engen den besagten Ausschnitt noch weiter ein.[255]

[255] Zu den Inhalten der Beispielskalen zum produktiven Schreiben sei auf die Skalenanalysen in Kapitel 3.4.3.2 dieser Arbeit verwiesen. Aus Gründen der Lesbarkeit wird das Verständnis des GER bezüglich der Schreibfertigkeit an dieser Stelle noch einmal erörtert, die Leserinnen und Leser mögen die dadurch entstehenden Redundanzen verzeihen.

Um sich dem Textbegriff zu nähern, der der Definition schriftlicher Produktion im GER zugrunde liegt, muss man GER-Abschnitt 4.6 *Texte* konsultieren: Das Textverständnis im GER ist produktorientiert und kommunikationskonstituierend: Texte umfassen alle „sprachlichen Produkte..., die Sprachverwendende/Lernende empfangen, produzieren oder austauschen – sei es eine gesprochene Äußerung oder Geschriebenes. Es kann demnach keine Kommunikation durch Sprache ohne einen Text geben“ (ebd.: 95). Dieses Verständnis entspricht *en gros* dem Textverständnis in DESI, doch als Entwicklungsbasis genügt es nicht, da wesentliche Charakteristika wie etwa der sozio-kulturelle Kontext, in dem ein Text situiert ist, oder der Zweck eines Textes, namentlich die Übermittlung einer bestimmten Sprechabsicht, nicht thematisiert werden. Zudem werden die Quellen des GER-Textbegriffs nicht offen gelegt. Die im GER folgenden Ausführungen zu Texten und Medien sind eher trivial und tragen nicht zur Definition der Schreibfertigkeit bei. In GER-Abschnitt 4.6.3 dann werden Textsorten aufgezählt, ohne dass ein Klassifizierungssystem erkennbar wäre und ohne dass benannt würde, auf welche Einteilungsschemata zurückgegriffen wurde. Es werden etwa *Zeitschriften* und *Zeitungen* neben *Verpackung von Waren* und *Datenbanken* als Textsorten bezeichnet (ebd.: 97). Die Liste erweckt den Eindruck einer arbiträren Aufzählung verschiedener medialer Auftrittsformen von Texten und illustriert bestenfalls die Vielfalt geschriebener Texte. Zumindest können die in DESI gewählten Formen des persönlichen Briefs und des Berichts dort wieder gefunden werden, ohne dass dies zu ihrer validen Verankerung beitragen könnte, da die Basis dieser Aufzählung nicht deutlich gemacht wird. Solch unvollständige und unsystematische Aufzählungen helfen nicht bei der Entscheidung, welche Textsorten bei einem Schreibtest erfasst werden können oder sollen.

Im GER lassen sich keine Aussagen zur Entwicklung der Schreibfertigkeit finden. Es wird zwar behauptet, dass die GER-Skalen die Fertigkeiten der Lernenden „auf verschiedenen, aufeinander folgenden Niveaus“ (ebd.: 131) beschreiben würden. Doch wie in Kapitel 3.4 dieser Arbeit gezeigt, gibt es keine hinreichende Begründung oder Validierung dieser Behauptung. Deshalb können Annahmen bezüglich der Entwicklung der Schreibfertigkeit weder mit den GER-Skalen noch mit theoretischen Ausführungen im GER belegt werden.

Aussagen zu Produktions- und Interaktionsstrategien, die bei den genannten Aktivitäten zum Einsatz kommen, finden sich in den GER-Abschnitten 4.4.1.3 respektive 4.4.3.5 des GER, wobei sich letzterer nur auf die mündliche Interaktion bezieht. Dieser Bruch zwischen den Aussagen zu schriftlicher Interaktion und dem Fehlen jeglicher Aussagen zu den dabei zum Einsatz kommenden Strategien ist nicht nachvollziehbar. Die Beschreibung der Produktionsstrategien hingegen ist detailliert, wenn auch nicht mit Quellenangaben belegt. Die Produktionsstrategien werden hinreichend charakterisiert und umfassen die im GER „Strategien“ genannten Prozesse der Planung, Ausführung, Kontrolle und Reparatur, ähnlich den Prozessen im oben diskutierten Modell von Börner.

Auf die am Schreiben beteiligten Prozesse wird im GER unter Abschnitt 4.5 *Kommunikative Sprachprozesse* eingegangen: Zunächst werden die am Schreiben beteiligten Fertigkeiten

benannt, namentlich kognitive, sprachliche und manuelle Fertigkeiten, um die Mitteilung organisieren und formulieren zu können und sie in schriftlicher Form festhalten zu können (GER: 93). Im Anschluss daran finden sich überraschenderweise die in GER-Abschnitt 4.4 erwähnten *Strategien* wieder, wobei sie nun als *Prozesse* beschrieben werden, ohne dass die beiden Ausführungen zueinander in Bezug gesetzt würden. Es drängt sich der Eindruck auf, verschiedene Autoren des GER hätten an verschiedenen Stellen dieselben Aspekte einmal als Strategien und einmal als Prozesse beschrieben. Zumindest finden sich in GER-Abschnitt 4.5 nähere Hinweise auf am Schreiben beteiligte Teilfertigkeiten bezogen auf die verschiedenen Prozesse: Bei der Planung sind „generelle und kommunikative Sprachkompetenzen" (ebd.: 95) beteiligt, um das kommunikative Ziel abzustecken bezüglich der zur Verfügung stehenden Ressourcen; bei der Ausführung werden die Ergebnisse der Planungsphase in sprachliche Form umgesetzt, wobei „lexikalische, grammatische und … orthographische Prozesse, die voneinander unterschieden werden können" (ebd.: 95), zum Einsatz kommen. Allerdings ist terminologisch nicht nachvollziehbar, wieso diese sprachlichen Wissensbestände oder Kompetenzen als „Prozesse" bezeichnet werden. Die darauf im GER folgenden Feststellungen zur „motorischen Anregung der Handmuskulatur" als Teil der Artikulation sind trivial und tragen nicht zur Charakterisierung des Produktionsprozesses bei; zudem sind sie auf einer viel grundsätzlicheren Ebene angesiedelt – sie sind Teil der biologischen Grundausstattung und nicht etwa ein Charakteristikum des Schreibprozesses selbst. Abschnitt 4.5 des GER schließt mit den Kontrollstrategien – dort werden sie wieder *Strategien*, und nicht *Prozesse* genannt –, die bei einer prozessorientierten Bewertung der Schreibfertigkeit durchaus als Bewertungskriterien angesetzt werden könnten, etwa als *kommunikative Strategien*, bei denen Kompensationsstrategien wie etwa Umschreibungen, Risikobereitschaft u. Ä. erfasst werden könnten.[256] Die Aussagen des GER-Abschnitts 4.5 sind die einzigen, die sich im GER zu den am Schreiben beteiligten Fertigkeitsteildimensionen finden lassen – als Basis der Ableitung von Bewertungskriterien sind sie jedoch nicht ausreichend, da im GER selbst nicht offen gelegt wird, in welchem Kompetenzmodell diese Dimensionen verankert sind (vgl. hierzu die obigen Ausführungen in den Kapiteln 2.5.2 und 3.4.1 der vorliegenden Arbeit) und keine Bezüge auf Schreibentwicklungs- oder Schreibprozessmodelle genommen werden.

Bezüglich der am Schreiben beteiligten Leistungsteildimensionen bietet das Referenzsystem des GER eine Vielzahl an möglichen Kategorien (die der horizontale Einteilung des GER-Skalensystems entsprechen), doch welche Kategorie inwieweit relevant für eine anstehende Bewertung ist, muss selbstverständlich aus dem gegebenen Kontext begründet werden. Ein Referenzrahmen kann aufgrund seiner Natur zwar Stellung dazu nehmen, welche seiner Kategorien bei welcher Art von kommunikativer Aktivität eine Rolle spielen, wie es in Abschnitt 4.5

[256] Das Kriterium der *kommunikativen Strategien* ist bis zur Pilotierungsphase in DESI auch angesetzt gewesen, doch hat es sich als nicht zuverlässig bewertbar erwiesen, da die Identifizierung von Merkmalen wie Kompensation oder Risikobereitschaft in den Textprodukten selbst i .d. R. auf Spekulationen beruht. Um valide beurteilen zu können, welche kommunikativen Strategien von den Lernenden verwendet werden, bräuchte es beispielsweise Methoden der Introspektion oder Techniken des lauten Denkens.

des GER auch geschieht, doch eine endgültige Auswahl der jeweils relevanten Bewertungskriterien muss der GER notwendigerweise seinen Nutzern überlassen, will er keine „Zwangsjacke" darstellen. Die „Freiheit" der Nutzer, andere Kompetenzmodelle als das des GER (das zudem im GER nicht offen gelegt wird) zu nutzen, wird beispielsweise in der Einleitung zum Kategoriensystem der *linguistischen Kompetenzen* in GER-Abschnitt 5.2 anerkannt, dort allerdings bezogen auf linguistische Beschreibungsmodelle:

> Ziel der folgenden Systematisierung ist es, einige Parameter und Kategorien als Klassifikationsinstrumente vorzuschlagen, die zur Beschreibung sprachlicher Inhalte, oder auch als Basis der Reflexion, hilfreich sein können. Praktikern, die ein anderes Bezugsmodell bevorzugen, steht es selbstverständlich – hier wie überall – frei, dies zu tun. Sie sollten jedoch ihre Theorien und Verfahren sowie die Traditionen, in denen sie arbeiten, offen legen. (GER 2001: 110)

Diese Freiheit gilt natürlich auch für die Wahl eines Bezugsmodells, aus dem Bewertungskriterien abgeleitet werden – allerdings muss von den Autoren des GER eingefordert werden, ihre Theorien und Traditionen, in denen sie arbeiten, ebenfalls offen zu legen.

Die Aussagen des GER zu Texten, Schreiben, Schreibprozessen und den daran beteiligten Wissensbeständen und Teilkompetenzen sind keinesfalls ausreichend, um darin ein Testkonstrukt zu verankern: Im GER fehlen, wie bereits in den vorangegangenen Kapiteln dieser Arbeit wiederholt festgestellt werden musste, jegliche Hinweise auf Theorien oder Modelle, die dem Schreib- und Textbegriff zugrunde liegen. Die Ausführungen erwecken allzu oft den Anschein von unsystematischen Aufzählungen und unbelegten Behauptungen. Um eine umfassende Sicht auf das Verständnis des GER hinsichtlich der am Schreiben beteiligten Prozesse und Wissensbestände zu erhalten, müssen die Nutzer verschiedene Stellen des GER konsultieren – kein Beitrag zur Benutzerfreundlichkeit. Letztlich können die Ausführungen im GER zu einem abschließenden Vergleich genutzt werden, wenn das betreffende Testkonstrukt definiert und fundiert in Fachliteratur und wissenschaftlichen Theorien verankert wurde – eventuell können durch den Vergleich relevanter Kategorien im GER mit entsprechenden Aspekten des Testkonstrukts Lücken im Konstrukt identifiziert werden, die dann mittels geeigneter Theorien oder Modelle geschlossen werden können. Doch welche Bedeutung soll man Fehlendem im GER zuschreiben? Ist beispielsweise ein Gedicht nicht als Textsorte geeignet, nur weil es bei den Textsorten nicht aufgelistet wird? Wieso finden sich keine Aussagen etwa zu den Strategien der schriftlichen Interaktion? Wie sollen sich die Nutzer im Referenzrahmen wiederfinden, wenn nicht Bezug genommen wird auf Fachliteratur und dort diskutierte Theorien und Modelle, die den Kategorien des GER zugrunde liegen? Aufgrund der hier diskutierten Defizite und offenen Fragen kann der GER keinesfalls als Ausgangspunkt der Testentwicklung, als Basis eines zu definierenden Testkonstrukts dienen.

4.3 Aufgabenbeschreibung und Instrumentenentwicklung

Im Folgenden werden auf Basis der in der Fachliteratur[257] diskutierten Merkmale „guter" Schreibtasks (wobei *Task* in der vorliegenden Arbeit als handlungsorientierte Aufgabenstellung verstanden wird) die Aufgaben charakterisiert, die im DESI-Projekt zum Einsatz kommen. Damit soll das Universum der Tasks definiert werden, auf das hin die Testergebnisse verallgemeinert werden können. Im Anschluss wird ein Beispieltask vorgestellt. Die nachstehende Tabelle gibt eine Übersicht über Merkmale guter Schreibtasks, welche bei der Operationalisierung helfen sollen, zu validen Schreibaufgaben zu kommen:

Tasks	- Sie sollen bedeutsam sein in Bezug auf Situationen aus der realen Welt der Probanden. - Sie sollen Performanzen elizitieren, die auf Wissen und Fertigkeiten basieren, welche integrativ im Kontext eines zweckbestimmten Problemlösens angesiedelt sind. - Sie sollen Lernressourcen darstellen und Lernanreize bieten. - Sie müssen konsistent sein mit den Lernzielen der Institutionen, in denen die Beurteilung stattfindet. - Sie dürfen eine Herausforderung darstellen, doch dann sollte genügend Hilfestellung gegeben werden, um allen eine Chance auf erfolgreiche Bearbeitung zu geben: Alle Probanden sollen den Task entsprechend ihrer Fähigkeiten lösen können.
Anweisung	- Die Arbeitsanweisung muss eindeutig und für alle verständlich sein: klare, angemessene Sprache, kurz, eindeutig, und für die Zielgruppe zugänglich. - Situation, Schreibanlass und Adressaten sollen klar umrissen werden. - Limitierung hinsichtlich Inhalten, Textsorten und Länge soll gegeben werden. - Es soll hinreichend Information gegeben werden, Redundanzen jedoch sollen vermieden werden.
Themen	- Sie sollen von potentiellem Interesse für Probanden (und Bewerter) sein. - Sie sollen konsistent sein mit den Vorgaben der getesteten Institution (z. B. curriculare Vorgaben) und nach didaktischen Gesichtspunkten ausgewählt werden. - Themen, die zu potentiellen Kontroversen führen könnten oder schwierig zu bewerten sind, sollen vermieden werden.
Bewertung	- Die Bewertungskriterien sollen öffentlich gemacht werden und allen Beteiligten bekannt sein. - Die Bewerter müssen geschult sein, informiert an die Bewertung treten und das Universum der Schreibenden und der Tasks kennen. - Alle Probanden sollen entsprechend ihrer Fähigkeiten bewertet werden können.
Ziel	- Die Beurteilung soll – abhängig vom jeweiligen Testkontext – zeigen, was die Probanden schon können und wo es noch Lernbedarf gibt. - Die Beurteilung soll die verschiedenen Dimensionen und Facetten der Schreibfertigkeit abbilden.

Tabelle 6: Merkmale guter Schreibtasks

[257] Vgl. beispielsweise Camp 1996, Cohen 1994, Lehmann 1990, Hamp-Lyons & Kroll 1996, Hughes 1986 u. a..

4.3.1 Aufgabenbeschreibung

Aufgabenmerkmale bilden den Rahmen der Testentwicklung in DESI. Genau definierte Aufgabenmerkmale helfen, das Universum der Tasks zu charakterisieren, um den Horizont zu bestimmen, auf den hin die Beurteilung generalisiert werden kann. Dabei können sie drei Funktionen übernehmen: Sie können zum ersten schwierigkeitsbestimmend sein und insofern variiert werden, um die Schwierigkeit angemessen auf die jeweilige Zielgruppe hin auszulegen; zum zweiten können sie dimensionsbestimmend sein, das heißt die Leistungsdimensionen bestimmen, die damit erfasst werden sollen, und zum dritten können die Merkmale kontingente Funktion haben, folglich der Kontrolle bestimmter Parameter dienen. Die jeweilige(n) Funktion(en) der einzelnen Merkmale werden bei den sich unten anschließenden Ausführungen deutlich gemacht, ebenso wie die unterschiedlichen Ausprägungen, die einige Merkmale annehmen können.

Schreiben wird i. A. charakterisiert durch die Sprachhandlung, die dabei ausgeführt wird. Diese kann man grundsätzlich nach Art der sprachlichen Aktivität einteilen: Der GER (2001: 25) unterscheidet beispielsweise Produktion, Rezeption, Interaktion und Mittlung. Man kann die Sprachhandlungen aber auch nach ihrer kommunikativen Funktion einteilen: Börner (1989: 350) etwa verweist auf expressive, epistemische und informationsverarbeitende Rollen des Schreibens, Hughes (1986: 76) zählt eine Reihe von Funktionen auf wie das Äußern von Dank, Bitten oder Meinungen oder das Elizitieren von Informationen. Ein weiteres Charakteristikum des Schreibens ist in der Botschaft, der Textaussage zu finden: Was wird zu welchen Themen in welchen Kontexten in welcher Komplexität versprachlicht? In der Literatur (vgl. etwa Börner 1989 oder de Beaugrande 1985) finden sich Hinweise, dass konkrete Themen bezogen auf alltägliche Situationen einfacher darzustellen sind als abstrakte Abhandlungen zu komplexen Themen. In diesem Kontext spielt auch die Textsorte, die die Botschaft vermitteln soll, eine Rolle, da jede Textsorte eigene Anforderungen stellt. Börner beispielsweise verweist auf den „inhärenten Schwierigkeitsgrad“ (Börner 1989: 358) verschiedener Textsorten und gibt eine Auflistung von Textsorten, die sich für die didaktische Textproduktion eignen. In einer Testsituation ist zudem besonderes Augenmerk zu richten auf die Aufgabenstellung und den Stimulus (vgl. beispielsweise Hughes 1986 oder Kroll 1998): Was wird in welcher Form gefordert und wie beeinflusst möglicherweise die Aufgabenstellung die Performanz? Welche Hilfestellungen, seien es nun Planungshilfen, Strukturierungsmuster, Formulierungsmuster oder die Nutzung von Wörterbüchern, werden gegeben? Wie beeinflussen Aufgabenstellung und Hilfen die Schwierigkeit?

Folgende Merkmale[258] sind als relevant in die Beschreibung der semikreativen Aufgabenstellung im DESI-Projekt eingegangen:

[258] Zur Basis der Merkmale vgl. u. a. Börner 1989, GER (2001: 155ff), Hughes 1986, Kroll 1998.
Die Verwendung des GER könnte gleich hier bei der Entwicklung der Aufgabenmerkmale dokumentiert werden. Doch ist es stringenter, die Beurteilung der Verwendbarkeit des GER wiederum in einem eigenen Unterkapitel 4.3.3 darzustellen, schon um den Lesern das Auffinden der GER-bezogenen Beurteilungen zu erleichtern. Zudem spiegelt diese Struktur das Vorgehen im Projekt wider: Zunächst wurden relevante Merkmale analysiert und definiert. Erst dann wurde der GER hinzugezogen, um zu sehen, ob er zusätzliche Informationen bietet.

- *Schreibmodus*: Kommunikatives Schreiben wird im DESI-Projekt unterscheiden in produktives und interaktives Schreiben: Beim interaktiven Schreiben müssen sich die Probanden neben dem Inhalt stärker als beim produktiven Schreiben auf einen (bei den entsprechenden Aufgaben genau charakterisierten) Rezipienten beziehen. Die Schwierigkeit einer interaktiven Aufgabe, bei der der Fokus auf Inhalt und Rezipient gleichermaßen liegt, unterscheidet sich von der Schwierigkeit einer produktiven Schreibaufgabe, bei der der Fokus stärker auf der spannenden oder interessanten Darstellung des Inhalts liegt. Allerdings handelt es sich bei den unterschiedlichen Schwierigkeiten nicht um Abstufungen der Schwierigkeit, sondern um unterschiedliche Anforderungsausprägungen, so dass dieses Merkmal im Hinblick auf die Aufgabenschwierigkeit nicht determinierend ist. Produktives und interaktives Schreiben unterscheiden sich aber nicht nur hinsichtlich der Anforderungen, sondern es handelt sich dabei um zwei unterschiedliche Facetten der Schreibfertigkeit: Es werden generalisierbare Aussagen über die Schreibfertigkeiten bezüglich dieser beiden Genres erwartet. Deshalb wird der Schreibmodus über die Vorgabe der zu erstellenden Textsorte als ein kontingentes Merkmal kontrolliert – jede Schülerin/jeder Schüler soll sowohl eine produktive als auch eine interaktive Aufgabe bearbeiten, um die Schreibfertigkeit möglichst breit zu erfassen. Das Merkmal „Schreibmodus“ ist demnach ein dimensionsbestimmendes und ein kontingentes. Es kann nur in einer Ausprägung vorliegen, d. h. ein gegebener Task ist entweder als interaktiv oder als produktiv zu klassifizieren.

- *Kommunikative Sprachhandlung*: Je nach Art der kommunikativen Handlung, die in einer Schreibaufgabe gefordert wird, werden unterschiedliche Teilkomponenten sprachlicher Handlungsfähigkeit getestet: Beispielsweise wird die narrative Funktion der Sprache mittels einer Erlebniserzählung überprüft. Bezogen auf die semikreativen Aufgaben des DESI-Projekts sollen folgende, durch Curriculaanalysen und Experteneinschätzung gestützte Sprachhandlungen unterscheiden und erfasst werden: Narrative Handlung, Rat geben, Selbst- bzw. Fremdbild darstellen, Problem lösen, Strategien beschreiben, Gefühle beschreiben, Fremdperspektive beschreiben. Die Funktion der kommunikativen Sprachhandlung bestimmt die Aufgaben einerseits in den Fertigkeitsteildimensionen und wird deshalb bei der Auswertung berücksichtigt: „Welche kommunikativen Handlungen kann die Schülerin/der Schüler sprachlich ausführen?“ Andererseits sollen möglichst viele Sprachhandlungen überprüft werden, um die Schreibfertigkeit unter Berücksichtigung möglichst vieler Facetten zu erfassen. Das Merkmal „kommunikative Sprachhandlung“ ist also dimensionsbestimmend und kontingent. Die Ausprägungen dieses Merkmals *narrative Handlung, Rat geben, Selbst- bzw. Fremdbild darstellen, Problem lösen, Strategien beschreiben, Gefühle beschreiben, Fremdperspektive beschreiben* können auch gleichzeitig auf eine Aufgabe zutreffen.

- *Versprachlichung*: Mit diesem Merkmal soll der Hintergrund erfasst werden, den die Schülerinnen und Schüler in ihren Texten versprachlichen: Handelt es sich um Erfahrungen, Fakten, fiktive Welten oder Abstraktionen, die versprachlicht werden sollen? Es ist anzunehmen, dass die Darstellung unmittelbarer Erfahrungen und Fakten leichter fällt und dass hierfür andere

Teilfertigkeiten aktiviert werden müssen als für die Versprachlichung fiktiver oder gar abstrakter Gedanken, denn dabei geht der Versprachlichung das Erfinden der Fiktion bzw. die Abstraktion voraus. Über das Merkmal „Versprachlichung“ können Aufgabenschwierigkeit und Fertigkeitsteildimensionen erfassen werden. Auch für die Gewährleistung der Testbreite wird dieses Merkmal herangezogen. Das bedeutet, dass zu jeder Merkmalsausprägung (Erfahrungen, Faktisches, Fiktives oder Abstraktionen) auch eine entsprechende Aufgabe entwickelt worden ist. Das Merkmal „Versprachlichung“ ist demnach schwierigkeitsbestimmend, dimensionsbestimmend und kontingent. Die Ausprägungen dieses Merkmals *Erfahrungen, Fakten, fiktive Welten* oder *Abstraktionen* können auch gleichzeitig auf eine Aufgabe zutreffen.

- *Thema*: Da auch die Themenauswahl aufgrund von Curriculumsanalysen und Experteneinschätzung getroffen wurde und in der Lebenswelt der Schüler angesiedelt ist, kann dieses Merkmal weniger Rückschlüsse auf Task-Schwierigkeit oder Fertigkeitsdimensionen zulassen; vielmehr handelt es sich um ein zu kontrollierendes kontingentes Merkmal. Folgende Themen sollen die Breite der Aufgabenstellung reflektieren: *Erlebnisse, Reisen und Abenteuer*, *(Lebens)-Strategien* (hierunter fallen beispielsweise Problemlösestrategien), *Sorgen und Probleme Jugendlicher*. Die Ausprägungen dieses Merkmals können auch gleichzeitig auf eine Aufgabe zutreffen.

- *Textsorte*: Die Curriculaanalysen haben ergeben, dass die Genres *persönlicher Brief* und *Schülerzeitungsbericht* schulformübergreifend eingesetzt werden können. Sie sind in allen Schulformen hinreichend bekannt und bieten die Offenheit, die eine semikreative Aufgabe verlangt. Über das dimensionsbestimmende Merkmal „Textsorte“ lassen sich genrespezifische Schreibfertigkeiten feststellen, die sich in den Bewertungskriterien wiederfinden. Die Textsorte wird aber auch zur Kontrolle der Testbreite als kontingentes Merkmal herangezogen, um zu generalisierbaren Aussagen zu kommen. Dieses Merkmal kann nur in einer Ausprägung vorliegen.

- *Stimulus*: Das Merkmal „Stimulus“ fällt üblicherweise unter die Kategorie „Hilfen“, zu der auch Aufgabenstellung und Zeit zählen. Die Zeit wird kontrolliert: Es stehen jeder Schülerin/jedem Schüler für jede semikreative Aufgabe 20 Minuten zur Verfügung. Alle Aufgaben sind gleichermaßen standardisiert und kontextualisiert: Es werden ausreichend Informationen auf Deutsch über Setting, Inhalte, Kommunikationspartner, Art der sprachlichen Handlung, Textsorte, Länge und Zeit gegeben. Allerdings unterscheiden sich die der Aufgabenstellung vorangehenden Stimuli hinsichtlich der Verwendung von englischsprachigen Texten, Bildern oder einer Kombination von Bild und Text. Der Stimulus soll als kontingentes Merkmal kontrolliert werden, doch lassen sich keine Aussagen in Bezug auf Schwierigkeitsbestimmung und Dimensionsbestimmung machen. Aufgrund der Standardisierung liegt die Schwierigkeitsdifferenzierung nicht im Stimulus selbst, sondern im unten beschriebenen Task-Zugang. Das Merkmal „Stimulus“ kann nur in einer der Ausprägungen *Text*, *Bild* oder *Text und Bild* vorliegen.

- *Task-Zugang*: Unter *Task-Zugang* werden die Erschließungsprozesse verstanden, die nötig sind, um die betreffende semikreative Aufgabe bearbeiten zu können. Dabei wird unterschieden

zwischen einem unmittelbar-rezeptiven, einem interpretativen, einem abstrahierenden und einem kreativen Zugang, je nach Anforderungen an die Verarbeitungskapazitäten der Schülerinnen und Schüler. Eine Aufgabe, die unmittelbar erschlossen werden kann, kann nach Rezeption des Stimulus und der Aufgabenstellung ohne weitere Verarbeitungsprozesse direkt bearbeitet werden, im Gegensatz zum interpretativen Zugang, bei dem Stimulus und Aufgabenstellung erst von den Schülerinnen und Schülern gedeutet werden muss, bevor sie entscheiden können, wie sie die Aufgabe bearbeiten möchten. Der abstrahierende Zugang verlangt darüber hinaus abstrakte Verarbeitungsprozesse, wie beispielsweise Perspektivenwechsel oder das Verknüpfen von Stimulus, Deutung desselben und Weltwissen, ohne welches die betreffende Aufgabe nicht zu bearbeiten ist. Während diese drei Zugänge die Aufgaben in ihren Schwierigkeiten bestimmen, ist der kreative Zugang ein rein kontingentes Merkmal, das die Aufgaben in ihrer Offenheit charakterisiert. Neben der Schwierigkeitsbestimmung dient das Merkmal *Task-Zugang* auch der Kontrolle der Testbreite. Deshalb wird bei der Aufgabenkonstruktion darauf geachtet, die hier genannten Erschließungsmöglichkeiten auch abzudecken. Die Ausprägungen dieses Merkmals *unmittelbar, interpretativ, abstrahierend* oder *kreativ* können auch gleichzeitig auf eine Aufgabe zutreffen.

4.3.2 Aufgabenentwicklung und Validierung

Die erwähnten Merkmale „guter“ Schreibaufgaben und die obigen Aufgabencharakteristika standen am Beginn der Aufgabenentwicklung im DESI-Projekt: Zunächst wurden 14 Aufgabenstellungen entwickelt. Diese wurden im Frühjahr 2001 präpilotiert an je zwei Haupt und Realschulen und an zwei Gymnasien im Raum Augsburg, wobei jeweils zwei Klassen beteiligt waren. Nach der Testdurchführung wurden Lehrkräfte und Lernende zu ihren Eindrücken befragt. Dabei zeichnete sich ab, dass zu freie Aufgabenstellungen in der knappen Zeit nicht bearbeitet werden können, ebenso wie sich zeigte, dass die anfänglich englischen Arbeitsanweisungen auf Deutsch gegeben werden müssen, da es unter den schwächeres Lernenden Probleme gibt, die englische Aufgabenstellung zu verstehen; dies würde jedoch die Erfassung der Schreibfertigkeit mit der der Lesefertigkeit konfundieren. Die Präpilotierung diente der Überprüfung der Aufgabenstellungen, der Machbarkeit und Verständlichkeit derselben, und der Gewinnung von Lernertexten. Letztere wurden auf relevante Merkmale hin analysiert, um Leistungsdimensionen und deren mögliche Abstufungen zu gewinnen und zu überprüfen. Auf diese Aufsatzanalysen wird unter Kapitel 4.4 *Bewertungsschema* eingegangen. Die Aufsatzanalysen dienten aber auch der inhaltlichen Validierung und der Konstruktvalidierung der Tasks: Es konnten diejenigen Aufgaben identifiziert werden, die nicht die gewünschten Leistungen elizitierten, etwa weil sie zu exzessivem Stimuluskopieren anregten oder keine zusammenhängenden Texte hervorriefen, da beispielsweise die Hilfen als Fragen aufgefasst wurden, die einzeln beantwortet wurden. In Anhang 22 wird ein Beispieltask vorgestellt, wie er zur Präpilotierung eingesetzt wurde.

Aufgrund der Erkenntnisse aus der Präpilotierung und der Analysen der Lernertexte wurden Testanleitungen, Stimuli, Arbeitsanweisungen und Hilfen revidiert und diejenigen Tasks für die Pilotierung ausgewählt, die sich als verständlich erwiesen, bewertbare Leistungen elizitierten, und von Lehrkräften und Lernenden als motivierend und in der 9. Klasse bearbeitbar bezeichnet wurden. Anhang 23 zeigt den überarbeiteten Task, wie er in der Pilotierung eingesetzt wurde. Dazu wurde die Aufgabenstellung revidiert, auf Deutsch gestellt und zwei der Bilder ausgetauscht, da sich geschlechtsspezifische Bevorzugungen zeigten.

Die Pilotierungsphase diente im Modul semikreatives Schreiben einerseits der Überprüfung und Revidierung des Bewertungsschemas. Auf alle Fragen im Zusammenhang mit der Entwicklung und Validierung desselbigen wird im Anschluss in Kapitel 4.4 eingegangen. Andererseits diente die Pilotierung der Validierung der Aufgaben selbst und der endgültigen Aufgabenauswahl: Die Validierung der DESI-Testmodule erfolgte mittels des so genannten EU-Tests (*Assessment of Pupils' Attainment in English in European Countries*) und entsprechenden Instrumenten aus der LAU 9 Studie, die ebenfalls in der Pilotierung eingesetzt wurden.[259] Auf Basis der Pilotierungsergebnisse wurden die Tasks ausgewählt, die bewertbare Leistungen im Sinne des oben beschriebenen Testkonstrukts elizitierten. Die Auswahl wurde unter Beachtung der oben genannten kontingenten Merkmale getroffen, wobei sie sich zusätzlich auf die Kriterien der Machbarkeit in allen Schulformen, auf die Kriterien der Verständlichkeit und Bearbeitbarkeit der Aufgabenstellung in der gegebenen Zeit und auf das Kriterium der Bewertbarkeit der Lernertexte stützte; dazu floss auch das Feedback der Bewerter der Pilotierungstexte mit ein. Ergänzend wurde die Datenlage der bewerteten Pilotierungstexte zusammen mit den Statistikern im DESI-Projekt geprüft, um die Aufgabenauswahl auch unter diesem Aspekt abzusichern.

Für die Hauptuntersuchung wurde der eigentlichen Aufgabenstellung eine Anleitung vorangestellt und der oben vorgestellte Task wie folgt überarbeitet:

Anleitung zur semikreativen Schreibaufgabe

Liebe Schülerin, lieber Schüler,

in diesem Teil des Tests findest du eine Schreibaufgabe, bei der du neben deinen Englischkenntnissen auch deine Kreativität und Phantasie nutzen sollst. Du bearbeitest hier die erste von zwei Aufgaben, wobei du je nach Aufgabenstellung einen Brief beziehungsweise einen Bericht in deinem besten Englisch schreiben sollst.

Für diese Aufgabe stehen dir 20 Minuten zur Verfügung.

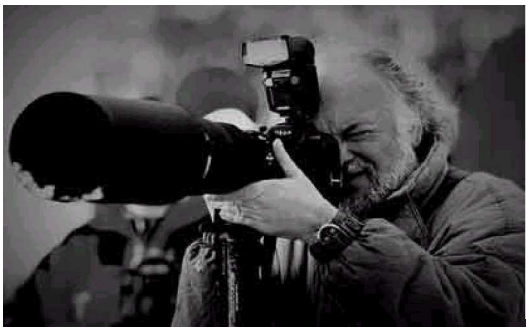
Wichtig ist, dass du dir die Aufgabenstellung genau durchliest und dich daran hältst.

Du darfst natürlich alle Namen, Ereignisse oder sonstigen Details erfinden, die du für deinen Brief beziehungsweise Bericht brauchst.

Schreibe deinen Brief bzw. Bericht in den dafür vorgesehenen Raum in deinem Testheft.

Abb. 24: Anweisung

[259] Zur Validierung vgl. Klieme, Eichler et al. 2003; zur LAU 9 Studie vgl. Lehmann et al. 2000.



Schreibe einen Artikel für die Schülerzeitung. Wähle dazu ein Bild aus und berichte über das Leben dieser Person. Du darfst natürlich alle notwendigen Einzelheiten zur Geschichte dieser Person erfinden.
Der Bericht über deine Person sollte ungefähr eine Seite lang werden.
Schreibe auf Englisch, so gut und interessant du kannst.

Hier ein paar Tipps:

Gib deiner Person einen Namen und erzähle, wer sie ist, wie sie sich fühlt und was sie erlebt hat.
Berichte, woher deine Person kommt, wo sie lebt und ob sie Familie hat.
Erwähne auch den Beruf, die Träume oder Wünsche deiner gewählten Person.

Lass dich von deinem gewählten Bild anregen und denke dir die Geschichte dieser Person aus.

Abb. 25: Task

Aufgrund der begrenzten zeitlichen wie finanziellen Kapazitäten und der Fülle der Testhefte war es unabdingbar, dieselben Schreibaufgaben an allen Schulformen einzusetzen, so dass vier Aufgabenstellungen in die Hauptuntersuchung eingingen: zwei Berichte für eine Schülerzeitung, davon einer zum Thema *Personenbeschreibung* (vgl. Abb. 25), der andere zum Thema *Erlebnisse auf Klassenfahrt*; zwei persönliche Briefe, davon einer zum Thema *Probleme Jugendlicher*, der andere zum Thema *Erlebnisse während eines Londonaufenthaltes*. Die Lernenden bearbeiteten je einen Brief und einen Bericht zu unterschiedlichen Themen. In einer Klasse kamen alle vier Aufgaben zum Einsatz.

Während der gesamten Entwicklungsphase wurden Tasks und Bewertungsinstrumente selbstverständlich fortlaufend evaluiert und revidiert. Dazu wurden die in Kapitel 2 dieser Arbeit vorgestellten Checklisten und Fragenkataloge genutzt, ebenso wie diejenigen aus dem UGE, wie im Folgenden noch ausführlicher dargestellt wird.

4.3.3 Die Bedeutung des GER bei der Aufgabenbeschreibung und Entwicklung

Welche Informationen finden sich im GER zu kommunikativen Schreibhandlungen, zu Schreibsituationen und Themen, zu Schreibaufgaben und relevanten Charakteristika derselben? Können der GER respektive der UGE bei der Aufgabenentwicklung helfen? In GER-Abschnitt 9.2 wird vorgeschlagen, Tests und Prüfungen inhaltlich (unter Rekurs auf die GER-Abschnitte 4.4 zu den *kommunikativen Aktivitäten*, 4.6 zu *Texten* und 7.3 zu *Schwierigkeiten kommunikativer Aufgaben*) zu beschreiben und sie in das Kategoriensystem des GER einzuordnen (GER 2001: 174).[260] Des Weiteren wird dort (ebd.) behauptet, man könne Testaufgaben auf Basis des GER-Abschnitts 5.2 (zu den *kommunikativen Sprachkompetenzen*) erstellen. Diese Aussagen zur Beschreibung, Einordnung und Testentwicklung haben sich als nicht haltbar erwiesen, wie im Folgenden bezogen auf das DESI-Projekt gezeigt wird.[261] Dazu werden relevante Aussagen des GER zu den Bereichen *kommunikative Schreibaktivitäten, Texte, kommunikative Schreibaufgaben, Aufgabencharakteristika* und *Aufgabenentwicklung* zusammengestellt und in ihrer Bedeutsamkeit bewertet.

Aussagen zu **kommunikativen Aktivitäten** lassen sich in GER-Abschnitt 4 finden: Unterabschnitt 4.1 beschäftigt sich mit allgemein gültigen Aussagen zu Domänen, Situationen, Bedingungen und Einschränkungen und den mentalen Kontexten der Kommunikationspartner. Die Aussagen dort sind jedoch so generell gehalten, dass sie zur konkreten Merkmalsentwicklung kommunikativer Schreibaufgaben wenig beitragen können. Die in GER-Abschnitt 4.2 folgende Auflistung verschiedener *Themen der Kommunikation* ist exemplarisch zu verstehen und muss in den konkreten Kontexten der Testentwicklung begründet werden (wie es etwa in DESI durch die erwähnten Curriculaanalysen geschehen ist). GER-Abschnitt 4.3 beschreibt *kommunikative Aufgaben und Ziele* näher, doch in Bezug auf kommunikative *Handlungen und Aktivitäten*, die man im realen Leben (im GER bezogen auf das Berufsleben und den privaten Bereich) ausführen können sollte, weniger bezogen auf (didaktische) *Aufgabenstellungen*.[262] Die sehr generell gehaltenen Aussagen zu den kommunikativen Handlungen helfen bei der Testaufgabenbeschreibung wenig, müssen doch die Handlungen, die ein Test erfassen soll, aus dem jeweiligen Kontext bezogen auf die jeweilige Testpopulation abgeleitet werden und etwa in Curriculums- oder Lehrwerkanalysen begründet werden. *Kommunikative Aktivitäten und Strategien* werden wie gesagt in GER-Abschnitt 4.4 beschrieben, bezogen auf Schreibaktivitäten insbesondere in den Unterabschnitten 4.4.1.2 *Schriftliche produktive Aktivitäten* und 4.4.3.4 *Schriftliche Interaktion*. Doch wie oben bereits festgestellt, sind die dort gegebenen Definitionen nicht hinreichend, um relevante Charakteristika

[260] Der Überblick auf S. 174 des GER über sein Kategoriensystem für kommunikative Aktivitäten ist allerdings unvollständig: Beispielsweise wird dort nichts zu Rezeption und Sprachmittlung ausgesagt, und in der Kategorie „schriftliche Interaktion“ bleibt eine Leerstelle. Warum dort nicht etwa auf Briefe eingegangen wird, ist nicht nachvollziehbar.
[261] In diesem Zusammenhang darf auch auf Kapitel 3.4.4 der vorliegenden Arbeit verwiesen werden, in welchem die Verwendbarkeit der GER-Skalen beurteilt wird.
[262] Informationen zu *Aufgabenstellungen* finden sich vielmehr in GER-Abschnitt 7, der in dieser Arbeit gleich im Anschluss besprochen wird.

guter Schreibaufgaben abzuleiten oder Tasks zu spezifizieren, wie es der GER (ebd.: 173)[263] vorschlägt. GER-Abschnitt 4.5 *Kommunikative Sprachprozesse* spielt bei der Aufgabenbeschreibung in DESI keine Rolle, da Prozesse in DESI nicht erfasst werden können.

Aussagen zu **Texten** und Textsorten[264] werden in GER-Abschnitt 4.6 getroffen. In Unterabschnitt 4.6.4 *Texte und Aktivitäten* findet sich wiederum die erwähnte grundsätzliche Einteilung der kommunikativen Aktivitäten in die vier Grundtypen Produktion, Rezeption, Interaktion und Vermittlung, diesmal bezogen auf (mündliche wie schriftliche) Texte. Es werden Schemata vorgestellt, die „das Beziehungsgefüge zwischen dem Sprachverwendenden/Lernenden … und dem oder den an der Kommunikation Beteiligten sowie den Aktivitäten und den Texten" darstellen (GER 2001: 100). Die beiden folgenden Schemata sind für das Schreiben in DESI relevant:

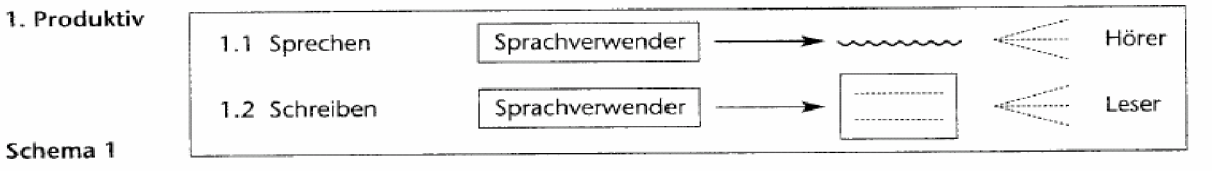


Abb. 26: Kommunikative Aktivitäten: Schema Produktion (GER 2001: 100)

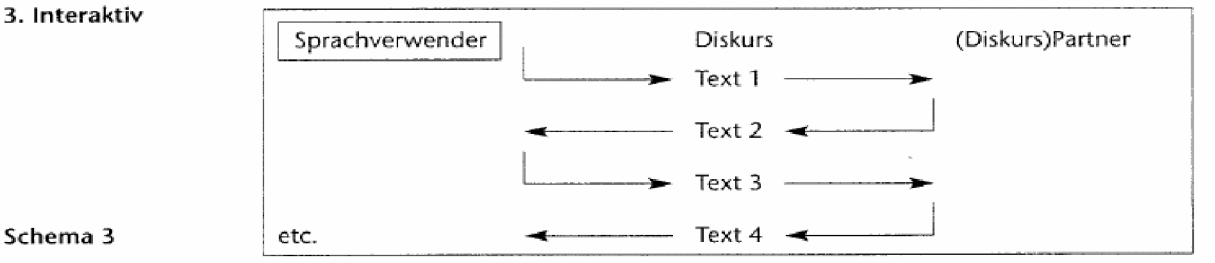


Abb. 27: Kommunikative Aktivitäten: Schema Interaktion (ebd.: 101)

Während bei Schema 1 ein Text im Mittelpunkt steht, der von den Sprachverendenden produziert wird und sich an einen Adressaten/Rezipienten richtet, von diesem jedoch nicht beantwortet werden muss (vergleichbar der produktiven Aufgabenstellung im DESI-Projekt), richtet Schema 3 das Augenmerk auf den Diskurs zwischen den Kommunikationsteilnehmern und damit auf mehrere Texte, die aufeinander bezogen sind. Wendet man dieses Schema auf die DESI-Aufgaben an, so gibt die interaktive Aufgabe entweder die Rahmenbedingungen zur Erstellung von Text 1 des obigen Schemas vor, oder die interaktive Aufgabe stellt einen Stimulustext (der Text 1 des Schemas darstellt) bereit, den die Probanden durch einen Text 2 beantworten sollen – letztere Aktivität wird im GER (ebd.: 101) allerdings jenseits der vier genannten Grundtypen beschrieben, im DESI-Projekt sieht man sie hingegen als einen Unteraspekt des interaktiven Schreibens.

[263] An dieser Stelle fällt eine weitere Übersetzung im GER negativ auf: In der deutschen Ausgabe heißt es, dass man Abschnitt 4.4 zu Rate ziehen kann, „wenn man eine *Testanleitung* für eine kommunikativ orientierte Beurteilung entwirft" (Herv. d. V.). Da im Zusammenhang mit inhaltlichen Beschreibungen von Tests oder Prüfungen unklar ist, was mit „Testanleitungen" gemeint sein könnte, hilft wieder einmal der Blick ins englische Originaldokument: Dort ist die Rede von "…drawing up a *task specification* for a communicative assessment" (Herv. d. V.). Es geht also, wie die Überschrift des betreffenden Abschnitts 9.2.1 besagt, um *Testspezifikationen*, inhaltliche Beschreibungen also, die durchaus der Erstellung von Testitems dienen können, doch handelt es sich dabei nicht um *Testanleitungen*, die im Allgemeinen den Tests vorgeschaltet sind und Handlungsanweisungen für die Probanden enthalten.
[264] Vgl. auch die Ausführungen unter Kapitel 4.2.5 dieser Arbeit zum Textbegriff im GER.

Aussagen zu **kommunikativen Aufgaben** finden sich in Abschnitt 7 des GER.[265] Kommunikative Aufgaben werden dort in einem umfassenden Sinn verstanden als „Merkmale des alltäglichen Lebens", die „zielgerichtete Handlungen mit einem klar definierten Ziel und einem speziellen Ergebnis" beinhalten, „ihrem Wesen nach sehr unterschiedlich sein" und „sprachliche Aktivitäten in unterschiedlichem Umfang enthalten" können (alle Zitate hier: GER 2001: 153, erster Absatz). Aufgaben werden im GER unterschieden in solche, die „reale Sprachverwendung widerspiegeln" und solche, die „im Wesentlichen didaktischer Art sind" (ebd.: 153, letzter Absatz) – obwohl man im GER bis zu dieser Stelle ausdrücklich die Sprachverwendenden und Sprachlernenden gleichsetzte (vgl. etwa ebd.: 51). Auch wenn diese Unterscheidung einen internen Bruch in der GER-Konzeption darstellt, so ist sie doch hilfreich, denn wie in der vorliegenden Arbeit in Kapitel 1.3 gezeigt, sind Sprachverwendung und Sprachlernen eben nicht gleichzusetzen. Didaktische Aufgaben sind selbstverständlich an der Realität auszurichten, dennoch unterscheiden sie sich zu realen Aufgaben etwa dahingehend, dass sie konstruiert werden und daher bestimmte Charakteristika gezielt kontrolliert werden können – konsequenterweise lassen sich im GER folgende Merkmale kommunikativer didaktischer Aufgaben ausmachen (vgl. ebd.: 153f):

- Sie ermöglichen eine aktive Beteiligung der Lernenden an sinnvoller Kommunikation;
- sie sind relevant für das „Hier und Jetzt";
- sie sind eine „Herausforderung", jedoch für die Lernenden „machbar";
- sie führen zu „erkennbaren… Ergebnissen";
- sie sind „in dem Maß kommunikativ, in dem sie von den Lernenden verlangen, Inhalte zu verstehen, auszuhandeln und auszudrücken, um ein kommunikatives Ziel zu erreichen";
- sie fokussieren auf die „erfolgreiche Bewältigung" des Ziels, „im Mittelpunkt steht folglich die inhaltliche Ebene".

Dieses Verständnis kommunikativer Aufgaben deckt sich mit dem der o. g. Fachliteratur und dem Verständnis im DESI-Projekt, dennoch ist ein Rekurs auf die Fachliteratur ratsam, da die Quellen der GER-Ausführungen nicht explizit genannt werden. Die im GER genannten Merkmale stecken aber einen sinnvollen Rahmen für die Entwicklung kommunikativer Aufgaben.

Bei der Entwicklung der **Aufgabencharakteristika** haben sich die Ausführungen in GER-Abschnitt 7.3 zu *Schwierigkeitsgraden kommunikativer Aufgaben* als informativ erwiesen. Dort wird ausgehend von der Erläuterung des Zusammenhangs zwischen Aufgabenanforderungen und Probandenmerkmalen (wie etwa Kompetenzen, Sprachvermögen und Persönlichkeitsmerkmalen) unter Abschnitt 7.3.1 auf die *Kompetenzen und Merkmale der Lernenden* und unter 7.3.2 auf die *Bedingungen und Einschränkungen kommunikativer Aufgaben* eingegangen. Insbesondere die Ausführungen in GER-Abschnitt 7.3.2.1 zu *Interaktion und Produktion* haben sich als hilfreiche Checkliste erwiesen, um zu überprüfen, ob alle relevanten Merkmale in der DESI-Aufgabenbeschreibung auch erfasst wurden. Der GER zählt als schwierigkeitsbestimmend im Hinblick auf interaktive und produktive Aufgaben folgende Bedingungen und Einschränkungen

[265] Die hier folgenden Aussagen sind in Kapitel 1.3.4 dieser Arbeit bereits getroffen worden, doch um der Lesbarkeit willen sollen sie an dieser Stelle noch einmal angeführt werden.

auf: Hilfen, Zeit, Ziel, Vorhersehbarkeit, materielle Bedingungen und die Teilnehmenden (ebd.:158f). Die Erläuterungen im GER zu diesen Aspekten sind einleuchtend, jedoch wieder nicht mit Quellenangaben belegt, so dass sie als alleinige Basis für Aufgabenmerkmale nicht ausreichen. Zudem gibt es weitere, oben genannte Merkmale, die im GER an dieser Stelle nicht aufgeführt werden, wie beispielsweise die Ansprüche der Themenstellung und Textsorte oder die Art der Aufgabenstellung. Wieder kommt man als Nutzerin des GER zu dem Schluss, dass er zwar eine Fülle teils hilfreicher Informationen zu vielfältigen Aspekten bietet, die meisten dieser Aspekte jedoch nicht umfassend abdeckt und die Informationen dazu oft nicht systematisiert darbietet.

Zusätzlich zu den Ausführungen im GER haben sich die Checklisten und Formulare, die im *Manual*[266] insbesondere in seinem Abschnitt 4 die Spezifizierung betreffend angeboten werden, als hilfreich bei der Beschreibung von Aufgabenmerkmalen erwiesen. Beispielsweise bieten *Form A12 Written Interaction* (*Manual* 2003: 46) und *Form A14 Written Production* (ebd.: 48f) eine gute Übersicht zu relevanten Merkmalen dieser Aktivitäten und sie nehmen explizit Bezug auf relevante Aussagen und Skalen im GER. Ein rückblickender Vergleich der unabhängig in DESI angesetzten Merkmale mit den im *Manual* genannten zeigt eine hohe Übereinstimmung (etwa bezüglich der Merkmale Themen, kommunikative Handlungen, Textsorten oder bezüglich der Aufgabenstellung), so dass dieser Vergleich als gegenseitige inhaltliche Validierung der jeweils angesetzten Merkmale betrachtet werden kann.

Bei der eigentlichen **Aufgabenentwicklung** in DESI hat der GER selbst keinen Beitrag leisten können. Die Aussage des GER (ebd.: 174), dass „Abschnitt 5.2. über ‚kommunikative Sprachkompetenzen' … die inhaltliche Basis für die Erstellung von Testaufgaben bzw. für die Phasen eines Tests mündlicher Fertigkeiten dar[stellt]", ist wie gesagt im Licht der Skalenanalysen des Kapitels 3.4 der vorliegenden Arbeit nicht nachvollziehbar.[267] Wie dort gezeigt wurde, sind die Skalen des betreffenden Abschnitts ausgerichtet auf generelle, dekontextualisierte Beschreibungen der sprachlichen Kompetenzen und deshalb nicht geeignet zur Testerstellung, da sie die betreffenden Merkmale eben nicht beschreiben. Insbesondere finden sich keine Deskriptoren, die Merkmale offener Schreibaufgaben abbilden und sich als Aufgaben operationalisieren lassen. Mag man noch Hinweise auf Themen und Funktionen kommunikativer Handlungen beispielsweise in der Skala „Spektrum sprachlicher Mittel" (ebd.: 110f) finden, die sich in zu konstruierenden Aufgaben widerspiegeln könnten, so fehlt deren Abstufung, wie ebenfalls in Kapitel 3.4 dieser Arbeit gezeigt wurde, die valide empirische Basis. Zudem lassen sich in den meisten der Skalen des GER-Abschnitts 5.2 keine solch expliziten Hinweise finden. Die Skala „Orthographie" etwa enthält zwar im Bereich A1 den Deskriptor „ Kann … einfache Schilder oder Anweisungen… abschreiben", welcher operationalisierbar ist – doch welcher Aspekt der Schreibfertigkeit wird dabei erfasst? Wie könnte dieser in ein Testkonzept eingefügt werden? Wie soll der C2-Deskriptor dieser Skala („Die schriftlichen Texte sind fei von orthographischen Fehlern")

[266] Vgl. Kapitel 3.5 dieser Arbeit.
[267] Dasselbe gilt für die Skalen des GER-Abschnitts 4.4 – für die Testentwicklung stellen sie keine valide Basis dar, wie in Kapitel 3.4 der vorliegenden Arbeit gezeigt wurde.

operationalisiert werden, wenn er in dasselbe Textkonzept gefügt werden soll wie der erwähnte A1-Deskriptor? Die im GER (ebd.: 174) erwähnten Lernzielbeschreibungen, die den genannten GER-Abschnitt und seine Skalen ergänzen, mögen zusätzliche Informationen und Beispiele geben (ähnlich wie die o. g. Curriculaanalysen), doch eine inhaltliche Basis für die Erstellung von Testaufgaben können sie nicht darstellen. Dazu gehört weit mehr, wie die obigen Ausführungen zu Testkonzept, Testkonstrukt und Aufgabenmerkmalen zeigen.

Bei der Aufgabenerstellung hat sich der UGE jedoch als hilfreich erwiesen. Dort wird, wie in Kapitel 2.6 dieser Arbeit erläutert, der Testentwicklungsprozess im Detail beschrieben und es werden Überblicke und Auflistungen zu den verschiedenen Phasen gegeben, die helfen, den Prozess fortlaufend zu evaluieren. Beispielsweise finden sich die wichtigsten Merkmale und Aspekte zur Testspezifizierung und zur Beschreibung von Aufgabenmerkmalen in UGE-Abschnitt 2.2, der für die Aufgabenbeschreibung im DESI-Projekt abschließend konsultiert wurde. Der Überblick in Abschnitt 2.3 über die unterschiedlichen Entwicklungsschritte von der Planung eines Tests bis hin zum eigentlichen Schreiben von Tasks ist ebenfalls nützlich, um sich während der Instrumentenentwicklung zu vergewissern, dass keine wesentlichen Schritte oder Dokumentationen von Entscheidungen vergessen werden. Beispielsweise wurden die Fragen im Kasten auf S. 17 des UGE als Checkliste für die entwickelten Tasks in DESI genutzt. Die Beschreibungen der Phasen des Prätestens im kleinen Rahmen und der Pilotierung mit größeren Stichproben in UGE-Abschnitt 2.4 sind aufschlussreich, unterscheiden sie doch zwischen objektiv auszuwertenden Items und subjektiv zu bewertenden Schreib- (und Sprech-) Tests: Bei den subjektiven Tests muss laut UGE eine Analyse der Testprodukte und ihrer Bewertungen zeigen, ob die Tasks die gewünschten Reaktionen elizitieren, ob das Bewertungsschema zu einer validen Bewertung führt, und wo es noch Verbesserungsbedarf gibt. Auch im DESI-Projekt wurde diesem Aspekt Rechnung getragen, wie im Folgenden in Kapitel 4.4 dieser Arbeit gezeigt wird. Die sich im UGE in den Abschnitten 2.5 und 2.6 anschließenden Ausführungen zu den Aspekten der Testkonstruktion und des Itemschreibens setzen einen hilfreichen Rahmen, der bei der Taskkonstruktion im DESI-Projekt immer wieder zu Rate gezogen wurde, sei es, um alle relevanten Aspekte zu bedenken und soweit möglich zu kontrollieren oder sei es, um anstehende Entscheidungen noch einmal kritisch zu überprüfen. Es muss aber festgehalten werden, dass bei der Entwicklung der DESI-Schreibaufgabe keine Entscheidung alleine auf Basis des UGE getroffen wurde – wie oben dokumentiert, sind alle Entscheidungen in Kontextanalysen und in wissenschaftlichen Theorien und Modellen verankertt worden, ehe GER und UGE konsultiert wurden.

4.4 Bewertungsschema und Skalenkonstruktion

Für das DESI-Testmodul *semikreatives Schreiben* bedarf es eines Bewertungsschemas, das der o. g. Breite der Antwortmöglichkeiten und den unterschiedlichen Entwicklungsständen der Lernenden gerecht wird. Die Bewertung der Aufsätze hat das Ziel, zu diagnostischen und generalisierbaren Aussagen zu kommen. Dazu muss die Bewertung eine Profilbildung der Kompetenzen der Lernenden ermöglichen, wobei sie selbstverständlich den Gütekriterien der Objektivität, Reliabilität und Validität genügen muss. Um zu einem validen Bewertungsinstrumentarium zu kommen, basiert die Entwicklung desselbigen zum einen in Analysen von Lernertexten und zum anderen in der Forschung zur Aufsatzbewertung.

Deshalb werden im Folgenden die Erkenntnisse der Bewertungsforschung dargestellt, die für das besagte DESI-Testmodul relevant sind. Darauf aufbauend wird die Entwicklung des Bewertungsschemas in der Praxis beschrieben und die eigentliche Skalenkonstruktion im DESI-Projekt dokumentiert. Am Ende dieses Unterkapitels werden wiederum der GER und sein Beitrag zur theoretischen Verankerung und praktischen Entwicklung des Bewertungsschemas betrachtet.

4.4.1 Forschung zur Aufsatzbewertung und Ableitung des DESI-Bewertungsschemas

Im DESI-Modul Textproduktion geht es um die Erfassung der Schreibkompetenzen anhand von fremdsprachlichen Textprodukten. Das dem Modul zugrunde gelegte Kompetenzmodell ist zum Teil aus Produkteigenschaften abgeleitet, doch werden diese nicht als „abstrakte Textnorm“ angesetzt, an der die Schülertexte gemessen werden. Vielmehr wird die Bewertung auf die Struktur von Lernertexten ausgelegt, wie dies auch Feilke (2005: 5) fordert: „Eine erwerbsorientierte Produktbewertung müsste auf die Struktur von Lernertexten bezogen sein. Sie setzt dafür eine Theorie nicht gelungener Texte bzw. mehr oder weniger gelungener Problemlöseversuche (attempts) voraus.“ Deshalb ist die Analyse von Lernertexten, die während der Präpilotierung gewonnen wurden, eine zentrale Säule der Entwicklung des DESI-Bewertungsschemas. Auf diese Analysen wird bei der Skalenentwicklung in Kapitel 4.4.2 dieser Arbeit noch näher eingegangen; die Ergebnisse der Analyse sind in Anhang 24 dargestellt.

Da im DESI-Modul *Schreiben Englisch* keine (idealisierte) Textnorm angesetzt wird, an denen die Lernertexte gemessen werden können und Fehlendes aufgezeigt werden kann, macht ein Herangehen an die Texte im Sinne der klassischen Negativkorrektur wenig Sinn. Ein Fehlerindex mag Aufschlüsse geben über das, was noch nicht korrekt (im Hinblick auf bestimmte Normen) beherrscht wird, er sagt aber nichts über das aus, was schon gekonnt wird. Beispielsweise kann ein einfacher, kurzer Text zur eigenen Person fehlerfrei sein, doch sagt diese Fehlerfreiheit nichts über die textuellen und sprachlichen Qualitäten des Produkts aus. Aus Sicht der Interimsprachentheorie (vgl. Kapitel 1.3.1 dieser Arbeit) stellen Fehler einen notwendigen Teil

der interimsprachlichen Entwicklung dar und geben für sich alleine genommen nur bedingt Aufschluss über das Sprachvermögen. Zudem haben unterschiedliche Fehler unterschiedliche Bedeutung und Auswirkung, so dass ein reines Zählen der Fehler ohne Klassifizierung nicht zum gewünschten Ergebnis führen kann. Hamp-Lyons & Kroll (1996) beispielsweise unterscheiden zwischen „globalen" Fehlern, die sie als kommunikationsbelastend beschreiben, und „lokalen" Fehlern, die keine Verständigungsprobleme schaffen. Ersteren sollte bei einer Bewertung dessen, was schon beherrscht wird, Aufmerksamkeit zukommen – dies wird im DESI-Bewertungsschema beachtet.

Deshalb (und aus den Gründen, die bereits in Kapitel 3.3 dieser Arbeit erwähnt wurden) wird im DESI-Projekt positiv an die Lernertexte herangetreten, um deren Qualität bewerten zu können und um einzuschätzen, welche Fertigkeiten und Kompetenzen bei den Lernenden im Bereich des Schreibens inwieweit ausgeprägt sind. Um die Qualität der Lernertexte einzuschätzen, genügen jedoch rein quantitative Verfahren nicht: „Einem sprachlichen Gefüge, in dem Ziele, Aufbau, Inhalt und Ausdruck eng aufeinander abgestimmt sind, wird man nicht gerecht, indem man Satzstrukturen auflistet oder den Wortschatz auszählt." (Börner 1989: 370). Deshalb werden in DESI die qualitative und die quantitative Seite erfasst, wie es in Kapitel 2.2.3 dieser Arbeit erläutert wurde: Der dort erwähnte Vorschlag (b) von Pollitt & Murray (1996: 75f) wird im DESI-Projekt wie folgt umgesetzt: In der Präpilotierung werden Lernertexte elizitiert, analysiert und in relevanten Merkmalen beschrieben. Auf dieser Basis werden die zu erwartenden Performanzen beschrieben. Diese Erwartungen werden unter Beachtung des Schreibentwicklungsmodells von Bereiter (1980) bestimmten Niveaus zugewiesen, wobei ein Interpretationsschema (vgl. das Handbuch in Anhang 27) für ein gemeinsames Verständnis der Merkmale und Abstufungen unter den Bewertern sorgt. Die Niveauzuweisungen werden zur Validierung mit Außenkriterien verglichen, zum einen mit bereits existenten Bewertungsskalen der *Cambridge-ESOL-Tests*, zum anderen mit relevanten Skalen aus dem GER. So können Testperformanzen unter quantitativen und qualitativen Aspekten beschrieben werden und im Hinblick auf Schreibkompetenzen, die vermutlich auch im realen Leben zur Verfügung stehen, generalisiert werden.

Angemessene Verfahren zur Einschätzung der Qualität von Lernertexten bei einer offenen Aufgabenstellung sind in *Rating*-Verfahren zu finden. An dieser Stelle darf auf die grundlegenden Ausführungen hierzu in Kapitel 3.3 dieser Arbeit verwiesen werden, welche auch dem Ansatz im DESI-Modul *Schreiben Englisch* zugrunde liegen: Die Einschätzung der globalen Schreibfertigkeit erfolgt holistisch. Um die in Kapitel 3.3 dieser Arbeit dargestellten Nachteile des holistischen Verfahrens zu minimieren und die Vorteile zu nutzen, wird das Globalurteil im Sinn des *multiple-trait* Verfahrens durch Vorgaben konkretisiert und durch analytische Kriterien (vgl. die Übersicht in Abb. 23 dieser Arbeit) ergänzt, die nach Lehmann (1990) die Reliabilität der Bewertung erhöhen. Damit ermöglicht die DESI-Studie auch Profilbildung in den einzelnen Teilfertigkeiten.

In der Literatur finden sich keine schlüssigen Hinweise, welche Kriterien in welcher Gewichtung bei der Aufsatzbeurteilung angesetzt werden sollen. Lediglich Hintergrundfaktoren

bezüglich textanalytischer Qualitäten (Inhalt und Aufbau), sprachlicher Qualitäten (Korrektheit und Stil) und bezüglich des Kontextbezugs (Originalität und Problembewusstsein) haben sich in verschiedenen Untersuchungen[268] durch Faktorenanalysen bestätigen lassen, doch je nach Untersuchungsdesign mit unterschiedlichen Gewichtungen. Milanovic, Saville & Shuhong (1996) stellen Analysen von *Rater*-Verhalten vor, wodurch sie einige der Elemente identifizieren, auf die sich die *raters* bei der Bewertung konzentrieren, wie etwa auf Länge, Lesbarkeit, Struktur, auf die traditionellen Elemente Grammatik, Lexik, Rechtschreibung und Zeichensetzung, oder auf Elemente wie Ton und kommunikative Effektivität. Bei diesen Untersuchungen stellen Milanovic et al. fest, dass diese Elemente sich gegenseitig beeinflussen und nicht unabhängig voneinander zu erfassen sind. Bezüglich der Gewichtung der angesetzten Kriterien finden sich keine eindeutigen Ergebnisse in der Forschung, wie Hamp-Lyons (1996b) oder Milanovic, Saville & Shuhong (1996) feststellen. Deshalb werden die Bewertungskriterien des DESI-Moduls *Schreiben Englisch* wie oben erläutert aus dem Testkonstrukt und dem Kompetenzmodell abgeleitet. Sie bleiben gleichgewichtet nebeneinander stehen; ihre internen Beziehungen werden bei der Skalierung der Bewertungen untersucht. Die Datenlage nach Bewertung der Hauptuntersuchung, auf die unter Kapitel 4.6 eingegangen wird, entscheidet, welches Kriterium in welcher Gewichtung in welche (Sub-)Skala einfließt.

Um zu einer zuverlässigen und damit generalisierbaren Bewertung zu kommen, müssen nach Lehmann (1990) die o. g. Testgütekriterien bezogen auf das Bewertungsinstrumentarium und die Bewertenden kontrolliert werden, so auch im DESI-Projekt. Wenden wir uns zunächst der Objektivität in der Beurteilung von Textprodukten zu: In der Literatur wird eine gewisse Subjektivität als unabdingbar anerkannt, deren Ursache in der Person der Bewertenden zu finden ist (vgl. dazu auch die Ausführungen oben in Kapitel 3.3.2.2). Shale (1996) etwa erkennt die Subjektivität als Teil der Natur des Bewertungsprozesses an, der durch geeignete Messtheorien begegnet werden kann. Die Skalierung der DESI-*Ratings* trägt den Strenge-/Milde-Tendenzen der *raters* Rechnung (vgl. dazu auch Kapitel 4.6 dieser Arbeit). Zusätzlich wird die Subjektivität über die Verfahren der Doppelt-Blind-Korrektur, der Schulung der *raters* und der Vorgabe von *benchmarks* limitiert.

Eng mit dem Kriterium der Objektivität hängt das Kriterium der Reliabilität der Bewertung zusammen; im Allgemeinen beziehen sich Reliabilitätskontrollen im Rahmen von Aufsatzbewertungen auf die Homogenität zwischen verschiedenen Bewertern (Inter-*Rater*-Reliabilitäten), auf den Zusammenhang zwischen Erst- und Zweit-*Ratings* und auf die Stabilität von Bewertungen eines *raters* über einen gewissen Zeitraum hinweg (Intra-*Rater*-Reliabilitäten). Diese Reliabilitäten werden auch bei der DESI-Auswertung fortlaufend kontrolliert. Zu geringe Reliabilitäten werden durch Nachschulungen aufgefangen. Zu den Werten der Hauptuntersuchung darf auf Kapitel 4.5 dieser Arbeit verwiesen werden.

[268] Vgl. hierzu die Übersicht in Lehmann 1990.

Das Kriterium der Validität bezieht sich bei Aufsatzbewertungen in der Regel auf die Validität des Bewertungsinstrumentariums, auf die Kriterien und deren Abstufungen also, vorausgesetzt den Aufsätzen liegen valide konstruierte Tasks zugrunde. Die *rating scales* des semikreativen Testmoduls, deren Konstruktion und Validierung gleich im Anschluss dokumentiert wird, müssen hinsichtlich ihres Beschreibungsgegenstands, ihrer Einteilung in Kategorien und Abstufungen und hinsichtlich der verwendeten Sprache validiert werden.[269] Wie oben dargestellt sind der Beschreibungsgegenstand und dessen Einteilung in Bewertungskriterien aus dem Testkonstrukt abgeleitet, welches in der wissenschaftlichen Forschung verankert ist. Zusätzlich werden die Kriterien durch Diskussion mit den *raters* während der Schulungen und in der Pilotierungsphase verfeinert. Die Abstufungen sind wie gesagt in Analysen von relevanten Lernertexten begründet und werden an Außenkriterien überprüft. Die Sprache der Deskriptoren wird ebenfalls während der *Rater*-Schulung und in Workshops analysiert und auf ihre Verständlichkeit hin überarbeitet.

Die bekannten *Ursachen für Messfehler* in der Aufsatzbewertung sind damit kontrolliert:

- Die *Aufgabenstellung* ist auf die Probanden ausgelegt; die Verteilung der vier Tasks wird innerhalb einer Klasse kontrolliert und als *random effect* behandelt; das Bewertungsinstrumentarium enthält taskspezifische Elemente; das Skalierungsmodell trägt der Aufgabenschwierigkeit Rechnung.

- Die *Probanden* erhalten zwei Tasks zur Bearbeitung, um die Momentaufnahme auf breitere Basis zu stellen und somit der Fehlerquelle der „persönlichen Verfassung“ am Testtag entgegen zu wirken.

- *Abweichungen der raters* werden aufgefangen durch Doppelt-Blind-Korrektur, durch Schulungen vor und während der Bewertung und durch ein Skalierungsmodell, das den *Rater*-Strenge-/Milde-Tendenzen Rechnung trägt.

- Die *Bewertungsverfahren* stellen eine Kombination aus holistischem *multi-trait* Verfahren und analytischem Vorgehen dar, so dass Nachteile der einzelnen Herangehensweisen möglichst minimiert und Vorteile genutzt werden können.

- Die *rating scales* werden auf die Bewertung der DESI-Tasks hin konstruiert und im Rahmen der Möglichkeiten des DESI-Projekts validiert: Ihre horizontale Einteilung ist in wissenschaftlichen Modellen verankert; die vertikalen Abstufungen werden aus Lernertextanalysen gewonnen, aus einem wissenschaftlichen Schreibentwicklungsmodell abgeleitet und an bereits existierenden Skalen validiert; zusätzlich werden Sortieraufgaben zur Validierung genutzt.

[269] Vgl. dazu die Ausführungen in Kapitel 3.2.4 dieser Arbeit.

4.4.2 Entwicklung des Bewertungsinstrumentariums

Um eine reliable und valide Bewertung zu gewährleisten, wurde ein Handbuch erstellt, in dem die Vorgehensweisen der Bewertung genau festgelegt und standardisiert sind. Dieses Handbuch enthält die Definition und Abgrenzung der oben dargestellten Bewertungskriterien. Jedes der o. g. Kriterien (bis auf das auszählbare Kriterium der Textlänge und das Kontrollkriterium der *swear words*) wird auf sechs Niveaus in je einer Skala beschrieben. Die Beschreibungen basieren auf Lernertextmerkmalen. Die Niveaus sind durch *Benchmark*-Texte illustriert, wobei es sich dabei um prototypische Lernertexte handelt, die von insgesamt 40 Bewertern und den Testentwicklern in gemeinsamer Diskussion bestimmt wurden und auf die im Handbuch Bezug genommen wird. Das Handbuch ist Grundlage des *Rater*-Trainings und der eigentlichen Bewertung. Ein Beispiel solch eines Handbuchs findet sich in Anhang 27; es bezieht sich auf den oben dargestellten Task. Die dazugehörigen *Benchmark*-Texte finden sich in Anhang 28 dieser Arbeit.

4.4.2.1 Skalenkonstruktion

Die Konstruktion von Skalen umfasst, wie in Kapitel 3 dieser Arbeit erläutert, drei Hauptaspekte: Zunächst muss die Einteilung in die horizontale Dimensionen in Theorien und Modellen begründet werden, um die Bereiche valide zu bestimmen, die die betreffenden Skalen abdecken sollen. Sind die horizontalen Teildimensionen gefunden, muss die Abstufung in angemessene Niveaus entwickelt werden – wiederum durch die Verankerung in Theorien und Modellen; durch die abgestufte Beschreibung relevanter Merkmale des Gegenstandsbereichs, den die Skalen abdecken sollen; und idealiter durch psychometrische Messmodelle, die die Skalierung der einmal entwickelten Deskriptoren validieren helfen. Dem dritten Aspekt, der Basis der Beschreibung des Skalengegenstands, kommt zentrale Bedeutung zu, bestimmt diese doch den Status der Deskriptoren und damit auch die Verwendungsmöglichkeiten der Skalen.

Die horizontale Skaleneinteilung des DESI-Bewertungsschemas resultiert aus den oben analysierten Leistungsdimensionen, die zu den genannten Bewertungskriterien führen, welche in je einer Skala beschrieben werden. Die vertikale Abstufung der Skalen basiert einerseits auf dem oben vorgestellten Schreibentwicklungsmodell von Bereiter (1980); andererseits gründet sie auf der genannten Analyse der Lernertexte aus der Präpilotierung, welche gleichzeitig die Basis der Beschreibungen darstellen. Bei der Analyse wurde wie folgt vorgegangen: Zunächst wurden die Texte nach mehrmaliger Lektüre von zwei Teammitgliedern (u. a. der Verfasserin der vorliegenden Arbeit) in drei grobe Kategorien eingeteilt: *basic* – *medium* – *advanced*. Dann wurden die Aufsätze unter folgenden Gesichtspunkten in Anlehnung an die oben entwickelten Kriterien beschrieben, um die Zuordnung aufgrund des ersten Eindrucks zu überprüfen und ggf. zu revidieren:

- *Textsorte*: Folgende Aspekte wurden beschrieben: Handelt es sich um einen Bericht? Welche Makrostruktur lässt sich ausmachen? Ist der Text in sinnvolle Absätze gegliedert?

- *Länge*: Hier wurde die Wortanzahl überschlagen.

- *Inhalt*: Der Inhalt wurde unter den nachstehenden Gesichtspunkten analysiert: Welche Ideen sind vorhanden? Wie viele unterschiedliche Ideen werden vorgebracht? Von welcher Relevanz sind die Ideen? In welcher Qualität sind sie entwickelt?

- *Sprache*: Die Bereiche der Orthographie, des Wortschatzes und der Grammatik wurden auf Umfang und Korrektheit der sprachlichen Mittel hin beschrieben. Dabei wurden herausstehende Merkmale festgehalten.

- *Kommunikative Wirkung*: Bei diesem Kriterium wurden nachstehende Facetten untersucht: Welchen Effekt hat der Text auf die Leser? Über welche Qualitäten – wie etwa Spannung, Witz, emotive oder atmosphärische Darstellungen – verfügt der Text? Ist ein Adressatenbezug erkennbar?

Die Analysen der Lernertexte zu dem oben vorgestellten Task finden sich in Anhang 24. Die dabei auf drei Stufen beschriebenen prototypischen Merkmale bilden die Grundlage der Skalenabstufung, wobei sich diese vertikale Kategorisierung wiederum am in Kapitel 1.2.1 dieser Arbeit vorgestellten Prototypenmodell orientiert, so dass die Beschreibungen jeweils auf die Mitte eines Niveaus abzielen:

- *Basic*: Formale Merkmale meist nicht umgesetzt (keine Absätze, keine Struktur); sehr kurze Texte; Inhalte nur zum Teil relevant, nicht elaboriert oder motiviert entwickelt; sehr einfache Sprache, fehlerhaft, starke muttersprachliche Interferenzen, teils auswendig gelernte Strukturen und *chunks*; Wirksamkeit nicht oder nur sehr begrenzt gegeben.

- *Medium*: Formale Merkmale teils umgesetzt (Absätze vorhanden, aber teils nicht angemessen, einfache, doch meist logische Makrostruktur); wesentliche Ideen vorhanden, doch wenig elaboriert; Beherrschung der Basisstrukturen und des Grundwortschatzes, Umschreibungen erkennbar, Sprache teils noch fehlerhaft; Verständlichkeit meist gegeben, Botschaft erkennbar, doch teils mit Einschränkungen.

- *Advanced*: Formale Merkmale korrekt (Einleitung und Schluss, angemessene Absatzeinteilung, stringente, teils komplexe Makrostruktur); Texte i. d. R. von geforderter Länge; relevante und interessante Ideendarstellung und Entwicklung; idiomatischer Sprachgebrauch, Fehler selten und i. d. R. nicht kommunikationsbelastend; hochwirksame Texte mit spannenden, witzigen, emotiven Qualitäten.

Da sich die Texte auf drei Stufen nicht hinreichenden differenziert einordnen ließen, wurden die relevanten Merkmale auf fünf Niveaus und einem untersten "*insufficient evidence"*-Niveau differenziert. Das unterste Niveau ist bedingt durch die Notwendigkeit, „valide" von „nicht validen" Texten zu unterscheiden: Nicht valide Texte sind solche, in denen die Aufgabenstellung nicht bearbeitet wird; Texte, in denen die Aufgabenstellung bearbeitet wird, doch in solch knapper

oder irrelevanter Weise, dass sie eine begründete Bewertung nicht zulassen, werden als *insufficient evidence* auf die unterste Stufe eingeordnet.

Die Abstufungen der Merkmale werden zusätzlich im o. g. Schreibentwicklungsmodell von Bereiter (1980) verankert. Der Zusammenhang zwischen den Phasen dieses Modells und den abgestuften Merkmalen lässt sich am Beispiel des Globalurteils veranschaulichen:

Die erste Entwicklungsphase des assoziativen Schreibens (vgl. oben: Bereiter: 1980) lässt sich in Schüleraufsätzen (die auf das vorläufige Niveau *basic* eingestuft wurden) daran erkennen, dass diese Aufsätze wenig Struktur aufweisen, sei es nun auf Makroebene oder auf der Ebene der einzelnen Propositionen. Häufig sind dabei auch Defizite bezüglich der Sprache und der Umsetzung der Aufgabenanforderungen zu beobachten: Es scheint, dass Schüler, die sich in dieser Phase befinden, noch nicht in der Lage sind, sprachliche, textuelle oder pragmatische Konventionen umzusetzen. Die nächste Phase des performativen Schreibens lässt theoretisch erwarten, dass die Lernenden, die diese Stufe erreicht haben, Konventionen bezüglich des Aufbaus und der Sprache beachten können. Tatsächlich lässt sich bei den Aufsätzen im Mittelbereich feststellen, dass der Aufbau zunehmend logisch erfolgt und die sprachlichen Leistungen weniger kommunikationsbelastende Fehler aufweisen. Basierend auf den Annahmen der dritten Phase, des kommunikativen Schreibens, müssten die besten Aufsätze auch die kommunikativ erfolgreichsten sein, bei denen der Aufbau das Lesen erleichtert, ein Adressatenbezug feststellbar ist und sprachliche Mittel wirkungsvoll eingesetzt werden können. Diese Annahmen lassen sich anhand der Charakteristika der Schüleraufsätze im oberen Bereich bestätigen.

4.4.2.2 Validierung

Die genannten theoretischen Annahmen und empirischen Beobachtungen wurden über den Vergleich mit den Bewertungsskalen der *Cambridge-ESOL-Tests CPE, CAE und FCE* und mit den Niveaubeschreibungen des Referenzrahmens zu produktivem und interaktivem Schreiben validiert. Dabei wurden folgende GER-Skalen genutzt: *Schriftliche Produktion* (GER 2001: 67), *Briefe und Aufsätze schreiben* (ebd.: 68), *Kreatives Schreiben* (ebd.: 67f), *Schriftliche Interaktion allgemein* (ebd.: 86), *Korrespondenz* (ebd.), *Schreiben* aus dem Selbstbeurteilungsraster (ebd.: 36); daneben wurden die Skalen zu den linguistischen Kompetenzen bezogen auf Umfang und Korrektheit von *Lexik*, *Grammatik* und *Orthographie* (ebd.: 112f, 114, 118), die Skala *Themenentwicklung* und die Skala *Kohärenz und Kohäsion* (ebd.: 125) genutzt.

Dieser Abgleich wird in der vorliegenden Arbeit wiederum am Beispiel des Globalurteils der *Biography*-Aufgabe dokumentiert. Dazu finden sich Tabellen in Anhang 25, die die Berührungspunkte der erwähnten *Cambridge*-Skalen mit den entsprechenden GER-Deskriptoren und den DESI-Merkmalen darstellen. Auf dieser Basis wurde eine vorläufige Globalskala entworfen. Diese wurde mit dem Wortlaut relevanter GER-Deskriptoren abgeglichen: In Anhang 26 sind die

Deskriptoren aus den dort benannten GER-Skalen der DESI-Globalskala gegenübergestellt. Dabei zeigt sich auch in der Praxis[270], dass einige der Formulierungen der GER-Skalen für die Ableitung von Deskriptoren zur Bewertung genutzt werden können, soweit sie Merkmale beschreiben, die sich auch in den Lernertexten identifizieren lassen. Beispielsweise finden sich Merkmale wie „einfache Wendungen über sich selbst und fiktive Menschen" (GER, Skala *Kreatives Schreiben*: A1), „Gefühle und Reaktionen … beschreiben" (ebd.: B1) oder „entscheidende Punkte hervorheben, Standpunkte ausführlich darstellen, durch … Beispiele stützen … und Text durch angemessenen Schluss abrunden" (GER, Skala *Schriftliche Produktion*: C1) auch in den Lernertexten wieder, so dass die DESI-Globalurteilsbeschreibung in diesen Punkten an die Formulierungen des GER angelehnt wird. Die für eine Adaption im DESI-Kontext geeigneten Formulierungen sind in Anhang 26 fettgedruckt. Andere GER-Deskriptoren sind für dieses Vorgehen nicht direkt geeignet, da sie selbst für ein Globalurteil zu dekontextualisiert und generell gehalten sind. Hierunter fallen Deskriptoren wie beispielsweise „Kann eine Geschichte erzählen" (GER, Skala *Kreatives Schreiben*: B1), „Kann unkomplizierte, zusammenhängende Texte zu mehreren vertrauten Themen… verfassen…" (GER, Skala *Schriftliche Produktion*, B1); hierunter fällt auch der Bereich der Textsorten und Themen, die in den GER-Skalen notwendigerweise nur generell beschrieben werden können. Mit solch dekontextualisierten Deskriptoren ist es nicht möglich, die in den GER-Skalen beschriebenen generellen Merkmale an konkreten zu bewertenden Performanzen festzumachen. Mit ihnen können lediglich Berührungspunkte lokalisiert werden als Basis für die nach der Bewertung zu konstruierenden Kompetenzskalen, die in gewissem Umfang Generalisierungen enthalten. Beispielsweise lassen sich Berührungspunkte zwischen GER- und DESI-Niveaus bezüglich der Themen von „vertraut" über „aus Interessengebiet" hin zu „komplex" ausmachen, welche auch bei der Beschreibung der DESI-Kompetenzniveaus genutzt werden. Doch dazu Näheres unter Kapitel 4.6 dieser Arbeit.

Die DESI-Deskriptoren des Globalurteils (Stufen eins mit fünf) sind während ihrer Entwicklung mehrmals durch fortgeschrittene Studierende der Anglistik auf die fünf Niveaus eingestuft worden.[271] Solche Deskriptoren, bei denen sich die Studierenden uneinig waren, sind überarbeitet worden; Deskriptoren, die hohe Einigkeit unter den Sortierern erzielten, sind i. d. R. beibehalten worden. Auch die *raters* nahmen an diesen Sortieraufgaben im Rahmen der Schulungen (vgl. Kapitel 4.5.1 dieser Arbeit) teil und gaben Feedback hinsichtlich der Verständlichkeit, Interpretierbarkeit und Anwendbarkeit der Deskriptoren, welches in den Überarbeitungsprozess rückfloss.

Hier verdeutlicht eine Tabelle exemplarisch die prozentualen Übereinstimmungen von 18 *raters* hinsichtlich der Einstufung der überarbeiteten Deskriptoren der Stufen 1 mit 5 des Globalurteils (vgl. Anhang 26, rechte Spalte oder das Globalurteil im Handbuch, Anhang 27); die Sortierung fand in der ersten Phase der letzten Schulung statt. Die Tabelle ist wie folgt zu lesen: Vertikal

[270] Es darf auf die theoretischen Ausführungen in den Kapiteln 3.4.3 und 3.4.4 dieser Arbeit zur Verwendbarkeit der GER-Skalen auf Basis der Skalenanalysen verwiesen werden, die durch die Praxiserfahrungen bestätigt werden.

[271] In diesem Zusammenhang darf auch auf den Workshop auf der Tagung *Standards in Language Learning and the Common European Framework* des *British Council* im März 2004 in Berlin verwiesen werden, bei dem die Teilnehmer ebenfalls Deskriptoren des Globalurteils sortierten, vgl. Harsch 2004.

sind die je vier Deskriptoren eines jeden DESI-Niveaus angegeben, die eingestuft werden sollen; horizontal sind die Niveaus I mit V angegeben, auf die die Deskriptoren eingestuft wurden:

	I	II	III	IV	V
I.1 Textsorte	67%	33%	0%	0%	0%
I.2 Inhalt	100%	0%	0%	0%	0%
I.3 Sprache	89%	11%	0%	0%	0%
I.4 Wirkung	100%	0%	0%	0%	0%
II.1 Textsorte	17%	83%	0%	0%	0%
II.2 Inhalt	6%	78%	11%	0%	6%
II.3 Sprache	6%	94%	0%	0%	0%
II.4 Wirkung	0%	67%	33%	0%	0%
III.1 Textsorte	0%	0%	100%	0%	0%
III.2 Inhalt	0%	6%	67%	28%	0%
III.3 Sprache	0%	0%	100%	0%	0%
III.4 Wirkung	0%	22%	78%	0%	0%
IV.1 Textsorte	0%	0%	0%	89%	11%
IV.2 Inhalt	0%	22%	11%	56%	6%
IV.3 Sprache	0%	0%	6%	94%	0%
IV.4 Wirkung	0%	0%	0%	94%	6%
V.1 Textsorte	0%	0%	0%	6%	94%
V.1 Inhalt	0%	0%	0%	6%	94%
V.3 Sprache	0%	0%	0%	11%	89%
V.4 Wirkung	0%	0%	0%	0%	100%

Tabelle 7: Übereinstimmung bei der Sortierung der DESI-Deskriptoren Globalurteil

Neben der vertikalen Abstufung der Deskriptoren wurde in DESI auch die oben beschriebene horizontale Einteilung der Bewertungskriterien an der Datenlage der Pilotierungsstudie überprüft: Die Aufsätze der Pilotierung wurden von sechs geschulten *raters* (ebenfalls fortgeschrittene Studierende der Anglistik) nach dem hier dargestellten Schema (allerdings noch nicht in der hier vorgestellten überarbeiteten Version) auf sechs Stufen eingeschätzt. Diese Daten wurden skaliert, um die statistischen Dimensionalitäten und die Zusammenhänge der Bewertungskriterien zu prüfen. Auf Basis der Skalierungen und aufgrund von Faktorenanalysen[272] können in Abhängigkeit von der Aufgabenstellung eine respektive zwei Dimensionen (bezüglich der inhaltlichen und der sprachlichen Kriterien) bestimmt werden. Faktorenanalysen ergeben bei allen Tasks einen Faktor, der die Varianz zu etwa 85% erklärt. Wird der Einfluss des Globalurteils statistisch herausgerechnet, so lassen sich bei allen Aufgabenstellungen zwei Faktoren identifizieren, auf die die inhaltlich-formalen respektive die sprachlichen Kriterien laden.

Da diese statistischen Analysen auf mehr als eine Dimension hinweisen und zudem die statistische Dimensionalität nicht mit der didaktischen verwechselt werden darf (vgl. Kapitel 2.1 dieser Arbeit), wurde in der Hauptuntersuchung an den detaillierten Kriterien festgehalten, auch um diagnostische Informationen über den Stand der Lernenden zu gewinnen.

[272] Die Skalierung erfolgte wie erwähnt an der Humboldt-Universität zu Berlin, die Berechnungen in Zusammenarbeit mit dem DIPF.

4.4.2.3 Aspekte der Beschreibung und Illustrierung

Die Basis der Beschreibungen der DESI *rating scales* ist wie erläutert in relevanten Lernertexten zu finden; relevant sind diese Lernertexte insofern, als sie eine Reaktion auf die DESI-Tasks darstellen und somit relevant für die Bewertung dieser Tasks sind. Da es sich wie gesagt um eine Positivbewertung handelt, sollten die Deskriptoren sowohl die Textmerkmale als auch das Können, das sich in den Lernertexten zeigt, positiv beschreiben. Bei der Entwicklung der Deskriptoren wurde allerdings von den in Kapitel 3.2.3 und 3.3.1 genannten Grundsätzen guter Deskriptoren in begründeten Fällen abgewichen: Die dort erwähnten Aspekte der Ausrichtung der Beschreibung auf den Zweck der Skala, der Benutzerfreundlichkeit, der Kontextualisierung von *rating scales*, der inhaltlichen und qualitativen Charakterisierung prototypischer Merkmale in klarer, präziser und verständlicher Sprache wurden soweit möglich bei der Konstruktion der DESI-Skalen beachtet. Allerdings wurde zum Teil auf durchgängig positive KANN-Formulierungen verzichtet, wenn sie zu unangemessenen oder nicht handhabbaren Formulierungen führten. Beispielsweise zeigt sich auf den unteren Niveaus Fehlendes als charakteristisch, das konsequenterweise negativ beschrieben wird; bei den Kriterien des *Inhalts* und der *Textsorte* liegt der Fokus nicht auf der Beschreibung des Könnens aus Lernerperspektive (welches sich an den Lernertexten zeigt), sondern auf der Beschreibung relevanter Textmerkmale, so dass diese in den zu bewertenden Schülertexten lokalisiert werden können; auch bei den sprachlichen Kriterien werden typische sprachliche Strukturen und Phänomene in den Deskriptoren benannt, um die Identifizierung in den zu bewertenden Performanzen zu erleichtern.

Die DESI-Deskriptoren beschreiben neben prototypischen Merkmalen von Lernertexten typisches, an den Texten beobachtbares Verhalten bei der Textproduktion, um den Bewertern möglichst viele Aspekte zu bieten, die sie mit den zu bewertenden Performanzen in Zusammenhang bringen können. Dabei müssen die Bewertenden die im Text zu beobachtenden Merkmale identifizieren und interpretieren – interpretieren insofern, als diese Merkmale ein Indiz für das Können sind, das sich an der zu bewertenden Performanz zeigt. Deshalb sind die Deskriptoren je nach Kriterium teils als KANN-Beschreibungen, teils als Beschreibungen von Textmerkmalen gehalten.

Zur Illustrierung der DESI-Niveaus dienen wie gesagt *Benchmark*-Texte. Die Auswahl dieser Texte erfolgte in mehreren Arbeitsgängen: Zunächst dienten Lernertexte der Präpilotierung, die sich bei den erwähnten Analysen als aussagekräftig erwiesen und typische Merkmale aufwiesen, als Orientierung für die Bewertung der Pilotierungstexte. Auf Basis dieser Bewertung wurden in Rücksprache mit den Bewertern solche Pilotierungstexte ausgewählt, die eindeutig einer der Stufen zugewiesen werden konnten, da sie für das jeweilige Niveau relevante und prototypische Merkmale trugen. Diese dienten in der Schulung für die Hauptuntersuchung (vgl. unten) als Illustrierung der typischen Mitte der jeweiligen DESI-Niveaus. Die endgültige Auswahl der *Benchmark*-Texte wurde in den Schulungen gemeinsam mit allen Beteiligten durch Bewertung der entsprechenden Texte und Diskussion der Texte und Bewertungen getroffen.

4.4.3 Die Bedeutung des GER bei der Ableitung und Konstruktion des DESI-Bewertungsschemas

Bei den theoretischen und praktischen Überlegungen des Kapitels 4.4.1 der vorliegenden Arbeit, die der eigentlichen Entwicklung des Bewertungsschemas vorangehen, spielt der GER lediglich eine untergeordnete Rolle. In GER-Abschnitt 9 lassen sich einige relevante Aussagen finden, die jedoch keine neuen Erkenntnisse bringen. Als Referenzmittel mögen sie hilfreich sein zur Prüfung, ob es noch Aspekte gibt, an die bisher nicht gedacht wurde. Doch auch für diese Aussagen gilt, dass sie keine tragfähige Basis darstellen, um daraus ein Bewertungsschema abzuleiten:

Beispielsweise besagt Abschnitt 9.2, der GER könne verwendet werden zur **Festlegung der Kriterien der Beurteilung oder Bewertung**: Um relevante Beurteilungs- bzw. Bewertungskriterien zu finden, könnten neben theoretischen Analysen wiederum die GER-Abschnitte 4 und 5 helfen, Kriterien für „kommunikative Aktivitäten“ und für die dabei involvierten „Aspekte der Sprachbeherrschung“ zu finden (GER: 174). Doch wie oben gezeigt, bieten die genannten Abschnitte des GER zwar eine Auswahl an möglichen Kategorien, doch keine ausreichende Begründung für eine Bestimmung von im jeweiligen Testkontext relevanten Bewertungskriterien.

Bezogen auf die **Wahl geeigneter Bewertungsverfahren** schlägt der GER Positivbewertung unter Zuhilfenahme von Skalen vor, wobei die Verknüpfung ganzheitlicher und analytischer Verfahren als sinnvoll betrachtet wird (ebd.: 185). Anerkannt wird dabei die Subjektivität dieser Herangehensweise; um diese zu reduzieren, finden sich folgende Möglichkeiten (ebd: 183):

> - Man entwickelt *inhaltliche Vorgaben* für die Beurteilung, z. B. basierend auf einem Referenzrahmen für den betreffenden Kontext;
> - man stützt sich bei der Auswahl von Inhalten und/oder der Beurteilung der Leistungen auf *gemeinsame Entscheidungen*;
> - man verwendet *Standardverfahren*, die festlegen, wie geprüft wird;
> - man stellt *verbindliche Bewertungsschlüssel* für indirekte Tests zu Verfügung und stützt die Urteile in direkten Tests auf *spezifische, klar definierte Kriterien*;
> - man fordert *mehrfache Beurteilung* und/oder die *Gewichtung verschiedener Faktoren*;
> - man bietet entsprechendes *Training* in Bezug auf die *Beurteilungsrichtlinien* an;
> - man kontrolliert die Qualität von Leistungsbeurteilungen (Validität, Reliabilität) durch eine *Analyse der Prüfungsdaten*.

Wie oben erläutert, werden in DESI diese Möglichkeiten bedacht und umgesetzt. Doch basieren sie im DESI-Projekt auf wissenschaftlicher Forschung. Denn ein reiner Bezug auf die nicht mit Quellen belegten, in diesem Punkt jedoch umfassenden und relevanten Aussagen im GER wäre keine hinreichende Begründung für ein wissenschaftliches Vorgehen.

Die **Skalen** der genannten GER-Abschnitte 4 und 5 stellen nach Aussagen des GER (ebd.: 174) eine „Quelle zur Entwicklung von Bewertungsskalen“ und die Deskriptoren eine „Hilfe bei der Formulierung von Kriterien“ dar. Wie in Kapitel 3.4 dieser Arbeit analysiert und oben am Beispiel der DESI-Globalskala exemplifiziert, sind die GER-Skalen alleine jedoch keine hinreichende Ausgangsbasis zur Entwicklung spezifischer, auf eine konkrete Bewertung hin ausgelegter *rating scales*. Vielmehr stellen sie ein Referenzsystem dar, mit dem Skalen abgeglichen

werden können, die in ihren spezifischen Verwendungskontexten entwickelt und validiert wurden. In diesem Zusammenhang können die GER-Skalen und Niveaus als Außenkriterium dienen, doch bei dem momentanen Stand der Dinge, wie er in Kapitel 3.4.4 dargestellt wird, kann es sich nur um eine gegenseitige Validierung der GER-Skalen und der neu konstruierten Skalen handeln. Bei gemeinsamen Berührungspunkten zwischen GER-Deskriptoren und den zu entwickelnden Skalen können manche der GER-Deskriptoren in der Tat bei der Formulierung Orientierung und Hilfe bieten, doch müssen Beschreibungsgegenstand, Abstufungen und Angemessenheit der Formulierungen der betreffenden GER-Deskriptoren immer im Hinblick auf den Kontext der zu entwickelnden Deskriptoren bedacht werden.

Im Hinblick auf die **Entwicklung von *rating scales*** finden sich im UGE in Abschnitt 2.6.5 Verweise auf entsprechende Ausführungen in GER-Abschnitt 3 und in den GER-Anhängen. Diese Ausführungen behandeln jedoch nicht die Entwicklung von rating scales: Beispielsweise stellt GER-Abschnitt 3 das Skalensystem des GER vor und gibt allgemeine Hinweise zur Orientierung und Nutzung der GER-Deskriptoren; Anhang A des GER (und des CEF) behandelt ebenfalls nicht die Konstruktion von *rating scales*, sondern vielmehr die „Entwicklung von Deskriptoren der Sprachkompetenz“ (respektive “the development of proficiency descriptors“). Wie jedoch in der vorliegenden Arbeit gezeigt, sind Skalen, die das Sprachvermögen und/oder kommunikative Kompetenzen beschreiben, nicht mit *rating scales* gleichzusetzen, so dass die Verweise im UGE nicht sehr hilfreich sind.

Bei der Bestimmung der erwähnten ***Benchmark*-Texte** kann der GER wenig beitragen, da diese auf das konkrete Testkonstrukt ausgelegt sein müssen. Dennoch wären *Benchmark*-Texte nützlich, die das Referenzsystem in seinen Niveaus und Kategorien illustrieren. Diese könnten dann beispielsweise genutzt werden, um testspezifische Texte abzugleichen. Die Ausführungen zum *benchmarking* im *Manual*[273] können dabei grundsätzlich hilfreich sein, doch beziehen sie sich auf die Einstufung von Performanzen auf die Niveaus des GER. Im DESI-Projekt war das *Manual* deshalb eher von untergeordneter Bedeutung, da eine Testanbindung an den GER nicht geplant war, ebenso wenig wie eine Einstufung der DESI-Lernertexte auf GER-Niveaus.

4.5 Die Bewertung in der Praxis

Die folgenden Ausführungen beziehen sich auf die Auswertung des *DESI-Moduls semikreatives Schreiben* der Hauptuntersuchung. Zunächst wird das *Rater*-Training skizziert, das der Bewertung vorgeschaltet ist. Anschließend wird wiederum beurteilt, welche Bedeutung der GER für solch ein Training hat. Am Ende wenden wir uns der eigentlichen Auswertung und den Gütekriterien der Bewertung zu.

[273] Vgl. dazu insbesondere die Abschnitte 5.3 und 5.6 des *Manual* 2003: 70f und 86f; vgl. auch Kapitel 3.5 dieser Arbeit.

4.5.1 *Rater*-Training und der GER

Für die Bewertung der semikreativen Schreibaufgabe im DESI-Projekt wurden 40 fortgeschrittene Studierende der Anglistik im Rahmen zweier Hauptseminare (im Wintersemester 2003/2004 und im Sommersemester 2004) über jeweils 15 Doppelstunden von der Verfasserin der vorliegenden Arbeit geschult. Dabei handelte es sich mehrheitlich um Lehramtsstudierende; alle Studierenden verfügten über sehr gute Englischkenntnisse. Dadurch soll sichergestellt werden, dass die *raters* über vergleichbare Hintergründe verfügen und einen vergleichbaren Horizont für die Bewertung herausbilden können.

Die folgende Übersicht stellt dar, welche Hintergrundkenntnisse vermittelt werden, wie diese in Bezug stehen zum DESI-Bewertungsschema, und welche Aspekte des Bewertungsschemas durch welche Aktivitäten der Studierenden eingeübt werden:

Phase	Hintergrund	DESI-Projekt	Studierende
1. Dauer: ca. 4h Hausaufgabe: ca. 2h	Testtheoretische Grundlagen; Zweck und Ziele einer Schulung; Aufgaben als *Rater*	DESI: Konzept – Schreibkonstrukt – Auswertungsschema; Situation der *Testees* Test = „künstliche“ Textsorte	*Homepage* lesen vorab Jede bearbeitet einen der vier Tasks – Fokus auf Erwartungshorizont; Einnehmen Schülerperspektive
2. Dauer: ca. 4h HA: ca. 2h	GER-Niveaus und relevante GER-Skalen; Ableitung von Bewertungskriterien; Konstruktion von Abstufungen; Rolle der Deskriptoren einer Skala	DESI-Schema: Kriterien Abstufungen Deskriptoren: prototypische Natur Basis: Lernertextanalysen	GER kennen lernen, Sortieraufgaben, Selbsteinschätzung Skalierung DESI-Globalurteil Herausarbeiten Merkmale und Abstufungen
3. Dauer: ca. 2h HA: ca. 2h	*Rating*-Prozesse: *Counting or Judging*; *Universe of Raters*; *Rating*-Strategien; Ablauf und Probleme	Handbücher: Fokus auf DESI-*Rating*-Modell und Ablauf	Handbücher: Ablauf *rating* in DESI herausarbeiten Textrezeption: Kintsch/vanDijk exzerpieren
4. Dauer: ca. 2h HA: ca. 2h	Modell zu Textrezeption und Bedeutungskonstruktion	Textrezeption in DESI; Handbücher: Fokus auf Bewertung impliziter Kriterien	Handbücher: Kriterien und Abstufungen herausarbeiten
5. Dauer: je Kriterium eine Sitzung: ca. 1,5h und HA:. ca. 2h	Kognitives Verständnis der Kriterien, ihrer Definition, Basis, Abstufungen und Abgrenzungen voneinander Schulung Verhaltensweisen	Parallel: - Jedes Kriterium erarbeiten anhand von *Benchmark*-Texten und Handbüchern; insbesondere: Inhalt und „Kohärenz“ (Inhalt, Aufbau, *linkage*) zu den vier Tasks - *Rating*-Prozesse einüben	Schulungssaufsätze im jeweiligen Kriterium bewerten und dokumentieren Mikro-/Makroanalysen der Struktur der Lernertexte Paarweise: *thinking aloud*
	Wenn genügend Sicherheit:	- Bewertung ganzer Aufsätze: exemplarisches Vorführen	Übungsaufsätze selbständig mit Dokumentation des Vorgehens
Exkurs	Schulung *Excel*: Dateneingabe, Formatierung, Speicherung, Dateinamen / E-Mail: *Attachments*		
Abschluss	*Essay* zum DESI-Bewertungsschema, zu den Aufgaben als *rater*, zum Vorgehen, zu möglichen Probleme und dem Umgang damit.		

Tabelle 8: Übersicht *Rater*-Training im DESI-Projekt, Modul semikreatives Schreiben

Die theoretische Basis der ersten vier Phasen des Trainings ist in Form eines Skripts zusammengestellt, das in Anhang 29 vorgestellt wird. Das Skript setzt die theoretischen Ausführungen des Kapitels 3.3 dieser Arbeit um und konkretisiert die Darstellungen in Bezug auf das DESI-Bewertungsschema. Diese Grundlagen vermitteln einen kognitiven Zugang zum nötigen theoretischen Wissen, in Anknüpfung an schon vorhandenes Wissen. Zugleich werden Möglichkeiten geboten, sich in die Probanden hineinzuversetzen und erste eigene Erfahrungen im Umgang mit Aufgabenstellung, Bewertungsschema und Schülertexten zu machen. In den sich anschließenden Sitzungen wird diese Basis bezüglich des Vorgehens in DESI in die Praxis umgesetzt und es werden *Rating*-Strategien und relevante Verhaltensweisen eingeübt, wobei den Studierenden ermöglicht wird, sich anfangs auf einen Aspekt und eine neu zu erlernende Strategie zu konzentrieren. Sukzessive werden die Übungen komplexer, bis schließlich die Bewertungsprozedur insgesamt zunächst an Beispieltexten demonstriert werden kann, um darauf aufbauend von den Studierenden erübt zu werden. Dabei wird darauf geachtet, Verunsicherungen frühzeitig aufzufangen und individuelles Feedback zu geben. Die Studierenden arbeiten anfangs in kleinen Gruppen, dann in Zweier-Teams, um sich auch gegenseitig Rückmeldung geben zu können. Im Verlauf der Schulung nimmt das eigenständige Arbeiten zu. Zudem werden alle zu Hause erledigten Übungen dokumentiert und von der Seminarleiterin kommentiert, um die dabei ablaufenden Prozesse zu kontrollieren und ggf. zu steuern.

Welche Bedeutung hat der **GER** bei solch einer Schulung? Welche Aspekte sind hilfreich, um Bewerter auf ein gemeinsames Verständnis des Bewertungsschemas und auf eine vergleichbare Anwendung desselbigen zu schulen? Welchen Beitrag kann das *Manual*, insbesondere die Ausführungen zur Familiarisierung, dazu leisten?

Der GER behauptet (GER 2001: 83), er könne die Basis für die Herausbildung eines gemeinsamen Verständnisses beispielsweise in einer *Rater*-Schulung bieten:

> […]erste Schritte in Richtung auf eine Verminderung der Subjektivität auf allen Stufen eines Beurteilungsverfahrens [bestehen] darin, ein gemeinsames Verständnis vom betreffenden Konstrukt herzustellen, d. h. einen gemeinsamen Bezugsrahmen. Der *Referenzrahmen* versucht, eine solche Basis für die *Beschreibung der Inhalte* und einen Fundus für die Entwicklung *genau definierter, spezifischer Kriterien* für direkte Tests zur Verfügung zu stellen.

Wie oben ersichtlich, spielt der GER bei der Schulung insofern eine Rolle, als er zunächst als Referenzmittel vorgestellt wird. Die Studierenden werden vertraut gemacht mit dem Kategorien- und Niveausystem, erarbeiten die Bedeutung der Niveaus durch Analysen und Sortieraufgaben und nutzen das Raster auf S. 36 des GER zur Selbstbeurteilung. Dennoch kann der GER auch bei einer *Rater*-Schulung nur einen generellen Rahmen bieten, da er nicht auf spezifische Testkontexte oder Probandengruppen ausgelegt ist, zumal nicht auf Jugendliche und deren schulische Situation. In dem Umfang, in dem der GER bei der theoretischen Verankerung und praktischen Entwicklung der Tasks und des Bewertungsschemas genutzt wurde, wurde er auch in der

Schulung eingeführt. Das Niveausystem bildet den größeren Rahmen, in dem sich auch die Niveaus des DESI-Projekts bewegen. Doch eine eindeutige Zuordnung[274] der DESI-*Rating*-Niveaus auf die Niveaus des GER ist, wie gerade gezeigt, aufgrund der nicht ausreichenden Berührungspunkte im Testkonstrukt und den Analysen der Lernertexte nicht möglich, so dass im Verlauf der Schulung die DESI *rating scales* in den Mittelpunkt gerückt werden.

Bezüglich der Auswahl und Bestimmung von *Benchmark*-Texten kann der GER wie gesagt keinen Beitrag leisten, da es bisher keine solchen Texte für das Referenzsystems gibt. Wenn es sie für die Niveaus und Kategorien des GER gäbe, könnten sie in einer Schulung genutzt werden, um die Bedeutung der GER-Niveaus greifbarer zu machen und einen Vergleich zwischen den Referenz-*Benchmarks* und den taskspezifischen *Benchmark*-Texten ziehen zu können. Dies würde auch eine eindeutigere Verortung im Referenzsystem ermöglichen, da die Niveaus durch solche Texte konkretisiert werden. Inwieweit die DESI-*Benchmarks* hierzu einen Beitrag leisten können, muss die Fachdiskussion zeigen.

Die Ausführungen in den Abschnitten 3 und 5 des *Manuals* zur Familiarisierung mit dem GER und zur Standardisierung (vgl. auch Kapitel 3.5 dieser Arbeit) bieten wie erwähnt hilfreiche Tipps, von denen sich einige auch im DESI-Training finden, doch die Zeitangaben des Manual müssten aufgrund der in den Seminaren gemachten Erfahrungen verdoppelt werden. Da das *Manual* noch nicht vorlag, als die DESI-Schulung konzipiert wurde, könnte man die Tipps im *Manual* und das Vorgehen im DESI-Training als sich gegenseitig stützend betrachten – eine mögliche Form der Validierung der Schulung.

4.5.2 Die Auswertung der Hauptuntersuchung

Die Stichprobengröße in der Hauptuntersuchung umfasste etwa 11000 Schülerinnen und Schüler; letztlich lagen ungefähr 20000 Aufsätze zur Bewertung vor. Durch die erwähnte Doppel-Blind-Korrektur waren ca. 40000 *ratings* zu bewältigen. Nach der Schulung standen bis zu 25 *raters* zur Verfügung, die im Zeitraum von Juni 2004 bis März 2005 die Lernertexte bewerteten. Die Texte lagen gescannt in elektronischer Form vor, so dass sie maschinell verteilt werden konnten. Zur Unterstützung der Auswertung standen drei Hilfskräfte zur Verfügung.

Die Verteilung der Aufsätze auf die *raters* und der Datenrücklauf wurden unter Kontrolle der folgenden Parameter durchgeführt:

- Um die genannten *Halo*-Effekte zu kontrollieren (die sich etwa dadurch ergeben können, dass man viele gute Texte liest und in der Bewertung der nachfolgenden Texte dadurch beeinflusst wird), werden die Aufsätze in Form von so genannten Paketen rotiert. Diese

[274] Vgl. hierzu auch die Ausführungen unter Kapitel 4.7.1 zu den im *Manual* beschriebenen Anbindungsprozeduren in Bezug auf die DESI-Schreibtests.

Pakete enthalten in der Regel 30 Aufsätze, die sich aus allen Schulformen und unterschiedlichen Klassen zusammensetzen. Die Pakete werden computergestützt[275] geschnürt.

- Die *raters* werden in Gruppen von je 5 bis 7 *raters* kombiniert. Innerhalb dieser Gruppen werden die Pakete auf Basis eines Rotationsplans je zwei *raters* zugeteilt. Die Gruppeneinteilung wird im Lauf der Auswertung mehrmals geändert und ist den *raters* nicht bekannt, ebenso wenig wie die *Rater*-Paarungen.

- Es wird kontrolliert, dass alle *raters* mindestens 200 Texte zu einer Aufgabenstellung bewerten, um stabile Werte für die Skalierung zu erhalten. Die *raters* werden in allen Aufgabenstellungen eingesetzt, um die Bewertung auch über die Tasks hinweg stabil und vergleichbar zu halten. Dabei wird darauf geachtet, dass nicht beide Texte eines Probanden durch dieselben *raters* bewertet werden.

- Die Sicherheit der Verteilung der Aufsätze an die *raters* wird dadurch garantiert, dass die Pakete auf einen passwortgeschützten *Download*-Bereich[276] des Servers der Universität Augsburg gestellt werden und dort von den *raters* heruntergeladen werden können. Die *raters* haben dabei lediglich Zugang zu ihren eigenen Paketen und können nicht sehen, welche Pakete von welchen *raters* bearbeitet werden.[277]

- Der Datenrücklauf in Form von *Excel-Score-Sheets* erfolgt per E-Mail. Es gibt standardisierte Rücklaufverfahren, um folgende Kriterien fortlaufend zu kontrollieren:

  • Korrekte Schüler-Identifizierungsnummer.

  • Valider Wertebereich der *scores*, um Tippfehler zu bereinigen.

  • Zusammenhänge zwischen den Erst- und Zweit*raters* (Inter-*Rater*-Reliabilitäten) und Höhe der Abweichungen: Bei Abweichungen von über 2 Punktwerten oder unterschiedlicher Kodierung hinsichtlich der Validität der Aufsätze werden diese Bewertungen von beiden *raters* ohne Absprache überprüft. Bei wiederholten Abweichungen und niedrigen Zusammenhängen werden die betroffenen *raters* nachgeschult.

  • Zusammenhänge zwischen den Erst- und Zweit-*Ratings*: Bei zu niedrigen Zusammenhängen finden begleitende Schulungssitzungen statt. Zudem gibt es ein E-mail Forum, in dem offene Fragen und Probleme diskutiert werden, nicht aber Absprachen zur laufenden Bewertung. Fragen dazu können direkt mit der Schulungsleiterin besprochen werden.

  • Konsistenz über die gesamte Zeit der Auswertung: Dazu finden begleitende Schulungen statt. Zusätzlich wird auf der Basis einer Zwischenskalierung individuelles Feedback zu den Strenge-/Milde-Tendenzen und der individuellen Angepasstheit an das der

[275] Die entsprechenden Programme sind von Stefan Langer entwickelt worden, vgl. auch www.stefanlanger.de.
[276] In diesem Zusammenhang untersützte uns Frau Kötterle aus dem Rechenzentrum der Universität Augsburg.
[277] Die *raters* müssen eine Verschwiegenheitserklärung und eine Versicherung unterschreiben, dass ihnen bekannt ist, dass Absprachen unzulässig sind und mit Vertragsstrafen versehen werden.

Skalierung zugrunde gelegte *Rater*-Modell[278] gegeben. Gegen Ende der Auswertung werden Aufsätze aus der Anfangszeit der Auswertung noch einmal bewertet, um die innere Konsistenz der *raters* zu kontrollieren (Intra-*Rater*-Reliabilität).

Folgende Tabellen exemplifizieren die kontrollierten Gütekriterien über die gesamte Auswertung hinweg, wiederum am Beispiel des *Biography*-Tasks. Die Reliabilitäten werden durch die Produkt-Moment-Korrelation nach Pearson auf Basis aller *ratings* (jedoch differenziert nach Aufgabenstellung und Bewertungskriterien) bestimmt:

Inter-*Rater*-Reliabilitäten (IRR):

Dabei werden die Zusammenhänge zwischen den Bewertungen der von zwei *raters* gemeinsam gelesenen Aufsätze betrachtet, um Abweichungen kontrollieren zu können, die durch individuelles *Rater*-Verhalten bedingt sind. Diese Korrelationen wurden fortlaufend geprüft. Die folgende Tabelle zeigt die Zusammenhänge am Ende der Auswertung:

Task	***Rater*-Paarung** (exemplarischer Auszug)	**Anzahl** gemeinsam gelesener **Aufsätze**	***Range* der IRR** (niedrigster und höchster Wert/Bewertungskriterium, Korrelation nach Pearson)
ST02	4 / 13	56	0,779 (Lexik) bis 0,861 (Inhalt)
	5 / 6	65	0,774 (Aufbau) bis 0,897 (Wirkung)
	5 / 8	152	0,697 (Lexik) bis 0,802 (Orthographie)
	6 / 12	117	0,741 (Kohäsion) bis 0,823 (Global)
	12 / 22	73	0,708 (Wirkung) bis 0,839 (Grammatik)
	16 / 23	201	0,712 (Inhalt) bis 0,790 (Global)

Tabelle 9: Inter-*Rater*-Reliabilitäten, Modul semikreatives Schreiben im DESI-Projekt

Die Reliabilitäten liegen im akzeptablen Bereich. Die hier vorgestellten Werte stehen stellvertretend für die weiteren *Rater*-Kombinationen und die anderen Tasks, bei denen sich die Korrelationen in ähnlichen Größenverhältnissen bewegen. Das Kriterium der Länge korreliert überall im Bereich über 0,9, da es sich wie gesagt um ein auszählbares und damit objektives Kriterium handelt.

Zusammenhänge zwischen Erst- und Zweit-*Ratings*:

Zunächst werden die Korrelationen zwischen Erst- und Zweit-*Ratings* zu jedem Bewertungskriterium betrachtet, unabhängig von der jeweiligen *Rater*-Paarung: Die Variablennamen sind wie folgt zu lesen: E_ST0201 bedeutet Erst-*Rating* des semikreativen Tasks 02 im Kriterium 01 (Biographie Globalurteil, vgl. Handbuch).

[278] Dieses Modell bestimmt auf mathematischer Basis die Schwellenübergänge zwischen den einzelnen Bewertungsstufen, so dass das individuelle *Rater*-Verhalten mit dem Modell verglichen werden kann. Die Zwischenskalierung wurde wiederum an der Humboldt-Universität zu Berlin durchgeführt.

	Z_ST0201	Z_ST0202	Z_ST0203	Z_ST0204	Z_ST0205	Z_ST0206	Z_ST0207	Z_ST0208	Z_ST0209	Z_ST0210
E_ST0201										
Korrelation nach Pearson	**0,807**	0,769	0,779	0,751	0,773	0,783	0,772	0,779	0,797	-0,144
Signifikanz (2-seitig)	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
N	**4483**	4483	4483	4483	4483	4483	4483	4483	4483	4538
E_ST0202										
Korrelation nach Pearson	0,774	**0,961**	0,774	0,716	0,726	0,728	0,701	0,733	0,759	-0,082
Signifikanz (2-seitig)	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
N	4483	**4483**	4483	4483	4483	4483	4483	4483	4483	4538
E_ST0203										
Korrelation nach Pearson	0,781	0,772	**0,769**	0,735	0,739	0,749	0,733	0,747	0,774	-0,152
Signifikanz (2-seitig)	0,000	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000	0,000
N	4483	4483	**4483**	4483	4483	4483	4483	4483	4483	4538
E_ST0204										
Korrelation nach Pearson	0,753	0,709	0,724	**0,742**	0,716	0,725	0,711	0,723	0,741	-0,147
Signifikanz (2-seitig)	0,000	0,000	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
N	4483	4483	4483	**4483**	4483	4483	4483	4483	4483	4538
E_ST0205										
Korrelation nach Pearson	0,783	0,730	0,744	0,724	**0,776**	0,777	0,763	0,763	0,776	-0,120
Signifikanz (2-seitig)	0,000	0,000	0,000	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000
N	4483	4483	4483	4483	**4483**	4483	4483	4483	4483	4538
E_ST0206										
Korrelation nach Pearson	0,791	0,730	0,749	0,726	0,771	**0,785**	0,774	0,769	0,780	-0,112
Signifikanz (2-seitig)	0,000	0,000	0,000	0,000	0,000	**0,000**	0,000	0,000	0,000	0,000
N	4483	4483	4483	4483	4483	**4483**	4483	4483	4483	4538
E_ST0207										
Korrelation nach Pearson	0,772	0,702	0,728	0,712	0,760	0,773	**0,772**	0,759	0,760	-0,109
Signifikanz (2-seitig)	0,000	0,000	0,000	0,000	0,000	0,000	**0,000**	0,000	0,000	0,000
N	4483	4483	4483	4483	4483	4483	**4483**	4483	4483	4538
E_ST0208										
Korrelation nach Pearson	0,786	0,746	0,750	0,738	0,763	0,777	0,768	**0,774**	0,776	-0,112
Signifikanz (2-seitig)	0,000	0,000	0,000	0,000	0,000	0,000	0,000	**0,000**	0,000	0,000
N	4483	4483	4483	4483	4483	4483	4483	**4483**	4483	4538
E_ST0209										
Korrelation nach Pearson	0,796	0,757	0,769	0,739	0,765	0,774	0,761	0,768	**0,789**	-0,148
Signifikanz (2-seitig)	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	**0,000**	0,000
N	4483	4483	4483	4483	4483	4483	4483	4483	**4483**	4538
E_ST0210										
Korrelation nach Pearson	-0,155	-0,090	-0,171	-0,148	-0,122	-0,124	-0,113	-0,122	-0,159	**0,913**
Signifikanz (2-seitig)	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	**0,000**
N	4520	4520	4520	4520	4520	4520	4520	4520	4520	**4848**

Tabelle 10: Korrelationen Erst-/Zweit-*Ratings*, Modul semikreatives Schreiben im DESI-Projekt

Die Zusammenhänge zwischen Erst- und Zweit-*Ratings* in den jeweiligen Bewertungskriterien sind hochsignifikant und hinreichend hoch, um die Auswertung als reliabel zu betrachten. Die hohen Korrelationen im Bereich der Länge (ST0202) und der *Swear Words* (ST0210) sind dadurch zu erklären, dass es sich dabei nicht um *ratings*, sondern um ein quantitatives respektive um ein dichotom zu kodierendes Kriterium handelt.

Des Weiteren werden die Abweichungen hinsichtlich der vergebenen *scores* in den einzelnen Bewertungskriterien (bei den validen Texten) betrachtet, um die Spannweite der Abweichungen zu kontrollieren. Folgende Übersicht zeigt die prozentualen Unterschiede in Bezug auf die vergebenen *scores*:

ST02 Kriterien	Unterschiede (in %)			
	0	1	2	>2
01 Global	56,1	42,5	1,4	0,0
03 Inhalt	51,2	44,6	4,0	0,2
04 Sorte/Aufbau	51,7	42,3	5,0	1,0
05 Orthographie	55,9	41,2	2,8	0,1
06 Lexik	55,8	41,5	2,7	0,0
07 Grammatik	55,0	42,2	2,7	0,1
08 Kohäsion	53,5	43,5	2,9	0,1
09 Komm. Wirkung	52,5	44,3	3,1	0,1

Tabelle 11: Unterschiede hinsichtlich der vergebenen *scores*

Bei allen Kriterien zeigt sich, dass in über 50% aller Fälle dieselben *scores* vergeben wurden; die *ratings* unterscheiden sich zu etwas mehr als 40% um einen *score*. Die Abweichungen von 2 und mehr *scores* liegen im Allgemeinen unter 4%, nur bei den Kriterien Inhalt und Textsorte/Aufbau liegen sie etwas höher; dies liegt jedoch in der gegebenen Subjektivität der Einschätzungen dieser Kriterien, da wie im Schulungsskript erläutert (vgl. Anhang 29) die Mikro- wie Makrostruktur eines Textes von den *raters* konstruiert werden muss und sich die Subjektivität bei der Bewertung des Inhalts am deutlichsten zeigt. Die Werte sprechen dafür, dass die Kriterien zuverlässig eingeschätzt wurden.

Intra-*Rater*-Reliabilitäten (Konsistenz):

Dabei werden die Zusammenhänge zwischen zwei Bewertungen derselben Aufsätze durch dieselben *raters* im Abstand von etwa 5 Monaten betrachtet, um die interne Konsistenz der *raters* zu kontrollieren. Die folgende Tabelle enthält exemplarisch die Daten der *raters*, die in Tabelle 9 bereits genannt wurden. Insgesamt haben 24 *raters* an der Konsistenzprüfung teilgenommen.

Rater-ID	Task	Anzahl Aufsätze	01 Global	02 Länge	03 Inhalt	04 Textsorte	05 Orthographie	06 Lexik	07 Grammatik	08 Kohäsion	09 Wirksamkeit	10 Swear
r05	ST04	23	0,943	0,980	0,872	0,822	0,814	0,784	0,757	0,694	0,913	1,000
r06	ST01	20	0,945	0,999	0,942	0,905	0,926	1,000	0,954	0,855	0,925	1,000
r08	ST03	23	1,000	0,987	1,000	1,000	0,983	0,981	0,980	1,000	0,977	1,000
r12	ST02	20	0,908	0,942	0,898	0,891	0,943	0,828	0,962	0,881	0,893	1,000
r13	ST03	23	0,968	0,994	0,958	0,942	0,919	0,951	0,928	0,907	0,843	1,000
r16	ST02	20	0,896	0,989	0,895	0,812	0,863	0,827	0,897	0,835	0,864	1,000
r22	ST01	20	0,688	0,925	0,847	0,662	0,769	0,842	0,820	0,729	0,838	1,000
r23	ST02	20	0,909	0,997	0,674	0,680	0,818	0,873	0,890	0,770	0,859	1,000

Tabelle 12: Intra-*Rater*-Reliabilitäten, Modul semikreatives Schreiben im DESI-Projekt

Die Konsistenz der *raters* ist gewährleistet, da die Korrelationen in der Regel im Bereich von über 0,8 liegen. Die etwas niedrigeren Werte des *rater* 22 bei Globalurteil und Textsorte/Aufbau sind dadurch zu erklären, dass dieser *rater* die Konsistenzprüfung einige Monate nach dem vorzeitigen Ausscheiden abgelegt hat. Die teils hohen Werte der *raters* 6 und 8 sind ein Hinweis auf die hohe Konsistenz dieser *raters* und nicht etwa dadurch entstanden, dass die ursprünglichen Bewertungen bei der Konsistenzprüfung „übernommen" wurden.[279]

4.6 Rückmeldung

In der letzten Phase des Testlaufs werden die Auswertungsergebnisse in dem Rahmen generalisiert, den Testkonzept, Konstrukt und Bewertungsschema erlauben. Oben wurden die wesentlichen Komponenten der Beurteilung der Schreibfertigkeit beschrieben, oder um es mit den Worten von Shale (1996) zu sagen, die *universes of tasks, writing samples and raters* definiert: Die DESI-Tasks sind in ihren Anforderungen charakterisiert und auf den Horizont der Probanden ausgelegt. Die möglichen Reaktionen darauf (die Lernertexte) bilden die Basis des Bewertungssystems, das durch aufgabenspezifische Kriterien und Abstufungen detailliert beschrieben ist. Die *raters* bilden eine interpretative Gemeinschaft, die die Kriterien und Abstufungen des Bewertungssystems in dem Sinn versteht und interpretiert, in dem sie vom Testkonstrukt intendiert sind; diese *Rater*-Gemeinschaft interpretiert und bewertet die Lernertexte in vergleichbarer Weise auf Basis der *rating scales*. In diesem Rahmen können nun Generalisierungen zur Rückmeldung der Testergebnisse an die Beteiligten abgeleitet werden.

Diese theoretischen Überlegungen werden im DESI-Projekt durch psychometrische Modelle abgesichert. Das Testmodul *Schreiben Englisch* wurde wie erwähnt an der Humboldt-Universität zu Berlin skaliert. Dabei wurden Aufgabenschwierigkeiten und *Rater*-Tendenzen berücksichtigt. Bei der Skalierung wurde u. a. geprüft, wie sich die Bewertungskriterien der vier Tasks zueinander verhalten. Dadurch kann bestimmt werden, welche Kriterien sich mit welcher Gewichtung auf welchen (statistischen) Dimensionen ausmachen lassen und durch wie viele (Sub-)Skalen sich die Schreibfertigkeit in DESI demnach darstellen lässt. Beispielsweise wird bei solchen Skalierungen untersucht, ob sich sprachliche von inhaltlich-formalen Faktoren unterscheiden lassen und ob sich solche Unterschiede taskspezifisch oder taskunabhängig verhalten. Die Skalierungen der Hauptuntersuchung[280] ergeben einen Globalfaktor Schreiben über alle Aufgabenstellungen und Bewertungskriterien hinweg. Regressions- und Faktorenanalysen bestätigen den engen statistischen Zusammenhang der Kriterien und lassen Rückmeldung in Form einer Globalskala „Schreiben" zu.

[279] Dies wurde u. a. dadurch sichergestellt, dass *Score-Sheets* und Daten in regelmäßigen Abständen von den *raters* vernichtet werden mussten.
[280] Details der Skalierungen und Ergebnisse können an dieser Stelle nicht veröffentlicht werden, da sie Bestandteil des DESI-Abschlussberichts für die KMK sind. Dieser Bericht wird voraussichtlich im Februar 2006 erscheinen.

Aufgrund dieser Ergebnisse wird die Globalskala oder Kompetenzskala, wie sie in DESI genannt wird, aus der Synopse der vier taskspezifischen *rating scales* (Globalurteil) abgeleitet. Zur detaillierteren Betrachtung der Schülerleistungen können zusätzlich inhaltlich begründete Subskalen zu den taskspezifischen Kriterien einerseits und zu den sprachlichen Kriterien andererseits aus den betreffenden *rating scales* abgeleitet werden. Die KANN-Deskriptoren dieser Kompetenzskala beschreiben nun nicht mehr primär Merkmale der Lernertexte, sondern sie stellen Generalisierungen des im DESI-Schreibtest gezeigten Schreibvermögens dar. Dabei wird auf das Wissen und Können rückgeschlossen, das in den Tests auch gefordert war und das im realen Leben vermutlich zur Verfügung steht. So wird beispielsweise nur auf die in DESI geforderten Textsorten (persönlicher) Brief und Bericht generalisiert, ebenso wie sich die Beschreibung der versprachlichten Ideen und Themen auf die in den DESI-Tasks geforderten bezieht.

Das Vorgehen wird wiederum am Beispiel der Überführung der vier taskbezogenen Globalurteile in die Kompetenzskala „Schreiben“ demonstriert: Die folgende Übersicht zeigt jeweils auf drei Niveaus (exemplarisch die DESI-Niveaus 1, 3 und 5) zunächst die Synopse der vier *rating scales* (Globalurteile zu den vier Tasks), wobei die adaptierten Formulierungen der GER-Niveaus A1, B1 respektive B2+ darin zur besseren Kenntlichmachung fett gedruckt sind. Darauf aufbauend werden die KANN-Deskriptoren vorgestellt, die aus der Synopse formuliert wurden:

Synopse der Merkmale der DESI-Bewertungsstufe 1 über alle Tasks hinweg / Adaptionen der GER-Formulierungen des Niveaus A1 fettgedruckt:

- **Kurzer, einfachster** Text; Mängel im Formalen; meist keine Makrostruktur erkennbar (assoziative Reihung).
- Ideen ansatzweise relevant, aber nicht entwickelt; Darstellung bleibt konkret; **einfache Wendungen und Sätze über sich selbst und andere (auch fiktive) Menschen**: wie sie zusammengehören, **wo sie leben**, wie sie leben, **was sie tun** oder was sie tun wollen, sollen oder können.
- Zeigt begrenzte sprachliche Mittel aus frequentem **Basisbereich**; Text kann bruchstückhaft und fehlerhaft sein (lexikalische, grammatische, syntaktische, orthographische Fehler; muttersprachliche Interferenzen).
- Gewünschte Botschaft wird nur ansatzweise und oft missverständlich vermittelt.

KANN-Formulierung DESI-Kompetenzniveau 1:

Kann kurze, einfachste, persönliche Briefe/Berichte schreiben, die Mängel im Formalen zeigen. Kann dabei die Gedanken assoziativ reihen.
Kann einfache Wendungen und Sätze über sich selbst und andere (auch fiktive) Menschen schreiben: wie sie zusammengehören, wo und wie sie leben, was sie tun oder was sie tun wollen, sollen oder können.
Zeigt begrenzte sprachliche Mittel aus dem hochfrequenten Basisbereich, wobei die Texte meist bruchstückhaft und fehlerhaft sind (lexikalische, grammatische, syntaktische, orthographische Fehler, starke muttersprachliche Interferenz), weshalb die gewünschte Botschaft nur ansatzweise und oft missverständlich vermittelt werden kann.

Synopse der Merkmale der DESI-Bewertungsstufe 3 über alle Tasks hinweg / Adaptionen der GER-Formulierungen des Niveaus B1 fettgedruckt:

- **unkomplizierter** Brief/Bericht in **üblichem Standardformat**; teils noch nicht alle Formalia erfüllt; logische Makrostruktur erkennbar, es kann aber zu Sprüngen kommen.
- **Unkomplizierter, zusammenhängender** Text über **reale wie fiktive Ereignisse**, Abenteuer und **Erfahrungen** und über **einfache Sachverhalte** im Rahmen eines Londonaufenthaltes/einer Klassenfahrt oder über zwischenmenschliche Konstellationen und Probleme, Erfahrungen und Strategien aus Lebensumwelt der Verfasser oder über **reale wie fiktive Personen und deren Umfeld und Erfahrungen**; dabei

werden einfache, doch begründete Ratschläge oder Lösungen angeboten. **Interessenschwerpunkte können sichtbar werden**. Zwischenmenschliche Beziehungen, **Gefühle und Reaktionen** darauf werden in **unkomplizierter** Weise beschreiben.
- Sprachliche Mittel werden bis zu einem gewissen Grad angemessen und in **hinreichendem Umfang** eingesetzt, um das Ziel zu erreichen (gute Beherrschung des Grundwortschatzes und der gängigen grammatischen Strukturen); gewisse narrative Grundqualitäten (wie etwa Textkohärenz) sind gegeben. Fehler und Interferenzphänomene können im Text vorhanden sein, in der Regel jedoch nicht kommunikationsbelastend.
- Botschaft wird grundsätzlich kommunikativ wirksam vermittelt, doch teils mit Einschränkungen.

KANN-Formulierung DESI-Kompetenzniveau 3:

Kann in einem üblichen Standardformat unkomplizierte Briefe/Berichte schreiben, die eine logische Struktur aufweisen.
Kann unkomplizierte, zusammenhängende Texte zu einer Reihe verschiedener Themen aus seiner/ ihrer Lebensumwelt, über reale wie fiktive Personen und Ereignisse, zwischenmenschliche Konstellationen und Probleme, persönliche Erfahrungen und einfache Sachverhalte schreiben. Interessenschwerpunkte können sichtbar werden. Kann zwischenmenschliche Beziehungen, Gefühle und Reaktionen darauf in unkomplizierter Weise beschreiben. Kann einfache, aber begründete Ratschläge geben und Lösungen zu alltäglichen Problemstellungen anbieten.
Kann sprachliche Mittel bis zu einem gewissen Grad angemessen und in hinreichendem Umfang einsetzten, sofern sie sich auf vorhersehbare Situationen beziehen. Zeigt eine gute Beherrschung des Grundwortschatzes und der gängigen grammatischen Strukturen. Gewisse narrative Grundqualitäten (wie etwa Textkohärenz) sind gegeben. Fehler können im Text vorhanden sein, sie schränken das Textverständnis jedoch nur gelegentlich ein.
Kann die Botschaft grundsätzlich kommunikativ wirksam vermitteln, doch teils mit Einschränkungen.

Synopse der Merkmale der DESI-Bewertungsstufe 5 über alle Tasks hinweg / Adaptionen der GER-Formulierungen der Niveaus B2+ / C1 fettgedruckt:

- Brief/Bericht ist entsprechend der **geltenden Konventionen** geschrieben; Aufbau logisch, stringent und konsistent, u. U. komplex entwickelt; die Makrostruktur erleichtert das Verständnis.
- **Klarer, detaillierter, gut strukturierter und ausführlicher Text** zu einem Londonaufenthalt/einer Klassenfahrt *oder* zu komplexen zwischenmenschlichen Problemstellungen, deren Ursachen und möglichen Auswegen *oder* Entwicklung einer umfassenden Biographie: Es sind **entscheidende Punkte hervorgehoben, Standpunkte und Ansichten ausführlich dargestellt und durch Unterpunkte oder geeignete Beispiele oder Begründungen gestützt; Text ist durch angemessenen Schluss abrundet.** Einstellungen und Gefühle sind versprachlicht; verschiedene Perspektiven werden ggf. deutlich gemacht
- **Großer Umfang und sichere Beherrschung** sprachlicher Mittel, Idiomatik und Korrektheit sind meist gegeben. Text ist **flüssig, in lesergerechtem, überzeugendem, persönlichem und natürlichem Stil** verfasst und besitzt durchgehend narrative (oder empathische) Qualität.
- Die kommunikative Wirkung wird umfassend erzielt.

KANN-Formulierung DESI-Kompetenzniveau 5:

Kann Briefe/Berichte entsprechend der geltenden Konventionen schreiben, wobei der logische, stringente und konsistente Aufbau das Verständnis erleichtert.
Kann klare, detaillierte, gut strukturierte und ausführliche Texte zu komplexen Themen verfassen wie beispielsweise eine Biographie schreiben, zwischenmenschliche Problemstellungen und deren Ursachen und Auswege erläutern oder Erlebnisse bei einem Auslandsaufenthalt beschreiben. Kann die eigene Ansicht darstellen. Kann dabei die entscheidenden Punkte hervorheben, Standpunkte ausführlich darstellen und durch Unterpunkte oder geeignete Beispiele oder Begründungen stützen und den Text durch einen angemessenen Schluss abrunden. Kann Einstellungen und Gefühle adäquat versprachlichen. Kann unterschiedliche Standpunkte einnehmen und verschiedene Perspektiven deutlich machen.
Verfügt über einen großen Umfang sprachlicher Mittel und kann sie angemessen und variiert verwenden; verfügt über Kollokationen und idiomatische Wendungen; Fehler sind selten, können rückblickend korrigiert werden und sind nicht kommunikationsbelastend. Kann flüssige Texte in lesergerechtem, überzeugendem, persönlichem und natürlichem Stil verfassen, welche narrative, emphatische, humoristische oder spannende Qualität aufweisen.
Kann die angestrebte kommunikative Wirkung umfassend erzielen.

Die DESI-Kompetenzskalen beschreiben in generalisierender Form die Schreibfertigkeit auf der Grundlage des DESI-Testkonstrukts, der Aufgabenstellungen und der dadurch elizitierten Lernertexte. Die dabei verwendeten Deskriptoren enthalten – wo angemessen – adaptierte Formulierungen der entsprechenden GER-Niveaus, ansonsten beruhen sie auf den DESI *rating scales*. Deswegen erheben die DESI-Kompetenzskalen nicht den Anspruch, an die betreffenden Niveaus des GER angebunden zu sein. Vielmehr zeigen sich einige Berührungspunkte mit entsprechenden GER-Skalen, die zur Validierung der DESI-Kompetenzskalen als Außenkriterium genutzt werden können. Doch aufgrund der oben gezeigten Schwierigkeiten bezüglich der Verortung von Textkonstrukt, Zielgruppe, Aufgabenstellungen und Bewertungsschema im GER bedarf es einer eigenen Studie, um das DESI-Testmodul *Schreiben Englisch* fundiert und empirisch abgesichert an die Niveaus des GER anzubinden. Diese Studie ist im Rahmen des DESI-Projekts nicht vorgesehen. Im folgenden Ausblick wird deshalb unter anderem ein Kurzüberblick über das zur Anbindung nötige und mögliche Vorgehen in Anlehnung an das *Manual* gegeben.

4.7 Ausblick

Wie in Kapitel 2.4 dieser Arbeit dargestellt, kommt dem Aspekt der systemischen Validität nach Camp (1996) gerade im Rahmen von Schulleistungsstudien eine nicht zu vernachlässigende Bedeutung zu: Solche Studien müssen, um den damit verbunden Aufwand zu rechtfertigen, das Lehren und Lernen derjenigen Fertigkeiten fördern, auf die die Beurteilungen abzielen. Auch der UGE erkennt die Bedeutung des *educational impacts* eines gegebenen Tests an und schlägt zur Evaluation desselbigen folgenden Fragenkatalog vor (Council of Europe 2002[2]: 36f):

- who is taking the test (i.e. profile of the candidates);
- who is using the test results and for what purpose;
- who is teaching towards the test and under what circumstances;
- what kinds of courses and materials are being designed and used to prepare candidates;
- what effect the test has on public perceptions generally (e.g. regarding educational standards generally);
- how the test is viewed by those directly involved in educational processes (e.g. by students, test-takers, teachers, parents, etc.);
- how the test is viewed by members of society outside education (e.g. politicians, businessmen, etc.).

Die beiden ersten Fragen sind bezogen auf die DESI-Studie klar zu beantworten: Die Schülerprofile sind bekannt; die Testergebnisse werden von Lehrern, Schulleitern und Bildungsqualitätsentwicklern genutzt, um den schulischen Unterricht zu verbessern. Die obigen Spiegelstriche 3 und 4 sind für DESI uninteressant, da es sich um einen einmaligen Testlauf handelt. Die Frage nach der Wahrnehmung des Tests im Bildungssektor wird in DESI beispielsweise dadurch erfasst, dass Mitglieder der *Fachgruppen Bildungsstandards Deutsch und erste Fremdsprache der KMK* zum Verhältnis der DESI-Testaufgaben zu den genannten Bildungsstandards befragt wurden. Die beiden letzten Fragen der obigen Aufzählung müssten in Bezug auf die DESI-Tests noch untersucht werden: Wenngleich in den testbegleitenden Fragebögen die „Beliebtheit“ der

Tests durch die Probanden erfragt wurde und die Lehrkräfte zum Bekanntheitsgrad der Aufgaben befragt wurden, hat eine Beurteilung der DESI-Tests im o. g. Sinn dennoch nicht stattgefunden. Eine Befragung der Öffentlichkeit außerhalb des Bildungssektors steht ebenfalls aus, könnte sich aber im Zuge der Veröffentlichung der DESI-Ergebnisse anschließen.

Um systemische Validität in einem gegebenen Bildungssystem zu erzielen, sollte die getestete Institution einen Erkenntnisgewinn aus der Leistungsstudie ziehen können, indem etwa quantitative Daten in qualitative Beschreibungen überführt werden, die den Beteiligten detaillierte Rückmeldung über Testanforderungen und Leistungsstand im Sinne bereits erreichter (Teil-)-Kompetenzen geben – wie es u. a. auch im oben dargestellten Testmodul *Schreiben* in DESI geschehen ist. Konkret müssen Testergebnisse den beteiligten Schulen und Klassen zugänglich gemacht werden, um an den Schulen zur Verbesserung des Lehrens und Lernens umgesetzt werden zu können.[281] Dadurch wird die individuelle Förderung aller Lernenden ermöglicht, wie es beispielsweise Torrance (1998) fordert.

Im Rahmen der Beurteilung des Schreibvermögens wäre es zur individuellen Förderung neben der Rückmeldung und Beschreibung des erreichten Kompetenzniveaus wünschenswert, den Rückfluss der Lernertexte in die getesteten Institutionen und dort an die getesteten Schülerinnen und Schüler zu ermöglichen, um den *instructional value* eines Leistungstests zu gewährleisten. Stellen wir momentan den Haupteinwand gegen dieses Vorgehen zurück, namentlich die Bedenken der Datenschützer, da sich die Anonymität der Probanden durchaus gewähren ließe: Die Namenslisten, die den Probanden ihre Identifizierungsnummern für den Testlauf zuweisen, verbleiben bei den Klassenlehrkräften – mittels dieser Listen wäre der Rücklauf der Aufsätze (die nur durch die Identifizierungsnummern gekennzeichnet sind) über die Lehrkräfte an die entsprechenden Schülerinnen und Schüler durchaus möglich. Daneben muss vermutlich der Einwand, dieser Rücklauf sei nicht finanzierbar und rechne sich nicht, entkräftigt werden: Sicherlich sind damit Kosten verbunden, doch der mögliche Nutzen dürfte die Kosten aufwiegen, denn die Aufsätze könnten als Lernanreiz dienen und Anlass zur Überarbeitung bieten, etwa im Rahmen eines Portfolio-*Assessments*, auf das gleich im Anschluss eingegangen wird. Im Folgenden sind einige der Vorteile aufgezählt, die mit einem Rücklauf der Aufsätze in die getestete Institution einhergehen könnten:

- Die Lernenden erhalten ihre Texte mit externer Bewertung zurück: Dies dürfte sich schon auf die Testleistung motivierend auswirken, denn wenn den Probanden bekannt ist, dass die Testleistungen wieder in den Unterricht einfließen, dürfte dies zu verminderter Verweigerungshaltung führen und könnte als Ansporn dienen, zu zeigen „was man schon kann“. Auf diese Weise lernen die Schülerinnen und Schüler zudem externe Bewertungsschemata kennen – diese können ihrerseits zum Anlass genommen werden, Sinn und Bedeutung von

[281] Im DESI-Projekt gibt es, eine zusätzliche Studie des Instituts für Schulentwicklung Dortmund, die den Umgang der Schulen mit den Rückmeldungen untersucht. Deshalb fokussiert dieser Ausblick auf konkrete Möglichkeiten des Umgangs mit den Ergebnissen der Bewertung der Schreibfertigkeit in der Fremdsprache Englisch und mit den Lernertexten selbst.

Bewertungskriterien und Merkmale guter Texte zu diskutieren und so die Beurteilung in größere Kontexte der Evaluation einzubinden.

- Die Lehrenden erhalten einerseits Einblick in die Performanzen ihrer Lernenden bei externen Aufgabenstellungen; andererseits werden sie mit externen Bewertungsschemata vertraut gemacht und können wertvolles Feedback geben für zukünftige Beurteilungen. Dadurch könnten sie im Zukunft stärker in externe Beurteilungsprozesse eingebunden werden.

- Die Lernertexte selbst gewinnen durch dieses Vorgehen an Bedeutung und können als Lernanreiz dienen, etwa indem sie in Anlehnung an die (zyklische) Textproduktion in der realen Welt überarbeitet und redigiert werden. Solche Lernertexte bieten zahlreiche Gelegenheiten, bestimmte Aspekte – seien es solche der sprachlichen Korrektheit, der Angemessenheit des Ausdrucks, oder der wirksamen Darstellung – in den Mittelpunkt zu rücken und so die Arbeit an den Testaufsätzen gezielt in den Unterricht einzubauen. Neben gezielter Textarbeit bieten sich Projektarbeiten an, etwa die Überarbeitung der in DESI elizitierten „Schülerzeitungs"-Berichte mit dem Ziel der Veröffentlichung der besten Texte in einer realen Schülerzeitung. Die Lernertexte können aber auch in ein Portfolio einfließen und auf diese Weise in die schulische Evaluation eingebunden werden – auf diese Möglichkeit wird unter Kapitel 4.7.2 näher eingegangen.

Nicht nur Torrance (1998) und Elbow (1996) erkennen im Zusammenhang mit Schulleistungsstudien die Notwendigkeit der Entwicklung der schulischen Evaluation, um beispielsweise die externen Beurteilungen in Verbindung zu bringen mit den schulinternen Beurteilungssystemen. Um jedoch schulinterne und externe Beurteilungssysteme aufeinander beziehen zu können, bedarf es eines gemeinsamen Bezugspunkts – dieser Bezugspunkt könnte im GER gefunden werden, allerdings im Rahmen seiner in dieser Arbeit analysierten Möglichkeiten. Der GER hat bereits Einzug gehalten in Lehrpläne und Bildungsstandards für die erste Fremdsprache, und über das Sprachenportfolio auch in die Klassenzimmer. Wenn nun Lehrkräfte und Lernende mit seinem System vertraut gemacht werden, so können die Kategorien und Niveaus des GER durchaus zur Selbstbeurteilung und zur Beurteilung des allgemeinen Sprachvermögens durch die Lehrkräfte genutzt werden, da den Lehrenden ihre Lernenden hinreichend bekannt sind.

Um jedoch die Ergebnisse von Schulleistungsstudien mittels des Systems des GER zu kommunizieren und in die schulische Evaluation einfließen zu lassen, bedarf es der validen Anbindung der dabei eingesetzten Tests und Kompetenzniveaus – erst dann können die aufgrund einer externen Beurteilung erzielten Leistungstestergebnisse in Bezug gesetzt werden zu schulinternen, ebenfalls am GER ausgerichteten Beurteilungen.

4.7.1 Anbindung des DESI-Moduls *Textproduktion Englisch* an die Niveaus des GER

Wie erwähnt ist die Anbindung der DESI-Kompetenzniveaus an die Niveaus des GER im DESI-Projekt weder vorgesehen noch finanzierbar. Zudem erschien das *Manual* erst im September 2003, zu einer Zeit als sich das DESI-Projekt bereits in der Durchführungsphase befand. Deshalb soll in dieser Arbeit zumindest theoretisch auf die Möglichkeiten der Anbindung eingegangen werden, wie sie im *Manual* dargestellt sind (für eine grundlegende Besprechung des *Manual* vgl. Kapitel 3.5 dieser Arbeit).

Zunächst müssen sich Testentwickler, Bewerter und diejenigen, die die Anbindung vornehmen, mit dem Referenzrahmen und dem Bewertungssystem in DESI **familiarisieren**. Bezogen auf das DESI-Team und die *raters* ist diese Familiarisierung gründlich erfolgt; die Personen, die bei den Anbindungsprozeduren mitarbeiten, müssten sich diesem Prozess noch unterziehen.

Daran schließt sich die Phase der **Spezifizierung** an. Die DESI-Tests sind, wie oben dokumentiert, in ihren charakteristischen Anforderungsmerkmalen beschrieben. Allerdings kann der GER, wie ebenfalls oben erörtert, dabei nicht Ausgangs- oder Bezugspunkt sein, da Testkonstrukt und relevante Charakteristika der Aufgaben weder im GER verortet noch aus seinem System abgeleitet werden können. Dazu müsste, ähnlich wie im *Dutch Grid Project*, zunächst ein solcher *grid* für die schriftlichen Aktivitäten und Aufgaben erstellt werden, mithilfe dessen Testaufgaben spezifiziert werden könnten im Hinblick auf die Beschreibungskategorien des GER. Ob sich der erwähnte *ALTE Grid*[282] hierfür eignet, muss die Praxis zeigen, sobald er denn in endgültiger Form vorliegt. Momentan sind offene Schreibaufgaben alleine auf Basis des GER-Systems nicht hinreichend zu spezifizieren.

Auch in der darauf folgenden Phase der **Standardisierung** kann der GER nicht als Bezugssystem genutzt werden: Gemäß *Manual* (2003: 71f und 76ff) sollen dazu Performanzen mithilfe entsprechender GER-Skalen auf die Niveaus des GER eingestuft werden und Minimalstandards für das Erreichen eines Niveaus festgelegt werden. Doch wie oben gezeigt können einzelne Performanzbeispiele nicht auf Basis der GER-Skalen alleine eingestuft werden, da es für den Bereich *Schreiben* keine Deskriptoren gibt, die auf konkrete Aufgaben bezogene Anforderungen oder Merkmale beschreiben und sich deshalb für das *rating* von Performanzen eignen würden (diese Aussage trifft auch auf das Beurteilungsraster auf S. 82 des *Manual* zu); zudem liegen bisher keine *benchmarks* vor, die gemäß *Manual* jedoch in dieser Phase genutzt werden sollten (vgl. ebd.: 71f). Deshalb werden im DESI-Projekt die Bewertungsstandards und *benchmarks* bestimmt aus dem konkreten Testkonstrukt, dem daraus abgeleiteten Kompetenzmodell und den der Bewertung zugrunde liegenden Lernertextanalysen. Die im *Manual* geforderte Schulung zur Bestimmung der Standards hat im DESI-Projekt im oben dokumentierten *Rater*-Training stattgefunden. Im Rahmen dieser Schulung wurden *benchmarks* bestimmt, die das

[282] Vgl. http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual/, Zugriff am 22.8.2005.

DESI-Testkonstrukt und die DESI-Niveaus illustrieren. Dabei wurden Standards bestimmt in Bezug auf die prototypische Mitte eines Niveaus, nicht jedoch bezogen auf Minimalstandards, die den Beginn eines Niveaus bezeichnen. Denn die Niveaus lassen sich nicht durch scharfe Grenzen (*cut-off points*) voneinander unterscheiden; vielmehr haben sie gerade in den Grenzbereichen fließende Übergänge – die Zugehörigkeit in das eine oder andere Niveau wird, wie im Verlauf dieser Arbeit an verschiedenen Stellen gezeigt, eher durch prototypische Merkmale bestimmt, die die Mitte eines Niveaus charakterisieren und sich in den Deskriptoren wiederfinden.

Aufgrund dieser Sachlage scheint es ratsam, die im *Manual* zur Standardisierung vorgeschlagene Prozedur im Hinblick auf eine mögliche Anbindung der DESI-Schreibaufgaben wie folgt abzuändern: Man administriert die DESI-Schreibtests in einer Lernergruppe und lässt die Performanzen von Experten, idealiter den Lehrkräften dieser Probanden, mithilfe des DESI-Bewertungsschemas bewerten. Diese Experten müssen auch mit dem System des GER vertraut sein, ebenso wie sie die Lernergruppe sehr gut kennen müssen. Dann können sie die Lernenden in deren globaler Schreibfertigkeit auf die Niveaus des GER einstufen. Da die beiden Einschätzungen sich auf dieselben Probanden beziehen und von derselben Expertengruppe vorgenommen werden, ist es mit psychometrischen Methoden möglich, die DESI-Niveaus den entsprechenden GER-Niveaus zuzuweisen, so wie es im *Manual* im Fallbeispiel auf S. 117ff beschrieben wird. Dann könnte man von empirischer Anbindung der DESI-Niveaus an die Niveaus des GER sprechen.[283]

4.7.2 Einbindung der Testergebnisse in ein Portfolio-*Assessment*

Die Idee, die Schreibfertigkeit im Rahmen eines Portfolios[284], einer Sammlung von Lernertexten also, zu beurteilen, entstammt der Diskussion Mitte der 80er Jahre um die Nachteile der Beurteilung des Schreibens in Testsituationen: In der Regel gibt es in einem Test zeitliche Beschränkungen, die Bewertung erfolgt meist nur aufgrund einer Momentaufnahme, die Lernenden haben keine Kontrolle über Themen und Schreibanlässe, sie erhalten meist kein individuelles Feedback zu ihrer Leistung und es ergeben sich keine Lernmöglichkeiten aus dem Test.[285] Das Erlernen der Schreibfertigkeit jedoch ist nach Kroll (1998) gekennzeichnet durch Lernen mittels Feedback, Zeitinvestitionen seitens Lernender und Lehrender, und Einsicht in Lern- und Schreibprozesse, die sich etwa über Positivkorrektur von Aufsätzen erzielen lassen. Um Lernprozesse zu fördern und auf diese Weise einen Beitrag zur systemischen Validität von

[283] Auf der EALTA-Tagung in Krakau im Mai 2006 wurde von F. Kaftandjieva und S. Takala im *Pre-conference Workshop* vorgeschlagen, die DESI-Lernertexte von Experten auf die GER-Niveaus einstufen zu lassen. Dazu könne das Raster zur Bewertung schriftlicher Produktionen im *Manual*, S.82 genutzt werden. Diese Einstufungen wiederum könnten genutzt werden, um die DESI-Kompetenzskalen auf die Niveaus des GER zu beziehen. In diesem Zusammenhang darf auf die Präsentation von Harsch unter http://www.ealta.eu.org/conference/2006/docs/Harsch_ealta2006.ppt verwiesen werden. Ob die direkte Einstufung mithilfe des Manual-Rasters valide erfolgen kann, wird sich in der Praxis zeigen. Die Anbindung wird derzeit in Kooperation des Lehrstuhls für Didaktik des Englischen der Universität Augsburg und dem Institut für Qualitätsentwicklung (IQB) der Humboldt-Universität zu Berlin geplant.

[284] An dieser Stelle darf noch einmal darauf hingewiesen werden, dass die vorliegende Arbeit sich nicht mit dem Europäischen Sprachenportfolio beschäftigt, da dies ein Instrument zur Selbstbeurteilung ist, welche nicht Gegenstand dieser Arbeit ist.

[285] Vgl. dazu auch Kroll 1998, Murphy & Grant 1996, Torrance 1998 u. a..

Beurteilungen zu leisten, bedarf es neuer Wege der Schreibbeurteilung, wie Camp (1996) feststellt. Eine Möglichkeit ist die des Portfolio-*Assessments*, das Kroll (1998: 231) in seinen Vorteilen wie folgt beschreibt:

> A portfolio containing multiple examples of a student's work provides multiple kinds of writing for evaluation. These materials are produced potentially free from the negative effects of test anxiety and without the typical time constraints of an exam. Thus, a primary impetus for the portfolio approach is the possibility of having evaluation relate more directly to pedagogy and to the goals of writing instruction.

Folgende lernförderliche Aspekte lassen sich im Rahmen eines Portfolio-*Assessments* ausmachen:[286]

- *Valide Beurteilungsgrundlage*: Die Beurteilung der Schreibfertigkeit findet über längere Zeiträume statt und ihr liegt im Idealfall eine Vielzahl verschiedener Schreib-*Samples* zu einer Vielzahl von Situationen, Themen, Schreibanlässen und Textsorten eines Probanden zugrunde; sie findet in Abstimmung zwischen Lehrenden und Lernenden statt, so dass Fremd- und Selbstbeurteilung miteinander verknüpft werden.

- *Eigenverantwortung und individuelle Passung*: Die Lernenden können gemeinsam mit den Lehrenden bestimmen, wie das Portfolio beschaffen sein soll (ob es im Sinn eines *show portfolio* nur die besten Arbeiten enthalten soll oder ob es daneben auch Textentwürfe, Aufgabenstellungen, Selbstreflexion etc. enthalten soll als so genanntes *full portfolio*[287]), welche Lernziele und welche Lernbeweise im gegebenen Kontext relevant sind, welche Arbeiten ins Portfolio kommen (seien es nun die „fertigen" Produkte oder aber auch die Dokumentation der Entwicklung eines Textes) und nach welchen Kriterien und Standards diese beurteilt werden. In diesem Zusammenhang können valide Bewertungsmaßstäbe und Merkmale guter Texte diskutiert werden, so dass das Bewusstsein der Lernenden auch im Hinblick auf diese Aspekte herausgebildet werden kann. Dadurch wird die Eigenverantwortung im Lern- und Beurteilungsprozess gefördert.

- *Individueller Fokus auf lernförderliche Schreibprozesse*: Durch die Möglichkeit, Texte zyklisch unter ausgewählten Gesichtspunkten zu überarbeiten, kann die Aufmerksamkeit gezielt auf bestimmte Prozesse und Strategien gelenkt werden. Diese können systematisch erprobt und geübt werden, ebenso wie lernerindividuell ausgewählte sprachlich-kommunikative Aspekte erarbeitet werden können. Die Überarbeitungsziele und die Bewertung der überarbeiteten Texte können auf diese Weise zusammen mit den Lernenden auf den gerade aktuellen Unterrichtsgegenstand und die momentanen Lernziele ausgerichtet werden und dem individuellen Lernstand angepasst werden. Die Lehrenden könnten dadurch im Korrekturaufwand entlastet werden, wenn sie an einem gegebenen Lernertext, der zyklisch überarbeitet wird, lediglich ausgewählte Aspekte betrachten und dazu gezielte Rückmeldung geben. Damit können Menge, Qualität und Nützlichkeit der diagnostischen Informationen für Lernende wie Lehrende vermehrt werden.

[286] Vgl. dazu etwa Camp 1996, Elbow 1996, Hamp-Lyons 1996, Kroll 1998, Murphy & Grant 1996 u. a..
[287] Vgl. Hamp-Lyons (1996: 238).

- *Einbindung in den größeren Kontext des Lehrens und Lernens*: Portfolios bieten die Möglichkeit, Lernertexte aus Klasszimmerprojekten, außerschulischen Aktivitäten oder eben auch aus Schulleistungsstudien mit aufzunehmen und so die gesamte Umgebung der Lernenden mit einzubeziehen und individuelle Schwerpunktsetzung zu ermöglichen. Gleichzeitig kommt dabei das Prinzip des selbstgesteuerten Lernens zum Tragen, das nach Börner (1989) zum Lernerfolg beiträgt und auf ein lebenslanges Lernen vorbereitet. Das lebenslange Lernen hat auf dem Gebiet des Schreibens besondere Bedeutung, denn Schreiben ist eine der basalen kulturellen Techniken und Schlüsselqualifikationen. Es trägt darüber hinaus auch zur Persönlichkeitsentwicklung bei, denn "… learning to write may largely be a process of 'personal growth in social context'." (Cumming 1998: 66), so dass die Schreibentwicklung auch jenseits der Schule eine besondere Stellung einnimmt.

Dem Portfolio-*Assessment* sind Grenzen gesetzt, innerhalb derer es sinnvoll entwickelt und eingesetzt werden muss: So sollte der Zweck des Portfolio-*Assessments* zunächst mit curricularen Vorgaben abgestimmt werden. Das Portfolio muss dann auf die Kontexte hin ausgelegt werden, in denen es zum Einsatz kommt – es kann nicht auf andere Kontexte übertragen werden. Inhalte und Verfahrensweisen müssen vor dem Einsatz spezifiziert werden, um die Bewertung der vielfältigen Arbeiten vergleichbar zu halten.[288] Doch im Rahmen ihrer Möglichkeiten tragen Portfolio-*Assessments* das Potenzial, Lernertexte aus den verschiedensten Kontexten (wie beispielsweise aus Schulleistungsstudien) in die schulische Beurteilung einfließen zu lassen und so der Entwicklung der Schreibfertigkeit umfassender gerecht zu werden, als es traditionelle Beurteilungsformate können. Darüber hinaus kann von einem Portfolio-*Assessment* eine nicht zu unterschätzende Wirkung ausgehen – lassen Sie mich dazu mit den Worten von White (1996: 303) schließen:

> Most significantly, portfolios return assessment to local control,
> shifting power from testing authority to student.

[288] Vgl. hierzu etwa Larson 1996 oder Murphy & Grant 1996.

Resümee

Im Folgenden werden die Analyseergebnisse zu Sprachbegriff, Lern- und Lehransätzen, Testbegriff, Beurteilungskonzepten und zum Skalenansatz im GER zusammengefasst. Diese Zusammenfassung basiert auf den Ergebnissen, die in der vorliegenden Arbeit jeweils am Ende der Analysen des GER auf diese Begrifflichkeiten hin in den Kapiteln 1.2.5.5, 1.3.4.6, 2.5.3, 2.5.5 respektive 3.4.4 zu finden sind. An diese Zusammenfassung schließt sich hier im Resümee die Darstellung kritischer wie positiver Aspekte des *Gemeinsamen europäischen Referenzrahmens* auf Basis der Analysen der vorliegenden Arbeit an. Den Abschluss bilden Desiderate zur Weiterentwicklung des Referenzsystems.

Schlüsselbegriffe im GER

Der Sprachbegriff, der sich im GER ausmachen lässt, ist von seiner theoretischen Konzeption her mehrdimensional und umfasst sprachliche Teilkompetenzen und kommunikative Aktivitäten. Sprache wird handlungs- und verwendungsorientiert betrachtet und theoretisch in ihre soziokulturellen Kontexte eingebettet, auch wenn sich diese Einbettung nicht in den Beispielskalen niederschlägt. Die Vielfalt der Sprachen und Kulturen Europas wird anerkannt, jedoch in den Skalen ebenfalls nicht adäquat operationalisiert. Beispielsweise wird die Problematik der Migrantensprachen in Europa nicht thematisiert, ebenso wenig wie der Bereich der interkulturellen Kommunikation, dem wesentliche Bedeutung in einem multilingualen und plurikulturellen Europa zukommt. Der Kommunikationsbegriff im GER ist daher als idealisiert zu betrachten, denn er ist auf den Idealfall der gelungenen Kommunikation hin ausgerichtet.

Der lerntheoretische Ansatz des GER ist handlungsorientiert und betrachtet deshalb Sprachverwendung als denselben Bedingungen unterworfen wie Sprachlernen. Auch Erwerb, Lernen und Anwendung werden nicht differenziert. Diese Undifferenziertheit in der theoretischen Konzeptionalisierung führt dazu, dass zwei für einen *europäischen Referenzrahmen für Sprachen* wesentliche Bereiche nicht thematisiert werden: Einerseits wird – wie auch schon beim Sprachbegriff – die Perspektive des ungesteuerten Erwerbs im multilingualen europäischen Kontext ignoriert; andererseits wird die Bedeutung der Interimsprachentheorie für das Erlernen einer Sprache nirgends erörtert. Damit fehlen wesentliche lerntheoretische Grundlagen im GER. Diskutiert werden jedoch lernförderliche Bedingungen wie die Bedeutung authentischer Sprachverwendung und Interaktion, die Bedeutung von Sprachbewusstheit und die Bedeutung der Lernerautonomie.

Ein kohärentes Vermittlungskonzept lässt sich im GER allerdings nicht ausmachen: Es finden sich keine Charakteristika eines „guten“ Fremdsprachenunterrichts, die einen Rahmen darstellen würden, innerhalb dessen sich die Nutzer des GER wiederfinden könnten. Vielmehr

werden unkommentierte Auflistungen verschiedener (nicht immer lernförderlicher) Optionen dargeboten, beispielsweise zu methodischen Ansätzen, Übungsformen oder dem Umgang mit Fehlern. Daneben finden sich einige generelle Aussagen zu Inhalten, Themen, Aktivitäten und kommunikativen Aufgaben. Wesentliche Bereiche eines Fremdsprachenunterrichts, der die Mehrsprachigkeit in Europa fördern will, werden jedoch nicht thematisiert: So wird etwa die Bedeutung des Input im Sprachlernprozess nicht anerkannt; dies zeigt sich daran, dass Aspekte wie etwa der *classroom discourse* und die ihm zugrunde liegenden Lehrerkompetenzen in der Zielsprache nicht diskutiert werden. Auch wird die europäische Dimension im Fremdsprachenunterricht (vgl. Kapitel 1.3.3.5 der vorliegenden Arbeit) nicht umgesetzt, wo dies doch ein zentraler Bereich des GER, eines Instruments zur Förderung der Mehrsprachigkeit in Europa, sein müsste.

Der Testbegriff des GER, der sich im Wesentlichen in GER-Abschnitt 9 niederschlägt, ist aufbauend auf Sprach- und Lernbegriff ebenfalls kommunikativ-handlungsorientiert. Um der Vielfalt sprachlich-kommunikativer Handlungen und der ihnen zugrunde liegenden Kompetenzen gerecht zu werden und das Sprachvermögen adäquat zu erfassen, schlägt der GER die Verknüpfung aller Wissens- und Kompetenzbereiche beim Testen vor. Der Testansatz des GER ist modell-basiert, denn diese Wissens- und Kompetenzbereiche, die durch das GER-Referenzsystem operationalisiert werden, lassen sich im Wesentlichen in Bachmanns Modell der kommunikativen Kompetenz verorten (vgl. Bachmann 1991a, wobei die Quellen des Kompetenzmodells im GER selbst nicht offen gelegt werden). Der GER schlägt, wo möglich, direktes Testen und die positive Bewertung von Leistungen vor; die Subjektivität in der direkten Bewertung wird anerkannt und es werden Möglichkeiten der Objektivierung aufgezeigt. Der GER wählt einen kriteriumsorientierten Ansatz, um Aussagen bezüglich des individuellen Sprachvermögens unabhängig von der jeweiligen Lernergruppe treffen und Beurteilungen über verschiedene Kontexte hinweg vergleichen zu können.

Der GER ist ausgelegt auf die Bildung eines gemeinsamen Referenzsystems zur Beurteilung von Sprachvermögen, wobei er die Bedeutsamkeit der Verknüpfung von Fremd- wie Selbstbeurteilung und der Vergleichbarkeit von Beurteilungen in verschiedenen Kontexten anerkennt. Um diese Verknüpfung und Vergleichbarkeit zu erzielen, ist das GER-Referenzsystem generell und dekontextualisiert gehalten. Es steht im Zentrum des GER, weshalb auch alle Beurteilungsaspekte, wie etwa die Ausführungen in GER-Abschnitt 9, auf das Referenzsystem und dessen Beispielskalen bezogen sind.

Der Skalenansatz des GER operationalisiert das besagte Konzept der kriteriumsorientierten modell-basierten Positivbewertung: Zu den (meisten) Kategorien des genannten Kompetenzmodells werden kriterienbezogene Deskriptoren zur Verfügung gestellt, die das Sprachvermögen abgestuft auf sechs Niveaus in der Regel in Form von positiven KANN-Formulierungen beschreiben. Allerdings erfüllen nicht alle Formulierungen der GER-Deskriptoren die im GER aufgestellten Anforderungen der Kürze, Klarheit, Unabhängigkeit, Positivformulierung und der Ermöglichung einer Ja/Nein-Entscheidung. Zudem ist der Beschreibungsgegenstand der Skalen

nicht transparent und nicht empirisch validiert, so dass der Status der Skalen nicht eindeutig zu bestimmen ist; darüber hinaus repräsentieren die Skalierungen des besagten Schweizer Konstruktionsprojekts die Konzeptionalisierung der am Projekt Beteiligten, so dass sie nicht ohne Weiteres auf den europäischen Kontext hin verallgemeinert werden können. Aufgrund dieser Beschränkungen und des undurchsichtigen Status der Deskriptoren können die Skalen nicht zu all den Funktionen eingesetzt werden, die im GER genannt werden (vgl. hierzu die ausführliche Erörterung in den Kapiteln 3.4.4.3 und 3.4.4.4 der vorliegenden Arbeit).

Nach dem momentanen Stand der Dinge geben die GER-Skalen Anlass zur Diskussion der Kategorisierung und Abstufung verschiedener Aspekte des Sprachvermögens; sie müssen sich im europäischen Kontext erst beweisen. Der GER stellt derzeit keine Ausgangsbasis dar, um beispielsweise Tests, Bewertungsschemata oder Lehrmaterialien auf Grundlage der Skalen zu beschreiben oder zu entwickeln. Anstehende Entscheidungen etwa im Kontext der Testentwicklung, Unterrichtsplanung oder curricularer Entwicklungen können ebenfalls nicht mit dem GER begründet werden, da die im GER aufgestellten Behauptungen nicht belegt werden. Man kann ihn jedoch im Sinn eines *Referenz*mittels nutzen, um solch anstehende Entscheidungen zu reflektieren und einmal getroffene Entscheidungen oder Entwicklungen posthoc mit dem System des GER zu vergleichen. Da die GER-Skalen Aspekte des Sprachvermögens aus Perspektive der Lernenden beschreiben, können sie bei der Beurteilung von den Beurteilern sehr gut bekannten Lernenden und bei der Selbstbeurteilung zum Einsatz zu kommen. Vor jedem Einsatz jedoch müssen die Deskriptoren auf die jeweiligen Verwendungskontexte, den Gegenstand, auf den sie abzielen sollen und auf adäquate Formulierungen hin überprüft und adaptiert werden.

Kritische Aspekte am GER

Die im GER geltend gemachten „Prinzipien einer pluralistischen Demokratie“ (GER 2001: 29) führen bei vielen brisanten Fragestellungen dazu, dass im GER keine eindeutige Stellung „für die eine oder andere Seite“ (ebd.) bezogen wird. Meist werden Begrifflichkeiten oder Konzepte zwar in existierenden unterschiedlichen Definitionen oder Konzeptionalisierungen vorgestellt, doch wird in der Regel nichts darüber ausgesagt, wie diese Konzepte im GER selbst verstanden und genutzt werden. Beispielsweise wird nicht definiert, wie im GER die Begriffe *Lernen* und *Erwerb* konzeptionalisiert werden (ebd.: 137f); wenngleich verschiedene methodische Ansätze im Fremdsprachenunterricht vorgestellt werden, so findet sich doch kein begründeter Kommentar, welche Ansätze in welcher Art von Unterricht angemessen oder empfehlenswert wären (ebd.: 140ff); der Begriff der Performanz etwa wird zwar in seinen möglichen Definitionen beschrieben, doch es wird nicht definiert, wie er im GER verwendet wird (ebd.: 182). Wenn der GER seinen Nutzern einen Rahmen stecken will, den sie zur Reflexion, Beschreibung oder Begründung ihres jeweiligen Vorgehens (vgl. ebd.: 8, 10, 15f) heranziehen können, so muss er

Stellung beziehen und seine Schlüsselkonzepte so transparent definieren, dass den Nutzern eine Begründung ihres Vorgehens auch ermöglicht wird.

Zu solch einer transparenten Definition von Kernkonzepten tragen im Allgemeinen das Offenlegen des eigenen Verständnisses hinsichtlich eines bestimmten Konzepts sowie dessen Verankerung in wissenschaftlicher Forschung und eine konsistente Terminologieverwendung bei. Beides ist jedoch im GER nicht im benötigten Ausmaß gegeben. Dies führt dazu, dass viele Konzepte und Begrifflichkeiten, wie etwa die der Kompetenz und Performanz in der Sprachbeurteilung oder der Sprachbegriff, der dem GER zugrunde liegt, durch die Ausführungen im GER verschleiert werden. Für die Nutzer ist nicht nachvollziehbar, aus welchen Forschungsergebnissen die Behauptungen im GER stammen, in welchen wissenschaftlichen Modellen sie verankert sind und welchen Status sie deshalb einnehmen: Nach dem momentanen Stand der Dinge handelt es sich eher um unbelegte Tatsachenbehauptungen, die zu einer fundierten Diskussion unter den Nutzern wenig beitragen können. Zum Beispiel werden bei der Thematisierung von Fehlern in GER-Abschnitt 6.5 Möglichkeiten des Umgangs mit Fehlern aufgelistet, jedoch ohne die einzelnen Verfahrensweisen zu bewerten oder wenigstens in ihrer Angemessenheit zu kommentieren – diese Auflistung kann zu Reflexion oder Standortbestimmung keinen Beitrag leisten. Die Haltung des GER, „nur Fragen“ zu stellen, doch „keine Antworten“ zu geben (ebd.: 8), ist für solch ein Instrument nicht angemessen. Man erwartet als Nutzerin eines Referenzrahmens sicherlich keine Vorgaben oder Handlungsanweisungen. Dennoch sollte ein solcher Rahmen die Bedingungen fremdsprachlichen Lernens und Lehrens diskutieren, begründete und mit Quellenangaben belegte Aussagen treffen und die jeweils relevanten Kernkonzepte und Schlüsselbegriffe in transparenter Terminologie definieren. Dadurch könnten die Eckpunkte des Referenzrahmens gesteckt, verschiedene Vorgehensweisen in ihren Vor- und Nachteilen diskutiert und Stellung bezogen werden bezüglich der Maßnahmen, die als lernförderlich angesehen werden.

Die inkonsistente Terminologieverwendung könnte sich teils auch auf Übersetzungsprobleme zurückführen lassen. Begrifflichkeiten des englischen Originaldokuments werden in der deutschen Übersetzung an verschiedenen Stellen mit je unterschiedlichen Begriffen wiedergegeben, was zu Verwirrung und im schlimmsten Fall zu sinnentstellenden Übersetzungen führt. Wie in Kapitel 2.5.2 der vorliegenden Arbeit gezeigt, werden beispielsweise die Konzepte *proficiency, competence, performance* und *achievement* im Kontext der Sprachbeurteilung mit je verschiedenen Termini ins Deutsche übersetzt. Hier müsste dringend eine europäische Vereinheitlichung der im GER verwendeten Terminologie geschaffen werden.

Ein weiterer kritischer Aspekt ist der im GER immer wieder postulierte Anspruch der Umfassendheit (vgl. etwa GER: 3, 9, 12, 14, 19, 20, 21, etc.). Auf S. 9 behauptet der GER, dass er dem Kriterium *umfassend* gerecht werden will und dass man deshalb „wirklich alles finden sollte“, was man zur Beschreibung seiner „Ziele, Methoden und Produkte“ benötige (dort bezogen auf die Parameter, Kategorien und Beispiele in GER-Abschnitt 2). Auf derselben Seite, zwei

Abschnitte darunter, liest man jedoch: „Weder die Kategorien noch die Beispiele können für sich in Anspruch nehmen, *vollständig* zu sein“ (Hervorh. d. V.). Es ist nicht nachvollziehbar, wieso im GER auf dem Begriff *umfassend* bestanden wird, wenn eine Vollständigkeit i. S. von *umfassend* gar nicht angestrebt wird, zumal sich die Autoren bewusst waren, dass das GER-System noch der Weiterentwicklung aus der Erfahrung bedarf (vgl. beispielsweise GER: 10 zum Entwicklungsbedarf des taxonomischen Systems). Eine *umfassende* Darstellung dürfte die Nutzer des Referenzrahmens erwarten lassen, wirklich alle relevanten Aspekte zu einer bestimmten Thematik behandelt zu finden. Doch dies trifft nicht immer zu. Mag man im GER auch viele relevante Aspekte zu verschiedenen sprachlernbezogenen Konzepten finden, so gibt es doch ebenso viele Leerstellen. Die mangelnde Thematisierung der Mehrsprachigkeit (hier insbesondere der Dreisprachigkeit, wie sie im *Weißbuch Lehren und Lernen. Auf dem Wege zur kognitiven Gesellschaft* der Europäischen Union 1995 gefordert wird) und der interkulturellen Kompetenzen möge hier als Beispiel genügen: Im Vorwort des GER ist dazu im ersten Satz zu lesen (ebd.: 3):

> *Mehrsprachigkeit* und *kulturelle Kompetenz* sind die zentralen Themen in dieser umfassenden Publikation zum Fremdsprachenlernen und zugleich Ergebnis einer langjährigen Diskussion unter Fremdsprachenexperten aus 40 Ländern. Sie fasst den aktuellen Stand der Fremdsprachendiskussion zusammen (…).

Wie in den Kapiteln 1.2.5 respektive 1.3.4 der vorliegenden Arbeit jedoch gezeigt, wird das Konzept der Mehrsprachigkeit im GER nicht transparent vom Konzept der Vielsprachigkeit differenziert, und das, obwohl diese beiden Begriffe in der Europäischen Gemeinschaft klar umrissen sind. Interkulturelle Kompetenzen werden im GER ebenfalls nicht angemessen thematisiert. Beispielsweise finden sich im Kategoriensystem der GER-Abschnitte 4 und 5 keine entsprechenden Kategorien oder Beispielskalen, der Bereich der interkulturellen Missverständnisse wird nirgends thematisiert, und auch bei den Ausführungen zu Fremdsprachenlernen und -lehren in GER-Abschnitt 6 sind keine entsprechenden Charakteristika der mehrsprachigen oder plurikulturellen Kompetenzen zu finden. Dies stellt bei einem Instrument des Europarats, das sich der Förderung der europäischen Mehrsprachigkeit verpflichtet fühlt (vgl. etwa GER: 12, 14 oder 16), ein großes Manko dar. Wenn der GER die „aktuelle Fremdsprachendiskussion“ (s. oben) zusammenfassen will, so muss dazu vermerkt werden, dass in diesem Bereich ebenfalls wichtige didaktische Grundlagen fehlen: Es darf auf die obigen Ausführungen zu Lernbegriff und Vermittlungskonzept im GER verwiesen werden. Ein kohärentes Vermittlungskonzept lässt sich, wie gerade dargestellt, nicht ausmachen.

Diese konzeptionellen Lücken tragen unter anderem zur in Kapitel 3.4 dieser Arbeit analysierten Beurteilungslastigkeit des GER bei. Diese Ausrichtung auf Beurteilungsaspekte des Sprachvermögens gibt insofern Anlass zur Kritik, als sich das Selbstverständnis des GER ganz anders liest: Er will ein Instrument sein zur Umsetzung sprachenpolitischer Ziele des Europarats, wie sie etwa in GER-Abschnitt 1.2 dargestellt sind, insbesondere der Förderung der Mehrsprachigkeit (GER 2001: 12). Dazu will er Praktiker ermutigen, über ihr Vorgehen zu reflektieren

und er will eine Basis zur Verfügung stellen, um die Kommunikation und den Erfahrungsaustausch unter Praktikern anzuregen, denn (vgl. GER 2001: 8):

> (…) der Europarat [hat es sich] zur Aufgabe gemacht […], die Qualität der Kommunikation unter Europäern mit unterschiedlichem sprachlichen und kulturellen Hintergrund zu verbessern. Dies geschieht, weil eine verbesserte Kommunikation zu größerer Mobilität führt und zu vermehrten direkten Kontakten, was wiederum zu einem besseren Verständnis und zu besserer Zusammenarbeit führt. Der Europarat unterstützt Lern- und Lehrmethoden, die jungen Menschen, aber auch älteren Lernenden helfen, Einstellungen, Kenntnisse und Fähigkeiten zu entwickeln, die notwendig sind, um im Denken und Handeln unabhängiger zu werden und in ihren Beziehungen zu anderen Menschen verantwortungsbewusst und kooperativ zu handeln. Auf diese Weise trägt die Arbeit auch zur Förderung eines demokratischen, staatsbürgerlichen Bewusstseins bei.

Wie jedoch besagtes demokratisches, staatsbürgerliches Bewusstsein mit einem Instrument gefördert werden soll, in dem wesentliche Bereiche des Fremdsprachenlernens nicht diskutiert werden und das sich bei genauer Betrachtung als Beurteilungsinstrument fremdsprachlichen Könnens erweist, sei dahingestellt.

Die Behauptungen des GER bezüglich der Verwendbarkeit seines Referenzsystems sind, wie oben angedeutet, nicht haltbar. In diesem System, so jedenfalls behauptet es der GER (vgl. ebd.: 8ff und 14), könnten Lernende, Lernziele, Inhalte, Aufgaben, Materialien, Curricula, Tests, Prüfungen und Prüfungsergebnisse eingeordnet werden. Wie allerdings in Kapitel 3.4 der vorliegenden Arbeit gezeigt, ist diese Verortung nicht problemlos möglich, da das Skalensystem nicht angemessen ausgelegt ist, all diesen unterschiedlichen Zwecken zu dienen. Dies ist zurückzuführen auf den oben erwähnten intransparenten Status der GER-Skalen – ehe also das Skalensystem als in allen denkbaren Kontexten verwendbar dargestellt wird, sollte der Status der Skalen klargestellt und die Deskriptoren in den im GER genannten Verwendungskontexten validiert werden. Ansonsten könnten die Behauptungen im GER zu missbräuchlichen Verwendungen der Skalen in Kontexten führen, für die sie nicht ausgelegt sind.

Positive Impulse des GER

Neben der gerade geäußerten Kritik dürfen die positiven Impulse nicht vernachlässigt werden, die vom GER ausgehen können:

Der GER gibt Anstöße zur Reflexion vieler Aspekte, Inhalte und Verfahrensweisen im Kontext des fremdsprachlichen Lernens und Lehrens, wenn er auch keine begründete Entscheidungsbasis zur Verfügung stellen kann (siehe oben). Weiterhin regt der GER die Diskussion und Entwicklung von Standards des Sprachenlernens, der Sprachbeherrschung und der Beurteilung derselbigen an. Die Bildungsstandards der KMK für die erste Fremdsprache sind nur ein Beispiel für solch eine Standardentwicklung – auch sie müssen sich nun der öffentlichen Diskussion stellen.

Das Skalensystem des GER regt den Dialog unter Praktikern wie Theoretikern an hinsichtlich der Angemessenheit der angesetzten Kategorien, der Abstufungen und der Merkmale, die

für die jeweiligen Niveaus charakteristisch sind. Er stellt ein breites Angebot an möglichen Kategorien und Abstufungen zur Verfügung; dieses Angebot kann und soll auf konkrete Kontexte hin geprüft und adaptiert werden. Der GER lädt seine Benutzer bewusst ein, sein Referenzsystem und die Deskriptoren „kritisch zu nutzen“ (GER 2001: 10) und dem Europarat über „Erfahrungen bei der Umsetzung in die Praxis“ zu berichten (ebd.) – insofern können alle Nutzer des GER zu seiner Weiterentwicklung beitragen, indem sie die Verwendbarkeit des GER in unterschiedlichen Kontexten und unter verschiedenen Bedingungen kritisch beurteilen, so wie es die vorliegende Arbeit versucht.

Die Konstruktion des Referenzsystems an sich ist eine innovative Entwicklung, die versucht, alle an der Sprachverwendung beteiligten Facetten und Teilkomponenten in ihren charakteristischen Merkmalen abgestuft zu beschreiben. Dabei knüpfen die Abstufungen, die Referenzniveaus also, an Vorarbeiten des Europarats an, wie etwa an dessen Lernzielbeschreibungen (vgl. etwa van Ek 1975 und 1980, oder van Ek & Trim 1990, 1991 und 1997). Auch wurden die Beispielskalen mit existenten Beurteilungssystemen, wie etwa dem Skalensystem der ALTE, verbunden. Diese Anbindung sorgt für Kohärenz im europäischen Rahmen der Sprachbeurteilung.

Wenngleich die Verwendbarkeit des GER-Skalensystems in der Beurteilung des Sprachvermögens derzeit wie erwähnt auf Beurteilungen von den Beurteilern sehr gut bekannten Lernenden beziehungsweise auf die Selbstbeurteilung durch Lernende beschränkt ist, stellt der Skalenansatz dennoch eine nicht zu unterschätzende Neuorientierung in der Leistungsbeurteilung dar: Zum einen ermöglicht es das Referenzsystem, Lernende europaweit an vergleichbaren Maßstäben zu messen, vorausgesetzt die Beurteilenden verfügen über ein vergleichbares Verständnis der Kategorien und Niveaus. Zum zweiten wird mittels der KANN-Deskriptoren positiv an das Sprachvermögen der zu Beurteilenden herangetreten: Es werden nicht mehr Defizite und Fehlleistungen beurteilt, sondern es wird betrachtet, was in welchen Bereichen in welchem Umfang und in welchem Korrektheitsgrad schon beherrscht wird. Diese sinnvolle Ergänzung der traditionellen Negativkorrektur kann die Lernenden im Idealfall motivieren für ein lebenslanges Lernen. Zum dritten verknüpft das Skalensystem des GER die Beurteilung durch Lehrende mit der Selbstbeurteilung der Lernenden – dies stellt einen wichtigen Anstoß zur Kultivierung der Selbstbeurteilung als motivierendes und steuerndes Element im Lernprozess dar und kann, wiederum idealiter, zur Übernahme von Eigenverantwortung im Lernprozess führen, ebenfalls eine Voraussetzung für lebenslanges Lernen.

Desiderate zur Weiterentwicklung des Referenzsystems im GER

Abschließend sei es gestattet, Vorschläge zusammenzustellen die Aspekte betreffend, in denen der GER und das ihm zugrunde liegende Beschreibungssystem erforscht, erweitert oder verbessert werden könnten. Wenden wir uns zunächst dem Beschreibungssystem zu:

Das Kategorien- und Skalensystem müsste analysiert werden, ähnlich wie es bei den Vorarbeiten zur Erstellung des *Dutch Grid* gemacht wurde, um Inkonsistenzen, Lücken und Widersprüchlichkeiten zu identifizieren und zu einem konsistenten Beschreibungsinstrumentarium zu kommen. Hierbei sollten die theoretischen Grundlagen und Modelle offen gelegt werden, in denen das Referenzsystem verankert ist, ebenso wie pragmatisch bedingte Klassifizierungsentscheidungen transparent dokumentiert werden sollten. Die verwendete Terminologie sollte definiert und einheitlich verwendet werden, um gemeinsames Verständnis zu erleichtern beziehungsweise zu ermöglichen.

Im Zusammenhang mit Terminologiefragen wäre es wie erwähnt ratsam, die Übersetzungen des GER zu untersuchen, um ein europäisches Terminologiesystem im Rahmen des GER zu entwickeln. Dabei soll solch ein Terminologiesystem keine „Zwangsjacke“ von Terminidefinitionen darstellen, sondern im Rahmen der Verwendung des GER praktikable Termini bieten, die von allen Beteiligten in den jeweiligen Sprachen, in die der GER übersetzt wurde, vergleichbar verstanden und verwendet werden können.

Eine empirische Absicherung der Beschreibungsgegenstände und Abstufungen der GER-Skalen wäre hilfreich, um das Referenzsystem zusätzlich zu validieren. Derzeit ist, wie oben ausgeführt, die Basis der Beschreibungen nicht transparent, da die Deskriptoren aus verschiedenen Quellskalen stammen. Könnte man den Beschreibungsgegenstand der Deskriptoren und die Niveaus des GER empirisch verifizieren, so hätte man eine Ausgangsbasis für Verwendungsbereiche jenseits des Berichterstattens geschaffen. Selbstverständlich muss den Bereichen, die nicht im genannten Konstruktionsprojekt skaliert werden konnten (wie etwa dem Schreiben), besondere Aufmerksamkeit zukommen.

Um die Niveaus des GER gemeinsam interpretieren und verstehen zu können, wäre die Entwicklung und Diskussion von *benchmarks*, von Performanzbeispielen, die ein Niveau illustrieren, von großer Bedeutung. Die GER-Deskriptoren sind relativ abstrakt und generell gehalten und können deshalb nicht adäquat eingesetzt werden, um beispielsweise eine Performanz auf einem bestimmten Niveau einzustufen. Wenn nun aber aus einer Vielzahl von Performanzen, etwa auf dem Weg, der im *Manual* zur Standardisierung (vgl. *Manual* 2003: Abschnitt 5 und die Ausführungen in Kapitel 3.5 dieser Arbeit) vorgeschlagen wird, aussagekräftige *benchmarks* gefunden werden könnten, so wäre dies ein Ansatzpunkt, die Niveaus leichter zugänglich zu machen und sie mit konkreten Performanzen in Verbindung zu bringen.

Daneben könnte das Referenzsystem erweitert und adaptiert werden auf den Kontext jugendlicher Sprachlernender und Sprachverwendender, denn derzeit ist es ausgelegt auf den Horizont Erwachsener.

Auch bezüglich der Implementierung des GER lassen sich einige Desiderate aufstellen:

Zunächst kommt dem Aspekt der Professionalisierung der Benutzer des GER eine nicht zu unterschätzende Bedeutung zu: Lehrende wie Lernende, Testkonstrukteure, Prüfende, Curriculumplaner und Lehrwerksentwickler müssen vertraut gemacht werden mit dem Kategorien- und Niveausystem des GER. Es muss, ähnlich wie im *Manual* unter „Familiarisierung" beschrieben, ein grundsätzliches Verständnis des Referenzsystems aufgebaut werden, ehe Wege des sinnvollen Umgangs mit diesem Referenzsystem und Grenzen des Referenzrahmens aufgezeigt werden können, um Missbrauch vorzubeugen. Diesem Punkt sollte beispielsweise in der Lehreraus- und Fortbildung gebührende Beachtung geschenkt werden, denn nur wenn die Lehrenden sicher im Umgang mit dem GER sind, können sie ihn im Unterricht umsetzen, sei es hinsichtlich der curricularen Vorgaben oder hinsichtlich der Selbstbewertung seitens der Lernenden.

In den verschiedenen Beurteilungskontexten können Selbst- und Fremdbeurteilungen vergleichend untersucht werden, um Hinweise auf eventuell unterschiedliche Interpretationen der Deskriptoren zu erhalten. Die Auswirkungen von *Rater*-Schulungen verdienen nähere Betrachtung, um solche Schulungen möglichst effektiv zu gestalten. Die Erprobung und der Einsatz des Europäischen Sprachenportfolios sollten ebenfalls wissenschaftlich begleitet werden, um Hinweise auf lernfördernden Umgang mit diesem Instrument der Selbstbewertung zu erhalten.

Die Skalen des GER sollen zu vielfältigen Zwecken eingesetzt werden, doch momentan ist der Status der Skalen wie gesagt eher der von *reporting scales* – wenn jedoch die Skalen in all den Kontexten benutzt werden sollen, die im GER angegeben werden, so müssten (jenseits der empirischen Validierung der schon existenten Deskriptoren) all die dafür relevanten Gegenstände auf empirischer Basis in Formulierungen beschrieben werden, die den Zwecken angemessen sind. In diesem Zusammenhang wäre eine Erweiterung der dem GER-System zugrunde liegenden Deskriptorenbank hilfreich, sei es nun hinsichtlich der Aufnahme bisher nicht beschriebener Aspekte oder sei es hinsichtlich der Kennzeichnung und Offenlegung der jeweiligen Basis der Beschreibungen, um die Deskriptoren als valide Quellen für eine konkrete, zweckgerichtete Skalenkonstruktion anbieten zu können.

Um Test und Prüfungen an das System des GER anbinden zu können, bietet sich eine Erweiterung der Spezifikationen für die Bereiche jenseits des Lese- und Hörverstehens an, vergleichbar dem Projekt, aus dem der *Dutch Grid* resultiert. Diese Erweiterung müsste sich natürlich an die oben geforderte Analyse des GER-Systems anschließen.

Schließlich können Untersuchungen, wie der GER in verschiedenen Kontexten verwendet und seine Niveaus und Kompetenzbeschreibungen interpretiert werden, aufzeigen, wo es noch Verbesserungsbedarf gibt. An dieser Stelle darf auf die Fallstudien in Alderson (2002), auf die Pilotierung des *Manual*[289], auf die Homepage des *Dutch CEF Construct Project*[290] sowie auf das

[289] Vgl. http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual/default.asp#TopOfPage, Zugriff am 27.01.2005.
[290] Vgl. http://www.ealta.eu.org/dutch/grid.htm, Zugriff am 27.01.2005.

DIALANG-Projekt[291] und dessen Selbstbeurteilungssystem verwiesen werden. Die Anwendung des GER in der Praxis könnte von der Sprachlehrforschung begleitet werden, um Hinweise auf lernfördernde Maßnahmen zu erhalten.

Der GER ist ein Rahmen, der noch nicht perfekt ist, doch eine innovative und kreative Entwicklung stellt er allemal dar. Er kann Reflexionsanstöße geben; er kann helfen beim Aufbau eines gemeinsamen Verständnisses dessen, was Sprachvermögen ausmacht und wie dieses beschrieben werden kann; er kann Anstöße in der Beurteilung geben, indem die traditionelle Negativkorrektur ergänzt wird um Positivansätze, die jene Aspekte in den Mittelpunkt stellen, welche schon beherrscht werden. Er kann jedoch keine professionelle Ausbildung in den Bereichen des Lehrens oder Beurteilens von Sprachen ersetzen – vielmehr baut er auf vorhandenem Wissen auf.

Rückmeldungen aus der Praxis, aus der Verwendung des GER können helfen, diesen zu verbessern und seine Skalen zu verfeinern. Dazu lädt der GER auf S. 10 auch explizit ein: Die Autoren sind sich also durchaus bewusst gewesen, dass dieses Instrument noch der Weiterentwicklung bedarf. Bisher gibt es jedoch kein vergleichbares Instrument, das sich so intensiv mit Sprache, Lernen, Lehren und Beurteilen befasst – das Potential dieses Werkzeugs sollte demnach auch im Vordergrund stehen, und weniger die Mängel, die erst durch Praxiserfahrungen erkannt und beseitigt werden können. Bis der GER sinnvoll in den vielfältigen Kontexten eingesetzt werden kann, die seine Autoren vorschlagen, ist es noch ein weiter Weg, doch der Anfang ist gemacht. North & Schneider (1998: 243) wussten, dass die Implementierung des Referenzrahmens und seiner Skalen ein komplexes Unterfangen ist:

> Experience of a scale over a period of time, the relationship of the scale to levels used by publishers and schools, training with standardised performance examples, collection of collateral information from tests, analyses of rater behaviour (…):
> all these can contribute to effective implementation of a framework.

[291] Vgl. http://www.dialang.org, Zugriff am 03.02.2005.

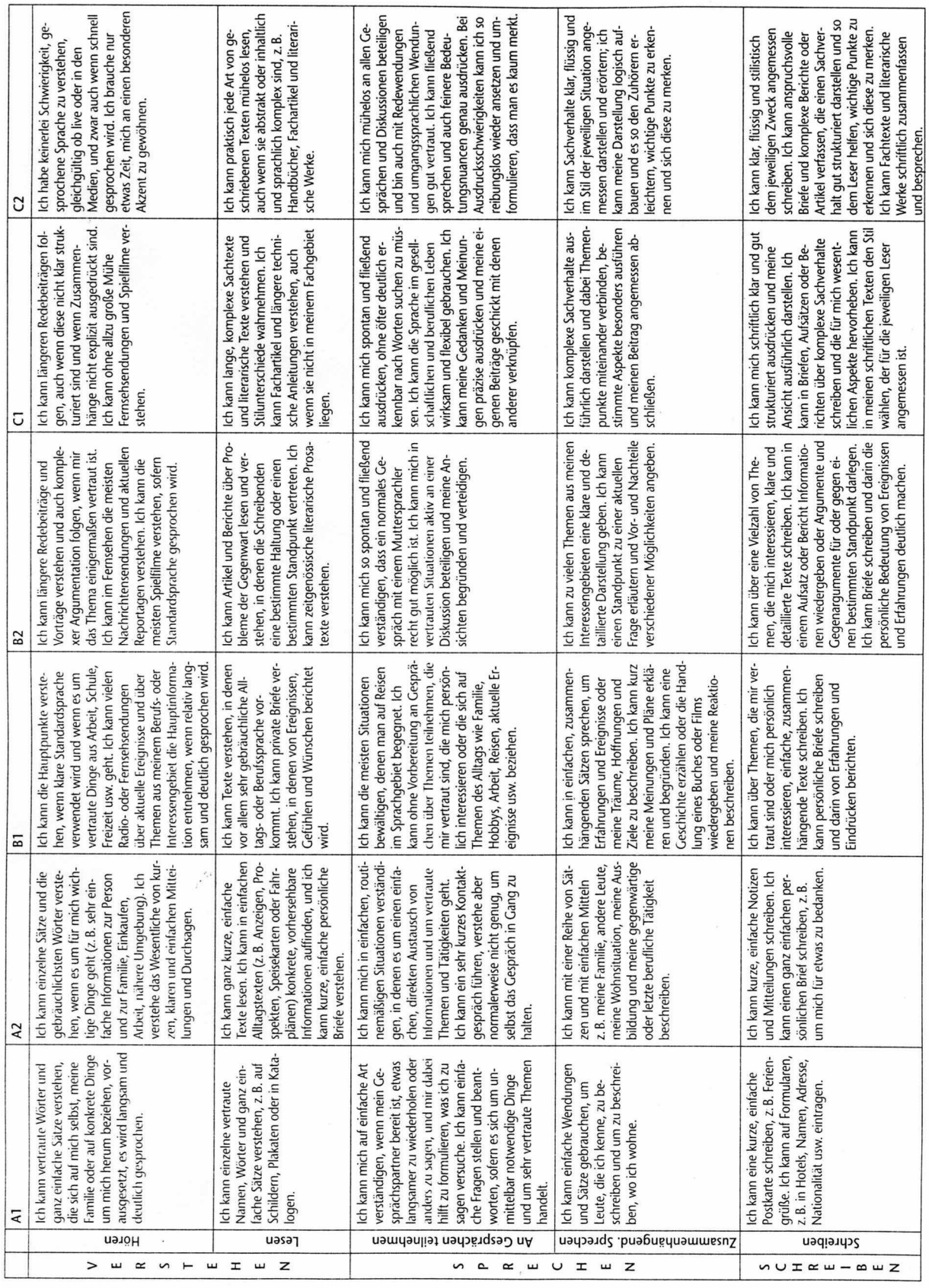
Anhänge

Anhang 1: *Globalskala*: GER (2001: 35)

Kompetente Sprachverwendung	**C2**	Kann praktisch alles, was er/sie liest oder hört, mühelos verstehen. Kann Informationen aus verschiedenen schriftlichen und mündlichen Quellen zusammenfassen und dabei Begründungen und Erklärungen in einer zusammenhängenden Darstellung wiedergeben. Kann sich spontan, sehr flüssig und genau ausdrücken und auch bei komplexeren Sachverhalten feinere Bedeutungsnuancen deutlich machen.
	C1	Kann ein breites Spektrum anspruchsvoller, längerer Texte verstehen und auch implizite Bedeutungen erfassen. Kann sich spontan und fließend ausdrücken, ohne öfter deutlich erkennbar nach Worten suchen zu müssen. Kann die Sprache im gesellschaftlichen und beruflichen Leben oder in Ausbildung und Studium wirksam und flexibel gebrauchen. Kann sich klar, strukturiert und ausführlich zu komplexen Sachverhalten äußern und dabei verschiedene Mittel zur Textverknüpfung angemessen verwenden.
Selbstständige Sprachverwendung	**B2**	Kann die Hauptinhalte komplexer Texte zu konkreten und abstrakten Themen verstehen; versteht im eigenen Spezialgebiet auch Fachdiskussionen. Kann sich so spontan und fließend verständigen, dass ein normales Gespräch mit Muttersprachlern ohne größere Anstrengung auf beiden Seiten gut möglich ist. Kann sich zu einem breiten Themenspektrum klar und detailliert ausdrücken, einen Standpunkt zu einer aktuellen Frage erläutern und die Vor- und Nachteile verschiedener Möglichkeiten angeben.
	B1	Kann die Hauptpunkte verstehen, wenn klare Standardsprache verwendet wird und wenn es um vertraute Dinge aus Arbeit, Schule, Freizeit usw. geht. Kann die meisten Situationen bewältigen, denen man auf Reisen im Sprachgebiet begegnet. Kann sich einfach und zusammenhängend über vertraute Themen und persönliche Interessengebiete äußern. Kann über Erfahrungen und Ereignisse berichten, Träume, Hoffnungen und Ziele beschreiben und zu Plänen und Ansichten kurze Begründungen oder Erklärungen geben.
Elementare Sprachverwendung	**A2**	Kann Sätze und häufig gebrauchte Ausdrücke verstehen, die mit Bereichen von ganz unmittelbarer Bedeutung zusammenhängen (z. B. Informationen zur Person und zur Familie, Einkaufen, Arbeit, nähere Umgebung). Kann sich in einfachen, routinemäßigen Situationen verständigen, in denen es um einen einfachen und direkten Austausch von Informationen über vertraute und geläufige Dinge geht. Kann mit einfachen Mitteln die eigene Herkunft und Ausbildung, die direkte Umgebung und Dinge im Zusammenhang mit unmittelbaren Bedürfnissen beschreiben.
	A1	Kann vertraute, alltägliche Ausdrücke und ganz einfache Sätze verstehen und verwenden, die auf die Befriedigung konkreter Bedürfnisse zielen. Kann sich und andere vorstellen und anderen Leuten Fragen zu ihrer Person stellen – z. B. wo sie wohnen, was für Leute sie kennen oder was für Dinge sie haben – und kann auf Fragen dieser Art Antwort geben. Kann sich auf einfache Art verständigen, wenn die Gesprächspartnerinnen oder Gesprächspartner langsam und deutlich sprechen und bereit sind zu helfen.

Tabelle 1 – Gemeinsame Referenzniveaus: Globalskala

Anhang 2: *Selbstbewertungsraster*: GER (2001: 36)



		A1	A2	B1	B2	C1	C2
VERSTEHEN	Hören	Ich kann vertraute Wörter und ganz einfache Sätze verstehen, die sich auf mich selbst, meine Familie oder auf konkrete Dinge um mich herum beziehen, vorausgesetzt, es wird langsam und deutlich gesprochen.	Ich kann einzelne Sätze und die gebräuchlichsten Wörter verstehen, wenn es um für mich wichtige Dinge geht (z. B. sehr einfache Informationen zur Person und zur Familie, Einkaufen, Arbeit, nähere Umgebung). Ich verstehe das Wesentliche von kurzen, klaren und einfachen Mitteilungen und Durchsagen.	Ich kann die Hauptpunkte verstehen, wenn klare Standardsprache verwendet wird und wenn es um vertraute Dinge aus Arbeit, Schule, Freizeit usw. geht. Ich kann vielen Radio- oder Fernsehsendungen über aktuelle Ereignisse und über Themen aus meinem Berufs- oder Interessengebiet die Hauptinformation entnehmen, wenn relativ langsam und deutlich gesprochen wird.	Ich kann längere Redebeiträge und Vorträge verstehen und auch komplexer Argumentation folgen, wenn mir das Thema einigermaßen vertraut ist. Ich kann im Fernsehen die meisten Nachrichtensendungen und aktuellen Reportagen verstehen. Ich kann die meisten Spielfilme verstehen, sofern Standardsprache gesprochen wird.	Ich kann längeren Redebeiträgen folgen, auch wenn diese nicht klar strukturiert sind und wenn Zusammenhänge nicht explizit ausgedrückt sind. Ich kann ohne allzu große Mühe Fernsehsendungen und Spielfilme verstehen.	Ich habe keinerlei Schwierigkeit, gesprochene Sprache zu verstehen, gleichgültig ob live oder in den Medien, und zwar auch wenn schnell gesprochen wird. Ich brauche nur etwas Zeit, mich an einen besonderen Akzent zu gewöhnen.
	Lesen	Ich kann einzelne vertraute Namen, Wörter und ganz einfache Sätze verstehen, z. B. auf Schildern, Plakaten oder in Katalogen.	Ich kann ganz kurze, einfache Texte lesen. Ich kann in einfachen Alltagstexten (z. B. Anzeigen, Prospekten, Speisekarten oder Fahrplänen) konkrete, vorhersehbare Informationen auffinden, und ich kann kurze, einfache persönliche Briefe verstehen.	Ich kann Texte verstehen, in denen vor allem sehr gebräuchliche Alltags- oder Berufssprache vorkommt. Ich kann private Briefe verstehen, in denen von Ereignissen, Gefühlen und Wünschen berichtet wird.	Ich kann Artikel und Berichte über Probleme der Gegenwart lesen und verstehen, in denen die Schreibenden eine bestimmte Haltung oder einen bestimmten Standpunkt vertreten. Ich kann zeitgenössische literarische Prosatexte verstehen.	Ich kann lange, komplexe Sachtexte und literarische Texte verstehen und Stilunterschiede wahrnehmen. Ich kann Fachartikel und längere technische Anleitungen verstehen, auch wenn sie nicht in meinem Fachgebiet liegen.	Ich kann praktisch jede Art von geschriebenen Texten mühelos lesen, auch wenn sie abstrakt oder inhaltlich und sprachlich komplex sind, z. B. Handbücher, Fachartikel und literarische Werke.
SPRECHEN	An Gesprächen teilnehmen	Ich kann mich auf einfache Art verständigen, wenn mein Gesprächspartner bereit ist, etwas langsamer zu wiederholen oder anders zu sagen, und mir dabei hilft zu formulieren, was ich zu sagen versuche. Ich kann einfache Fragen stellen und beantworten, sofern es sich um unmittelbar notwendige Dinge und um sehr vertraute Themen handelt.	Ich kann mich in einfachen, routinemäßigen Situationen verständigen, in denen es um einen einfachen, direkten Austausch von Informationen und um vertraute Themen und Tätigkeiten geht. Ich kann ein sehr kurzes Kontaktgespräch führen, verstehe aber normalerweise nicht genug, um selbst das Gespräch in Gang zu halten.	Ich kann die meisten Situationen bewältigen, denen man auf Reisen im Sprachgebiet begegnet. Ich kann ohne Vorbereitung an Gesprächen über Themen teilnehmen, die mir vertraut sind, die mich persönlich interessieren oder die sich auf Themen des Alltags wie Familie, Hobbys, Arbeit, Reisen, aktuelle Ereignisse usw. beziehen.	Ich kann mich so spontan und fließend verständigen, dass ein normales Gespräch mit einem Muttersprachler recht gut möglich ist. Ich kann mich in vertrauten Situationen aktiv an einer Diskussion beteiligen und meine Ansichten begründen und verteidigen.	Ich kann mich spontan und fließend ausdrücken, ohne öfter deutlich erkennbar nach Worten suchen zu müssen. Ich kann die Sprache im gesellschaftlichen und beruflichen Leben wirksam und flexibel gebrauchen. Ich kann meine Gedanken und Meinungen präzise ausdrücken und meine eigenen Beiträge geschickt mit denen anderer verknüpfen.	Ich kann mich mühelos an allen Gesprächen und Diskussionen beteiligen und bin auch mit Redewendungen und umgangssprachlichen Wendungen gut vertraut. Ich kann fließend sprechen und auch feinere Bedeutungsnuancen genau ausdrücken. Bei Ausdrucksschwierigkeiten kann ich so reibungslos wieder ansetzen und umformulieren, dass man es kaum merkt.
	Zusammenhängend. Sprechen	Ich kann einfache Wendungen und Sätze gebrauchen, um Leute, die ich kenne, zu beschreiben und um zu beschreiben, wo ich wohne.	Ich kann mit einer Reihe von Sätzen und mit einfachen Mitteln z. B. meine Familie, andere Leute, meine Wohnsituation, meine Ausbildung und meine gegenwärtige oder letzte berufliche Tätigkeit beschreiben.	Ich kann in einfachen, zusammenhängenden Sätzen sprechen, um Erfahrungen und Ereignisse oder meine Träume, Hoffnungen und Ziele zu beschreiben. Ich kann kurz meine Meinungen und Pläne erklären und begründen. Ich kann eine Geschichte erzählen oder die Handlung eines Buches oder Films wiedergeben und meine Reaktionen beschreiben.	Ich kann zu vielen Themen aus meinen Interessengebieten eine klare und detaillierte Darstellung geben. Ich kann einen Standpunkt zu einer aktuellen Frage erläutern und Vor- und Nachteile verschiedener Möglichkeiten angeben.	Ich kann komplexe Sachverhalte ausführlich darstellen und dabei Themenpunkte miteinander verbinden, bestimmte Aspekte besonders ausführen und meinen Beitrag angemessen abschließen.	Ich kann Sachverhalte klar, flüssig und im Stil der jeweiligen Situation angemessen darstellen und erörtern; ich kann meine Darstellung logisch aufbauen und es so den Zuhörern erleichtern, wichtige Punkte zu erkennen und sich diese zu merken.
SCHREIBEN	Schreiben	Ich kann eine kurze, einfache Postkarte schreiben, z. B. Feriengrüße. Ich kann auf Formularen, z. B. in Hotels, Namen, Adresse, Nationalität usw. eintragen.	Ich kann kurze, einfache Notizen und Mitteilungen schreiben. Ich kann einen ganz einfachen persönlichen Brief schreiben, z. B. um mich für etwas zu bedanken.	Ich kann über Themen, die mir vertraut sind oder mich persönlich interessieren, einfache, zusammenhängende Texte schreiben. Ich kann persönliche Briefe schreiben und darin von Erfahrungen und Eindrücken berichten.	Ich kann über eine Vielzahl von Themen, die mich interessieren, klare und detaillierte Texte schreiben. Ich kann in einem Aufsatz oder Bericht Informationen wiedergeben oder Argumente und Gegenargumente für oder gegen einen bestimmten Standpunkt darlegen. Ich kann Briefe schreiben und darin die persönliche Bedeutung von Ereignissen und Erfahrungen deutlich machen.	Ich kann mich schriftlich klar und gut strukturiert ausdrücken und meine Ansicht ausführlich darstellen. Ich kann in Briefen, Aufsätzen oder Berichten über komplexe Sachverhalte schreiben und die für mich wesentlichen Aspekte hervorheben. Ich kann in meinen schriftlichen Texten den Stil wählen, der für die jeweiligen Leser angemessen ist.	Ich kann klar, flüssig und stilistisch dem jeweiligen Zweck angemessen schreiben. Ich kann anspruchsvolle Briefe und komplexe Berichte oder Artikel verfassen, die einen Sachverhalt gut strukturiert darstellen und so dem Leser helfen, wichtige Punkte zu erkennen und sich diese zu merken. Ich kann Fachtexte und literarische Werke schriftlich zusammenfassen und besprechen.

Anhang 3: *Beurteilungsraster mündliche Kommunikation*: GER (2001: 37)

	Spektrum	Korrektheit	Flüssigkeit	Interaktion	Kohärenz
C2	Zeigt viel Flexibilität, Gedanken mit verschiedenen sprachlichen Mitteln zu formulieren, um feinere Bedeutungsnuancen deutlich zu machen oder um etwas hervorzuheben, zu differenzieren oder um Mehrdeutigkeit zu beseitigen. Verfügt auch über gute Kenntnisse umgangssprachlicher und idiomatischer Wendungen.	Zeigt auch bei der Verwendung komplexer Sprachmittel eine durchgehende Beherrschung der Grammatik, selbst wenn die Aufmerksamkeit anderweitig beansprucht wird (z. B. durch vorausblickendes Planen oder Konzentration auf die Reaktionen anderer).	Kann sich spontan und mit natürlichem Sprachfluss in längeren Redebeiträgen äußern und dabei Schwierigkeiten so glatt umgehen oder neu ansetzen, dass die Gesprächspartner es kaum merken.	Kann sich leicht und gewandt verständigen, wobei er/sie auch Mittel der Intonation und nicht-sprachliche Mittel offenbar mühelos registriert und verwendet. Kann eigene Redebeiträge ins Gespräch einflechten, indem er/sie ganz natürlich das Wort ergreift, auf etwas Bezug nimmt, Anspielungen macht usw.	Kann kohärente, zusammenhängende Redebeiträge machen; verwendet dabei in angemessener Weise unterschiedliche Mittel zur Gliederung sowie ein breites Spektrum von Verknüpfungsmitteln.
C1	Verfügt über ein breites Spektrum von Redemitteln, aus dem er/sie geeignete Formulierungen auswählen kann, um sich klar und angemessen über ein breites Spektrum allgemeiner, wissenschaftlicher, beruflicher Themen oder über Freizeitthemen zu äußern, ohne sich in dem, was er/sie sagen möchte, einschränken zu müssen.	Behält durchgehend ein hohes Maß an grammatischer Korrektheit; Fehler sind selten, fallen kaum auf und werden in der Regel selbst korrigiert.	Kann sich beinahe mühelos spontan und fließend ausdrücken; nur begrifflich schwierige Themen können den natürlichen Sprachfluss beeinträchtigen.	Kann aus einem ohne weiteres verfügbaren Repertoire von Diskursmitteln eine geeignete Wendung auswählen, um seine/ihre Äußerung angemessen einzuleiten, wenn er/sie das Wort ergreifen oder behalten will, oder um die eigenen Beiträge geschickt mit denen anderer Personen zu verbinden.	Kann klar, sehr fließend und gut strukturiert sprechen und zeigt, dass er/sie Gliederungs- und Verknüpfungsmittel beherrscht.
B2+					
B2	Verfügt über ein ausreichend breites Spektrum von Redemitteln, um in klaren Beschreibungen oder Berichten über die meisten Themen allgemeiner Art zu sprechen und eigene Standpunkte auszudrücken; sucht nicht auffällig nach Worten und verwendet einige komplexe Satzstrukturen.	Zeigt eine recht gute Beherrschung der Grammatik. Macht keine Fehler, die zu Missverständnissen führen, und kann die meisten eigenen Fehler selbst korrigieren.	Kann in recht gleichmäßigem Tempo sprechen. Auch wenn er/sie eventuell zögert, um nach Strukturen oder Wörtern zu suchen, entstehen kaum auffällig lange Pausen.	Kann Gespräche beginnen, die Sprecherrolle übernehmen, wenn es angemessen ist, und das Gespräch beenden, wenn er/sie möchte, auch wenn das möglicherweise nicht immer elegant gelingt. Kann auf vertrautem Gebiet zum Fortgang des Gesprächs beitragen, indem er/sie das Verstehen bestätigt, andere zum Sprechen auffordert usw.	Kann eine begrenzte Anzahl von Verknüpfungsmitteln verwenden, um seine/ihre Äußerungen zu einem klaren, zusammenhängenden Beitrag zu verbinden; längere Beiträge sind möglicherweise etwas sprunghaft.

Anhang 4: Skala *Texte verarbeiten*: GER (2001: 98)

	Texte verarbeiten
C2	Kann Informationen aus verschiedenen Quellen zusammenfassen und die Argumente und berichteten Sachverhalte so wiedergeben, dass insgesamt eine kohärente Darstellung entsteht.
C1	Kann lange, anspruchsvolle Texte zusammenfassen.
B2	Kann ein breites Spektrum von Sachtexten und fiktiven Texten zusammenfassen und dabei die Hauptthemen und unterschiedliche Standpunkte kommentieren und diskutieren. Kann Auszüge aus Nachrichten, Interviews oder Reportagen, welche Stellungnahmen, Erörterungen und Diskussionen enthalten, zusammenfassen. Kann die Handlung und die Abfolge der Ereignisse in einem Film oder Theaterstück zusammenfassen.
B1	Kann kurze Informationen aus mehreren Quellen zusammenführen und für jemand anderen zusammenfassen.
	Kann kurze Textpassagen auf einfache Weise zusammenfassen, indem er/sie dabei den Wortlaut und die Anordnung des Originals benutzt.
A2	Kann im Rahmen seiner/ihrer Erfahrungen und begrenzten Kompetenz aus einem kurzen Text Schlüsselwörter, Wendungen und kurze Sätze heraussuchen und wiedergeben.
	Kann kurze Texte in Druckschrift oder klarer Handschrift abschreiben.
A1	Kann einzelne Wörter und kurze Texte, die in gedruckter Form vorliegen, abschreiben.

Anhang 5: Skala *Schriftliche Produktion*: GER (2001: 67)

	Schriftliche Produktion allgemein
C2	Kann klare, flüssige, komplexe Texte in angemessenem und effektivem Stil schreiben, deren logische Struktur den Lesern das Auffinden der wesentlichen Punkte erleichtert.
C1	Kann klare, gut strukturierte Texte zu komplexen Themen verfassen und dabei die entscheidenden Punkte hervorheben, Standpunkte ausführlich darstellen und durch Unterpunkte oder geeignete Beispiele oder Begründungen stützen und den Text durch einen angemessenen Schluss abrunden.
B2	Kann klare, detaillierte Texte zu verschiedenen Themen aus seinem/ihrem Interessengebiet verfassen und dabei Informationen und Argumente aus verschiedenen Quellen zusammenführen und gegeneinander abwägen.
B1	Kann unkomplizierte, zusammenhängende Texte zu mehreren vertrauten Themen aus seinem/ihrem Interessengebiet verfassen, wobei einzelne kürzere Teile in linearer Abfolge verbunden werden.
A2	Kann eine Reihe einfacher Wendungen und Sätze schreiben und mit Konnektoren wie *und*, *aber* oder *weil* verbinden.
A1	Kann einfache, isolierte Wendungen und Sätze schreiben.

Anhang 6: Skala *Kreatives Schreiben*: GER (2001: 67f)

	Kreatives Schreiben
C2	Kann klare, flüssige und fesselnde Geschichten und Beschreibungen von Erfahrungen verfassen, und zwar in einem Stil, der dem gewählten Genre angemessenen ist.
C1	Kann klare, detaillierte, gut strukturierte und ausführliche Beschreibungen oder auch eigene fiktionale Texte in lesergerechtem, überzeugendem, persönlichem und natürlichem Stil verfassen.
B2	Kann klare, detaillierte, zusammenhängende Beschreibungen realer oder fiktiver Ereignisse und Erfahrungen verfassen dabei den Zusammenhang zwischen verschiedenen Ideen deutlich machen und die für das betreffende Genre geltenden Konventionen beachten.
	Kann klare, detaillierte Beschreibungen zu verschiedenen Themen aus seinem/ihrem Interessengebiet verfassen. Kann eine Rezension eines Films, Buchs oder Theaterstücks schreiben.
B1	Kann unkomplizierte, detaillierte Beschreibungen zu einer Reihe verschiedener Themen aus seinem/ihrem Interessengebiet verfassen. Kann Erfahrungsberichte schreiben, in denen Gefühle und Reaktion in einem einfachen, zusammenhängenden Text beschrieben werden. Kann eine Beschreibung eines realen oder fiktiven Ereignisses oder einer kürzlich unternommenen Reise verfassen. Kann eine Geschichte erzählen.
A2	Kann in Form verbundener Sätze etwas über alltägliche Aspekte des eigenen Umfelds schreiben, wie z. B. über Menschen, Orte, einen Job oder Studienerfahrungen. Kann eine sehr kurze, elementare Beschreibung von Ereignissen, vergangenen Handlungen und persönlichen Erfahrungen verfassen.
	Kann in einer Reihe einfacher Sätze über die eigene Familie, die Lebensumstände, den Bildungshintergrund oder die momentane oder vorige berufliche Tätigkeit schreiben. Kann kurze, einfache, fiktive Biographien und einfache Gedichte über Menschen schreiben.
A1	Kann einfache Wendungen und Sätze über sich selbst und fiktive Menschen schreiben: wo sie leben und was sie tun.

Anhang 7: Skala *Briefe und Aufsätze schreiben*: GER (2001: 68)

	Berichte und Aufsätze schreiben
C2	Kann klare, flüssige, komplexe Berichte, Artikel oder Aufsätze verfassen, in denen ein Argument entwickelt oder ein Vorschlag oder ein literarisches Werk kritisch gewürdigt wird. Kann den Texten einen angemessenen, effektiven logischen Aufbau geben, der den Lesenden hilft, die wesentlichen Punkte zu finden.
C1	Kann klare, gut strukturierte Ausführungen zu komplexen Themen schreiben und dabei zentrale Punkte hervorheben. Kann Standpunkte ausführlich darstellen und durch Unterpunkte, geeignete Beispiele oder Begründungen stützen.
B2	Kann einen Aufsatz oder Bericht schreiben, in dem etwas systematisch erörtert wird, wobei entscheidende Punkte angemessen hervorgehoben und stützende Details angeführt werden. Kann verschiedene Ideen oder Problemlösungen gegeneinander abwägen.
	Kann in einem Aufsatz oder Bericht etwas erörtern, dabei Gründe für oder gegen einen bestimmten Standpunkt angeben und die Vor- und Nachteile verschiedener Optionen erläutern. Kann Informationen und Argumente aus verschiedenen Quellen zusammenführen.
B1	Kann einen kurzen, einfachen Aufsatz zu Themen von allgemeinem Interesse schreiben. Kann im eigenen Sachgebiet mit einer gewissen Sicherheit größere Mengen von Sachinformationen über vertraute Routineangelegenheiten und über weniger routinemäßige Dinge zusammenfassen, darüber berichten und dazu Stellung nehmen.
	Kann in einem üblichen Standardformat sehr kurze Berichte schreiben, in denen Sachinformationen weitergegeben und Gründe für Handlungen angegeben werden.
A2	Keine Deskriptoren verfügbar
A1	Keine Deskriptoren verfügbar

Anhang 8: Skala *Themenentwicklung*: GER (2001: 125)

	Themenentwicklung
C2	Wie C1
C1	Kann etwas ausführlich beschreiben oder berichten und dabei Themenpunkte miteinander verbinden, einzelne Aspekte besonders ausführen und mit einer geeigneten Schlussfolgerung abschließen.
B2	Kann etwas klar beschreiben oder erzählen und dabei wichtige Aspekte ausführen und mit relevanten Details und Beispielen stützen.
B1	Kann recht flüssig unkomplizierte Geschichten oder Beschreibungen wiedergeben, indem er/sie die einzelnen Punkte linear aneinander reiht.
A2	Kann eine Geschichte erzählen oder etwas beschreiben, indem er/sie die einzelnen Punkte in Form einer einfachen Aufzählung aneinander reiht.
A1	Keine Deskriptoren verfügbar

Anhang 9: Skala *Orthographie*: GER (2001: 118)

	Beherrschung der Orthographie
C2	Die schriftlichen Texte sind frei von orthographischen Fehlern.
C1	Die Gestaltung, die Gliederung in Absätze und die Zeichensetzung sind konsistent und hilfreich. Die Rechtschreibung ist, abgesehen von gelegentlichem Verschreiben, richtig.
B2	Kann zusammenhängend und klar verständlich schreiben und dabei die üblichen Konventionen der Gestaltung und der Gliederung in Absätze einhalten. Rechtschreibung und Zeichensetzung sind hinreichend korrekt, können aber Einflüsse der Muttersprache zeigen.
B1	Kann zusammenhängend schreiben; die Texte sind durchgängig verständlich. Rechtschreibung, Zeichensetzung und Gestaltung sind exakt genug, so dass man sie meistens verstehen kann.
A2	Kann kurze Sätze über alltägliche Themen abschreiben – z. B. Wegbeschreibungen. Kann kurze Wörter aus seinem mündlichen Wortschatz 'phonetisch' einigermaßen akkurat schriftlich wiedergeben (benutzt dabei aber nicht notwendigerweise die übliche Rechtschreibung).
A1	Kann vertraute Wörter und kurze Redewendungen, z. B. einfache Schilder oder Anweisungen, Namen alltäglicher Gegenstände, Namen von Geschäften oder regelmäßig benutzte Wendungen abschreiben. Kann seine Adresse, seine Nationalität und andere Angaben zur Person buchstabieren.

Anhang 10: Skala *Wortschatzspektrum*: GER (2001: 112)

	Wortschatzspektrum
C2	Beherrscht einen sehr reichen Wortschatz einschließlich umgangssprachlicher und idiomatischer Wendungen und ist sich der jeweiligen Konnotationen bewusst.
C1	Beherrscht einen großen Wortschatz und kann bei Wortschatzlücken problemlos Umschreibungen gebrauchen; offensichtliches Suchen nach Worten oder der Rückgriff auf Vermeidungsstrategien sind selten. Gute Beherrschung idiomatischer Ausdrücke und umgangssprachlicher Wendungen.
B2	Verfügt über einen großen Wortschatz in seinem Sachgebiet und in den meisten allgemeinen Themenbereichen. Kann Formulierungen variieren, um häufige Wiederholungen zu vermeiden; Lücken im Wortschatz können dennoch zu Zögern und Umschreibungen führen.
B1	Verfügt über einen ausreichend großen Wortschatz, um sich mit Hilfe von einigen Umschreibungen über die meisten Themen des eigenen Alltagslebens äußern zu können wie beispielsweise Familie, Hobbys, Interessen, Arbeit, Reisen, aktuelle Ereignisse.
A2	Verfügt über einen ausreichenden Wortschatz, um in vertrauten Situationen und in Bezug auf vertraute Themen routinemäßige, alltägliche Angelegenheiten zu erledigen.
	Verfügt über genügend Wortschatz, um elementaren Kommunikationsbedürfnissen gerecht werden zu können. Verfügt über genügend Wortschatz, um einfache Grundbedürfnisse befriedigen zu können.
A1	Verfügt über einen elementaren Vorrat an einzelnen Wörtern und Wendungen, die sich auf bestimmte konkrete Situationen beziehen.

Anhang 11: Skala *Wortschatzbeherrschung*: GER (2001: 112)

	Wortschatzbeherrschung
C2	Durchgängig korrekte und angemessene Verwendung des Wortschatzes.
C1	Gelegentliche kleinere Schnitzer, aber keine größeren Fehler im Wortgebrauch.
B2	Die Genauigkeit in der Verwendung des Wortschatzes ist im Allgemeinen groß, obgleich einige Verwechslungen und falsche Wortwahl vorkommen, ohne jedoch die Kommunikation zu behindern.
B1	Zeigt eine gute Beherrschung des Grundwortschatzes, macht aber noch elementare Fehler, wenn es darum geht, komplexere Sachverhalte auszudrücken oder wenig vertraute Themen und Situationen zu bewältigen.
A2	Beherrscht einen begrenzten Wortschatz in Zusammenhang mit konkreten Alltagsbedürfnissen.
A1	Keine Deskriptoren verfügbar

Anhang 12: Skala *Grammatische Korrektheit*: GER (2001: 114)

	Grammatische Korrektheit
C2	Zeigt auch bei der Verwendung komplexer Sprachmittel eine durchgehende Beherrschung der Grammatik, selbst wenn die Aufmerksamkeit anderweitig beansprucht wird (z. B. durch vorausblickendes Planen oder Konzentration auf die Reaktionen anderer).
C1	Kann beständig ein hohes Maß an grammatischer Korrektheit beibehalten; Fehler sind selten und fallen kaum auf.
B2	Gute Beherrschung der Grammatik; gelegentliche Ausrutscher oder nicht-systematische Fehler und kleinere Mängel im Satzbau können vorkommen, sind aber selten und können oft rückblickend korrigiert werden.
	Gute Beherrschung der Grammatik; macht keine Fehler, die zu Missverständnissen führen.
B1	Kann sich in vertrauten Situationen ausreichend korrekt verständigen; im Allgemeinen gute Beherrschung der grammatischen Strukturen trotz deutlicher Einflüsse der Muttersprache. Zwar kommen Fehler vor, aber es bleibt klar, was ausgedrückt werden soll.
	Kann ein Repertoire von häufig verwendeten Redefloskeln und von Wendungen, die an eher vorhersehbare Situationen gebunden sind, ausreichend korrekt verwenden.
A2	Kann einige einfache Strukturen korrekt verwenden, macht aber noch systematisch elementare Fehler, hat z. B. die Tendenz, Zeitformen zu vermischen oder zu vergessen, die Subjekt-Verb-Kongruenz zu markieren; trotzdem wird in der Regel klar, was er/sie ausdrücken möchte.
A1	Zeigt nur eine begrenzte Beherrschung einiger weniger einfacher grammatischer Strukturen und Satzmuster in einem auswendig gelernten Repertoire.

Anhang 13: Skala *Kohärenz und Kohäsion*: GER (2001: 112)

	Kohärenz und Kohäsion
C2	Kann einen gut gegliederten und zusammenhängenden Text erstellen und dabei eine Vielfalt an Mitteln für die Gliederung und Verknüpfung angemessen einsetzen.
C1	Kann klar, sehr fließend und gut strukturiert sprechen und zeigt, dass er/sie die Mittel der Gliederung sowie der inhaltlichen und sprachlichen Verknüpfung beherrscht.
B2	Kann verschiedene Verknüpfungswörter sinnvoll verwenden, um inhaltliche Beziehungen deutlich zu machen.
	Kann eine begrenzte Anzahl von Verknüpfungsmitteln verwenden, um seine/ihre Äußerungen zu einem klaren, zusammenhängenden Text zu verbinden; längere Beiträge sind möglicherweise etwas sprunghaft.
B1	Kann eine Reihe kurzer und einfacher Einzelelemente zu einer linearen, zusammenhängenden Äußerung verbinden.
A2	Kann die häufigsten Konnektoren benutzen, um einfache Sätze miteinander zu verbinden, um eine Geschichte zu erzählen oder etwas in Form einer einfachen Aufzählung zu beschreiben.
	Kann Wortgruppen durch einfache Konnektoren wie *und*, *aber* und *weil* verknüpfen.
A1	Kann Wörter oder Wortgruppen durch sehr einfache Konnektoren wie *und* oder *dann* verbinden.

Anhang 14: Tabellen zur Analyse der *Globalskala* (GER 2001: 35)

A1:

Operation	Was/Wie	Einschränkungen	Situationen	Themen
verstehen und verwenden	Ausdrücke; Sätze	vertraut, ganz einfach, konkret	Befriedigung konkreter Bedürfnisse	
vorstellen; (Fragen) stellen, (Antwort) geben	sich und Andere; Fragen und Antwort zur Person		Vorstellen von Personen	wo man wohnt; was für Leute man kennt; was für Dinge man hat
sich verständigen	auf einfache Art	Partner sprechen langsam und deutlich und sind bereit zu helfen		

A2:

Operation	Was/Wie	Einschränkungen	Situationen	Themen
verstehen	Sätze und Ausdrücke	häufig gebraucht, von *ganz* unmittelbarer Bedeutung	Bereiche von *ganz* unmittelbarer Bedeutung	Informationen zur Person, Familie; Einkaufen, Arbeit, *nähere* Umgebung
sich verständigen	Austausch von Informationen	*einfach*, direkt, routinemäßig, vertraut, geläufig	einfache, routinemäßige Situationen	vertraute, geläufige Dinge
beschreiben		mit einfachen Mitteln	im Zusammenhang mit unmittelbaren Bedürfnissen	eigene Herkunft und Ausbildung, *direkte* Umgebung, Dinge im Zusammenhang mit unmittelbaren Bedürfnissen

B1:

Operation	Was/Wie	Einschränkungen	Situationen	Themen
verstehen	Hauptpunkte	klare Standardsprache, *vertraute* Dinge		Arbeit, Schule, Freizeit usw.
bewältigen		die *meisten* Situationen	auf *Reisen* im Sprachgebiet	
äußern	*einfach* und zusammenhängend	*einfach* und zusammenhängend zu Vertrautem		vertraute Themen und persönliche Interessensgebiete
berichten; beschreiben; geben	Begründungen, Erklärungen	kurz		Erfahrungen, Ereignisse; Träume, Hoffnungen, Ziele; zu Plänen und Ansichten

B2:

Operation	Was/Wie	Einschränkungen	Situationen	Themen
verstehen	Hauptinhalte komplexer Texte Fachdiskussionen	 im eigenen Spezialgebiet		konkrete und abstrakte Themen
verständigen	*spontan* und fließend	ohne größere Anstrengung auf beiden Seiten	*normales* Gespräch mit Muttersprachlern	
ausdrücken; erläutern; angeben	klar und detailliert; Standpunkt; Vor-, Nachteile			breites Spektrum; aktuelle Frage;

C1:

Operation	Was/Wie	Einschränkungen	Situationen	Themen
verstehen; erfassen	anspruchsvolle, längere Texte implizite Bedeutungen			
ausdrücken	*spontan*, fließend	ohne deutlich/öfter nach Worten suchen zu müssen		
Sprachgebrauch	wirksam und flexibel		im gesellschaftlichen oder beruflichen Leben, in Ausbildung und Studium	
äußern; anwenden	klar, strukturiert, ausführlich verschiedene Mittel der Textverknüpfung			komplexe Sachverhalte

C2:

Operation	Was/Wie	Einschränkungen	Situationen	Themen
verstehen	alles, was gelesen, gehört wird; mühelos			alles
zusammenfassen; wiedergeben	Informationen aus verschiedenen schriftlichen, mündlichen Quellen Begründungen, Erklärungen in zusammenhängender Darstellung			
ausdrücken; deutlich machen	*spontan*, sehr flüssig, genau; feinere Bedeutungsnuancen			 auch bei komplexeren Sachverhalten

Anhang 15: Tabellen zur Analyse der Skalen aus dem Bereich der kommunikativen Aktivitäten: *Schriftliche Produktion* (SP, GER 2001: 67), *Kreatives Schreiben* (KS, ebd.), *Briefe und Aufsätze schreiben* (BA, ebd.: 68), *Schreiben* (S, aus Selbstevaluationsraster, ebd.: 36)

A1:

Skala	Operation	Was/Wie	Einschränkungen	Themen
SP	schreiben	Wendungen und Sätze	einfach, isoliert	
KS	schreiben	Wendungen und Sätze	einfach, isoliert	über sich selbst und fiktive Menschen: wo sie leben und was sie tun.
BA	-	-	-	-
S	schreiben	Postkarte	kurz, einfach	z. B. Feriengrüße

A2:

Skala	Operation	Was/Wie	Einschränkungen	Themen
SP	schreiben verbinden	eine Reihe Wendungen und Sätze mit *Konnektoren*	einfache wie ‚und', ‚aber' oder ‚weil'	
KS	schreiben schreiben schreiben verfassen	eine Reihe Sätze Biographien und *Gedichte* in Form verbundener Sätze Beschreibung	einfach kurz, einfach, fiktiv einfach *sehr* kurz, elementar	über die eigene Familie, die Lebensumstände, den Bildungshintergrund oder die momentane oder vorige berufliche Tätigkeit über Menschen über alltägliche Aspekte des eigenen Umfelds, wie z. B. über Menschen, Orte, einen Job oder Studienerfahrungen von Ereignissen, vergangenen Handlungen und persönlichen Erfahrungen
BA	-	-	-	-
S	schreiben schreiben	Notizen und Mitteilungen Brief	kurz, einfach *ganz* einfach, persönlich	 z. B. um sich für etwas zu bedanken

B1:

Skala	Operation	Was/Wie	Einschränkungen	Themen
SP	verfassen	zusammenhängende Texte	unkompliziert; einzelne kürzere Teile in linearer Abfolge verbunden	zu mehreren vertrauten Themen aus seinem/ihrem Interessengebiet
KS	verfassen	*detaillierte* Beschreibungen	unkompliziert	zu einer Reihe verschiedener Themen aus seinem/ihrem Interessengebiet
	(be)schreiben	zusammenhängende Erfahrungsberichte	einfach	Gefühle und Reaktion
	verfassen	Beschreibung		eines realen oder fiktiven Ereignisses; einer kürzlich unternommenen Reise
	erzählen	Geschichte		
BA	schreiben – dabei:	Berichte in üblichem Standardformat	*sehr kurz*	
	weitergeben angeben			Sachinformationen *Gründe* für Handlungen
	schreiben – dabei:	Aufsatz	*kurz*, einfach	zu Themen von allgemeinem Interesse
	zusammenfassen, darüber berichten und dazu Stellung nehmen	größere Mengen von Sachinformationen	im eigenen Sachgebiet mit einer gewissen Sicherheit	über vertraute Routineangelegenheiten und über *weniger* routinemäßige Dinge
S	schreiben	zusammenhängende Texte	einfach	zu vertrauten Themen/aus Interessengebiet
	schreiben berichten	persönliche Briefe darin [in Briefen]		von Eindrücken, Erfahrungen

B2:

Skala	Operation	Was	Wie	Themen
SP	verfassen – dabei:	Texte	klar, detailliert	zu verschiedenen Themen aus seinem/ihrem Interessengebiet
	zusammenführen, (gegeneinander) abwägen	Informationen und Argumente aus verschiedenen Quellen		
KS B2-	verfassen	Beschreibungen	klar, detailliert	zu verschiedenen Themen aus seinem/ ihrem Interessengebiet
	erstellen	Rezension		zu Film, Buch, Theaterstück
KS B2+	verfassen – dabei:	Beschreibungen	klar, detailliert, *zusammenhängend*	reale oder fiktive Ereignisse und Erfahrungen
	deutlich machen	Zusammenhang zwischen verschiedenen Ideen		
	beachten	die für das betreffende Genre geltenden Konventionen		

Fortsetzung B2

BA B2-	erörtern – dabei: angeben erläutern zusammenführen	in einem Aufsatz oder Bericht Gründe für oder gegen einen bestimmten Standpunkt Vor- und Nachteile verschiedener Optionen Informationen und Argumente aus verschiedenen Quellen		
BA B2+	schreiben – dabei: erörtern – wobei: hervorheben anführen; abwägen	einen Aufsatz oder Bericht (etwas) entscheidende Punkte, stützende Details; verschiedene Ideen oder Problemlösungen	systematisch angemessen	
S	schreiben; wiedergeben darlegen; schreiben – dabei: deutlich machen	Texte Informationen in einem Aufsatz, Bericht (Gegen)Argumente, Standpunkte; Briefe	klar, detailliert;	über Vielzahl von Themen von persönlichem Interesse; persönliche Bedeutung von Ereignissen und Erfahrungen

C1:

Skala	Operation	Was	Wie	Themen
SP	verfassen – dabei: hervorheben darstellen – diese: stützen abrunden	Texte entscheidende Punkte Standpunkte, (den Text) durch Schluss	*klar*, gut strukturiert ausführlich durch Beispiele angemessen	zu komplexen Themen
KS	verfassen	Beschreibungen oder auch eigene fiktionale Texte	*klar, detailliert*, gut strukturiert und ausführlich in lesergerechtem, überzeugendem, persönlichem und natürlichem Stil	
BA	schreiben – dabei: hervorheben; darstellen – diese: stützen	Ausführungen zentrale Punkte; Standpunkte	*klar*, gut strukturiert ausführlich durch Unterpunkte, geeignete Beispiele oder Begründungen	zu komplexen Themen;
S	(schriftlich) ausdrücken darstellen; schreiben – dabei: hervorheben; wählen	Ansicht; Briefe, Aufsätze oder Berichte wesentliche Aspekte; Stil in schriftlichen Texten	*klar*, gut strukturiert ausführlich; angemessen für die jeweiligen *Leser*	über komplexe Sachverhalte

C2:

Skala	Operation	Was	Wie	Themen
SP	schreiben	Texte Stil Struktur	*klar*, flüssig, komplex; angemessen, effektiv; logisch – erleichtert Lesern das Auffinden der wesentlichen Punkte	
KS	verfassen	Geschichten und Beschreibungen von Erfahrungen; Stil	*klar*, flüssig, fesselnd dem gewählten Genre angemessen	
BA	verfassen – dabei: entwickeln würdigen geben	Bericht, Artikel, Aufsatz Argument Vorschlag oder literarisches Werk (dem Text) Aufbau	*klar*, flüssig, komplex kritisch angemessen, effektiv, logisch – hilft Lesenden, die wesentlichen Punkte zu finden	Literatur
S	schreiben; verfassen – dabei: darstellen; schriftlich zusammenfassen und besprechen	Briefe, Berichte, Artikel Sachverhalte; Fachtexte und literarische Werke	*klar*, flüssig, stilistisch dem Zweck angemessen; anspruchsvoll, komplex gut strukturiert – hilft Leser, wichtige Punkte zu erkennen und sich zu merken;	Fachtexte und Literatur

Anhang 16: Tabellen zur Analyse der Skalen aus dem Bereich der sprachlichen Kompetenzen: Skala *Orthographie* (GER 2001: 118)

A1:

Operation	Was/Wie	Einschränkungen/ Bedingungen	konkrete Beispiele	Themen
abschreiben	Wörter und Redewendungen	kurz, vertraut	einfache Schilder oder Anweisungen, Namen alltäglicher Gegenstände, Namen von Geschäften oder regelmäßig benutzte Wendungen	
buchstabieren	Adresse, Nationalität, andere Angaben zur Person			

A2:

Operation	Was/Wie	Einschränkungen/Bedingungen	konkrete Beispiele	Themen
abschreiben	Sätze	kurz	Wegbeschreibungen	alltägliche Themen
(schriftlich) wiedergeben	Wörter aus mündlichem Wortschatz	kurze Wörter, 'phonetisch' einigermaßen akkurat		
benutzt		nicht notwendigerweise übliche Rechtschreibung		

B1:

Operation	Was/Wie	Einschränkungen/Bedingungen	konkrete Beispiele	Themen
schreiben	*zusammenhängend*	*Texte durchgängig verständlich*		
	Rechtschreibung, Zeichensetzung und *Gestaltung*	exakt genug, so dass man sie *meistens* verstehen kann		

B2:

Operation	Was/Wie	Einschränkungen/Bedingungen	konkrete Beispiele	Themen
schreiben	*zusammenhängend und klar verständlich*	klar		
einhalten	übliche Konventionen der Gestaltung und der *Gliederung in Absätze*	üblich		
	Rechtschreibung und Zeichensetzung	hinreichend korrekt, können aber Einflüsse der Muttersprache zeigen		

C1:

Operation	Was/Wie	Einschränkungen/Bedingungen	konkrete Beispiele	Themen
	Gestaltung, *Gliederung in Absätze,* Zeichensetzung	konsistent und hilfreich		
	Rechtschreibung	richtig, abgesehen von gelegentlichem Verschreiben		

C2:

Operation	Was/Wie	Einschränkungen/Bedingungen	konkrete Beispiele	Themen
	schriftliche Texte	frei von orthographischen Fehlern		

Anhang 17: Tabellen zur Analyse der Skalen aus dem Bereich der sprachlichen Kompetenzen: Skalen *Wortschatzspektrum* (WS) und *Wortschatzbeherrschung* (WB) (GER 2001: 112f)

A1:

Skala	Operation	Was/Wie	Einschränkungen/Bedingungen	Situationen	Themen
WS	verfügt	einzelne Wörter und Wendungen	elementarer Vorrat	bezogen auf bestimmte konkrete Situationen	
WB					

A2:

Skala	Operation	Was/Wie	Einschränkungen/Bedingungen	Situationen	Themen
WS A2-	verfügt (über) (um zu) erledigen	Wortschatz Angelegenheiten	*ausreichend* routinemäßige, alltägliche	in vertrauten Situationen	in Bezug auf vertraute Themen
WS A2+	verfügt (über) (um) gerecht zu werden befriedigen	Wortschatz	*genügend*	elementare Kommunikationsbedürfnisse elementare Grundbedürfnisse	
WB	beherrscht	Wortschatz	*begrenzt*	konkrete Alltagsbedürfnisse	

B1:

Skala	Operation	Was/Wie	Einschränkungen/Bedingungen	Situationen	Themen
WS	verfügt (über) (um sich zu) äußern	Wortschatz	ausreichend groß; mit Hilfe einiger Umschreibungen		die meisten Themen des eigenen Alltagslebens (z.B. Familie, Hobbys, Interessen, Arbeit, Reisen, aktuelle Ereignisse)
WB	zeigt macht	Beherrschung Grundwortschatz Fehler	gute Beherrschung elementare	in wenig vertrauten Situation	bei wenig vertrauten Themen oder komplexen Sachverhalten

B2:

Skala	Operation	Was/Wie	Einschränkungen/Bedingungen	Situationen	Themen
WS	verfügt (über) (kann) variieren	großer Wortschatz Formulierungen	 zur Vermeidung von Wiederholungen Lücken im Wortschatz können zu Zögerungen und Umschreibungen führen		eigenes Sachgebiet, in den meisten allgemeinen Themenbereichen
WB	Verwendung vorkommen	Wortschatz einige Verwechslungen und falsche Wortwahl	Genauigkeit i. A. groß ohne die Kommunikation zu behindern		

C1:

Skala	Operation	Was/Wie	Einschränkungen/Bedingungen	Situationen	Themen
WS	beherrscht gebrauchen Suchen Rückgriff Beherrschung	*großer* Wortschatz Umschreibungen nach Worten auf Vermeidungsstrategien idiomatische Ausdrücke, umgangssprachliche Wendungen	 problemlos (bei Wortschatzlücken) selten gute Beherrschung		
WB		 Wortgebrauch	gelegentlich kleine *Schnitzer*; keine größeren *Fehler*		

C2:

Skala	Operation	Was	Wie	Situation	Themen
WS	beherrscht ist sich bewusst	sehr reichen Wortschatz einschließlich umgangssprachlicher und idiomatischer Wendungen; der jeweiligen Konnotationen			
WB	Verwendung	Wortschatz	durchgängig korrekt und angemessen		

Anhang 18: Tabellen zur Analyse der Skalen aus dem Bereich der pragmatischen Kompetenzen: Skala *Kohärenz und Kohäsion* (GER 2001: 125)

A1:

Operation	Was	Wie	konkrete Beispiele	zu welchem Zweck
verbinden	Wörter oder Wortgruppen	durch sehr einfache Konnektoren	*und* oder *dann*	

A2-:

Operation	Was	Wie	konkrete Beispiele	zu welchem Zweck
verknüpfen	Wortgruppen	durch *einfache* Konnektoren	*und*, *aber* und *weil*	

A2+:

Operation	Was	Wie	konkrete Beispiele	zu welchem Zweck
benutzen	die häufigsten Konnektoren			einfache Sätze miteinander verbinden, Geschichte erzählen, Beschreibung in Form einer einfachen Aufzählung

B1:

Operation	Was	Wie	konkrete Beispiele	zu welchem Zweck
verbinden	eine Reihe kurzer und einfacher *Einzelelemente*			Verbindung einer linearen, zusammenhängenden Äußerung

B2-:

Operation	Was	Wie	konkrete Beispiele	zu welchem Zweck
verwenden	begrenzte Anzahl von Verknüpfungsmitteln längere Beiträge	möglicherweise etwas sprunghaft		Verbindung Äußerung zu klarem, zusammenhängendem Text

B2+:

Operation	Was	Wie	konkrete Beispiele	zu welchem Zweck
verwenden	verschiedene Verknüpfungswörter	sinnvoll		deutlich machen inhaltlicher Beziehungen

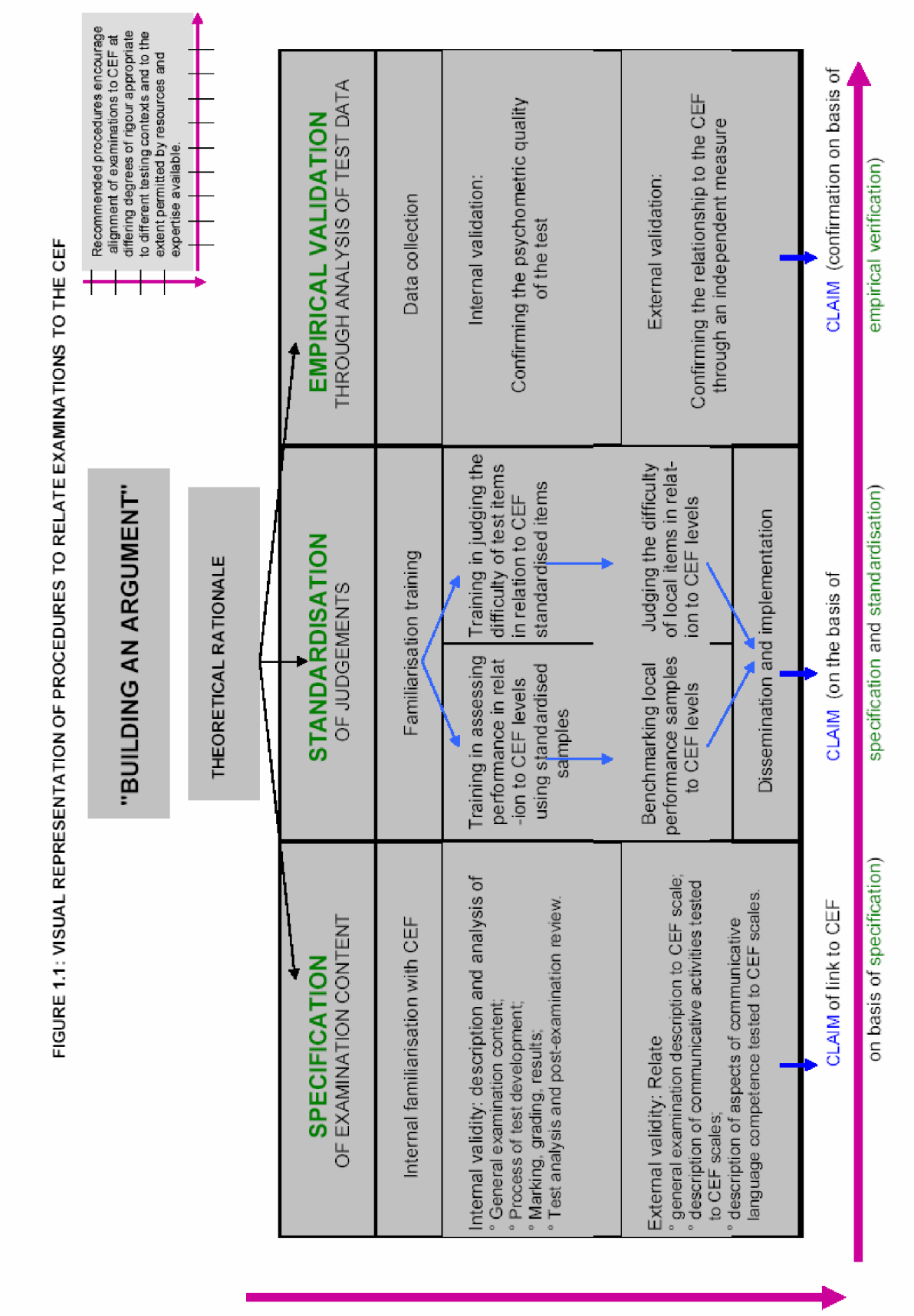
C1:

Operation	Was	Wie	konkrete Beispiele	zu welchem Zweck
sprechen zeigt	 Beherrschung der Mittel der Gliederung/Verknüpfung	klar, sehr fließend und gut strukturiert		 Gliederung sowie inhaltliche und sprachliche Verknüpfung

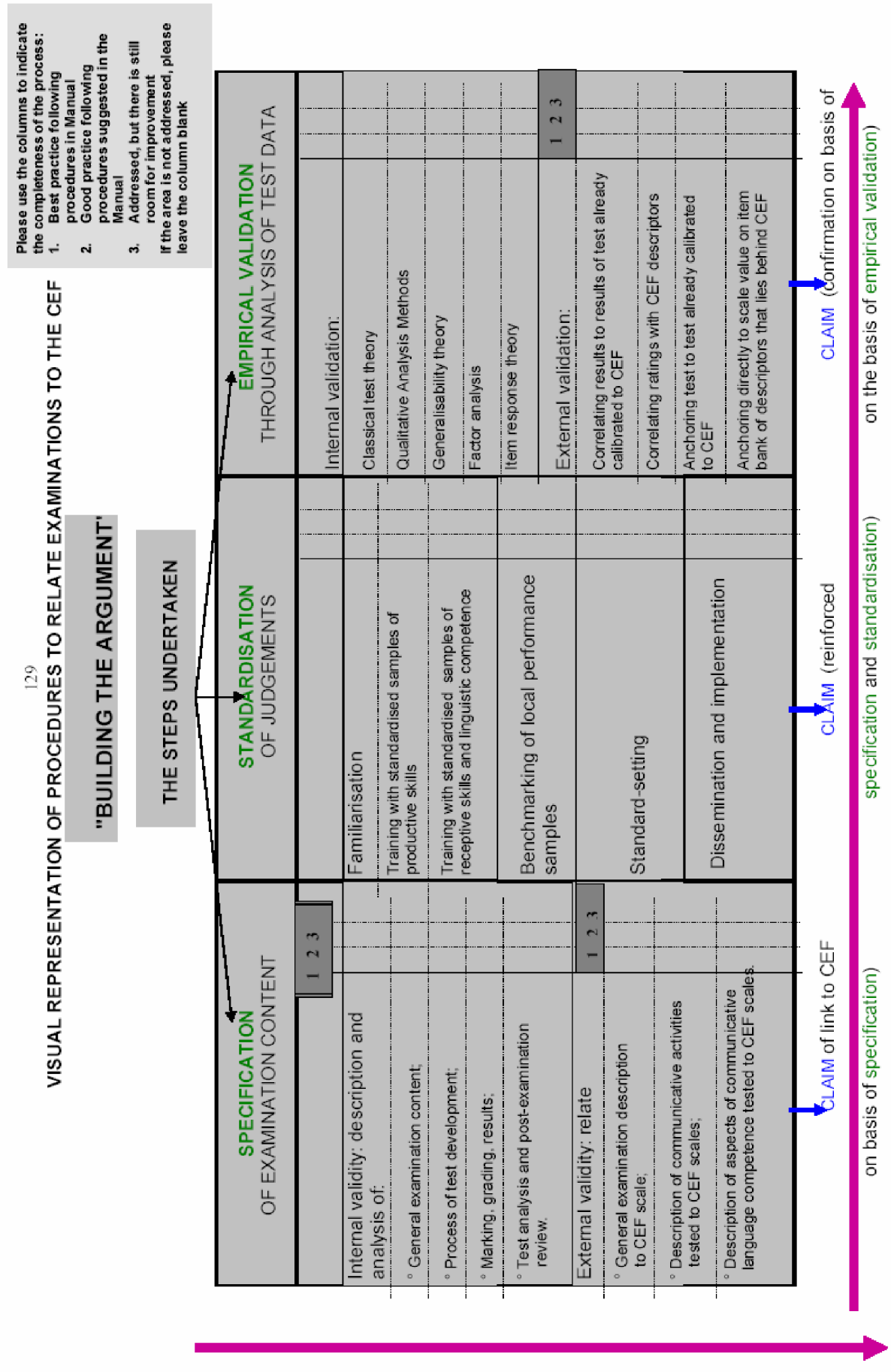
C2:

Operation	Was	Wie	konkrete Beispiele	zu welchem Zweck
erstellen einsetzen	Text Vielfalt an Mitteln	gut gegliedert und zusammenhängend angemessen		 Gliederung und Verknüpfung

Anhang 19: Theoretische Aspekte der Anbindungsprozeduren im *Manual* (ebd. 2003: 4)



Anhang 20: Checkliste: Konkrete Schritte der Testanbindung (*Manual* 2003: 129)



Anhang 21: DESI-Zeitplan

DESI - ZEITPLAN

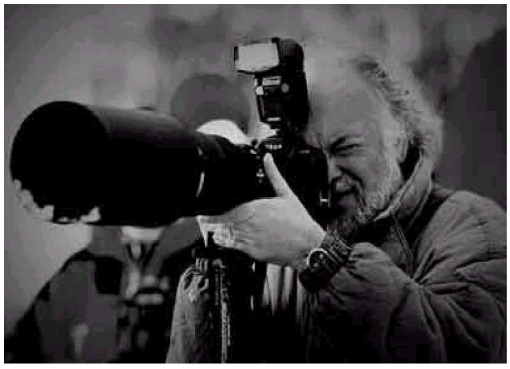
April 2001 bis Juli 2001	Bildung eines Projektbeirates und Konstituierung von Fachgruppen KMK-Expertentreffen für die Fächer Englisch und Deutsch Analyse der Curricula Beginn der Entwicklung von Tests und Fragebögen
Januar 2002 bis Mai 2002	Durchführung der Präpilotierungen an lokalen Schulen Auswertung der Präpilotierung
September 2002 bis März 2003	Durchführung der Pilotuntersuchungen an den Schulen Auswertung der Voruntersuchung
April 2003 bis August 2003	Vorbereitung der Haupterhebung: Stichprobenziehung Genehmigungsverfahren Druck der Instrumente
Oktober 2003	Haupterhebung (440 Klassen) - 1. Messzeitpunkt
Oktober 2004 bis April 2004	Video-Unterrichtsstudie (100 Klassen)
Mai 2004 bis Juni 2004	Haupterhebung (440 Klassen) - 2. Messzeitpunkt
Juli 2003 bis September 2004	Vorbereitung und Durchführung von Fallstudien (Schulportraits) sowie Erprobung von Interventionen
September 2004	Expertentagung zum Thema "Konzeptualisierung und Messung sprachlicher Kompetenzen"
November 2004	Dateneingabe, Datenaufbereitung und Basisauswertung
Frühjahr 2005	Feinauswertung der Daten und Berichtserstellung
Oktober 2005	Abgabe des Endberichts, 1. Teil
Dezember 2005	Abgabe des Endberichts, 2. Teil Wissenschaftliche Fachtagung

In Anlehnung an die Quelle: http://www.dipf.de/desi/zeitplan.html, Zugriff 9.6.2005

Layout: Hartig / Jude

Anhang 22: Semikreative Aufgabe Stand Präpilotierung

Entwurf 4:



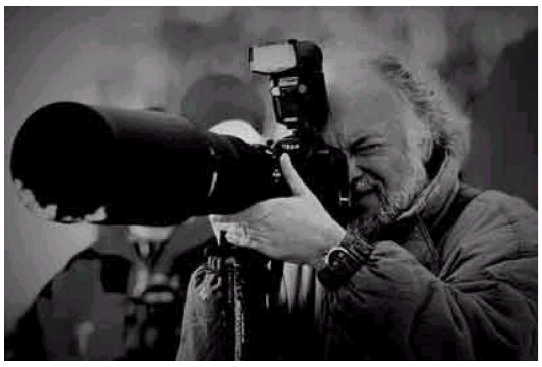
Look at the pictures. Use your imagination to write about **one** of these persons and his/her life.
Maybe these questions can help you:

What do the faces of the people express? Do they look sad, happy, content or angry?
What might the persons have experienced in their lives?
Where do they live? Think of their families, friends and home.
In which country were they born? Where did they grow up? Where do they live now and why?
What name and age would you give them?

Now **choose one picture** for your description and write a report about the life of this person for your local school news.

Anhang 23: Semikreative Aufgabe Stand Pilotierung

FST03



Schreibe einen Artikel für die Schülerzeitung. Wähle dazu ein Bild aus und berichte über das Leben dieser Person. Du darfst natürlich alle notwendigen Einzelheiten zur Geschichte dieser Person erfinden.
Der Bericht über deine Person sollte ungefähr eine Seite lang werden.
Schreibe in deinem besten Englisch.

Hier ein paar Tipps:

Gib deiner Person einen Namen und erzähle, was sie erlebt hat.
Berichte, woher deine Person kommt, wo sie lebt und ob sie Familie hat.
Erwähne auch den Beruf, die Träume oder Wünsche deiner gewählten Person.

Lass dich von deinem gewählten Bild anregen und denke dir die Geschichte dieser Person aus.

Anhang 24: Analysen der Lernertexte aus der DESI-Präpilotierung

Biography SK 4 FST03 ST02	**Textsorte** -Biographie -Makrostruktur -Absätze	**Länge** Anzahl W.	**Inhalt** -Ideen -Relevanz -Qualität	**Sprache** Umfang / Korrektheit: -Orthographie -Wortschatz -Grammatik	**kommunikative Wirkung** -Effekt auf Leser -spannend, witzig, emotiv, ... -Adressatenbezug	**Zuordnung** *-basic* *-medium* *-advanced*
BO9	nicht valide		Kommentar: Aufgabe zu schwer	deutsch-russisch		nicht valide
BO10	-Personenbeschreibung -logischer Aufbau -kein Ende	36	-Herkunft, Name, Alter -Wohnort -unverständliche Zusatzinformation -sehr knapp	-Lexik Basisbereich -Parataxe -Probleme bei Verlaufsform -3.Person-s korrekt	-begrenzt aufgrund knapper Ausführungen -letzter Satz unverständlich	*basic*
BO11	- Personenbeschreibung -logischer Aufbau -kein Ende	65	-Name, Herkunft knapp -Familie etwas detaillierter -Traum: glückliches Leben	-Probleme im Basisbereich (Orthographie, Satzstellung, Lexik, Kongruenz) -Lexik Basisbereich -Parataxe -risikobereit	-verständlich, wenn auch mit Mühe -Bild der Person entsteht, wenn auch wenig detailliert -Gefühle knapp angesprochen	*basic* *(- medium)*
BO12	-Beginn Beschreibung, dann Bruch: Autobiographie -unlogisch, Sprünge -kein Ende	70	-knapp: Name, Alter, Herkunft, Beruf -Schwerpunkt: *Model Agency* und *Competition* -wenig motiviert	-grobe Fehler im Basisbereich (Kongruenz, Pronomen, Verben, Satzstellung) -Lexik Grundwortschatz; Orthographie ok -Parataxe -risikobereit	-Holprigkeit und Bruch verlangen viel Aufmerksamkeit vom Leser -Bild der Person entsteht nicht	*basic* *(- medium)*
ST1	-Personenbeschreibung -sprunghaft, aber sprachlich verbunden -kein Ende	54	-Herkunft, Alter, Name -Wohnort und Familie -Beruf: Model -knapp, wenig Entwicklung, keine Details	-Grundwortschatz (GWS) sitzt -Satzstellung ok, Para- und Hypotaxe -Kohäsion: *when, now, and, usually, today*, Pronomen	-ganz knappes Bild entsteht, aber nicht vom Charakter der Person -zu kurz, um zu wirken, zudem ohne Schluss	Inhalt *basic* Sprache *medium-basic*
ST2	-Beginn Personenbescheibung -nicht zusammenhängend -kein Ende	31	-Name, Alter, Herkunft -Familie (widersprüchliche Informationen -sehr knapp, keine Details oder Entwicklungen	-Lexik Basisbereich -2 korrekte Sätze, ein Nebensatz (*because*)	-zu kurz, um zu wirken -gezeigte Sprache ist aber flüssig	wie bewerten bei der Kürze? *basic* (Sprache *medium*)

Biography SK 4 FST03 ST02	**Textsorte** -Biographie -Makrostruktur -Absätze	**Länge** Anzahl W.	**Inhalt** - Ideen -Relevanz -Qualität	**Sprache** Umfang / Korrektheit: -Orthographie -Wortschatz -Grammatik	**kommunikative Wirkung** -Effekt auf Leser -spannend, witzig, emotiv, … -Adressatenbezug	**Zuordnung** *-basic* *-medium* *-advanced*
ST3	-Personenbeschreibung -Makrostruktur: Kurzvorstellen - Rückblick auf Karrierebeginn – Sprung zu Herkunft und Wohnorten – kein Ende -keine Absätze	73	-Kurzvorstellung ohne Details -Detail: Karrierebeginn -Herkunft -Bewertung, Gefühle	-orthographische Probleme -Lexik innerhalb GWS, variiert, nicht immer treffend -Parataxe, teils deutsche (dt.) Wortstellung -Probleme *tenses* - verwendet sie aber -risikobereit (direkte Rede) - Verwendung dt. Wörter	-knappes Bild entsteht -positiv: Anekdote der Entdeckung zum Model -trotz Orthographieproblemen verständlich -fehlendes Ende vermindert die Wirkung etwas	*basic - medium* (da so ambitioniert trotz geringer Sprachkenntnisse)
ST5	-Autobiographie -keine Struktur, assoziativ gereiht	94	-Name, Alter, Job -Familie mit Namen, Alter, Job und Wohnort -keine Gefühle, Bewertungen (wo er doch Mafia-Killer war…)	-Probleme Lexik bei Dingen jenseits „Überlebensbereich" -parallele Parataxe -Probleme Plural, Kongruenz -keine Kohäsion	-Kernaussage fehlt -verlangt stellenweise Anstrengung seitens Leser (*agend* soll wohl *Agent* bedeuten, etc.) -Satzbau langweilig -kann sich aber im Grunde verständlich machen	*basic - medium*
ST7	-Personenbeschreibung -keine Struktur – eher assoziative Reihung -ohne Schluss	87	-viele Ideen, wenn auch teils unmotiviert gereiht -berufliche Entwicklung -Herkunft, Familie -Beschreibung Äußeres, aber keine Gefühle oder Charakter -viele kleine Details	-Reihung auch sprachlich: keine richtigen Sätze, keine Interpunktion, alles parataktisch gereiht -Lexik umfangreich (*silver hair*), wenn auch nicht immer korrekt (*make pictures*)	-Fülle an Informationen lässt doch Bild entstehen -trotz fehlender Satzzeichen verständlich und lesbar	*basic - medium*
ME1	-Biographie -Beginn am Bild aufgehängt, mit Ende/Ausblick -keine Absätze -kohärente Struktur erkennbar: Entwicklung Karriere – Reisen und Job – Familie und Freunde – offenes Ende	209	-Schwerpunkt Fotograf: Beginn als Hobby, Entwicklung, Hintergrund Erfolg detailliert -Verhalten beschrieben: alles Geld fürs Fotografieren, Tiere in ihrer natürlichen Umgebung -Familie: keine als Folge des Jobs, dafür gute Freunde -Bewertungen, die Person charakterisieren: Schwierigkeiten Job, Meinung der Freunde	-orthographische Probleme -Lexik im GWS, oft Interferenzen, nicht idiomatisch, aber verständlich; Nomen oft modifiziert *-tenses* unsicher (Bildung) -dt. Wortstellung, *that-clauses*, Relativsätze -> Versuche, aber noch fehlerlastig -Kohäsion gegeben: Pronomina, *this, it began… when he was…, because, also, this is the reason that…, and, but* - risikofreudig – umscheibt alles	-Leserfreundlich strukturiert, alle relevanten Informationen -aber: etwas mühsam zu lesen, da so fehlerlastig -detailliertes Bild von *Albeter* und dessen Charakter entsteht -wirkungsvoll: Details und Ende	Inhalt und Wirkung *advanced* Sprache *medium*

Biography SK 4 FST03 ST02	**Textsorte** -Biographie -Makrostruktur -Absätze	**Länge** An-zahl W.	**Inhalt** - Ideen -Relevanz -Qualität	**Sprache** Umfang / Korrektheit: -Orthographie -Wortschatz -Grammatik	**kommunikative Wirkung** -Effekt auf Leser -spannend, witzig, emotiv, … -Adressatenbezug	**Zuordnung** -*basic* -*medium* -*advanced*
ME2	-Personenbeschreibung -keine Absätze -teils Struktur: Kind – Erwachsener, der Farm aufbaut – Rückschlag – Nachschub Familie -Ende: Bezug auf Bild	177	-Schwerpunkt: Farm gegen Willen Vater, Traum schon als Kind -Aufbau Farm und Rückschlag -Familie nur angehängt -wenig Emotionen, Zwischenmenschliches, Wertungen (eher gegen Ende)	-Probleme: orthographischer oder lexikalischer Art z. B. Verwechslung *whant* mit *want* oder *won't* -Lexik auf GWS beschränkt - Probleme *tenses* (Verwechslung) -*aspect* korrekt -Versuch Hypotaxe (Relativsatz) -Risikobereitschaft	-teils schwer zu dekodieren wegen Lexik/Orthographie -Bild bleibt auf Farmerinteresse beschränkt – Charakter wird nicht deutlich -wenig interessante Aspekte -starkes Ende im Vgl. zum Text insgesamt	Sprache: Umfang *medium* Korrektheit *basic* Inhalt *medium*
ME3	-Bericht über Personen -Lebensalter als roter Faden innerhalb der Absätze -Text erscheint nicht als Netzwerk, eher Reihung	130	-knapp: Kindheit in *Great Britain*, Umzug nach *Germany* wegen Beruf Vater -Berufswunsch wie Vater -im Alter zurück nach GB -Familie angehängt -wenig jenseits bloßer Fakten, wenig motiviert, eher Ideen gereiht	-begrenzte Lexik, *false friends* -Probleme *tenses* (Verwechslung) -vorwiegend Parataxe, *when-clauses*, Relativsatz -Kohäsion: *at first… but when, when, before, because, then, and, but, there*	-Sprache schwach – nicht leicht und flüssig zu lesen -Bild auf Fakten beschränkt – keine Gefühle, nichts Zwischenmenschliches, keine Wertungen -Charakter entsteht nicht -nichts Witziges oder Überraschendes	*lower medium*
ME4	-Bechreibung Person auf Bild -keine Absätze -Struktur eher gereiht: wohnen, Familie, Job, Rückblick Krieg -endet wieder bei Bildbeschreibung	191	-Name, Alter, Wohnort -Schwerpunkt Familie mit Namen und Alter, ohne Charaktere oder Details -Nachschub: Job, Krieg früher, jetzt Lebensabend genießen - Gefühle: Trauer Tod Frau, Angst um Enkel, *terrible war, enjoy rest of life* -wenig Hintergrund oder Motivation erkennbar	-Lexik GWS, teils kommunikationsbelastende Fehler (*he has been a family*) -Probleme *tenses*: Bildung und Einsatz -*aspect* ok: *now he is sitting on this bank…* -meist Parataxe, Nebensätze: *when, that, who* -dt. Wortstellung - Kohäsion: Pronomina, *now, this, when, but* -risikofreudig	-lesbar, aber anfangs langweilig aufgrund reiner Aufzählung Familienmitglieder + Parataxe -die wenigen Bewertungen erhöhen die Wirksamkeit -aber: Kernaussage? -Bild der Person entsteht nicht detailliert, Charakter nur marginal	*lower medium* da belastende Fehler

Biography SK 4 FST03 ST02	**Textsorte** -Biographie -Makrostruktur -Absätze	**Länge** Anzahl W.	**Inhalt** - Ideen -Relevanz -Qualität	**Sprache** Umfang / Korrektheit: -Orthographie -Wortschatz -Grammatik	**kommunikative Wirkung** -Effekt auf Leser -spannend, witzig, emotiv, … -Adressatenbezug	**Zuordnung** *-basic* *-medium* *-advanced*
ME6	-Biographie bis zum jetzigen Zeitpunkt -Struktur: Familie, Sarahs Entwicklung, Karriere als Model -Höhepunkt = Ende: Gewinn Wettbewerb -auf Textebene nicht kohärent	165	-Hintergrund, Familie – Details und Gründe -Karriere mit Details, motiviert -Charakter wird deutlich über ganzen Text hinweg -Bewertungen und Einschätzungen	-Lexik breit (*famous, crime, popular, attitude, competition*), manchmal unangemessen (*attitude*), Umschreibungen aber problemlos möglich -*tenses* meist korrekt und angemessen -Satzbau Parataxe, *when-clauses, that-clauses, cleft sentence* -Probleme bei komplexeren Versuchen -Kohäsion teils gegeben: *when, and, always, this, but, first*	-gut zu lesen -unkomplizierte Beschreibung mit Hintergrund, Details und Höhepunkt -Bild von Sarah und Charakter entsteht -Sprache noch nicht idiomatisch, aber teils flüssige Umgangssprache	*medium*
ME8	-Fakten über Person -assoziativ, wenig Struktur -keine Absätze -nicht zu Ende geführt	144	-Name, Alter, Wohnorte -Familie knapp -Krieg und Hintergrund -Frau und Tod -keine Gefühle, Wertungen	-orthographische Probleme -Lexik: GWS und Anschlüsse sitzen -teils flüssige Umgangssprache (*The man I will tell you about…*) -*tenses*: ok, Vorzeitigkeit -bricht nicht ab, selten dt. Wort (*Mittelmeer*) -Kohäsion: Pronomina, *then, but, because, now –then, after, just before*	-Adressatenbezug vorhanden -manche Aussage missverständlich -„trocken" wegen faktischer Aussagen -Charakter ist nicht erkennbar	Sprache / Inhalt *medium* Textsorte / Struktur *below medium*
PKG39	-Biographie -Struktur chronologisch und logisch, ohne Brüche, mit Einleitung und rundem Ende -keine Absätze	253	-Kindheit in *Scotland* (lebendige Beispiele) -Probleme mit Farm Eltern -Tante und London als Sprungbrett (Ausbildung) -Neffe nimmt ihn nach USA mit, wo er Traum, Fotograf zu werden, realisieren kann -Gefühle, Bewertungen, Träume -alles Relevante vorhanden	-Lexik angemessen und breit, idiomatisch (*this text is about his life and the single chapters of it*) -Grammatik ausgefeilt, Satzbau variiert (temporale, kausale, verkürzte Nebensätze, Relativsätze -flüssige Sprache, Kohäsion gegeben, Mittel gut eingesetzt	-detailliertes Bild seines Lebens und seines Traums -wenig zu Charakter (kann man aber inferieren) -interessant zu lesen, abwechslungsreiche Geschichte, trotz vieler Ideen motiviert und nachvollziehbar -flüssige Sprache stützt Wirksamkeit	*advanced*

Biography SK 4 FST03 ST02	**Textsorte** -Biographie -Makrostruktur -Absätze	**Länge** Anzahl W.	**Inhalt** - Ideen -Relevanz -Qualität	**Sprache** Umfang / Korrektheit: -Orthographie -Wortschatz -Grammatik	**kommunikative Wirkung** -Effekt auf Leser -spannend, witzig, emotiv, ... -Adressatenbezug	**Zuordnung** *-basic* *-medium* *-advanced*
PKG44	-Biographie zum Model-Foto -Struktur zeigt Ansätze von Komplexität: jetzt Model – Rückblick auf Kindheit bis Hochzeit – Mann bringt sie zum Modelling (Referenz auf oben) – Karriere und Familie – *happy end* -keine Absätze	266	-viele Ideen, motiviert entwickelt (mit Beispielen und Details) -einige Wertungen und Gefühle, Traum vom „Rauskommen" -alles Wesentliche vorhanden für Lebensgeschichte	-Lexik: GWS sitzt sicher, etwas darüber hinaus, aber nichts Elaboriertes; wenig Modifikationen, eher frequente Lexeme -Grammatik: Zeitenbildung ok, wenn auch Verwechslungen; Satzbau teils noch Interferenzen; meist Parataxe; Relativsätze, *that-clause, compound* -Kohäsion über Wortfelder, wenig Kohäsionsmittel (*and, that, as, later, after that, when, then*, Pronomina)	-runde Geschichte ohne Brüche, aber auch ohne Witz, Spannung etc. -alles logisch und motiviert -Bild von Alyssa entsteht trotzdem nur bedingt – Charakter muss inferiert werden	*upper medium* Sprache *medium*
PKG66	-Biographie zum Fotografen -logische Struktur: Person vorstellen – Familie und Charakter – Job – abruptes Ende	285	-Name, Wohnort, Beruf (immer mit Details und motiviert entwickelt) -Begründung, warum er trotz viel Geld lieber bescheiden auf dem Land mit Familie lebt (auch hier wieder Details und Hintergründe) -Infos zu seinem Job und Episode mit Jackson (*car crash*) = abruptes Ende	-Lexik angemessen, breit, Modifikationen, Idiomatik (*he's always in a hurry/ where the housing prizes are ok/ one of the best tennis players of his age*) -Grammatik: *tenses* sicher, Satzbau sicher, NS: verkürzt (*to an area he supposes the stars to be*), temporal, kausal, *that-clauses*, Relativsätze, oft *compound sentences* -Kohäsion: zwischen Sätzen und im Text: Netzwerk; Mittel: Pronomen, *instead of, because, nearly all of his life, always, but, where, there…*	-Bild von Fotograf, seiner Familie, seinem Job entsteht -Charakter wird gezeichnet (*proud of his boys*, bescheiden…) -interessant zu lesen durch die vielen Beispiele und Details	*advanced*

Biography SK 4 FST03 ST02	**Textsorte** -Biographie -Makrostruktur -Absätze	**Länge** Anzahl W.	**Inhalt** - Ideen -Relevanz -Qualität	**Sprache** Umfang / Korrektheit: -Orthographie -Wortschatz -Grammatik	**kommunikative Wirkung** -Effekt auf Leser -spannend, witzig, emotiv, … -Adressatenbezug	**Zuordnung** *-basic* *-medium* *-advanced*
PKG83	-Biographie -komplexe Makrostruktur mit Rückblenden und Vorschau (teils assoziative Einschübe) -angemessene Absätze	>350	-biographischer Abriss von Kindheit bis zu „Miss Paris" -motivierte Entwicklung der Inhaltselemente -Gefühle, Wünsche, Bewertungen	-Umfang angemessen, Korrektheit meist gegeben: -Lexik im Themengebiet *modelling* ausgefeilt, viele Modifikationen (*Paris, the city of glamour, money and all models*) -Grammatik: Zeitenbildung sitzt, aber gelegentliche Verwechslungen; Satzbau meist Parataxe, teils Hypotaxe (temporal, kausal, Relativsätze)	-umfassendes Bild von Janet und ihrem bisherigen Leben, mit Gefühlen, Einstellungen, Charakter -interessanter, detaillierter Bericht mit teils komplexer Struktur, gut zu lesen, auch durch die vielen Hintergrundinformationen -Sprache noch nicht sehr ausgefeilt, aber keine Kommunikationsbelastung, alles verständlich	Sprache: *upper medium* sonst: *advanced*
PKG88	-Ausschnitt aus Leben von Paul -wenig Struktur, erweckt eher assoziativen Eindruck -Absätze teils angemessen -Ende wirkungsvoll	189	-Schwerpunkt: Parkbank -Infos zu Alter, Krieg, Familie (Bruder) -Name, Beruf (jetzt nicht mehr: viel Freizeit), -Frau tot, Vorlieben, Bewertung Leben aus Sicht Dritter (Nachbarn denken, es sei langweilig, Paul aber ist zufrieden) -Wiederaufnahme Bruder und alter Streit wegen *girlfriend* -Paul weiß um sein nahendes Lebensende	-Lexik: nichts ausgefeiltes, aber GWS sehr sicher und einiges darüber hinaus -Grammatik: Satzbau variiert, *that*, *whether*, Relativ, *compound*, verkürzte NSe, *tenses* ok, *modals* sicher - Konnektoren eher einfach (*but, before, and, then, after*)	-Ausgangs- und Endpunkt *Parkbank* sehr wirkungsvoll – alles rankt sich, eher assoziativ, darum -Bild von Paul entsteht, könnte aber elaborierter sein; Rückschlüsse auf Charakter möglich -interessant und gut zu lesen	*upper medium*

Anhang 25: Konstruktion der DESI-*Rating Scales* I: Synopse der Berührungspunkte der *Cambrige Assessment Scales zu CPE, CAE und FCE (Arial kursiv)*, der analysierten Lernertextmerkmale (Arial normal), und der GER-Skalen „Schriftliche Produktion allgemein“, „Briefe und Aufsätze“, „Schreiben“ aus Selbstevaluationsraster und „Spektrum sprachlicher Mittel allgemein“ (Times New Roman). Die Kategorien (vgl. die Spalteneinteilung) entstammen zum Teil den *Cambridge*-Skalen, zum Teil dem DESI-Analyseschema.

Stufen	Inhalt/Entwicklung/ Kohärenz	Emotive Sprache	Textsorte	Sprache	Komm. Wirkung	Wirkung auf Leser
C:inadequate attempt DESI: 0	*less than 50 words* irrelevant oder unverständlich oder zu wenig	nicht erkennbar	nicht erkennbar	sehr missverständlich oder unverständlich	*negative effect* keine Wirkung	
C: poor attempt at task DESI: 1 GER A1	*little relevance* teils relevant, Ideen nicht elaboriert, keine Struktur; einfache isolierte Wendungen und Sätze, über sich selbst und fiktive Menschen	Zusammengehörigkeit von Personen	einfachst gestrickt; sehr kurze einfachste Texte	*severly limited, inaccurate range* sehr begrenzte Mittel, nur frequenter Basisbereich; kann sehr fehlerhaft sein sehr elementares Spektrum; persönliche Bedürfnisse konkreter Art	*very negative effect* ansatzweise, kann missverständlich sein	*requires excessive effort by the reader*
C: some attempt at task, not adequate DESI: 2 GER A2	*inadequate development* gewisse Relevanz, knappe Entwicklung; zusammenhängende Sätze; sehr kurze, elementare Beschreibung von Ereignissen, vergangenen Handlungen und persönlichen Erfahrungen	Umreißen zwischenmenschlicher Beziehungen in elementarster Form	Mängel im Formalen kurze, einfache, persönliche Briefe/ Berichte	*limited, inaccurate use* relativ einfache Sprache, kann fehlerhaft sein, doch Ansätze von Kohärenz; gebräuchliche Ausdrücke, elementares Spektrum, begrenztes Repertoire	*negative effect* in groben Zügen	*requires considerable effort by the reader*
C: task reasonably achieved DESI: 3 GER B1	*adequate coverage* relevante Punkte abgedeckt, teils noch zu wenig elaboriert, einfache, aber logische Struktur unkomplizierter, zusammenhängender Text zu vorhersehbaren Ereignissen; Interessenschwerpunkte sichtbar	Zwischenmenschliches, Gefühle und Reaktionen in unkomplizierter Weise	Konventionen erkennbar unkompliziert, in üblichem Standardformat	*CPE: adequate use* *CAE: sufficiently adequate, good range* *FCE: adequate range* gewisse narrative Grundqualitäten (Kohärenz); Fehler noch vorhanden, doch nicht kommunikationsbelastend hinreichendes Spektrum, um zurechtzukommen	*achieves desired effect* grundsätzlich wirksam, doch teils mit Einschränkungen	*requires some effort*

Stufen	Inhalt/Entwicklung/ Kohärenz	Emotive Sprache	Textsorte	Sprache	Komm. Wirkung	Wirkung auf Leser
C: good realisation of task	*good development*			*CPE: competent use* *CAE: sufficiently natural, some evidence of range* *FCE: good range*	*positive effect*	*requires only a little effort*
DESI: 4	alles abgedeckt, aber evtl. noch Mängel in Entwicklung, oder Schwerpunkte auf Kosten anderer Punkte	Einstellungen und Gefühle versprachlicht	geltende Konventionen meist beachtet	insgesamt flüssig, einige Schwachstellen; narrative (humoristische, spannende) Qualität	ohne große Einbußen	
GER B2-	klar, detailliert, zusammenhängend, über vielseitige Erfahrungen…; Entwicklungen, Gründe, etc. benennen			hinreichend breites Spektrum auch in unvorhersehbaren Situationen		
C: task fully completed	*excellent development of topic*			*CPE: impressive use of language* *CAE: resourceful, controlled, natural use of language* *FCE: wide range*	*very positive, impressive*	*requires no effort*
DESI: 5	Relevanz, Entwicklung, Elaboration; Beispiele, Begründungen; stringenter Aufbau	DESI: Einstellungen und Gefühle adäquat und kompetent versprachl.; Perspektivenwechsel	durchgängig entsprechend der geltenden Konventionen	große, angemessene Breite sprachlicher Mittel sicher beherrscht, überzeugend, lesergerechter Stil	umfassend	
GER B2+	klar, detailliert, gut strukturiert;	unterschiedliche Standpunkte	Beachtung der geltenden Konventionen; angemessener Schluss	klarer Ausdruck ohne Eindruck von Einschränkungen		
GER C1	ausführlich; wesentliche Punkte hervorgehoben, durch Unterpunkte, Beispiele gestützt					

Anhang 26: Konstruktion der DESI-*Rating Scales* II: Abgleich der DESI-Skalen (auf Basis der Merkmale der Lernertexte) mit relevanten Skalen aus dem GER am Beispiel der Niveaus 1, 3 und 5 des DESI-Globalurteils zur *Biography*-Aufgabe und den Skalen *Schriftliche Produktion* (SP aus GER 2001: 67), *Kreatives Schreiben* (KS, ebd.: 67f), *Briefe und Aufsätze schreiben* (BA, ebd.: 68), *Schreiben* aus dem Selbstbeurteilungsraster (SB, ebd.: 36). Berührungspunkte sind in den GER-Deskriptoren fettgedruckt, soweit sie für die Formulierung der DESI-Deskriptoren genutzt wurden.

Stufe	Skala	Deskriptoren aus GER	DESI GLOBALURTEIL – Bericht *Biography* (ST02)
DESI: 1 GER: A1	SP	Kann einfache, isolierte Wendungen und Sätze schreiben.	• Kurzer, einfachster Text; Mängel im Formalen; meist keine Makrostruktur erkennbar (assoziative Reihung). • Ideen ansatzweise relevant, aber nicht entwickelt; Darstellung bleibt konkret; einfache Wendungen und Sätze über sich selbst und andere (auch fiktive) Menschen; wie sie zusammengehören, wo sie leben, wie sie leben, was sie tun oder was sie tun wollen. • begrenzte sprachliche Mittel aus frequentem Basisbereich; Text kann bruchstückhaft und fehlerhaft sein (lexikalische, grammatische, syntaktische, orthographische Fehler; muttersprachliche Interferenzen). • Gewünschte Botschaft wird nur ansatzweise und oft missverständlich vermittelt.
	KS	**Kann einfache Wendungen und Sätze über sich selbst und fiktive Menschen schreiben: wo sie leben und was sie tun.**	
	BA	Keine Deskriptoren verfügbar	
	SB	Kann kurze, einfache Postkarte schreiben, z. B. Feriengrüße	
DESI: 3 GER: B1	SP	Kann **unkomplizierte**, zusammenhängende Texte zu mehreren vertrauten Themen aus seinem/ihrem **Interessengebiet** verfassen, wobei einzelne kürzere Teile in linearer Abfolge verbunden werden.	• Unkomplizierter Bericht in üblichem Standardformat; teils nicht alle Formalia erfüllt; logische Makrostruktur erkennbar, es kann aber zu Sprüngen kommen. • Beschreibung realer wie fiktiver Personen und deren Umfeld und Erfahrungen. Interessenschwerpunkte können sichtbar werden. Zwischenmenschliche Beziehungen, Gefühle und Reaktionen darauf werden in unkomplizierter Weise beschreiben. • Sprachliche Mittel werden bis zu einem gewissen Grad angemessen und in hinreichendem Umfang eingesetzt, um das Ziel zu erreichen; gewisse narrative Grundqualitäten (wie etwa Textkohärenz) sind gegeben. Fehler und Interferenzphänomene können im Text vorhanden sein, in der Regel jedoch nicht kommunikationsbelastend. • Botschaft wird grundsätzlich kommunikativ wirksam vermittelt, doch teils mit Einschränkungen.
	KS	Kann **unkomplizierte**, detaillierte Beschreibungen zu einer Reihe verschiedener Themen aus seinem/ihrem Interessengebiet verfassen. Kann Erfahrungsberichte schreiben, in denen **Gefühle und Reaktion** in einem einfachen, zusammenhängenden Text beschrieben werden. Kann eine Beschreibung eines **realen oder fiktiven Ereignisses** oder einer kürzlich unternommenen Reise verfassen. Kann eine Geschichte erzählen.	
	BA	Kann in einem üblichen **Standardformat** sehr kurze Berichte schreiben, in denen Sachinformationen weitergegeben und Gründe für Handlungen angegeben werden. Kann einen kurzen, einfachen Aufsatz zu Themen von allgemeinem Interesse schreiben. Kann im eigenen Sachgebiet mit einer gewissen Sicherheit größere Mengen von Sachinformationen über vertraute Routineangelegenheiten und über weniger routinemäßige Dinge zusammenfassen, darüber berichten und dazu Stellung nehmen.	
	SB	Kann einfache, zusammenhängende Texte zu **vertrauten Themen/aus Interessengebiet** schreiben. Kann persönliche Briefe schreiben und darin von **Eindrücken, Erfahrungen** berichten.	

Stufe	Skala	Deskriptoren aus GER	DESI GLOBALURTEIL – Bericht *Biography* (ST02)
DESI: 5 GER: B2+ / C1	SP (kein B2+) C1	Kann **klare, gut strukturierte** Texte zu komplexen Themen verfassen und dabei die **entscheidenden Punkte hervorheben, Standpunkte ausführlich darstellen und durch Unterpunkte oder geeignete Beispiele oder Begründungen stützen und den Text durch einen angemessenen Schluss abrunden**.	• Bericht ist durchgängig entsprechend der geltenden Konventionen geschrieben; Aufbau logisch, stringent und konsistent, u. U. komplex entwickelt; Makrostruktur erleichtert das Verständnis. • Klare, detaillierte, gut strukturierte und ausführliche Beschreibungen einer Biographie: Es sind entscheidende Punkte hervorgehoben, Standpunkte und Ansichten ausführlich dargestellt und durch Unterpunkte oder geeignete Beispiele oder Begründungen gestützt; Text ist durch angemessenen Schluss abrundet. Einstellungen und Gefühle sind adäquat und kompetent versprachlicht; ggf. werden verschiedene Perspektiven deutlich gemacht. • Großer Umfang und sichere Beherrschung sprachlicher Mittel, Idiomatik und Korrektheit sind meist gegeben. Text ist durchgehend flüssig, in lesergerechtem, überzeugendem, persönlichem und natürlichem Stil verfasst und besitzt durchgehend narrative (oder humoristische oder spannende) Qualität. • Die kommunikative Wirkung wird umfassend erzielt.
	KS B2+	Kann klare, **detaillierte**, zusammenhängende Beschreibungen **realer oder fiktiver Ereignisse und Erfahrungen** verfassen, dabei den Zusammenhang zwischen verschiedenen Ideen deutlich machen und die für das **betreffende Genre geltenden Konventionen** beachten.	
	C1	Kann klare, detaillierte, gut strukturierte und **ausführliche** Beschreibungen oder auch eigene fiktionale Texte in **lesergerechtem, überzeugendem, persönlichem und natürlichem Stil** verfassen.	
	BA B2+	Kann einen Aufsatz oder Bericht schreiben, in dem etwas systematisch erörtert wird, wobei **entscheidende Punkte** angemessen hervorgehoben und **stützende Details** angeführt werden. Kann verschiedene Ideen oder Problemlösungen gegeneinander abwägen.	
	C1	Kann klare, gut strukturierte Ausführungen zu komplexen Themen schreiben und dabei zentrale Punkte hervorheben. Kann Standpunkte ausführlich darstellen und durch Unterpunkte, geeignete Beispiele oder Begründungen stützen.	
	SB (kein B2+) C1	Kann sich schriftlich klar und gut strukturiert ausdrücken und Ansicht ausführlich darstellen. Kann in Briefen, Aufsätzen oder Berichten über komplexe Sachverhalte schreiben und **wesentliche Aspekte** hervorheben. Kann in schriftlichen Texten den Stil wählen, der für die jeweiligen Leser angemessen ist.	

Anhang 27: Handbuch zum Task *Biography*, das bei der Auswertung der DESI-Hauptuntersuchung zum Einsatz kam.

Kodierhandbuch

Textproduktion Englisch

ST02 Semikreative Aufgabe: *Report Biography*

Liebe Kodiererin, lieber Kodierer,

Ihre Aufgabe ist es, englische Texte von Schülerinnen und Schülern der 9. Jahrgangsstufe zu bewerten. Das Kodierhandbuch enthält die genauen Anweisungen, wie Sie dabei vorgehen sollen.

Zuerst werden die Kriterien im Überblick vorgestellt, nach denen die Arbeiten zu bewerten sind. Dann folgt die Kodieranweisung, die Sie bitte genauestens befolgen. Eingebettet in diese Anweisung finden sich die Kriterien in ihren Merkmalen und Abstufungen detailliert beschrieben und mit Beispielen belegt.

Bewertungskriterien „semikreative Schreibaufgabe" im Überblick:

Globalurteil

1. Hier geben Sie Ihren **Gesamteindruck** wieder, der nicht dem Durchschnitt der folgenden acht Kriterien entsprechen muss.

Task Performance

2. **Textlänge**: Entspricht die Länge dem in der Aufgabenstellung geforderten Maß?
3. **Inhalt**: Werden die inhaltlichen Vorgaben aus der Aufgabenstellung umgesetzt?
4. **Textsorte und Aufbau**: Entspricht die Textsorte der Aufgabenstellung? Entsprechen Aufbau und Organisation der Inhaltselemente der generellen Erwartung an diese Textsorte?

Language Performance

5. **Orthographie**: Was kann schon korrekt geschrieben werden? Kommt es zu sinnentstellenden, kommunikationsbelastenden orthographischen Fehlern?
6. **Lexik und Lexiko-Grammatik**: Welcher Wortschatz ist vorhanden und wie angemessen wird er eingesetzt? Hierunter fallen auch Kollokationen und *complementations*.
7. **Grammatik**: Welche grammatikalischen Strukturen sind vorhanden und wie korrekt werden sie eingesetzt?
8. **Kohäsion**: Welche *linking language* ist vorhanden und wie werden Kohäsionsmittel (*cohesive ties*) eingesetzt?

Communicative Performance

9. **Kommunikative Wirkung**: Ist der Text flüssig geschrieben und gut lesbar? Ist der Stil natürlich und idiomatisch? Kommt an, was die Schüler ausdrücken wollten (beim fiktiven Adressaten, in den Sie sich hineinversetzen sollen)?

10. **Fäkalsprache**: Sind im Text *Swear Words* festzustellen?

Kodieranweisung:

Allgemeines zur Kodierung und zu den zu vergebenden *Scores*:

Die Bewertung des Globalurteils und der Kriterien 3 mit 9 erfolgt in **6 Stufen** von 0 bis 5, wobei 5 die höchste Bewertungsstufe ist. Den Stufen 0 bis 5 entsprechen die *Scores* 0 bis 5. Die einzelnen Kriterien werden im Detail vorgestellt, und jede Stufe eines Kriteriums ist in ihren charakteristischen Elementen in so genannten Deskriptoren beschrieben.

Das Kriterium 10 *Swear Words* wird immer **dichotom** kodiert: 0, wenn das Merkmal nicht vorhanden ist, 1, wenn es vorhanden ist. Auch bei *missing values* (s. u.) wird mit 0 respektive 1 kodiert!

Nur Kriterium 2 Textlänge wird nicht kodiert. Hier wird lediglich die **Wortanzahl** eingetragen, wenn es sich um einen validen Text handelt. **Kein Eintrag** bei *missing values*!

***Missing values* oder *default codes*:**

- Wenn die Schülerantwort **nicht valide** ist (z. B. weil der Text in einer anderen als der englischen Sprache verfasst wurde, oder weil statt einem Text eine Zeichnung angefertigt wurde), werden **alle** Kriterien (außer Textlänge) mit **8** kodiert. Hierunter fällt auch eine sog. „Themaverfehlung": Wenn das gestellte Thema **inhaltlich nicht** beantwortet ist (z. B. kein Bezug zur Aufgabenstellung erkennbar ist), oder ein anderes als das gestellte Thema in Form einer anderen als der geforderten Textsorte bearbeitet wurde, muss die Schülerleistung **insgesamt mit 8** (nicht valide) bewertet werden, und zwar bezogen auf **alle** Kriterien, denn der Schüler könnte ja auch einen auswendig gelernten Text niederschreiben: Eine „Themaverfehlung" fällt also aus dem Bewertungsschema heraus. Auch wenn der Schülertext nur aus einer (evtl. sogar adäquaten) Anrede besteht, gilt er als nicht valide!

Beispiel für Kodierung 8: Wenn z.B. in einem englischen Satz ausgedrückt wird, dass die Aufgabe zu schwer war, wird die gesamte Arbeit mit 8 kodiert, da es sich dann nicht um die Bearbeitung des Themas handelt, sondern lediglich um einen Kommentar. Auch Arbeiten, die keinen erkennbaren Bildbezug aufweisen, sind nicht valide, denn es könnte sich ja um einen auswendig gelernten Text handeln.

Handelt es sich aber erkennbar um eine Bearbeitung des gestellten Themas, jedoch beispielsweise in einer anderen als der geforderten Textsorte, so wird die Kategorie „Textsorte" mit **0** kodiert, die übrigen Kategorien gemäß Schülerleistung.

- Wenn die Aufgabe gar **nicht bearbeitet** worden ist („leeres Blatt"), wird mit **9** kodiert.

Vorsicht: Eine Vermischung der *missing values* mit den validen *Scores* ist **nicht** zulässig!

Vorgehensweise beim *Rating*:

Vorab einige Hinweise zu den Kriterien und deren Deskriptoren:

Die Kriterien in ihren Abstufungen **1 mit 5** und die dazugehörigen Deskriptoren sind auf der Basis der Analyse von Lernertexten entwickelt worden und in wissenschaftlichen Modellen verortet, die Sie in der Schulung im Detail kennen gelernt haben. Zusätzlich wurden das Globalurteil und die sprachlichen Kriterien mit den Niveaus A1 mit B2+ des „Gemeinsamen europäischen Referenzrahmen für Sprachen"[1] abgeglichen.

Die Elemente, die Sie in den Beschreibungen der verschiedenen Stufen eines Kriteriums finden, sind als prototypische Elemente zu verstehen, die diese Stufen kennzeichnen. Das heißt also, dass Sie nicht alle diese Elemente in den jeweiligen Aufsätzen finden werden, sondern durchaus auch Elemente der Nachbarstufen. Das hängt u.a. mit der Interimsprachenentwicklung zusammen, denn jeder Lerner erwirbt sprachliche Phänomene nach einer ihm eigenen Ordnung. Globalisierend sind nur bestimmte Phänomene auf einer Stufe zu beobachten. Deswegen werden Sie abwägen müssen, ob eine Arbeit eher in diese oder eher in jene Stufe fällt, wobei Ihnen die Beschreibung der prototypisch in eine bestimmte Stufe fallenden Elemente dabei helfen soll. Um also einen Text auf eine bestimmte Stufe zu setzen, muss dieser nicht alle Elemente der betreffenden Stufe aufweisen, es sollten sich aber mehrheitlich Elemente dieser Stufe im Text finden. Bei Zweifelsfällen ist ein Schülertext eher auf der unteren Stufe einzuordnen.

Zum **Vorgehen beim *Rating*** finden Sie hier eine Tabelle der *rating steps* im Überblick[2]:

[1] Europarat: Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen. Berlin: Langenscheidt 2001.

[2] In Anlehnung an Tom Lumley: "Assessment criteria in a large-scale writing test: what do they really mean to the raters?" In: *Language Testing* 2002 19 (3), 246-276.

Stage	Rater's focus	Observable behaviours
1. Pre-scoring	code-number text and task	• Identify script
2.1 First reading holistic scoring	Overall impression of text: global and local features and holistic scale	• Read task and text • Comment on salient features • Refer to scale descriptors • Articulate and justify scores
2.2 Consider score given	Scale and text: consistency of score given	• Compare with benchmark texts or other rated texts • Confirm or revise existing score
3.1 Rate the nine analytic categories in turn	Scale and text: focus only on the very category being rated	• Refer to scale descriptors • Reread text • Articulate and justify scores
3.2 Consider score given for each of the analytic categories	Scale and text: consistency of score given	• Compare with benchmark texts or other rated texts • Confirm or revise existing score

1. Identifizieren Sie die Codenummer des Schülertextes und der Aufgabe. Tragen Sie diese ins Kodierblatt ein. **Vorsicht**: nur die reinen Ziffern eintragen, keine Querstriche!
2. Lesen Sie sich die Aufgabenstellung und den Schülertext einmal aufmerksam durch. Geben Sie nach dieser ersten Lektüre Ihr Globalurteil auf einer Skala von 0 bis 5, wobei Sie sich mit den Deskriptoren und dem Schülertext auseinandersetzen, um zu diesem Urteil zu kommen. Notieren Sie sich die auffälligsten Merkmale, um Ihren ersten Eindruck zu verbalisieren. Nutzen Sie die *Benchmark*-Texte und ziehen Sie Vergleiche zwischen den Texten, um zu einem konsistenten Ergebnis zu kommen.
   Tragen Sie den *Score* ins Kodierblatt ein. Diesen *Score* dürfen Sie **nicht** mehr ändern, sobald Sie die folgenden neun Kategorien bewertet haben!
3. Bewerten Sie dann die übrigen Kriterien in der hier angegebenen Reihenfolge, wobei Sie sich fortwährend mit Schülertext und Deskriptoren auseinandersetzen. Versuchen Sie dabei, fortlaufend Ihre vergebenen *Scores* zu verbalisieren und zu rechtfertigen. Nutzen Sie die *Benchmark*-Texte und ziehen Sie Vergleiche zwischen den Texten, um zu einem konsistenten Ergebnis zu kommen.
   Tragen Sie Ihre *Scores* ins Kodierblatt ein.
4. Überprüfen Sie danach noch einmal die von Ihnen ermittelten *Scores*: Sind Ihre *Scores* gerechtfertigt und haltbar? Oder müssen Sie noch Änderungen vornehmen? Falls Ihre Beurteilung der Schülerarbeit stark vom Globalurteil abweicht: Formulieren Sie Gründe und Argumente für die Abweichungen.
   Sie dürfen Ihr Globalurteil **nicht** mehr ändern, nachdem Sie den Schülertext detailliert analysiert haben! Denn das Globalurteil sollte aus statistischen Gründen **unabhängig** vor der Analyse des Schülertextes entstehen und darf daher natürlich nach der Analyse nicht mehr geändert werden.
   Einzige Ausnahme: Sie haben sich tatsächlich geirrt. Dann aber müssen Sie die **gesamte** Bewertung noch einmal vornehmen!

Beschreibung der Kriterien:

1. Globalurteil (ST0201)

Hier sollen Sie Ihren **Gesamteindruck** des Schülertextes wiedergeben. Dieser muss **nicht** dem Durchschnitt der folgenden neun Kategorien entsprechen, sondern es handelt sich hierbei um eine **holistische** Kategorie, die **vor** der detaillierten Analyse des Textes bewertet wird. In diese Kategorie fließen der Gesamteindruck der Versprachlichung und der kommunikativen Botschaft, textuelle Kriterien wie Textsortenadäquanz oder Kohärenz, Kriterien der Flüssigkeit und Idiomatizität der Sprache sowie Natürlichkeit des Stils (*fluency and idiomaticity*), und der Gesamteindruck der emotiven und narrativen Sprache ein.
Als „Kurzformel" gilt:
"Qualität/Relevanz des Inhalts – Textsorte – Qualität der Sprache – kommunikative Wirkung."

Die Bewertung erfolgt auf **6 Stufen** von 0 bis 5, wobei 5 die höchste Bewertungsstufe ist. Hier werden alle 6 Stufen vorgestellt und mit Ganztextbeispielen belegt, um eine differenzierte Abgrenzung zu ermöglichen.

0 – Schülertext liegt unter Stufe 1: ☞ Bspl. 19960204

- Ohne Aufgabenstellung nicht verständlich, was zum Ausdruck gebracht werden sollte bzw. Aufgabenstellung nicht so umgesetzt, dass sie bewertet werden könnte.
- Irrelevanz des Inhalts oder sprachliche Gründe oder Text zu kurz.
- Auf dieser Stufe kann kein Bericht verfasst werden.

1 – *poor attempt*: ☞ Bspl. 19940110

- Ideen ansatzweise relevant, aber nicht entwickelt; Darstellung bleibt konkret; einfache Wendungen und Sätze über sich selbst und andere (auch fiktive) Menschen: wie sie zusammengehören, wo sie leben, wie sie leben, was sie tun oder was sie tun wollen.
- Kurzer, einfachster Text; Mängel im Formalen; meist keine Makrostruktur erkennbar (assoziative Reihung).
- Zeigt begrenzte sprachliche Mittel aus frequentem Basisbereich; Text kann bruchstückhaft und fehlerhaft sein (lexikalische, grammatische, syntaktische, orthographische Fehler; muttersprachliche Interferenzen).
- Gewünschte Botschaft wird nur ansatzweise und oft missverständlich vermittelt.

2 – *some attempt, but inadequate*: ☞ Bspl. 23930110

- Knappe, elementare Beschreibung von Personen, Ereignissen oder Erlebnissen aus dem alltäglichen Umfeld; Entwicklung ist der Aufgabenstellung nicht angemessen; zwischenmenschliche Beziehungen sind in elementarer Form umrissen.
- Einfacher, persönlicher Bericht; formale Anforderungen oft nur teilweise erfüllt; Makrostruktur meist als einfache Reihung oder Aufzählung erkennbar.
- Darstellung in relativ einfacher, begrenzter Sprache, die teils fehlerhaft (z.B. Interferenzen) sein kann, doch Ansätze von Kohärenz zeigt.
- Die kommunikative Zielsetzung wird in groben Zügen erreicht.

3 – *task reasonably achieved*: ☞ Bspl. 24970220

- Unkomplizierter, zusammenhängender Bericht über reale wie fiktive Personen und deren Umfeld und Erfahrungen; Interessenschwerpunkte können sichtbar werden. Zwischenmenschliche Beziehungen, Gefühle und Reaktionen darauf werden in unkomplizierter Weise beschreiben.
- Unkomplizierter Bericht in üblichem Standardformat; teils noch nicht alle Formalia erfüllt; logische Makrostruktur erkennbar, es kann aber zu Sprüngen kommen.
- Sprachliche Mittel werden bis zu einem gewissen Grad angemessen und in hinreichendem Umfang eingesetzt, um das Ziel zu erreichen; gewisse narrative Grundqualitäten (wie etwa Textkohärenz) sind gegeben. Fehler und Interferenzphänomene können im Text vorhanden sein, in der Regel jedoch nicht kommunikationsbelastend.
- Botschaft wird grundsätzlich kommunikativ wirksam vermittelt, doch teils mit Einschränkungen.

4 – *good realisation of task*: ☞ Bspl. NE9

- Klarer, detaillierter, zusammenhängender Bericht über reale wie fiktive Personen, deren Leben, Erfahrungen und Erlebnisse. Beispielsweise sind Zusammenhänge zwischen verschiedenen Ideen verdeutlicht, Entwicklungen aufgezeigt, Gründe benannt, wird auf Unterschiede eingegangen oder es werden Vor- und Nachteile verschiedener Optionen herausgestellt. Einstellungen und Gefühle sowie Zwischenmenschliches werden angemessen versprachlicht.
- Die für das betreffende Genre (Bericht) geltenden Konventionen werden weitgehend beachtet; die Makrostruktur (roter Faden) ist logisch, der Aufbau stringent.
- Umfang und Korrektheit der sprachlichen Mittel sind dem Task angemessen; der Text ist insgesamt flüssig geschrieben und besitzt, sieht man von einzelnen Schwachstellen ab, narrative (u. U. auch humoristische, empathische oder spannende) Qualität.
- Die kommunikative Wirkung wird ohne große Einbußen erreicht.

5 – *task fully completed*: ☞ Bspl. PKG83

- Klare, detaillierte, gut strukturierte und ausführliche Biographie: Es sind entscheidende Punkte hervorgehoben, Standpunkte und Ansichten ausführlich dargestellt und durch Unterpunkte oder geeignete Beispiele oder Begründungen gestützt; der Text ist durch einen angemessenen Schluss abrundet. Einstellungen und Gefühle sind adäquat versprachlicht; verschiedene Perspektiven werden ggf. deutlich gemacht.
- Der Bericht ist entsprechend der geltenden Konventionen geschrieben; der Aufbau ist logisch, stringent und konsistent, u.U. komplex entwickelt; die Makrostruktur erleichtert das Verständnis.
- Großer Umfang und sichere Beherrschung sprachlicher Mittel, Idiomatik und Korrektheit sind meist gegeben. Der Text ist flüssig, in lesergerechtem, überzeugendem, persönlichem und natürlichem Stil verfasst und besitzt durchgehend narrative (oder humoristische, empathische oder spannende) Qualität.
- Die kommunikative Wirkung wird umfassend erzielt.

Task Performance

Die folgenden drei Kriterien beziehen sich auf die Art und Weise, wie auf die Aufgabenstellung inhaltlich und formal reagiert worden ist.
Die Deskriptoren der folgenden drei Kriterien beschreiben die empirisch ermittelten Erwartungen an die Performanz bei genau dieser Aufgabenstellung.

2. Textlänge (ST0202)

Erwartet wird ungefähr **eine Seite** oder ca. 250 Wörter, wobei die Schriftgröße mit bedacht werden muss. In dieses Kriterium fließen überschlagsmäßig die Anzahl der Wörter, und in Zweifelsfällen die Anzahl der Sätze mit ein. Für Sie bedeutet das: Überschlagen Sie die Wortanzahl (und ggf. die Satzanzahl), indem Sie die durchschnittliche Wortanzahl aus fünf repräsentativen Zeilen (meist aus Textmitte) mitteln und auf die Zeilenanzahl hochrechnen. Halbe Zeilen zählen auch nur halb.

Ins Kodierblatt tragen Sie bitte die hochgerechnete Wortanzahl ein.
Wichtig: Im Falle von *missing values* tragen Sie **nichts** ein.

3. Inhalt (ST0203)

Erwartet wird ein **Bericht** über eine Person und deren Biographie im Stil eines Artikels für die Schülerzeitung. Die Person sollte mit Namen, Alter, Herkunft etc. kurz vorgestellt werden (Element 1); es sollte sich eine Beschreibung des persönlichen Umfelds finden (Freunde, Familie etc., Element 2); das bisherige Leben der Person sollte erwähnt werden (Erlebnisse, Jobs etc., Element 3). Des Weiteren sollte von den Träumen, Wünschen und Hoffnungen dieser Person berichtet werden (Element 4) und auf die Gefühle der Person und die zwischenmenschlichen Beziehungen in ihrem Umfeld eingegangen werden (Element 5; dieses Element kann sich durch den ganzen Text ziehen). Natürlich kann es passieren, dass die genannten

Elemente komplex verknüpft werden – dies ist positiv zu bewerten (Beispielsweise können Gefühle mit dem bisherigen Leben verknüpft werden).Der Bericht soll inhaltlich zu Ende geführt sein.

Die Biographie darf phantasievoll sein und muss nicht unbedingt realistisch bleiben. Den Schülerinnen und Schülern muss größtmögliche Freiheit zugestanden werden hinsichtlich dessen, was sie als relevant und interessant empfinden.

Im Folgenden werden die genannten fünf Inhaltselemente anhand von Schülerbeispielen vorgestellt:

1. Kurzvorstellung Person (Alter, Name, Herkunft etc.)
   PKG40 *Victoria Pearson was born in New York in the year 1980.*
   PKG63 *His name is Henning Stensrud, he is from Sweden…*
   PKG87 *Roland Mayer, who lives in Bielefeld, was born in Berlin in 1920.*

2. Persönliches Umfeld (Freunde, Familie etc.)
   PKG84 *His father died when Sean was twelve in a car accident.*
   PKG88 *He has no children and only one brother, who lives in Austria*
   PKG44 *...her family wasn't very rich and her father was alcohol-addicted.*

3. Bisheriges Leben (Erlebnisse, Jobs, Erfahrungen etc.)
   PKG66 (Unfall mit Michael Jackson bei Paparazzi-Verfolgungsjagd)
   PKG83 *She lived from parttime jobs in some coffes or bars.*
   PKG43 *When she was 14 she get married and lived together with her husband.*

4. Träume, Wünsche, Hoffnungen
   PKG90 *He wants to enjoy his life as long as he can, because he often thinks about the death, and what's after it.*
   PKG39 *…Steve McManaman´s dream come true: To live in Harlem and work for a successful magazine…*
   PKG42 *Since he was a little boy, he wanted to write.*

5. Gefühle und zwischenmenschliche Beziehungen (darf sich durch den ganzen Text ziehen)
   PKG66 *Thomas is very proud of his two boys because…*
   PKG86 (letzter Abschnitt: Beschreibung der Gefühle des Mannes, wenn er alleine auf der Bank sitzt und an die Zeit mit seiner Frau zurückdenkt)
   PKG88 *Paul´s neighbors think that his life must be very boring, but Paul himself likes his life.*

Bewertet werden die inhaltliche Qualität, die Motiviertheit der Ideen, begründete Entwicklungen und die Ausführlichkeit der Darstellungen.

Wichtiger als das quantitative Vorhandensein der oben genannten fünf Inhaltselemente ist demnach die **qualitative** Umsetzung, d. h. auch nicht alle der oben vorgestellten fünf Elemente sprachlich umgesetzt sind, die vorhandenen Ideen jedoch ausführlich und angemessen umgesetzt werden, kann die volle Punktzahl erreicht werden (z. B. weil die *language of emotion* überzeugend in die Biographie eingebaut ist).

☞ PKG40 – Bewertung mit 5 Punkten (zwar nicht alle Elemente vorhanden, aber eine ausführliche Darstellung des Lebenswegs bis zur Misswahl)

Umgekehrt kann es vorkommen, dass zwar quantitativ alle Inhaltselemente vorhanden sind, doch qualitativ so mangelhaft umgesetzt sind, dass die Maximalpunktzahl nicht vergeben werden kann:

☞ Bspl. PKG87 (es sind zwar alle Elemente vorhanden, aber zu kurz und zu wenig ausführlich – 3 Inhaltspunkte)

Die makrostrukturelle Anordnung der Inhaltselemente wird hier jedoch nicht bewertet – dies erfolgt bei Kriterium 3 **Textsorte und Aufbau**.

Die Qualität der Sprache wird hier ebenfalls nicht bewertet, dies erfolgt bei den **Kriterien 5 mit 8**.

Handelt es sich um eine **Themaverfehlung**, muss die Arbeit in **allen** Kategorien mit „**8**" kodiert werden.

Wird das geforderte Thema inhaltlich korrekt bearbeitet, doch in einer anderen als der geforderten Textsorte, so erfolgt nur die Bewertung der **Textsorte** mit „**0**", die **inhaltliche** Bewertung (sowie die Bewertung der anderen Kriterien) erfolgt jedoch nach Schülerleistung.

Die Bewertung erfolgt auf **6 Stufen** von 0 bis 5, wobei 5 die höchste Bewertungsstufe ist. Hier werden alle 6 Stufen vorgestellt, um eine differenzierte Abgrenzung zu ermöglichen.

Bewertung des inhaltlichen Kriteriums:

0 – Es ist zwar ein inhaltlicher Bezug erkennbar, doch ist die Arbeit inhaltlich so knapp gehalten (z. B. Aufgabenstellung nicht umgesetzt: „I want to write about this man, but it is too difficult to explain"), *oder* die umgesetzten Themenbereiche sind von solcher Irrelevanz, dass eine Bewertung des Inhalts **nicht** erfolgen kann. ☞ Bspl. 19960204

1 – Die inhaltlichen Themenbereiche sind so mangelhaft umgesetzt (Aufgabenstellung mangelhaft umgesetzt, z. B. keine zusammenhängende Biographie zu der betreffenden Person geschrieben), und von solch einer Kürze und Qualität, dass die Aufgabe inhaltlich **nicht zufrieden stellend** gelöst ist. ☞ Bspl. 19940110

2 – Der Inhalt ist von ausreichender Qualität: Es handelt sich entweder um eine äußerst knappe Versprachlichung der oben genannten Themenbereiche (knappster Abriss der Lebensgeschichte) *oder* es fehlen relevante Aspekte, so dass die Aufgabe inhaltlich als **gerade ausreichend** bewertet werden muss. ☞ Bspl. 23930110

3 – Es sind alle Themenbereiche knapp angesprochen und von befriedigender Qualität *oder* es werden thematische Schwerpunkte etwas ausführlicher ausgeführt, doch fehlen dabei relevante Aspekte (z. B. wenig Gefühle, Träume, Wünsche versprachlicht), so dass die Aufgabenstellung **als befriedigend erfüllt** angesehen werden kann. Auf dieser Stufe wird erwartet, dass der Inhalt eine gewisse Relevanz zur Aufgabe widerspiegelt. ☞ Bspl. 24970220

4 – Der Inhalt ist von guter Qualität: Entweder sind Schwerpunkte inhaltlich ausführlich umgesetzt (ohne dass etwas Relevantes fehlen würde) *oder* es sind alle oben geforderten Themenbereiche realisiert, doch nicht vollständig und ausführlich umgesetzt. Die Aufgabe ist inhaltlich **erfüllt**. ☞ Bspl. NE9

5 – Es sind die oben geforderten Inhaltselemente vollständig und elaboriert in sehr guter Qualität angeführt *oder* es werden weniger Ideen angesprochen, doch ist die Qualität der vorhandenen Elemente dergestalt, dass die Aufgabe inhaltlich als **voll erfüllt** betrachtet werden kann. ☞ Bspl. PKG83

Bei **Grenzfällen** kann es helfen, nach der Umsetzung der Personendarstellung („Entsteht ein Bild der betroffenen Person?“) zu suchen, oder nach der emotiven Sprache oder sonstigen auffälligen Stärken oder Schwächen der Arbeit: Je nachdem erfolgt die Einordnung eher nach oben oder nach unten.

Folgendes Vorgehen kann bei der Bewertung helfen:

- Ungenügende Umsetzung von Ideen, Aufgabe nicht erfasst und nicht erfüllt => 0
- Aufgabe inhaltlich nur mangelhaft erfasst und nur mangelhaft erfüllt (z. B. entsteht kein Bild der Person; die Darstellung bzw. die Ideen sind äußerst knapp) => 1
- Knappe Darstellung der Ideen:
  - Ideen knapp und kurz dargestellt, von ausreichender Qualität, jedoch z. B. Darstellung der zu beschreibenden Person dürftig => 2
  - Alles inhaltlich vorhanden und in befriedigender Qualität => 3
- Ausführlichere Darstellung der Ideen:
  - aber einzelne Themenbereiche nicht abgedeckt:
    Qualität neigt eher zu befriedigend, Darstellung könnte ausführlicher sein => 3
    Qualität der vorhandenen Elemente ist aber gut => 4
  - alles abgedeckt (und in guter Qualität) => 4

  [Abgrenzung zu 5: Hier erwartet man auch die Versprachlichung von Einstellungen, Gefühlen und zwischenmenschlichen Beziehungen, die auf Stufe 4 noch nicht so elaboriert vorhanden sein muss.]
- Elaborierte Darstellung der Ideen:
  Dazu müssen Person und deren Lebenshintergründe ausführlich dargestellt sein (es dürfen aber aus obigen fünf Themenbereichen individuelle Schwerpunkte gesetzt werden, die entsprechend ausgeführt sein müssen) und die Qualität muss insgesamt sehr gut sein => 5

4. Textsorte und Aufbau (ST0204)

Erwartet wird ein **Bericht** über eine Person und deren Biographie im Stil eines Artikels für die Schülerzeitung. Der Bericht soll hinsichtlich des **makrostrukturellen** Aufbaus und der **Themenentwicklung** den üblichen Gepflogenheiten folgen; d.h. erwartet wird eine Einleitung sowie eine inhaltlich logisch gegliederte Makrostruktur (z. B. könnte die Person vorgestellt werden, ehe auf Einzelheiten ihres Lebens eingegangen wird). Der Bericht soll zu Ende geführt sein.

Nicht in diese Bewertungskategorie fallen die sprachliche Organisation und die Verknüpfung durch *linking language*; diese werden unter der Kategorie 8 **Sprachliche Organisation** bewertet.

Nur wenn es sich eindeutig **nicht** um einen Bericht über das Leben einer Person (Biographie) für einen Schülerzeitungsartikel handelt, der Text sonst aber die Aufgabenstellung widerspiegelt, kodieren Sie den

Text in dieser Kategorie mit **0** (beispielsweise wenn eine Autobiographie zu einem der Bilder verfasst wurde), bewerten den Text aber in den anderen Kriterien.

Die Bewertung erfolgt auf **6 Stufen von 0 bis 5**, wobei 5 die höchste Bewertungsstufe ist. Hier werden alle 6 Stufen vorgestellt, um eine differenzierte Abgrenzung zu ermöglichen.

0 – Es handelt sich nicht um einen Schülerzeitungsbericht *oder* der Text ist nur mithilfe der Aufgabenstellung als solcher erkennbar. ☞ Bspl. 19960204

1 – Es handelt sich um einen Text, der nur bedingt als Schülerzeitungsbericht erkennbar ist. Oft fehlen gebräuchliche Elemente *und*/*oder* der Text ist nicht zu Ende geführt. Beim Erzählen oder Berichten kommt es zu Brüchen, eine Themenentwicklung ist nicht erkennbar. Die Inhaltselemente sind nicht logisch miteinander verknüpft und folgen keiner erkennbaren Struktur; dies kann zu Verständnisschwierigkeiten führen. ☞ Bspl. 19940110

2 – Der Text ist als Schülerzeitungsbericht erkennbar, doch sind die fremdsprachlichen Gepflogenheiten fehlerhaft umgesetzt (z.B. fehlt ein abgerundetes Ende, oder die Inhaltselemente sind unlogisch angeordnet). Der Text ist meist nicht zu Ende geführt. Es kommt zu Sprüngen in der Themenentwicklung, die i. d. R. aus einer einfachen Aufzählung besteht *oder* die Themenentwicklung ist teils unlogisch. Der Text ist nicht in Absätze gegliedert *oder* die Absatzgliederung folgt keinem sinnvollen Schema. ☞ Bspl. 23930110

3 – Der Text ist als Schülerzeitungsbericht erkennbar, doch sind die fremdsprachlichen Gepflogenheiten teils fehlerhaft umgesetzt. Beispielsweise sind (nicht immer angemessene) Absätze vorhanden, doch Mängel in der Themenentwicklung erkennbar *oder* die Themenentwicklung ist zufrieden stellend bei fehlenden Absätzen. Auf dieser Stufe erfolgt die Themenentwicklung meist linear, doch sind Ansätze komplexerer Verknüpfungen erkennbar; der Text wird i. d. R. ansatzweise in Absätze gegliedert, die aber nur bedingt den Gepflogenheiten entsprechen. Der Text sollte auf dieser Stufe zu Ende geführt sein; doch gute Themenentwicklung und Absatzeinteilung können ein fehlendes Ende aufwiegen. ☞ Bspl. 24970220

4 – Der Text entspricht im Großen und Ganzen den Gepflogenheiten eines Schülerzeitungsartikels, doch sind kleinere Mängel erkennbar: z. B. keine Absätze trotz angemessener Themenentwicklung *oder* angemessene Absatzeinteilung bei nicht ganz zufrieden stellender Themenentwicklung *oder* Anordnung der Inhaltselemente lässt Mängel erkennen, obwohl Absatzeinteilung den Gepflogenheiten entspricht, etc. Die Anordnung der Inhaltselemente ist logisch. ☞ Bspl. NE9

5 – Es handelt sich um einen Schülerzeitungsbericht, der den fremdsprachlichen Gepflogenheiten **voll** entspricht. Der Aufbau erleichtert das Verständnis und folgt den üblichen Gepflogenheiten (z. B. führt die Einleitung den Leser auf die Aussage hin und der Bericht besitzt ein abgerundetes Ende). Die Anordnung der Inhaltselemente ist komplex, doch logisch. Die Themenentwicklung ist klar und konsistent; relevante Punkte werden z.B. mit Details oder Beispielen ausgeführt. ☞ Bspl. PKG83

Language Performance

Bei den folgenden fünf Kriterien werden Umfang und Korrektheit der Schülersprache bewertet.
Die Deskriptoren dieser Kriterien beschreiben typische sprachliche Elemente und Phänomene, die anhand von Lernertextanalysen zu dieser Aufgabenstellung identifiziert wurden. Die KANN-Formulierungen sind, wo angemessen, an relevante Formulierungen des GER auf den Niveaus A1 mit B2+ angelehnt.

5. Orthographie (ST0205)

Bei diesem Kriterium geht es vorwiegend darum festzustellen, was schon korrekt geschrieben werden kann, **nicht** jedoch darum, Fehler im Bereich der Rechtschreibung oder Zeichensetzung quantitativ zu erfassen. Erst in zweiter Linie geht es um solche Fehler, die sich sinnentstellend und damit kommunikationsbelastend auswirken.

Die Auswirkung solcher ***impeding errors*** soll auf den **ganzen** Text bezogen bewertet werden. Zudem spielt der Umfang des demonstrierten Wortschatzes eine Rolle, denn ein fehlerfreier Text, der sich nur im sprachlichen Basisbereich bewegt, liegt auf einer niedrigeren Stufe als einer, der komplexe Strukturen und Wörter fehlerfrei wiedergeben kann.

Die Bewertung erfolgt auf **6 Stufen** von 0 bis 5, wobei 5 die höchste Bewertungsstufe ist. Hier werden alle 6 Stufen vorgestellt, um eine differenzierte Abgrenzung zu ermöglichen.

0 – Schreibt einen so kurzen Text, dass eine Bewertung auf Stufe 1 nicht möglich ist *oder* schreibt einen so fehlerhaften Text, dass der Text fast unverständlich ist. ☞ Bspl. 19960204

1 – Kann vertraute, meist hochfrequente Wörter und kurze Redewendungen schreiben. Jenseits dieser engen Bereiche sind Rechtschreibfehler häufig und von einer Art, dass sie die Verständlichkeit der Mitteilung belasten und sinnentstellend wirken. *Oder*: Schreibt einen so kurzen Text mit Wörtern des Elementarwortschatzes, dass der Schülertext eine Beurteilung jenseits dieser Stufe nicht zulässt. ☞ Bspl. 19940110

2 – Kann Mitteilungen zu alltäglichen Themen sowie Lexeme des Elementarwortschatzes im Allgemeinen korrekt wiedergeben; jenseits dieser engen Bereiche werden Lexeme so geschrieben, dass die reichlich vorhandenen Rechtschreibfehler die phonetische Gestalt der Eintragungen einigermaßen korrekt wiedergeben. Deutliche muttersprachliche Einflüsse sind gegeben. Die Rechtschreibfehler wirken sich oft kommunikationsbelastend aus. ☞ Bspl. 23930110

3 – Kann orthographisch im Ganzen konsistent schreiben, doch zeigen sich Einflüsse der Muttersprache. Rechtschreibung und Zeichensetzung sind exakt genug, so dass der Text bis auf einzelne, isolierte Lexeme und Passagen durchgängig verständlich ist. ☞ Bspl. 24970220

4 – Kann orthographische Konventionen auch bei Wörtern jenseits des Grundwortschatzes weitgehend einhalten, die Zeichensetzung ist hinreichend korrekt. Muttersprachliche Einflüsse werden vereinzelt deutlich. ☞ Bspl. PKG83

5 – Kann auch orthographisch anspruchsvolle Wörter korrekt schreiben und ist sicher in der Zeichensetzung, abgesehen von gelegentlichen Inkonsistenzen und wenigen muttersprachlichen Einflüssen im Bereich der Zeichensetzung. ☞ Bspl. PKG 27 (Insel-*Story*)

6. Lexik und Lexiko-Grammatik (ST0206)

Betrachtet wird hier der produktive Wortschatz, sein Umfang („Verfügbarkeit“) und das Maß an Korrektheit („Beherrschung“), mit dem er eingesetzt wird. In diese Kategorie fallen aber nicht nur lexikalische Elemente, sondern auch alles, was zur Wortbedeutung beiträgt: Wortanschlüsse, die Bildung von Wörtern und Phrasen, Wortvalenzen, Kollokationen und die idiomatische Verwendung der lexikalischen Mittel.

Die Bewertung erfolgt auf **6 Stufen** von 0 bis 5, wobei 5 die höchste Bewertungsstufe ist. Hier werden alle 6 Stufen vorgestellt, um eine differenzierte Abgrenzung zu ermöglichen.

0 – Zeigt nur unzureichende Beherrschung des Vokabulars (*insufficient evidence*). ☞ Bspl. 19960204

1 – Verfügt über einen elementaren Vorrat an einzelnen, meist hochfrequenten Wörtern und Wendungen, die sich auf eng begrenzte konkrete Situationen beziehen. ☞ Bspl. 19940110

2 – Beherrscht einen begrenzten Wortschatz in Zusammenhang mit einer konkreten Personenbeschreibung. Verfügt über genügend Wortschatz, um elementaren Kommunikationsbedürfnissen gerecht zu werden. ☞ Bspl. 23930110

3 – Zeigt eine gute Beherrschung des Grundwortschatzes, macht aber elementare Fehler, wenn es darum geht, komplexere Sachverhalte auszudrücken oder wenig vertraute Situationen und Themen zu bewältigen. Verfügt über einen ausreichend großen Wortschatz, um sich mit Hilfe von einigen Umschreibungen über eine Person und deren Leben äußern zu können; Fehler in Bezug auf Kollokationen und *complementations* kommen vor. ☞ Bspl. 24970220

4 – Verfügt über einen großen Wortschatz, um sich über auch komplexere Themen im Rahmen einer Biographie äußern zu können. Der Wortschatz wird i. d. R. angemessen verwendet, obgleich einige Verwechslungen und falsche Wortwahl vorkommen, ohne jedoch die Kommunikation zu behindern. Kann Formulierungen variieren, um häufige Wiederholungen zu vermeiden; Lücken im Wortschatz können dennoch zu Umschreibungen führen. Die gebräuchlichsten Kollokationen und Anschlüsse sind vorhanden. ☞ Bspl. PKG83

5 – Beherrscht einen breit ausgefächerten Wortschatz; selten kann es zu kleineren Fehlleistungen kommen. Gute Beherrschung idiomatischer Ausdrücke und Kollokationen, umgangssprachlicher Wendungen und der Wortbildung. Modifikationen sind zahlreich und angemessen. ☞ Bspl. PKG 27 (Insel-*Story*)

7. Grammatik (ST0207)

Betrachtet werden hier die grammatischen Strukturen, deren Umfang und das Maß an Korrektheit, mit dem sie eingesetzt werden. Zu den zu bewertenden Strukturen zählen u. a. die Satzmuster (z.B. Parataxe, Hypotaxe, verkürzte Nebensätze, etc.), Flexions- und Kongruenzphänomene, *voice*, *tenses* und *aspect*, die Wortstellung und die Artikelverwendung; auch Pluralbildung und -verwendung sowie unregelmäßige Verben gehen in diese Kategorie ein.

Die Bewertung erfolgt auf **6 Stufen** von 0 bis 5, wobei 5 die höchste Bewertungsstufe ist. Hier werden alle 6 Stufen vorgestellt, um eine differenzierte Abgrenzung zu ermöglichen.

0 – Zeigt nur unzureichende Beherrschung der Grammatik (*insufficient evidence*). ☞ Bspl. 19960204

1 – Zeigt eine begrenzte Beherrschung einiger weniger einfacher grammatischer Strukturen und parataktischer Satzmuster in einem auswendig gelernten Repertoire. Es sind gravierende muttersprachliche Einflüsse vorhanden. Das Verständnis kann stark eingeschränkt werden. ☞ Bspl. 19940110

2 – Kann einige einfache Strukturen und Satzmuster (Parataxe und wenige auswendig gelernte komplexere Satzmuster) korrekt verwenden, macht aber systematisch elementare Fehler, hat z. B. die Tendenz, Zeitformen zu vermischen bzw. zu verwechseln, oder Endungen (z. B. bei Adverbien oder 3.Person-*s*) nicht zu setzen; hat z. B. Probleme mit Nebensätzen. Es kommt zu Fehlern in der Wortstellung. Es zeigen sich deutliche muttersprachliche Interferenzen. Trotzdem wird in der Regel klar, was ausgedrückt werden soll. ☞ Bspl. 23930110

3 – Kann ein Repertoire von häufig verwendeten Redefloskeln und von Wendungen, die an eher vorhersehbare Situationen gebunden sind, ausreichend korrekt verwenden. Es zeigen sich einfache Variationen in den Satzmustern; hypotaktische Sätze (Nebensätze oft mit *when, but* eingeleitet) sowie Relativsätze und *if-clauses* sind zwar vorhanden – genau wie Passivkonstruktionen oder indirekte Rede – doch i. d. R. werden sie fehlerhaft eingesetzt. Im Allgemeinen hinreichende Beherrschung der grammatischen Strukturen trotz gelegentlicher Einflüsse der Muttersprache. Zwar kommen Fehler vor, aber es bleibt klar, was ausgedrückt werden soll. ☞ Bspl. 24970220

4 – Zeigt gute Beherrschung von Grammatik und komplexeren Satzstrukturen; Fehler, die zu Missverständnissen führen, kommen nicht vor. Beispielsweise werden die *tenses* korrekt verwendet; verkürzte Nebensätze angewandt; das Passiv verwendet. Gelegentliche Ausrutscher oder nicht-systematische Fehler und kleinere Mängel im Satzbau sind vorhanden, jedoch selten und können oft rückblickend korrigiert werden. ☞ Bspl. NE9

5 – Verfügt über ein hohes Maß an grammatischer Korrektheit und kann auch komplexe Strukturen angemessen verwenden; Fehler sind selten und fallen kaum auf. Beispielsweise zeigt sich eine sichere Beherrschung der Relativsätze, sämtlicher *if-clause*-Typen und der Zeitenfolge in komplexen Satzgefügen; die *ing*-Form wird adäquat eingesetzt; syntaktische Strukturen werden variiert.
☞ Bspl. PKG 27 (Insel-*Story*)

8. Kohäsion (ST0208)

Bewertet werden hier die **sprachliche** Gliederung des Textes, die **sprachliche** Verknüpfung der Inhalte und Ideen, der Einsatz von Kohäsionsmitteln (z.B. *referencing, linking language, cohesive ties*, Wortfelder, deiktische Elemente, Pronomina und Konjunktionen, etc.) und das Vorhandensein (oder eben Nichtvorhandensein) eines textuellen Netzwerks von sprachlichen Verknüpfungen, das zur Kohärenz eines Textes entscheidend beiträgt.

Nicht bewertet werden hier die inhaltlich-logische Gliederung sowie die Themenentwicklung, denn dies wird bei Kriterium 4 **Textsorte und Aufbau** bewertet.

Die Bewertung erfolgt auf **6 Stufen** von 0 bis 5, wobei 5 die höchste Bewertungsstufe ist. Hier werden alle 6 Stufen vorgestellt, um eine differenzierte Abgrenzung zu ermöglichen.

0 – Kohäsionsmittel werden nicht eingesetzt. ☞ Bspl. 19960204

1 – Kann Wörter und Wortgruppen durch sehr einfache Konnektoren wie *and* oder *then* verbinden. Ansonsten sind die Sätze ohne *cohesive ties* aneinander gereiht. Kohäsionsmittel werden nur mangelhaft eingesetzt. Dadurch kann es zu Verständnisschwierigkeiten und missverständlichen Aussagen kommen, oder der Text erscheint als *patchwork*. ☞ Bspl. 19940110

2 – Kann Wortgruppen durch einfache Konnektoren wie *and, but*, *because* verknüpfen. Kann die häufigsten Konnektoren wie z.B. *after, so, then* benutzen, um einfache Sätze miteinander zu verbinden, um etwas zu berichten oder zu beschreiben. Kann Kohäsionsmittel nur bedingt einsetzen, wodurch die kommunikative Absicht belastet werden kann. Es zeigen sich *cohesive ties* zwischen einzelnen Sätzen, doch noch nicht im Text insgesamt. ☞ Bspl. 23930110

3 – Kann eine Reihe kurzer und einfacher Einzelelemente zu einer meist linearen, zusammenhängenden Äußerung verbinden, indem gebräuchliche Kohäsionsmittel meist angemessen eingesetzt werden. Ein textuelles Netzwerk zeichnet sich ab, ebenso wie ansatzweise Kohärenz entsteht. ☞ Bspl. 24970220

4 – Kann eine Anzahl weniger frequenter sprachlicher Verknüpfungsmitteln (z.B. *instead of, not only...but also,* etc.) sinnvoll verwenden, um inhaltliche Beziehungen deutlich zu machen. Die Äußerungen werden durch den Gebrauch verschiedener Kohäsionsmittel zu einem kohäsiven, zusammenhängenden Text verbunden; längere Beiträge sind möglicherweise etwas sprunghaft. ☞ Bspl. NE9

5 – Kann klar, fließend und sprachlich gut strukturiert schreiben und zeigen, dass die Mittel der sprachlichen Gliederung sowie der sprachlichen Verknüpfung beherrscht werden. *Cohesive ties* werden durchgehend und konsistent eingesetzt. Der angemessene Gebrauch der sprachlichen Kohäsionsmittel trägt dazu bei, dass der Text ein kohärentes Ganzes darstellt. ☞ Bspl. PKG 27 (Insel-*Story*)

Communicative Performance

Das folgende Kriterium spiegelt das oberste Ziel im Fremdsprachenunterricht wider: die kommunikative Kompetenz. Diese soll über die Bewertung des kommunikativen Erfolgs erfasst werden. Das Kriterium des kommunikativen Erfolgs ist notwendigerweise ein hochkomplexes (und damit von anderen Bewertungskategorien abhängiges), denn kommunikativer Erfolg kann über eine Vielzahl von Faktoren erreicht werden, die im Folgenden auf sechs Niveaus beschrieben werden.

9. Kommunikative Wirkung (ST0209)

Bei diesem komplexen und von den obigen Kriterien abhängigen Kriterium werden die kommunikative **Wirkung** des Textes auf die (fiktiven) Rezipienten und damit einhergehend der **Adressatenbezug** und die kommunikativ erfolgreiche Versprachlichung der **Redeabsicht** bewertet. Auch der **Stil** des Berichts beeinflusst die kommunikative Wirkung und soll hier bewertet werden. Allerdings ist Stil hier nicht im Sinne von formal vs. informell zu verstehen, sondern eher im Bezug auf lebhafte, fesselnde oder emotive Sprache, wie man sie in einem Schülerzeitungsbericht erwarten kann.

Sie müssen sich also einmal in die Rolle der Jugendlichen versetzten, an die sich der Schülerzeitungsbericht wendet, und zum anderen bedenken, dass es sich bei den (fiktiven) Rezipienten um Muttersprachler des Englischen handeln soll. Gehen Sie dabei von Rezipienten aus, die äußerst *broad-minded* sind.

Bewertet werden einerseits die **Versprachlichung der Redeabsicht** und der (hier sehr breit auszulegende) **Adressatenbezug**: Fühlt sich der Rezipient vom Text angesprochen? Wird klar, was der Autor ausdrücken wollte? Andererseits wird der kommunikative Erfolg auch davon bestimmt, ob der Autor im Bericht ein **Bild von der betreffenden Person und deren Leben** zeichnen konnte. Dabei fließt der oben ausgeführte Stil mit ein: Ist die Sprache so **lebendig, fesselnd, emotiv und interessant**, dass man weiterlesen möchte?

Die Bewertung erfolgt auf **6 Stufen** von 0 bis 5, wobei 5 die höchste Bewertungsstufe ist. Hier werden alle 6 Stufen vorgestellt, um eine differenzierte Abgrenzung zu ermöglichen.

0 – Eine kommunikative Wirkung ist nicht erkennbar; die Äußerung ist (nahezu) unverständlich; ein Adressatenbezug ist (fast) nicht erkennbar (z.B. auf Englisch nur knapper Hinweis auf Beschreibung einer Person und der Hinweis, dass die Aufgabe zu schwer sei). ☞ Bspl. 19960204

1 – Die kommunikative Wirkung ist durch den begrenzten Umfang der Redemittel und den oft wenig kohärenten Aufbau stark eingeschränkt. Es kommt zu missverständlichen bzw. teils unverständlichen Äußerungen. Ein Adressatenbezug ist oft nur ansatzweise erkennbar. Ein Bild der betreffenden Person und von deren Leben entsteht nicht. ☞ Bspl. 19940110

2 – Kann sich in bekannten, alltäglichen Situationen verständlich machen und die Schreibabsicht in groben Zügen vermitteln. Ein Adressatenbezug ist gegeben, jedoch ist die Effizienz eingeschränkt. Ein Bild von der betreffenden Person und deren Leben ist ansatzweise erkennbar. Hat Mühe, z. B. Emotionen zu versprachlichen. ☞ Bspl. 23930110

3 – Kann sich einigermaßen präzise ausdrücken und deutlich machen, was bedeutsam ist, doch ist die effektive Versprachlichung i. d. R. eingeschränkt. Ein Adressatenbezug ist gegeben. Ein Bild von der betreffenden Person und deren Leben entsteht, ist aber nicht elaboriert. Die Darstellung ist insgesamt interessant und ansprechend. Emotionen und persönliche Einstellungen können ansatzweise versprachlicht werden. ☞ Bspl. 24970220

4 – Kann Redeabsichten, Standpunkte, Gefühle oder Einstellungen meist effektiv ausdrücken und sich auf die Absichten und Standpunkte von anderen beziehen: Die Mitteilung ist insgesamt rezipientenbezogen und für diese verständlich. Ein detailliertes Bild von der betreffenden Person und deren Leben ist gegeben. Schreibt in lebhafter, fesselnder Sprache und kann die eigenen Gedanken interessant darstellen. ☞ Bspl. NE9

5 – Kann sich so klar, präzise und angemessen ausdrücken, dass die Redeabsicht umfassend rezipierbar ist. Bezieht sich dabei durchgehend flexibel und effizient auf die Adressaten und gibt eine elaborierte Darstellung von der betreffenden Person und deren Leben. Der Stil ist lebhaft, fesselnd und interessant; Emotionen werden angemessen versprachlicht. ☞ Bspl. PKG83

10. Swear Words (ST0210)

In dieser Kategorie soll erfasst werden, ob sich im Schülertext *swear words* oder sonstige Fäkalsprache finden. Dieses Vorgehen soll es Ihnen erleichtern, auch „fäkalsprachliche“ Texte neutral zu bewerten.

Sie kodieren mit **1**, wenn solche sprachlichen Elemente **vorhanden** sind, und mit **0**, wenn sich **keine** solchen Belege im Text finden lassen.

Code 0: Keine Fäkalsprache erkennbar.

Code 1: Fäkalsprache vorhanden. ☞ Bspl. 19940208

Abschluss des *Ratings*:

Überprüfen Sie nun noch einmal die von Ihnen ermittelten *Scores* in den **analytischen *Rating*-Kriterien**: Sind Ihre *Scores* gerechtfertigt, plausibel und haltbar? Oder müssen Sie noch Änderungen vornehmen? Sie **müssen** dazu Vergleiche mit den *Benchmark*-Texten oder anderen schon bewerteten Texten ziehen, um zu einer konsistenten Bewertung zu kommen.
Falls Ihre Beurteilung der Schülerarbeit stark vom Globalurteil abweicht: Formulieren Sie Gründe und Argumente für die Abweichungen.

Sie dürfen aus Gründen der Unabhängigkeit Ihr Globalurteil **nicht** mehr ändern, nachdem Sie den Schülertext detailliert analysiert haben!

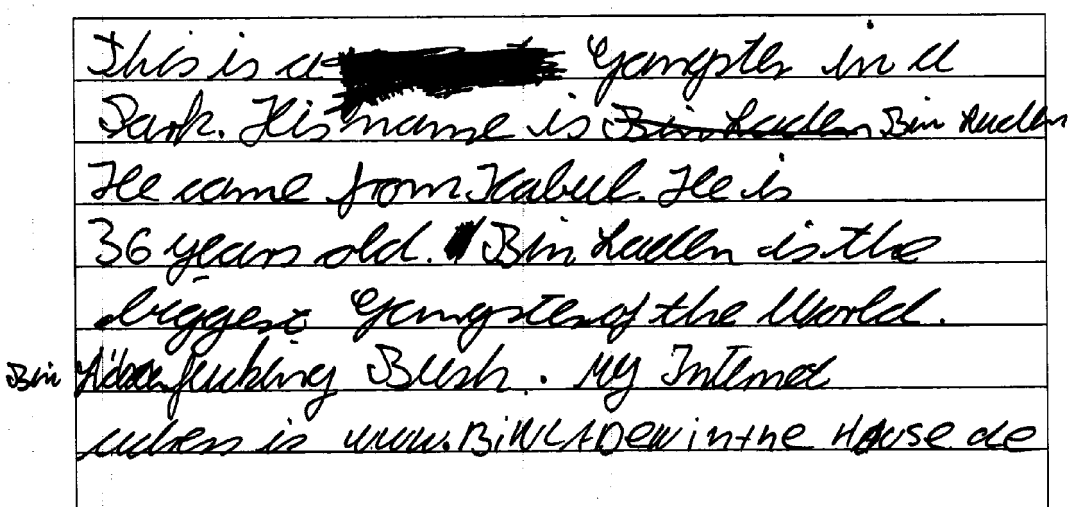
Einzige Ausnahme: Sie haben sich tatsächlich geirrt - in diesem Fall beginnen Sie bitte von vorne und bewerten den Schülertext noch einmal in **allen** Kriterien.

Anhang 28: *Benchmark*-Texte zum Task *Biography*

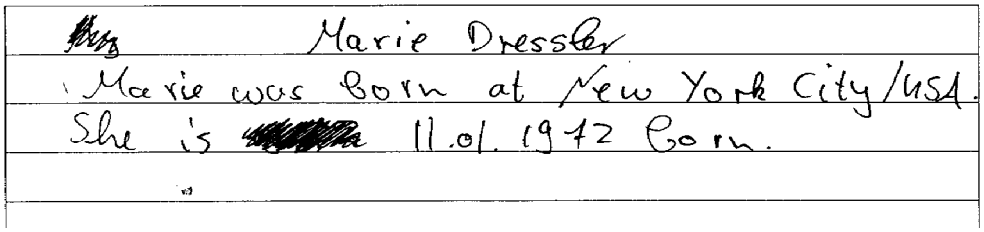
Tabelle, die die Zuordnung der *Benchmark*-Texte auf die Bewertungsstufen 0 mit 5 verdeutlicht:

Schülercode	1. Globalurteil	2. Länge	3. Inhalt	4. Sorte + Aufbau	5. Orthographie	6. Lexik	7. Grammatik	8. Kohäsion	9. Komm. Wirkung	11. Swear Words
15910112 Bin Laden	8		8	8	8	8	8	8	8	1
19960204 Marie Dressler	0	15	0	0	0	0	0	0	0	0
19940110 director of stadion	1	55	1	1	1	1-	1-	1-	1	0
23930110 Neumann	2	162	2	2	2	2	2	2	2	0
24970220 Marlene	3	140	3	3	3	3	3	3	3	0
NE9 Steward Brown	4	210	4	4-	3-4	3-4	4	4	4	0
PKG 83 Model	5	>250	5	5-	4	4-	3-4	4	5	0
PKG 27 (Vorsicht: anderer Task!) **Insel-*Story***					5	5	5	5		0

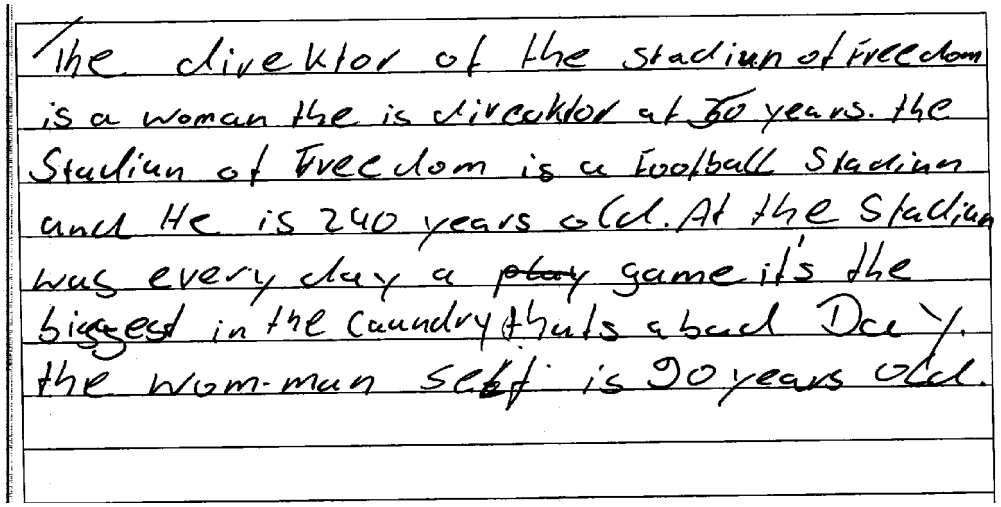
Nicht valider Text: 15910112:



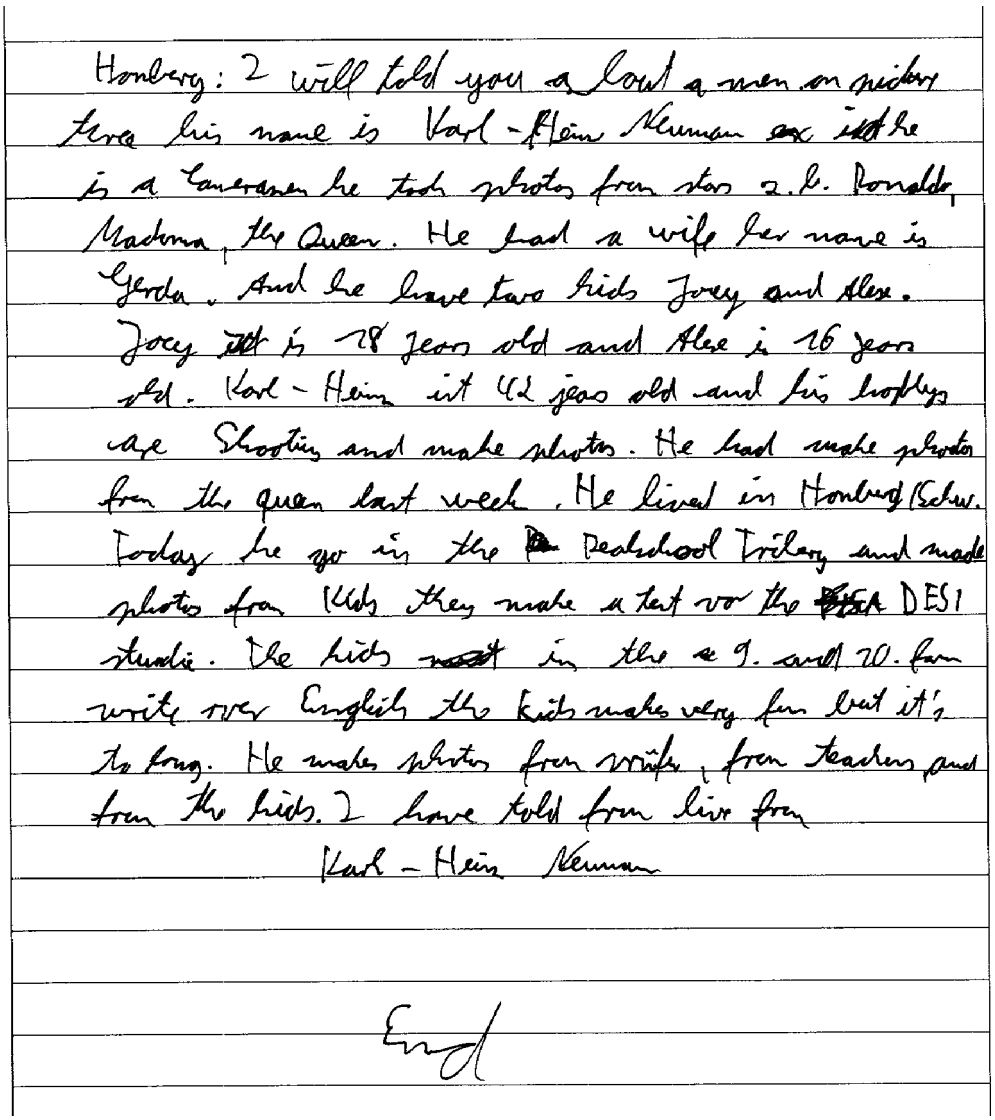
Globalurteil Niveau 0: 19960204:



Globalurteil Niveau 1: 19940110:



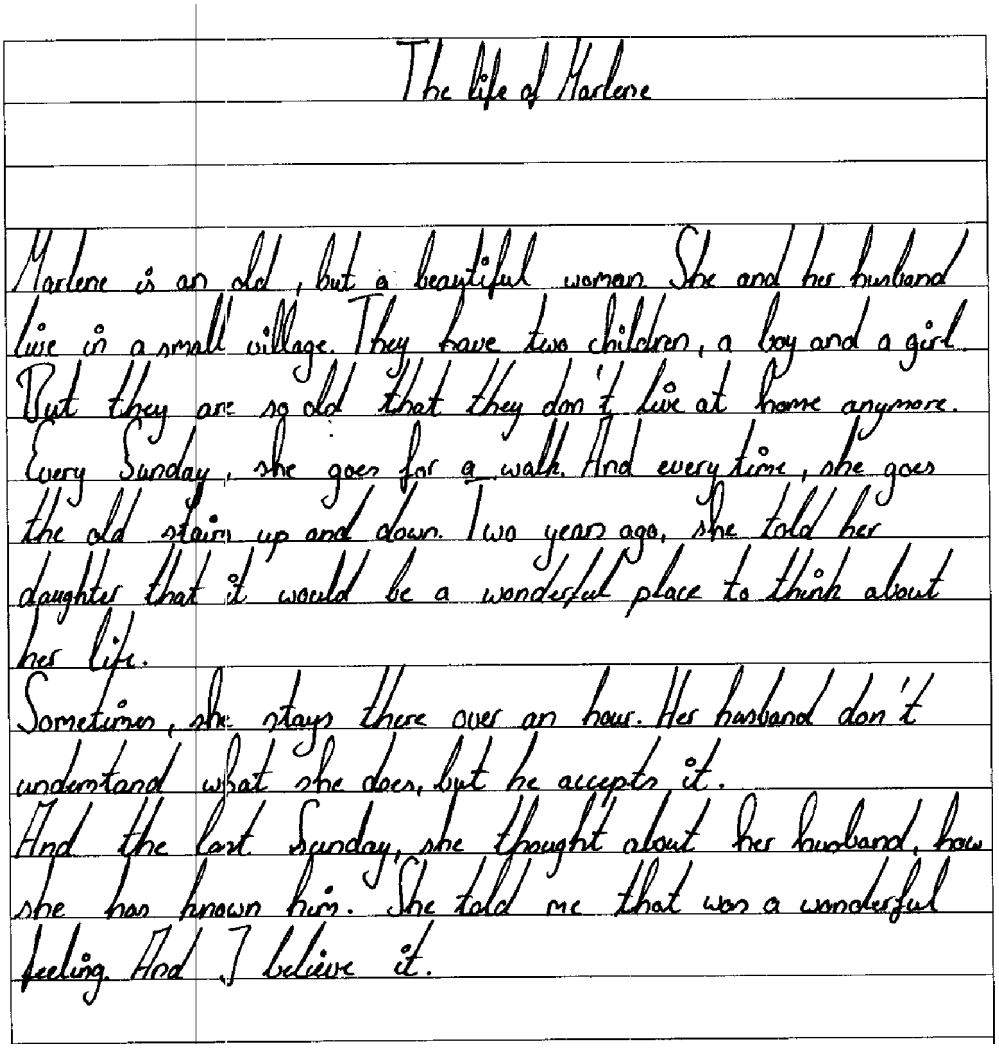
Globalurteil Niveau 2: 23930110:



Homberg: I will told you about a man an picture
there his name is Karl-Heim Kleman ~~sex~~ ~~ist~~ he
is a Cameraman he took photos from stars z.B. Ronaldo,
Madonna, the Queen. He had a wife her name is
Gerda. And he have two kids Joey and Alex.
Joey ~~ist~~ is 18 years old and Alex is 16 years
old. Karl-Heim ist 42 years old and his hobbys
are Shooting and make photos. He had make photos
from the queen last week. He lived in Homburg (Schw.
Today he go in the ~~B~~ Realschool Tribey and made
photos from kids they make a test vor the ~~PISA~~ DESI
studie. The kids ~~wait~~ in the 9. and 10. form
write over English the kids make very fun but it's
to long. He make photos from writer, from teachers and
from the kids. I have told from live from
Karl-Heim Kleman

End

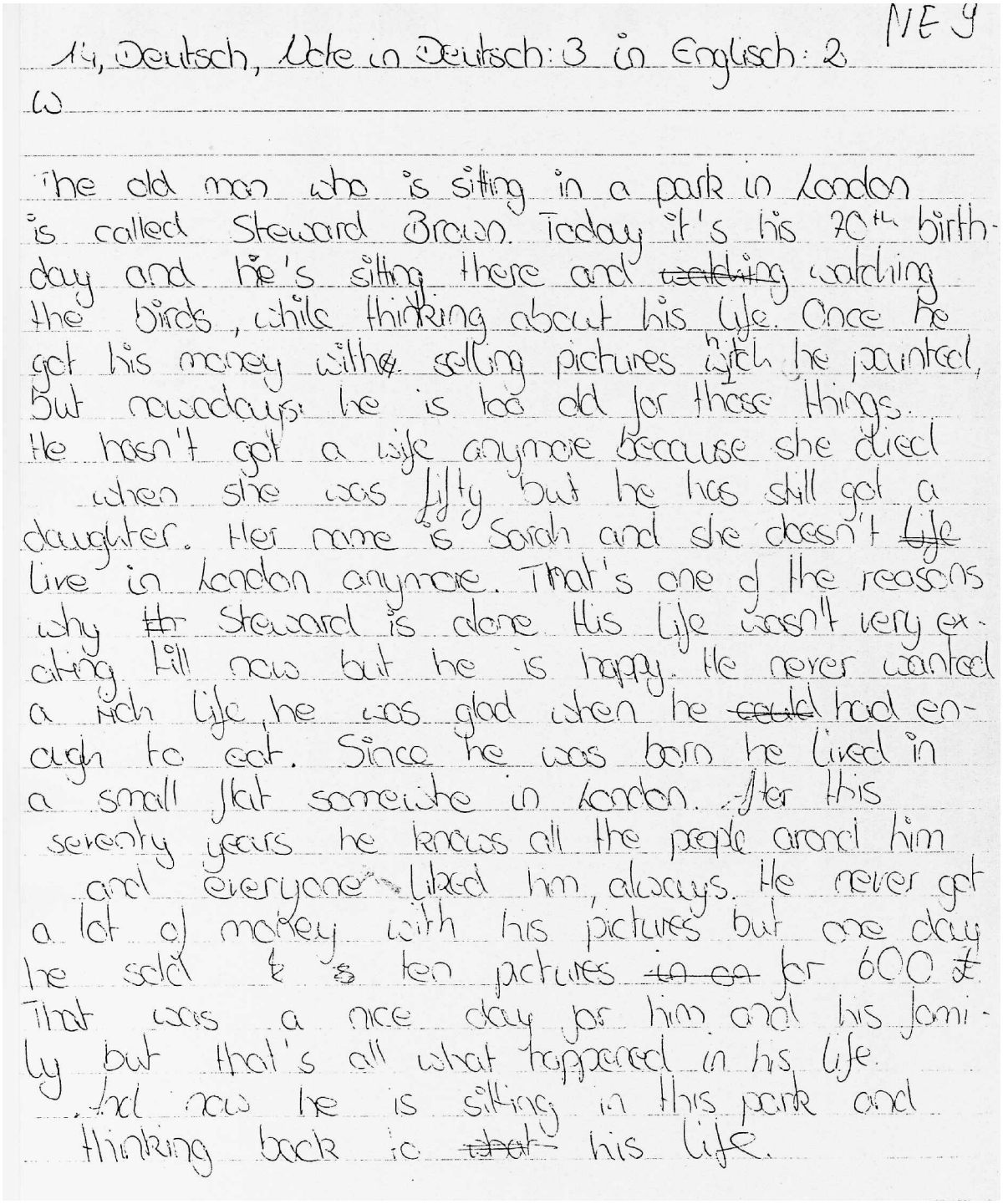
Globalurteil Niveau 3: 24970220:



The life of Marlene

Marlene is an old, but a beautiful woman. She and her husband live in a small village. They have two children, a boy and a girl. But they are so old that they don't live at home anymore. Every Sunday, she goes for a walk. And every time, she goes the old stairs up and down. Two years ago, she told her daughter that it would be a wonderful place to think about her life.
Sometimes, she stays there over an hour. Her husband don't understand what she does, but he accepts it.
And the last Sunday, she thought about her husband, how she has known him. She told me that was a wonderful feeling. And I believe it.

Globalurteil Niveau 4: NE9:



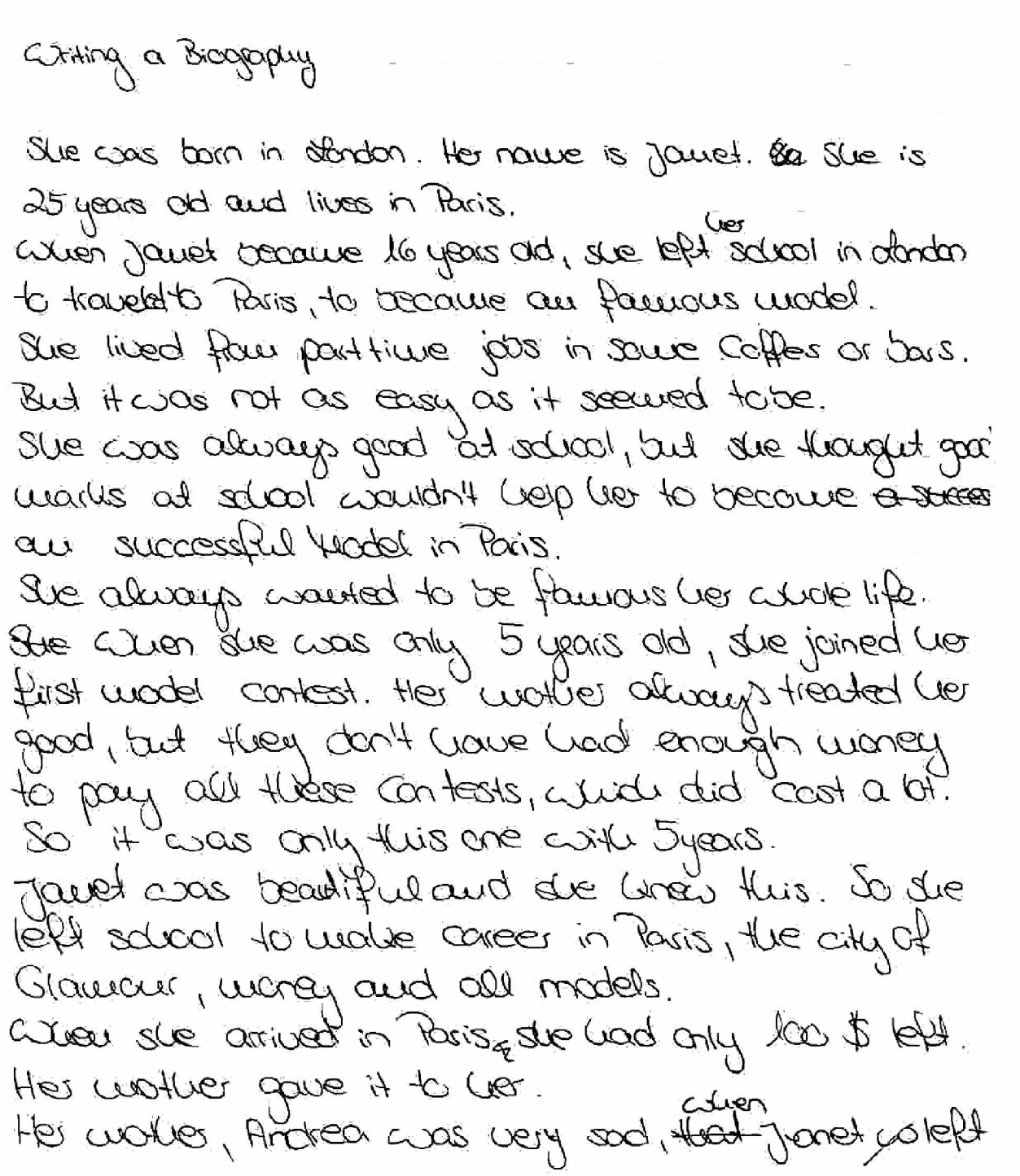
NE 9

14, Deutsch, Note in Deutsch: 3 in Englisch: 2

(2)

The old man who is sitting in a park in London is called Steward Brown. Today it's his 70th birthday and he's siting there and ~~feelding~~ watching the birds, while thinking about his life. Once he got his money with selling pictures which he painted, but nowadays he is too old for these things. He hasn't got a wife anymore because she died when she was fifty but he has still got a daughter. Her name is Sarah and she doesn't ~~life~~ live in London anymore. That's one of the reasons why ~~th~~ Steward is done His life wasn't very exciting till now but he is happy. He never wanted a rich life, he was glad when he ~~could~~ had enough to eat. Since he was born he lived in a small flat somewhere in London. After this seventy years he knows all the people around him and everyone liked him, always. He never got a lot of money with his pictures but one day he sold 6 of his ten pictures ~~in a~~ for 600 £. That was a nice day for him and his family but that's all what happened in his life. And now he is sitting in this park and thinking back to ~~that~~ his life.

Globalurteil Niveau 5: PKG83:

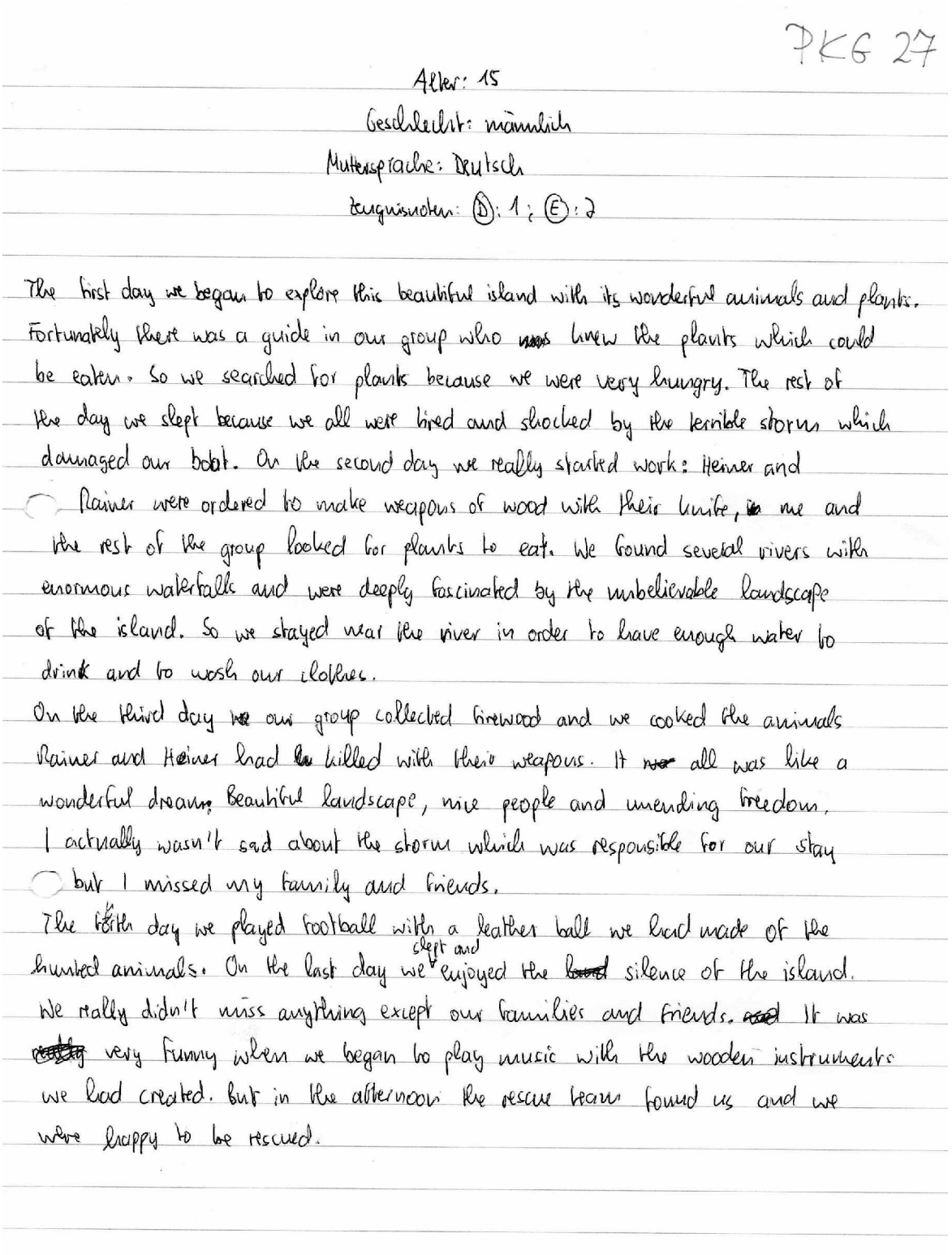


Writing a Biography

She was born in London. Her name is Janet. & She is 25 years old and lives in Paris.
When Janet became 16 years old, she left her school in London to travel to Paris, to become an famous model.
She lived from part time jobs in some Coffes or bars.
But it was not as easy as it seemed to be.
She was always good at school, but she thought good marks at school wouldn't help her to become ~~a success~~ an successful model in Paris.
She always wanted to be famous her whole life.
She when she was only 5 years old, she joined her first model contest. Her mother always treated her good, but they don't have had enough money to pay all these contests, which did cost a lot.
So it was only this one with 5 years.
Janet was beautiful and she knew this. So she left school to make career in Paris, the city of Glamour, money and all models.
When she arrived in Paris, she had only 100 $ left.
Her mother gave it to her.
Her mother, Andrea was very sad, ~~that~~ when Janet so left

wave, because she was only been 16 years old. But she thought, she will managed that.
The first years Janet lived by her aunt, who lived here, too. She got work in an cinema. She worked hard and kept all her money. She always got to castings of unserious model agency's. But all this doesn't work. No one wanted to give her a model job.
She lived there more than two years, when she became a call by one of the agency's. The man on the phone told her about the vote for Miss Paris and that his agency would manage this for her, that she can join the election.
She was more than surprised and very happy.
When the election began, she couldn't believe really to join it.
She became better and better and even became Miss Paris.
After that all serious model agency's wanted to have her and she became richer and richer and took her mother to her to Paris, to buy them a house and live there together.

Sprachliche Kriterien Niveau 5:
PKG27 (zum Task *Insel-Aufenthalt*, da es zum Task *Biography* in der Pilotierung keine Schülertexte gab, die in den sprachlichen Kriterien Niveau 5 entsprochen hätten):



PKG 27

Alter: 15
Geschlecht: männlich
Muttersprache: Deutsch
Zeugnisnoten: (D): 1; (E): 2

The first day we began to explore this beautiful island with its wonderful animals and plants. Fortunately there was a guide in our group who ~~was~~ knew the plants which could be eaten. So we searched for plants because we were very hungry. The rest of the day we slept because we all were tired and shocked by the terrible storm which damaged our boat. On the second day we really started work: Heiner and Rainer were ordered to make weapons of wood with their knife, ~~he~~ me and the rest of the group looked for plants to eat. We found several rivers with enormous waterfalls and were deeply fascinated by the unbelievable landscape of the island. So we stayed near the river in order to have enough water to drink and to wash our clothes.
On the third day ~~we~~ our group collected firewood and we cooked the animals Rainer and Heiner had ~~the~~ killed with their weapons. It ~~was~~ all was like a wonderful dream. Beautiful landscape, nice people and unending freedom. I actually wasn't sad about the storm which was responsible for our stay but I missed my family and friends.
The fourth day we played football with a leather ball we had made of the hunted animals. On the last day we slept and enjoyed the ~~found~~ silence of the island. We really didn't miss anything except our families and friends. ~~And~~ It was ~~really~~ very funny when we began to play music with the wooden instruments we had created. But in the afternoon the rescue team found us and we were happy to be rescued.

Anhang 29: Skript zum Hauptseminar „*Rating*-Prozesse in einer Schulleistungsstudie“ © Claudia Harsch

1. Grundlagen

- **Zweck einer Schulung**

Bei der Bewertung offener Aufsätze mittels *Rating*-Verfahren (s. u.) spielt die **Subjektivität** der Bewerter eine nicht zu unterschätzende Rolle. Sie hat Auswirkungen auf die Reliabilität und Validität der Bewertung. Um nun zu validen und reliablen Auswertungen zu kommen, und um die Subjektivität zu minimieren, bedarf es einer Schulung auf die anstehende Bewertung hin, während der das Bewertungsschema, die Kriterien und das Vorgehen geklärt und eingeübt werden. Letztlich wird aber immer eine „Restsubjektivität“ unter den *raters* bleiben. Diese wird man akzeptieren müssen, wie Shale (1996:93) bemerkt:

> There is no logical or philosophical basis to support such a proposition [i. e. the concept of identical marker behaviour, Anm. d. V.]. Does it not make more sense to accept that markers *naturally* vary in their judgments of texts and to settle on a measurement theory that allows us to accommodate this reality? Generalizability theory provides this structure, permitting us to specify markers as a facet in a study and to estimate the variation that is due to them – and to remove the effect of this variation from our considerations of other factors that may be of more direct interest.

Cumming (1990) schreibt, dass Training von *novices* wichtig sei, um zu gemeinsamen Kriterien und Abstufungen zu kommen, und um gemeinsame, vergleichbare Strategien einzusetzen, die das Ergebnis konsistent und reliabel machen. Laien hätten zwar Potential, müssten aber erst Erfahrungen sammeln im Umgang mit Texten, um von der Textoberfläche in die Texttiefe zu gelangen. Denn *rating* besteht überwiegend aus Abstraktionsprozessen, um relevante *features* in ein Gesamtbild des Textes zu integrieren, um die jeweils relevanten *features* der jeweiligen Kriterien unter den jeweils relevanten Gesichtspunkten zu *raten*, um letztlich zur „Gesamtwirkung“ eines Textes zu kommen. Diese Abstraktions- und Bewertungsprozesse müssen erfahren und eingeübt werden.

Shohamy *et al.* (1992) fanden heraus, dass Training eine wichtige Rolle spiele, besonders wenn Kriterien und Abstufungen geklärt werden und somit transparent und handhabbar für die *raters* werden; wenn die Terminologie definiert wird, so dass über Kriterien, Interpretation (der Kriterien und der Texte) und Vorgehen (*rating steps*) Konsens herrscht => solch ein Training könne die Validität und Reliabilität der Auswertung erhöhen.

Nach Shale (1996) sollen *raters* im Training ein gemeinsames Verständnis der Kriterien und Prozesse, der Ziele der Bewertung und des „Universums“ der *writing sample*s herausbilden. Sie stellen im Idealfall eine **interpretative Gemeinschaft** dar: Diese ist nicht nur gekennzeichnet durch ein “common agreement about how to read texts”, sondern auch durch ein “agreement about how they (the raters) will in fact ’write’ for themselves“. (Shale 1996:93) Auf dieser Basis kann man zu stabilen Interpretationen kommen, die auch den Anforderungen an Inter- wie Intra-*Rater*-Übereinstimmung genügen.

In Lumley (2002) finden sich folgende Implikationen für ein *Rater*-Training:
- Training und *reorientation sessions* (=begleitende Schulung) sind aus folgenden Gründen nötig:

> Rating is certainly possible without training, but in order to obtain reliable ratings, both training and reorientation are essential in order to allow raters to learn or (re)develop a sense of what the institutionally sanctioned interpretations are of task requirements and scale features, and how others relate personal impressions of text quality to the rating scale provided.(ebd.: 267)

- Der Zweck eines Trainings liegt darin, gemeinsames Verständnis, gemeinsame Interpretation und Übereinstimmung zu schaffen hinsichtlich der *task requirements*, der relevanten Kriterien und deren Interpretation, hinsichtlich der Abstufungen und der Interpretation der Deskriptoren, hinsichtlich der Textinterpretation (beispielsweise gibt es in DESI kein Auszählen der *features*, sondern ihr Auftreten wird interpretiert), hinsichtlich der Anwendung von *Rating*-Strategien und Prozessen, die vorgestellt und eingeübt werden müssen.
- Ein mögliches Problem bei der Schulung sei, dass Training und Skalen lediglich helfen, die *scores* zu rechtfertigen, nicht unbedingt, sie zu finden. Wir wollen die Schulung (und die Skalen) aber nutzen, um Ihnen bei der *Score*-Findung zu helfen, nicht nur bei der nachträglichen Rechtfertigung intuitiv gefundener *scores*!

Sie als **SeminarteilnehmerInnen** stellen als Studierende mit Hauptfach Englisch eine Gemeinschaft mit vergleichbarem Hintergrund dar, die über die in DESI getesteten Fertigkeiten selbst verfügen sollte (z. B. sollten die sprachpraktischen Kurse *Writing English* und *Composition* & *Style* mindestens mit der Note 2 abgelegt worden sein). Zudem sind Sie mehrheitlich Lehramtsstudierende. Während der Schulung werden

Testkonzept, Kriterien, Erwartungen an Schüleraufsätze und *Rating*-Prozesse genauestens erklärt und gemeinsam geübt, so dass Sie am Ende der Schulung ein "universe of raters" (vgl. Shale 1996) bilden.

- **Ziele eines *Rater*-Trainings** (nach Cohen 1994: 336)

1. Make sure that the raters gain the ability to give each assessment category the designated focus, whether or not it be equal focus.
2. Make sure that the raters use the **same** criteria for rating and that they all have the **same** understanding of what these criteria mean.
3. Strive to have novice raters approximate expert raters in terms of their rating behaviour.
4. If possible and if appropriate, provide for all raters training that will help them be sensitive to the rhetorical strategies of writers from other language and cultural backgrounds.

- **Aufgaben während der Schulung und als zukünftige *Raters***

- DESI-Konzept und Testkonstrukt internalisieren und gemeinsamen Erwartungshorizont herausbilden
- Auswertungsschema internalisieren und eine *interpretive community* bilden, die die Skalen und deren Deskriptoren in vergleichbarer Weise versteht, interpretiert und anwendet
- *Rating*-Prozesse einüben und in vergleichbarer Weise anwenden (z.B. vergleichbare Lesestrategien einsetzen; Fähigkeit zur Selbstreflexion und Kritik entwickeln; Abstraktions- und Interpretationsprozesse erlernen; etc.)
- Grundlagenwissen im Umgang mit *excel* und *E-mail*-Programmen erwerben und anwenden
- Aufsätze zuverlässig nach dem hier erlernten Schema interpretieren und bewerten; *Scores* in *excel*-Tabellen eintragen; Ergebnisse termingerecht per *E-mail* versenden
- Selbständiges Arbeiten „an einem hellen, ruhigen Ort", „nicht zu lange Sitzungen wegen Ermüdungserscheinungen" (nach Hughes 1986)

- **DESI: Semikreatives Schreiben**

DESI- **Testkonzept**: Vorstellung im Seminar, Internet: DESI-*Homepage*
DESI- **Schreibkonstrukt** und **Auswertungsschema**: Vorstellung im Seminar, Internet: Lehrstuhl-Seiten

⇨ Relevante Stichpunkte dazu werden von Ihnen als Hausaufgabe festgehalten
⇨ Jede/r bearbeitet einen DESI-Task in der vorgegebenen Zeit.
Gruppenarbeit: Erwartungshorizont, Auffälliges, Probleme herausarbeiten, im Seminar zur Diskussion stellen. Gemeinsame Bewertung Ihrer Aufsätze.

Wichtig: „Schreiben in einer Testsituation und für einen Test" ist eine künstliche Textsorte mit ganz eigenen Regeln: Zweck ist *Assessment*; Schreibziel unterliegt nicht der Kontrolle der Probanden und deren Lebenswelt; es handelt sich um ein nicht-authentisches Produkt (wenn es auch so weit als möglich authentisch gehalten wird) unter nicht-authentischen Bedingungen.
Dieses Kunst-Produkt müssen Sie beurteilen => Sie sollten Sensibilität entwickelt haben für solche Schreibsituationen und deren besondere Bedingungen (z. B. in *Writing English* oder *Composition*&*Style*)
Anmerkung zu Schreibsituation und „Verweigerern":
Elbow (1996:128) bemerkt hierzu: "We should not be too pure about taking 'the real reactions of real readers' as our only standard for judgement."
=> Verweigerer dürfen **nicht** bestraft werden! Vielmehr setzen wir differenzierte *Rating*-Skalen ein, um auch bei solchen Arbeiten noch Stärken und Schwächen herauszufinden.

- **Testgrundlagen**

Eingehen auf Testgrundlagen: Brainstorming (Grundwissen aus Grundkurs Fachdidaktik)
⇨ Referat „Testanforderungen" mit Klärung der folgenden Begrifflichkeiten:
- Testkonstrukt und Operationalisierung
- Anforderungen an „gute" Tests
- Konstruktion von Bewertungsschemata
⇨ Sie notieren sich relevante Stichpunkte

2. Skalen und der GER

- **Der *Gemeinsame Europäische Referenzrahmen für Sprachen* (GER)**

GER als Instrument des Europarats, als Publikation zum Fremdsprachenlernen, um ein mehrsprachiges und plurikulturelles Europa zu fördern.
Kernstück des GER: ein differenziertes System von sechsstufigen Skalen zur Beschreibung von kommunikativen Aktivitäten und von Sprachkompetenzen. Erfasst werden darin jenes Wissen und jene Fertigkeiten, mit denen Sprachlernende im öffentlichen, beruflichen und privaten Bereich sprachlich handlungsfähig werden. (Vgl. GER: 3)
Der GER wendet sich „an alle, die professionell im Bildungsbereich tätig sind." (GER: 3)

⇨ Referat „Ziele und Funktionen des GER"
⇨ Selbsteinschätzung anhand des entsprechenden GER-Rasters (GER: 36)
⇨ Rekonstruktion der zerlegten GER-Globalskala (GER: 35)

- **Der Skalenansatz in der Beurteilung**

Skalen bieten Vergleichsmöglichkeiten von Abschlüssen, Kursstufen und Prüfungsniveaus;
können in Binnen- wie Außenevaluation die subjektive Notengebung objektivieren;
schaffen Kohärenz, Transparenz und Vergleichbarkeit in der Leistungsmessung.
Konzept der Positivkorrektur – weg vom Fehlerzählen und Feststellen, was Lerner noch nicht können hin zur Bewertung dessen, was sie schon können, in Form von KANN-Beschreibungen (Deskriptoren).

Wie sollen gute Skalen beschaffen sein?
Nach Schneider & North (2000: 88):
- Die Niveaubeschreibungen machen für sich alleine genommen Sinn
- Sie ermöglichen eine Ja/Nein-Entscheidung.
- Das Können ist positiv formuliert.
- Sie sind konkret, klar und kurz.
- Sie enthalten wenig Fachterminologie.
- Sie sind kleinstufig.

Funktionen von Skalen
Skalen werden im Idealfall auf die Funktionen hin konstruiert, die sie auch übernehmen sollen. Aus Validitätsgründen sollte man Skalen nicht zu anderen als den intendierten Funktionen einsetzen.
Alderson (1991: 72ff) unterscheidet drei Skalentypen nach ihren jeweiligen Funktionen:
- (a) *User-oriented* mit *reporting function*: Solche Skalen beschreiben Kompetenzen, Verhalten, Performanzen etc., mit dem Ziel, über diese zu berichten.
- (b) *Assessor-oriented scales* mit der Funktion *guiding the rating process*: Diese Skalen fungieren als *common standards*, dienen der Validität und Reliabilität der Bewertung, und können im *Rater*-Training eingesetzt werden; fürs Training muss der Task spezifiziert werden, der die Performanz elizitieren soll, und *guidance* gegeben werden, wie diese Performanz zu bewerten ist.
- (c) *Test construction-oriented scales* mit der Funktion *providing guidelines for test construction*: Diese Skalen geben Spezifikationen in Bezug auf Tasks, Items, Schwierigkeiten, Inhalte, Texte.

Alderson bemerkt (ebd.: 74), dass es negative Folgen haben könnte, wenn man Funktionen vermischt, oder Skalen für andere als die intendierten Zwecke nutzt: Beispielsweise lässt die Performanz bei einem bestimmten Task (die in einer *rating scale* beschrieben wird) nur bedingt Rückschlüsse auf die *overall competence* zu (die in einer *reporting scale* beschrieben wird); eine Konstruktionsskala etwa, die Testmerkmale beschreibt, kann nicht zur Bewertung eingesetzt werden, denn dazu müssen Performanzen beschrieben werden. Bewertungsskalen sind meist viel zu detailliert, um gleichzeitig als Skalen zur Berichterstattung dienen zu können. Dazu müssen sie erst generalisiert werden auf das Universum, auf das Test und Bewertung ausgelegt sind.
Brindley (1998) sieht noch eine vierte Funktion von Skalen:
- (d) Diagnose-orientierte Skalen (nach Pollitt & Murray 1996), die Lehrern und Lernern detaillierte diagnostische Informationen geben.

Je nach Funktion der Skalen beschreiben deren Deskriptoren ganz unterschiedliche **Gegenstände**: Verhaltensbasierte Skalen beschreiben Performanzen oder sprachliches Verhalten; Kompetenzskalen beschreiben Wissensbestände und Fertigkeiten; *proficiency scales* beschreiben das Sprachvermögen, das Können, die Anwendbarkeit der Kompetenzen.
Die Beschreibung des Skalengegenstands, ebenso wie die Abstufungen, sollten validiert werden.

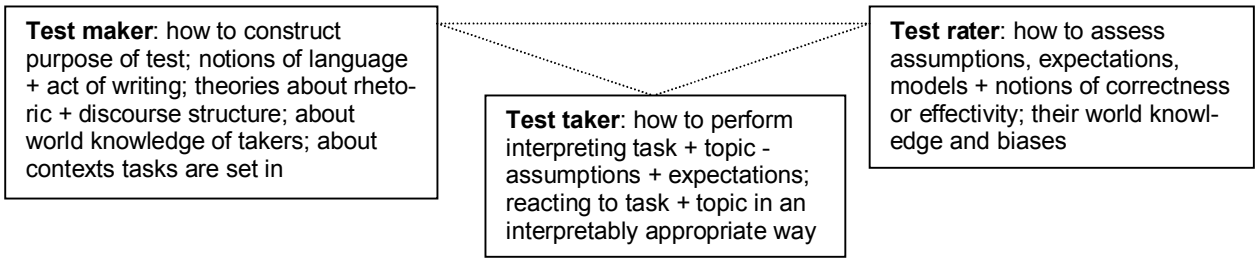
- **Konstruktion von Skalen**

Der Beschreibungsgegenstand einer Skala oder eines Skalensystems sollte in Theorien und Modellen verortet werden, die diesen Gegenstand auch beschreiben: Kompetenzskalen etwa können in Bachmanns Modell der kommunikativen Kompetenz (s. u.) verortet werden. Auf dieser Basis können begründete Entscheidungen hinsichtlich der horizontalen Einteilung des Gegenstands (im Falle von mehreren Skalen, die diesen Gegenstand erfassen) getroffen werden.
Die vertikalen Abstufungen der Deskriptoren sollten ebenfalls in relevanten Theorien und Modellen verortet werden: In unserem Fall etwa wurden die Niveaus verortet in Bereiters Modell der Entwicklung der Schreibfertigkeit (s. u.). Zusätzlich können die Deskriptoren einer Skala mithilfe von psychometrischen Messmodellen skaliert werden, wie es etwa bei den Deskriptoren des GER geschehen ist.

Bedeutung des GER bei der Skalenkonstruktion
GER-Skalen sind in ihrem Status (was beschreiben sie zu welcher Funktion?) und ihrem Beschreibungsgegenstand nicht transparent, weshalb es in jedem Fall notwendig und ratsam ist, Bewertungsskalen hinsichtlich der Bewertungskriterien und ihrer Abstufungen auf die jeweiligen Kontexte hin zu konstruieren und sie im Nachhinein mit relevanten Skalen des GER abzugleichen.
Der GER kann nur einen Rahmen geben, mit dem man einmal entwickelte Kriterien der Bewertung und Bewertungsskalen mit ihren Niveaubeschreibungen (Deskriptoren), die ausgelegt auf den jeweiligen Kontext konstruieren worden sind, vergleichen kann – **GER als Mittel der Außenvalidierung**.

- **Konstruktion des Bewertungsschemas in DESI**

Bei der Konstruktion eines Bewertungsschemas muss bedacht werden, dass nicht beliebig viele Aspekte bewertet werden können. Begrenzte Ressourcen, die Relevanz bestimmter Kriterien (immer bezogen auf das jeweilige Testkonstrukt) und die Handhabbarkeit des Bewertungsschemas zwingen zu einer Auswahl der zu bewertenden Leistungsdimensionen. Folgende drei Aspekte sind nach Cohen (1994: 307) wichtig bei der Konstruktion valider Bewertungsschemata:



Die Konzepte der Beteiligten innerhalb dieses Kommunikationsdreiecks sollten sich möglichst angleichen, die Beteiligten sollten gemeinsame Erwartungen an Testperformanz haben, um zu interpretierbaren und verwertbaren Ergebnissen zu kommen, um ein valides Bewertungsinstrument zu konstruieren!
Zweck dieser Schulung ist es, die Erwartungen von Testentwicklern und Testbewertern abzustimmen, zu „kalibrieren“, hinsichtlich der Kriterien und ihrer Auslegung, hinsichtlich der Bewertungsprozesse, und hinsichtlich der erwartbaren Schülerleistung.

Bewertungskriterien:
Relevante Merkmale der in einem Test elizitierten Leistung schlagen sich in den Bewertungskriterien nieder. Oft fließen mehrere Merkmale in ein Kriterium ein. In eine Skala können wiederum ein oder mehrere Kriterien einfließen, je nach dem, welche Merkmale sich als relevant erweisen und wie sie voneinander abgegrenzt werden (können). Diese Entscheidungen müssen auf Grundlage des Testrahmenkonzepts und des Testkonstrukts getroffen werden, um die Testvalidität zu sichern. Als Außenkriterium kann der GER (als Referenzmittel) nützlich sein.
Im DESI-Bewertungsschema fließen meist mehrere Merkmale in ein Kriterium ein (etwa Umfang und Korrektheit der grammatischen Mittel); es gibt je eine **Skala** für je ein Kriterium. Diese Skalen bestehen aus je fünf Niveaus sowie einem darunter angesiedelten Niveau 0. Die Deskriptoren basieren auf Analysen von Lernertexten, um die relevanten Merkmale und ihre Ausprägungen auf valide Basis zu stellen.

⇨ Folie „Leistungsdimensionen DESI“ und Folie „Merkmale Lernertexte“

⇨ Referat „Bachmanns Modell der kommunikativen Kompetenz (1991)“ zur Verortung der Kriterien

⇨ Referat „Schreibentwicklungsmodell von Bereiter (1980)“ zur Verortung der Niveaus

Gewichtung der Kriterien
Es gibt keine wissenschaftlich fundierten Hinweise, wie Kriterien zu gewichten sind.
Wir könnten uns auf ein Globalurteil beschränken, doch da auch bei einem holistischen Urteil implizit mehrere Aspekte bewertet werden, verlagert sich das Gewichtungsproblem dadurch in die Köpfe der *raters*. Eine Aufteilung in (analytische) Einzelkriterien reduziert dieses interne Gewichtungsproblem.
Wenn sich aus den Pilotierungsdaten keine statistischen Hinweise auf die Gewichtung ergeben, so lassen wir die einzelnen Kriterien gleichgewichtet in ein Leistungsprofil einfließen.

Aufgabenentwicklung und Konstruktion der *rating scales*:
Bei der Entwicklung der semikreativen Aufgabenstellung und der *rating scales* wurde wie folgt vorgegangen: Wir wussten aufgrund curricularer Analysen um die Kontexte, Textsorten, Inhalte etc., die geprüft werden sollten; die Taskentwicklung erfolgte daher nach bestimmten Vorgaben. Mit diesen Tasks wurden Lernertexte elizitiert, die auf ihre typischen und relevanten Merkmale hin analysiert wurden. Die Bewertungsskalen wurden verortet in o. g. relevanten Modellen und entwickelt auf Basis der Analysen der Lernertexte, in Anlehnung an Erwartungen und Standards in den 9. Klassen des deutschen Schulsystems (Curriculaanalysen). Die Deskriptoren wurden validiert an bereits existierenden Skalen (*Cambridge ESOL Tests* und relevante Skalen aus dem GER).

Status der Deskriptoren
Die DESI-Deskriptoren beschreiben die Merkmale eines Kriteriums abgestuft auf verschiedenen Niveaus. Ein Niveau ist jedoch nicht mit Lernstufen (im Sinn von Entwicklungsstufen, die alle Lerner durchlaufen) gleichzusetzen. Vielmehr werden prototypische Merkmale der Performanzen bei genau dieser Aufgabenstellung abgestuft beschrieben (im Sinn der unterschiedlichen Entwicklunsgsstände innerhalb der Probandengruppe). Jedes Niveau hat eine bestimmte Ausdehnung und ist jeweils in seiner **prototypischen** Mitte beschrieben. Die Deskriptoren sind als **Hilfsmittel** zu verstehen, um die für ein Niveau typischen Merkmale in den Schülertexten zu identifizieren. Es werden sich aufgrund der prototypischen Natur der Niveaubeschreibungen nie alle Merkmale eines Niveaus in einem Aufsatz wiederfinden, noch wird ein Aufsatz nur über die Merkmale genau eines Niveaus verfügen. Zur Zuordnung eines Aufsatzes auf ein bestimmtes Niveau bedarf es Identifizierungs- und Interpretationsprozesse, die in diesem Seminar geschult werden.
Wichtig: Deskriptoren sollen helfen, die *Scoring*-Entscheidung zu treffen, die Schülerleistung einzuschätzen. Leider dienen sie häufig als Rechtfertigung von intuitiv gefundenen Scores – das ist **nicht** zulässig! Wir nutzen die Deskriptoren, um zu unseren Entscheidungen zu kommen, d. h. eine Entscheidung, die nicht über die Deskriptoren gefunden wurde und somit nicht begründbar ist, ist keine valide Entscheidung!

- **Probleme, die sich im Zusammenhang mit Skalen ergeben können**

Skala – Kriterium: Oft müssen mehrere Kriterien in einer Skala zusammengefasst werden (z. B. Textsorte/makrostruktureller Aufbau: Die Textsorte bestimmt die Textstruktur entscheidend mit – es war eine pragmatische Entscheidung, Textsorte und Struktur zusammen zu bewerten, da es nur wenige formale Merkmale gibt, die wir bei den Textsorten „Schülerzeitungsbericht“ bzw. „informeller Brief“ ansetzen könnten).
Wichtig: Sollte sich innerhalb einer Skala ein Gewichtungsproblem ergeben, so gilt: Alle relevanten Merkmale, die in einem Bewertungskriterium zusammengefasst sind, werden als **gleichwertig** behandelt.

Abgrenzung der Kriterien: Sprache ist ein komplexes Netzwerk aus interagierenden Teilfertigkeiten, die sich nicht „sauber“ voneinander trennen lassen; deshalb können auch die Kriterien der Bewertung nicht als völlig unabhängig betrachtet werden. Wir sind uns der Abhängigkeit bewusst, doch müssen wir zu einem Verständnis der Kriterien kommen, das es uns ermöglicht, diese als „weitgehend unabhängig“ zu betrachten. Für die Bewertung bedeutet dies, dass wir uns innerhalb eines Bewertungskriteriums auf ganz bestimmte Aspekte konzentrieren und auf genau definierte Merkmale fokussieren.
Wichtig: Es kann passieren, dass ein Merkmal in einem Aufsatz alle anderen überlagert, oder dass man ein „besonders schlechtes Merkmal“ einer Schülerarbeit „bestrafen“ will – das ist **nicht** zulässig.
Es dürfen nur die in den Deskriptoren angeführten Merkmale bei den entsprechenden Kriterien bewertet werden; **es gibt keine Bestrafung**.

Interpretationsschwierigkeiten: Deskriptoren und Schülerleistungen müssen interpretiert werden. Da bei offenen Aufgaben eine breite Reaktionsmöglichkeit auf den Task gegeben ist, sind Interpretationsprozesse nötig, auf die später detailliert eingegangen wird. Wir werden anhand authentischer Schülertexte genau besprechen, was als ‚angemessene Reaktion’ auf welcher Stufe erwartet werden kann. Wir werden die DESI-Kriterien genau auslegen, um zu einem gemeinsamen Verständnis derselbigen zu kommen.
Wichtig: Sie müssen Ihre eigenen Konzepte (z. B. zu inhaltlicher Relevanz oder kommunikativer Wirksamkeit) hinten anstellen und sich auf die DESI-Konzepte einlassen, um die Deskriptoren im von DESI intendierten Sinn zu lesen und anzuwenden. Persönliche Einstellungen und Ansichten dürfen **nicht** in die

Bewertung einfließen. Es darf beispielsweise keine Bestrafungen für *foul language* geben, die man persönlich als verletzend empfindet, oder es dürfen gerade beim Kriterium des Inhalts keine persönlichen Konzepte zur Relevanz herangezogen werden!

Abstufungen der Skalen: Es existiert das grundsätzliche Problem der Verbalisierung von Schwierigkeitsabstufungen, da Merkmale zur Beschreibung derselbigen variieren können zwischen quantitativen Merkmalen („ein großes Repertoire“), qualitativen Merkmalen („korrekt“, „relativ leicht“), Einschränkungen („sofern Gesprächspartner langsam spricht“), Merkmalen von Themen („vertraute Themen“, „Alltagsbedürfnisse“) und Merkmalen von Textsorten oder Situationen („klar strukturiert“) – diese Merkmale sollten in einer Skala/einem Skalensystem in systematischer Weise zur Abstufung verwendet werden. Beispielsweise wird bei den Levels der britischen *National Language Standards* versucht, die Niveaus von *predictable* über *routine*, *varied* zu *complex* und *complex and specialized* durchzudeklinieren – das bringt aber wieder eigene Interpretationsschwierigkeiten mit sich.
Ein System rein verbaler Abstufungen (sehr einfach – einfach – mittelschwer – schwer) erleichtert die Bewertung nicht unbedingt: Wie sind die verbalen Abstufungen zu interpretieren? Hier bieten sich eher qualitative Abstufungen an, wie wir sie beispielsweise im Globalurteil verwendet haben.

Illustration der Niveaus: Aufgrund der prototypischen Natur der Deskriptoren kann es passieren, dass nicht alle Merkmale, die sich in den Schüleraufsätzen finden, in den Deskriptoren präzise genug verbalisiert wurden (Prototypen können nicht alles abdecken). In solchen Fällen könnte die Bewertung durch andere Faktoren beeinflusst werden. Beispielsweise kann die Reihenfolge der Aufsätze die Auswertung beeinflussen – je nach der gerade bewerten Leistung kann die Bewertung der nächsten Arbeit beeinflusst werden, wenn man sich alleine auf die Deskriptoren verlässt. Deshalb arbeiten wir mit sog. ***Benchmark***-Texten. Das sind Arbeiten, die als „Prototypen“ in eine bestimmte Stufe eingeordnet worden sind und die als Vergleichspunkt dienen. Auch werden alle *gerateten* Aufsätze zusätzlich *gerankt*, um die Konsistenz innerhalb der Kategorien und über die Aufsätze hinweg zu garantieren (s. unten).

Vorsicht: Zwar werden die Niveaus durch numerische *Scores* repräsentiert, doch dürfen diese numerischen *Scores* nicht mit präzisen Werten verwechselt werden. Denn die Niveaus einer Skala sind nicht punktuell zu verstehen, sondern haben eine gewisse Bandbreite auf der kontinuierlichen Skala. Auch sind die Abstände zwischen den *Scores* nicht gleich: Ein „unterer Dreier“ ist weiter von einem „unteren Vierer“ entfernt als ein „oberer Dreier“.
Grenzfälle: Mancher Aufsatz ist **nicht eindeutig** einer Stufe zuzuordnen, da er zwar schon einige *features* der nächsthöheren Stufe aufweist, doch eben auch viele der darunter liegenden. In solchen Fällen wird der Aufsatz auf der unteren Stufe eingeordnet, da sich der Schüler sicher auf dieser befindet.

⇨ Anwendung des Obigen: **Skalierung DESI-Globalurteil** und Herausarbeiten der Merkmale und deren Abstufungen im Globalurteil.

- **DESI-Handbücher**

⇨ austeilen und gemeinsam besprechen:
Kriterien – Skalen: relevante Merkmale; Skalen – Deskriptoren: Abstufung der Merkmale.
⇨ Hausaufgabe: Durchlesen der Kriterien und ihrer Abstufungen; Notieren der Merkmale und ihrer Abstufungen. Diskussion offener Fragen.

3. *Rating*-Prozesse:

- **Aufsatzbewertung:**

⇨ *Brainstorming* der verschiedenen Möglichkeiten, Aufsätze zu bewerten
⇨ Folie der Tabelle nach Pollitt (1991: 52) und Pollitt & Murray (1996), die die beiden grundsätzlichen Herangehensweisen *Counting* – *Judging* kontrastiert

Counting sei laut Pollitt sinnvoll bei rezeptiven Tasks, bei denen man beispielsweise die Lese- bzw. Hörverstehensprozesse nicht direkt bewerten kann. Dabei können die Testitems hinreichend beschrieben werden in ihren Anforderungen und Schwierigkeiten. Dann können Punktwerte mit der Aufgabenbeschreibung zusammengeführt werden, um auch von qualitativer Seite her zu beschreiben, was die Probanden können (Zusammenhang Aufgabenanforderung – Probandenfähigkeit).
Judging dagegen hält Pollitt bei allen produktiven Task für angemessen, wenn es um die direkte Bewertung der Performanz geht; dabei konzentriert man sich eher auf die Antwort als auf den Stimulus (der bei

offenen Aufgaben nicht definitiv in seiner Schwierigkeit bestimmt werden kann). Die abgestufte Beschreibung der Bewertungskriterien bildet die Grundlage der Beschreibung der Probandenfähigkeiten: Dazu müssen die *rating scales* in Kompetenzskalen überführt werden (i. d. R. durch Generalisierungen auf dasjenige „Universum" hin, das der Test erfassen soll, im Testkonstrukt festgelegt). Hier sieht man, dass z. B. Konstruktionsskalen, die den Schreibtask beschreiben, nicht als *rating scales* genutzt werden können, da sie eben nicht die zu bewertende Performanz beschreiben, vgl. oben die Ausführungen zu den unterschiedlichen Funktionen, die Skalen erfüllen können).

- **Bezug zu DESI**:

Bei der Bewertung werden wir *counting strategies* und *judging strategies* angemessen einsetzen: Beispielsweise wird die Länge gezählt, während das Globalurteil ein *judgement* darstellt; beim Kriterium des Inhalts werden Quantität und Qualität abgewogen und fließen beide in die Bewertung ein.

Welche Rolle kommt Ihnen als Bewerter/*Rater* zu?
Sie als *raters* stehen im Mittelpunkt des Bewertungsprozesses, da Sie es sind, die wahrnehmen, interpretieren und beurteilen – selbst nach dem Training werden Sie die Komplexität Ihres Denkens und Ihres Hintergrunds beibehalten.
Deshalb ist es von großer Bedeutung, eine interpretative Gemeinschaft zu schaffen, um die Beurteilung möglichst auf *common ground* zu basieren. Daher auch dieses Seminar, das sich an Studierende des Englischen im Hauptstudium wendet, um möglichst vergleichbare Voraussetzungen zu schaffen.

- **Charakterisierung der *Rating*-Prozesse**

Cumming (1990) hat in einer Untersuchung des *Rating*-Verhaltens 28 *decision making behaviours* identifiziert, die sich aus interpretativen *reading strategies* und evaluativen *judgement strategies* zusammensetzen:
- Interpretative Strategien: Aufgabenstellung und Text lesen, Situation (des Lerners, der Aufgabenstellung, der Textproduktion, des gedachten Rezipienten) vorstellen, Wirkfaktoren identifizieren und in ihrer Funktion bzw. Wirksamkeit interpretieren (beispielsweise Auflösung von Mehrdeutigkeiten).
- Evaluative Strategien: umfassen z. B. "personal response to quality", das Lesen der Kriterien, um eben diese Qualität festzustellen, das Vergleichen von Aufsätzen untereinander (und mit dem Handbuch), um zu evaluieren, welche Bedeutung die interpretierten *features* haben.

Wichtig: Die **Variabilität** der angewandten Strategien sei aber nach Cumming von *rater* zu *rater* sehr hoch. Nur die klassischen Negativkorrekturstrategien sind von allen gleichermaßen angewandt worden.
=> Diese Schulung soll Sie auf eine vergleichbare Anwendung von Positivkorrekturstrategien vorbereiten!

Zusammenfassung *Rating*-Prozesse:
Zuerst Fragliches **identifizieren** und **interpretieren** (z. B. kommunikationsbelastende Fehler, Wortschatzbreite, Auftreten bestimmter Strukturen…); dann Gebrauch, Einsatz des Fraglichen **bewerten** nach Verständlichkeit, Korrektheit, Angemessenheit, Relevanz, Breite, etc.: beispielsweise *linking language* identifizieren, deren Einsatz interpretieren (Was tritt wo auf? Angemessen? Fehlt was?), dann Angemessenheit und Effizienz in Abgleichung mit Handbuch, *Benchmarks* und anderen Aufsätzen bewerten.

Welche Arten von *Rating*-Verfahren gibt es?
Cohen (1994) nennt Vor- und Nachteile von **4 Arten von *rating/scoring***:
holistisch – analytisch – *primary trait* – *multi trait*
- ⇨ Referat zu Vor- und Nachteilen der vier Verfahrensweisen (In Anlehnung an die Abbildungen 9.1 mit 9.4 aus Cohen 1994)
- ⇨ Gruppenarbeit: Welches DESI-Kriterium entspricht welcher Art des *rating*?

- ***Rating* in DESI:**

Aufsatzbewertung in unserem Sinn ist ein komplexer, interaktiver mentaler Prozess:
Interaktiv zwischen Aufsatz und Handbuch, Aufsatz und *Benchmark*-Texten, Aufsatz und Bewertungskategorien im Kopf; interaktiv zwischen den Kriterien (man muss sich ständig die Abgrenzung und Interaktion zwischen den Kriterien klarmachen); interaktiv zwischen den Abstufungen innerhalb eines Kriteriums; Interaktiv zwischen Skalen des Handbuchs und Skalen im Kopf; etc.

Bei der Bewertung gibt es zwei grundsätzliche Vorgehensweisen:
Ranking: Beim Ranking werden die Texte in aufsteigende Reihenfolge gebracht, ohne dass wir sie primär den Niveaus zuordnen wollten. Es geht vordringlich um die Feststellung, welcher Text besser ist.

Rating: Beim Rating geht es primär um die Zuweisung von Niveaus (wobei natürlich auch hier die Frage, welcher Text der bessere ist, nicht ganz untergeht).
=> Beide Prozesse werden miteinander **kombiniert**, um Vorteile zu nutzen und Nachteile zu minimieren: Beim Globalurteil etwa bietet es sich an, die Texte – schon um einen ersten Eindruck zu erhalten – erst in eine grundsätzliche Reihenfolge zu bringen (*ranking*). Auf Basis dieser Reihenfolge dürfte das Bewertung (und damit das Zuweisen der Niveaus) leichter fallen und konsistenter sein (*rating*).
Dann aber ist es ratsam, jeden Text in den zehn analytischen Kriterien zu *raten* (dabei helfen ja die Vergleichsmöglichkeiten mit den *Benchmark*-Texten und anderen schon bewerteten Texten), um dem Text als „Ganzes" gerecht zu werden. Erst im Anschluss an die Bewertung einiger Texte werden die Kriterien textübergreifend betrachtet: Sie *ranken* die bewerteten Texte innerhalb der Kriterien, um dadurch die innere Konsistenz (innerhalb eines Kriteriums) zu garantieren.

Zuweisung der jeweiligen Bewertungsstufe:

1. **Identifizieren** der für das jeweilige Kriterium relevanten *features* (vgl. Deskriptoren) an Textoberfläche bzw. in tieferen Textstrukturen.
Hierbei auch Abgrenzung zu anderen Kategorien klarmachen: Beispielsweise „Inhalt": Auf die jeweiligen Propositionen sehen und wie die in sich entwickelt, ausgeführt sind – „Sorte und Aufbau": Hier die Themenentwicklung auf den Text (und nicht mehr auf die Propositionen) bezogen bewerten, z.B. Einteilung Propositionen-Absätze, Spannungsbogen, inhaltlich-logische Entwicklung, etc.
2. **Interpretieren** der identifizierten *features*: weg von Textoberfläche hin zur (abstrakter) Interpretation der „Wirksamkeit" der relevanten *features* bezogen auf den Text (aber nicht auf Textoberfläche) und die jeweilige Kategorie: z. B. „Inhalt": Ambiguitäten interpretieren („was könnte Schüler mit dem und jenem gemeint haben"?), Ideenumsetzung in Relevanz und Wirksamkeit bezogen auf das kommunikative Ziel interpretieren, usw.
Beispielsweise interpretieren der Fehler, um auf Stand der Interimsprache, systematische Fehler, *impeding errors*, etc. schließen zu können.
3. **Bewerten** der identifizierten und interpretierten *features* nach Vorgaben Handbuch für die betreffenden Kategorien bezogen auf den Ganztext (d.h. wir bewerten nicht einzelne *features*!), z.B. nach Korrektheit, Umfang, Wirkung, etc.
Dabei sollen Quantität wie Qualität (der jew. *features*) als zwei sich ergänzende Facetten in Relation zueinander erfasst werden, denn es gibt kein quantitatives System, mit dem man Sprache erfassen könnte – und Qualität ist etwas, das man nur über solche Bewertungsprozesse fassen kann.
Dabei spielen auch situative Faktoren mit herein: Wer schreibt das in welcher Situation an wen und was bedeutet das für die Bewertung? (Wie gehen wir beispielsweise mit „Verweigerern" um?)
4. **Selbstevaluation** als ständiger Prozess: "Do I do what I'm supposed to do?"

Beispiel Globalurteil:
1. Identifizieren der Merkmale aus den Deskriptoren (Inhalt, Textsorte, sprachliche Qualität, kommunikative Wirkung) im Schülertext
2. Interpretieren der im Schülertext auftretenden Merkmale hinsichtlich ihrer qualitativen Umsetzung, Relevanz, Wirksamkeit bezogen auf den Gesamttext
3. Bewerten der Umsetzung dieser gefundenen Merkmale – Abgleich der Skalen aus Handbuch mit der „Intuition", dem ersten Gesamteindruck
=> Merkmale fließen gleichgewichtet in das Urteil ein, denn es handelt sich hierbei um eine Kombination aus *holistischem* und *multi-trait* Urteil.

⇨ Besprechen der Vorgehensweise beim *Globalurteil* am Beispiel von *Benchmark*-Texten:
Erst grobes *ranking* (welcher Aufsatz ist „besser"): Verbalisierung der Begründung, warum welcher besser ist. Dann *rating*: Vergleich der Deskriptoren der Globalskala, die prototypischen Merkmale enthält, mit der Verbalisierung der entscheidenden textuellen Merkmale.

Modell des Entscheidungsprozesses beim holistischen *rating*

⇨ Folie des Modells der Entscheidungsprozesse bei holistischem *rating* nach Milanovic, Saville & Shuhong (1996: 95), basierend auf der Studie von Cumming (1990):
Das Modell zeigt typisches Rater-Verhalten und typische Elemente, auf die *Rater* bei **holistischem** *Rating* fokussieren, d. h. dieses Modell kann beim **Globalurteil** relevant werden. Es muss aber auf unsere Kriterien hin adaptiert werden, die ins Globalurteil einfließen sollen.

⇨ Gruppenarbeit/Hausaufgabe: Adaptieren Sie dieses Modell auf das Vorgehen in DESI
⇨ Vergleichen Sie Ihr Ergebnis mit den *rating steps* nach Lumley (2002) aus dem Handbuch

Noch zu beachten:

Milanovic, Saville & Shuhong (1996) stellen fest, dass das ***sequencing*** der Aufsätze ebenfalls Einfluss hätte auf die Bewertung. Deshalb muss jeder Aufsatz nicht nur mit den *Benchmark*-Texten, sondern auch mit den schon *gerateten* Texten vergleichen werden!

Milanovic, Saville & Shuhong (1996) beobachteten ***Halo*-Effekte** (d. h. ein Merkmal überlagert alle anderen: bei besseren Aufsätzen zählte eher der Inhalt, bei schlechteren zählte eher der kommunikative Effekt) Diesen *Halo*-Effekten begegnen wir durch die Vorgabe der analytischen Kriterien, die in immer derselben Reihenfolge bewertet werden müssen.

Das Kriterium der *swear words* soll Ihnen helfen, auch solche Texte neutral zu bewerten, die **unangemessene *foul language*** aufweisen. Auch „Fäkalsprache" muss, soweit die Antwort sich im validen Bereich bewegt, „neutral" bewertet werden:
Vgl. Arbeit 19920227: die Arbeit bewegt sich inhaltlich und in den kommunikativen Kriterien auf Stufe 0, sprachlich jedoch muss sie nach der vorhandenen Sprache bewertet werden – da Fäkalsprache nicht in den Deskriptoren beschrieben wird, werden beim Bewerten diese Ausdrücke ignoriert!
Deswegen gibt es das Kriterium *swear words*, das in solchen Fällen mit **1** kodiert wird.

4. Vorgehen bezüglich einzelner DESI-Kriterien

- **Konkrete Fragen und Probleme des *rating* aus den Pilotierungserfahrungen**

⇨ An Beispieltexten illustrieren und besprechen:
- Abgrenzung valide – nicht valide
- Texte zwischen zwei Stufen
- Neutrales Herangehen auch bei „unangemessenem" Inhalt
- Vorbeugen von *Sequencing*- und *Halo*-Effekten

- **Vorgehen bezüglich impliziter Kriterien**

Kriterien wie „Inhalt" oder „Wirksamkeit" sind nicht molekular an der Textoberfläche auszumachen, sondern eher durch Abstraktion und Interpretation in der Text-„Tiefe" => besser beurteilbar über die „Wirkung" der *features* denn über die Bewertung der an Textoberfläche beteiligten *features*. Molekular beobachtbare *features* der Textoberfläche kann man „unabhängig" bewerten (z. B. Textlänge), doch implizite *features*, die nur vernetzt-interagierend beobachtbar sind (z. B. Wirksamkeit des Texts) werden eher über ihre Auswirkungen bewertet.
=> Vorgehen: Lesen – Identifizieren der relevanten *features* (auf Basis Deskriptoren) – Interpretieren der identifizierten *features* auf das jeweilige Kriterium hin – Bewerten i. Sinne v. Einstufen durch Abgleich mit Skala und Vergleich mit *benchmarks* und schon bewerteten Texten.

Wichtig: Während des Lesens wird die Bedeutung des Textes vom Rezipienten konstruiert – die Rezeption eines Textes wird durch die Ziele und Zwecke des Lesens beeinflusst:
⇨ Referat „Textrezeption und Bedeutungskonstruktion" (nach Kintsch & van Dijk 1978)
⇨ Referat zur Unterscheidung Mikro-/Makro-Ebene

Anwendung
Bedeutungskonstruktion im Hinblick auf die „Leseziele", die das DESI-Bewertungsschema vorgibt.
Die Unterscheidung Mikro-/ Makroebene ist hilfreich für die Bewertung von „Inhalt" (Motiviertheit und Kohärenz auf Mikroebene) und für „Textsorte/Aufbau" (Gliederung und Struktur auf Makroebene nach konventionellen Erwartungen). Dazu müssen die Lernertexte analysiert werden, von Oberfläche hin zu Tiefenstruktur, von Mikro- zu Makroebene, von Inhalt zu Textsorte/Aufbau. Die sprachlichen Kohäsionsmittel werden eigens bewertet.

- **Bewertung Kriterium 3 „Inhalt"**

Dazu können die Überlegungen von Kintsch & van Dijk (1978) zu Textrezeption und Bedeutungskonstruktion hilfreich sein (vgl. Referate und die Ausführungen unten zu „Kohärenz").

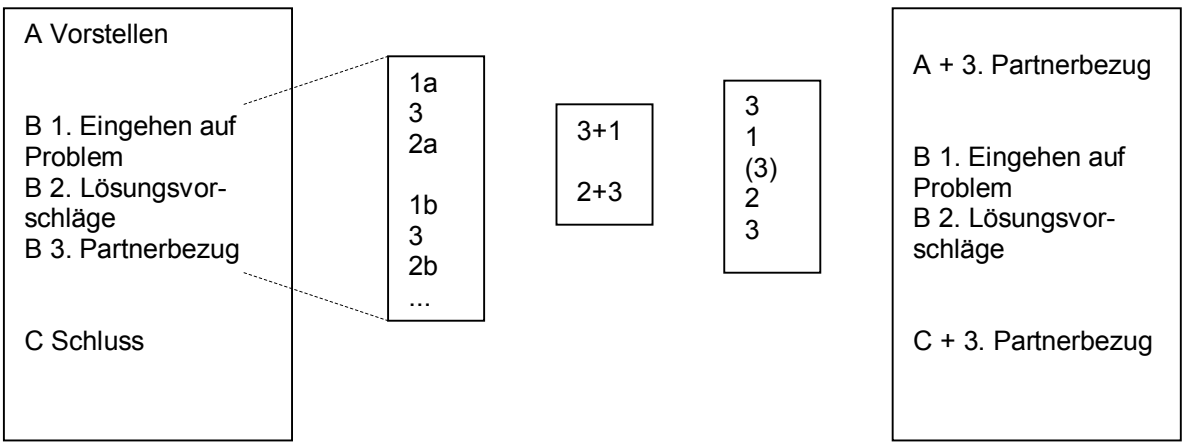
Vorgehen:
1. Identifizieren von Propositionen (Mikrostruktur) und evtl. Mehrdeutigkeiten (an Textoberfläche)
2. Interpretieren der Mehrdeutigkeiten und Zusammenfassen der Propositionen (weg von Textoberfläche, abstrahieren, *key features* gruppieren); interpretieren der Motiviertheit der Propositionen
3. Bewerten der qualitativen Umsetzung der Propositionen (Relevanz, Interesse-Faktor, Entwicklung Propositionen) und in zweiter Linie der Anzahl der Propositionen/Ideen
=> Von *features* an Textoberfläche über Abstraktion/Interpretation tieferer Strukturen hin zu Gesamteindruck, der bewertet werden kann.

Hintergrund:
Ein bloßes Abzählen der Elemente vernachlässigt oft implizite, komplex eingebaute Inhaltselemente bzw. würdigt nicht die Ausführlichkeit oder Entwicklung einer Idee (vgl. etwa Beispieltext zu „Pen Friend“: Die Erwähnung der Schwimmhalle in Tokio kann als impliziter Partnerbezug interpretiert werden).
Qualität ist eine schwer fassbare Eigenschaft: Ausgangspunkt ist eine Idee, die versprachlicht wurde; zur Bewertung müssen wir aber über die bloße Identifikation der Idee/Proposition an der Textoberfläche hinausgehen: Wie motiviert sind die einzelnen Propositionen? (Die Motiviertheit wird u. a. bestimmt durch das Verhältnis, in dem Propositionen stehen. Dieses Verhältnis kann auch funktional benannt werden: z. B. Spezifikation, Generalisierung, Erläuterung, Beispiel, Erklärung, Korrektur, etc.) Welche Modifikationen kommen vor? Wie tief wird auf relevante Details eingegangen? Wie relevant ist die Idee? Werden Beispiele zur Untermauerung gegeben? Werden Wertungen und Einstellungen deutlich? Werden evtl. sogar verschiedene Perspektiven angedeutet? Oder wird beispielsweise statt bloßer Beschreibung der Person eine implizite Charakterisierung geboten?
All diese Merkmale machen die Qualität des Inhalts aus und bestimmen deren Relevanz bezogen auf die jeweiligen Aufgaben!
Unser Vorgehen: Wir haben mögliche Themenbereiche aus Schülerarbeiten identifiziert und in Aufgabenstellung und Handbuch benannt, um den Erwartungshorizont einzugrenzen. Die Schüler sollen dennoch möglichst offen herangehen dürfen, d. h. wir erwarten uns keine vollständige Abdeckung aller möglichen Themenbereiche, sondern eine Umsetzung und sprachliche Realisierung einer für die jeweiligen SchülerInnen relevanten Auswahl an Themen! Wir wollen prüfen, in welcher Qualität diese Versprachlichung geschieht.
Wichtig: Wir wollen **nicht** den Einfallsreichtum der SchülerInnen testen! Es müssen nicht alle Themenbereiche abgedeckt sein; entscheidend ist die Relevanz und qualitative Versprachlichung! Das bedeutet, dass die angegebenen Themenbereiche oder das Abzählen der vorhandenen Propositionen zwar in Zweifelsfällen helfen können, eine fragliche Arbeit einzustufen, doch die Bewertungsgrundlage müssen Relevanz und Qualität der sprachlichen Umsetzung der Ideen sein.
Die Anzahl der Ideen tritt in den Hintergrund: Es geht um die Ausführlichkeit der Versprachlichung des Inhalts, um Qualität, um Relevanz, um „interessante“ Ideenumsetzung, nicht so sehr um Quantität.

- **Bewertung der Kohärenz:** Kriterien 3 „Inhalt“, 4 „Textsorte/Aufbau“, 8 „linking language“

Das Modell von Kintsch & van Dijk (*Textcomprehension* 1978) liefert einige interessante Gedanken in Bezug auf die Beurteilung von Kohärenzphänomenen (vgl. Referate oben).

Beispiel *Agony Aunt*: Makrostruktur (Textaufbau) – es gibt viele gleichwertige Lösungen:



Mikrostruktur (der inhaltlichen Propositionen): Hier betrachtet man z. B. die Umsetzung von B1-B2: Wie werden Lösungen entwickelt (aus Problemstellung heraus oder völlig irrelevant?)? Wie werden Beispiele und Details angeknüpft (stehen sie unverbunden oder motiviert, d.h. beispielsweise in referentieller

Beziehung zu schon Genanntem)? In welcher Beziehung stehen die Propositionen und Sätze?
Neben semantischen Relationen gibt es natürlich noch die faktischen Relationen in den und über die Propositionen hinweg, um Kohärenz zu entwickeln: Gibt es inhaltlich einen roten Faden?

⇨ In der Schulung und als Hausaufgabe entwickeln Sie solche Diagramme am Beispiel von Übungstexten, um Gespür für Makro- wie Mikroebene zu bekommen.

Vorgehen dabei: Sie erstellen eine Liste mit gefundenen Propositionen und ziehen Satzgrenzen ein: Je mehr Propositionen in einem Satz, desto komplexer dürfte er sein und, die Kohärenz vorausgesetzt, umso komplexer, dichter dürfte auch der Inhalt sein.
Das Herausarbeiten der Makrostruktur ist ein Konstruktionsprozess, den Sie erlernen müssen (hat aber starke Ähnlichkeit mit den Prozessen, die beim Exzerpieren ablaufen – die sind Ihnen bekannt): Dabei werden Mikropropositionen generalisiert (wenn relevant), weggelassen (wenn irrelevant), abstrahiert oder zusammengefasst (in übergeordnete faktische oder semantische Einheiten), um zu den Makropropositionen zu kommen, die ihrerseits die Struktur des Textes implizieren.
Wichtig: Diese Verstehensprozesse unterliegen der Kontrolle des Leseziels/Zwecks, welches in der Schulung explizit erläutert wird. Denn wenn man ohne Anweisung einen Text liest, trifft man, so legen es jedenfalls Vergleichsuntersuchungen von Kintsch/van Dijk nahe, keine Unterscheidung zwischen Mikro- und Makropropositionen und behandelt diese bei *recall* oder *summary* gleichwertig (vgl. ebd.: 387) – das darf beim *rating* nicht passieren, genauso wenig wie beim Exzerpieren.

Was macht einen Text kohärent?
Kohärenz als propositionales und sprachliches Netzwerk auf Mikro- wie Makroebene (ebd.: 389) setzt sich aus verschiedenen Komponenten zusammen – aus referentieller und faktischer Kohärenz auf beiden Ebenen, die sich in sprachlichen Kohäsionsmitteln ausdrückt und in Propositionen versprachlicht ist, die entweder in referentieller oder faktischer Beziehung stehen auf Mikroebene, und sich auf Makroebene in übergeordneten Propositionen oder "higher order fact units" konstruieren lassen (390):

- verbundene Propositionen/Fakten auf Mikroebene (Kriterium „Inhalt)
- übergeordnete Propositionen/*Units*, abgeleitet aus Mikropropositionen („Textsorte/Aufbau")
- Sprachliche Kohäsionsmittel („Kohäsion), die (idealiter) die Verbundenheit auf Mikro- wie Makroebene widerspiegeln (Komplexe Texte zeigen evtl. weniger explizite sprachliche Verknüpfungen, schaffen Kohärenz über beispielsweise Implikationen – nicht negativ zu werten!)

Wichtig: Der kommunikative wie pragmatische Kontext, in dem die Textrezeption abläuft, bestimmt mit, ob man einen Text als kohärent akzeptiert oder nicht – beim *raten* der DESI-Aufsätze setzen wir keine strikt-formalen Merkmale an, da die Zeit zu kurz bemessen ist und dieses textuelle Wissen in der Fremdsprache so noch nicht vorhanden sein kann. Stattdessen richten wir unser Augenmerk eher auf kommunikativ-pragmatische Verständlichkeit: Ist der Text dergestalt, dass er trotz dieser Kunstsituation ein „kohärentes Ganzes" ergibt?

- **Anwendung des Obigen auf den *Rating*-Prozess in DESI**

Man geht vom ersten Eindruck des Textes als „Ganzes" zu Propositionen auf Mikroebene, erarbeitet sich den Textaufbau auf Makroebene in aktiven Konstruktionsprozessen; nach dieser Analyse der taskbezogenen Kriterien widmet man sich den sprachlichen Kriterien; im Anschluss wird die kommunikative Wirksamkeit bewertet.
=> Man geht von Oberfläche zu Tiefenstruktur, von inhaltlichen Propositionen zu Makrostruktur/Textaufbau, so wie wir es im Handbuch beschrieben haben:

1. Alle Propositionen auf Mikroebnen wahrnehmen und auf Relevanz, Motiviertheit (Verbundenheit semantischer/referentieller oder faktischer Art), Qualität abklopfen => wird unter **Inhalt** bewertet.
2. Die relevanten Mikropropositionen zu Makropropositionen kondensieren und Makrostruktur herausarbeiten => Textaufbau, formal-logische Gliederung wird unter **Textsorte/Aufbau** bewertet, da Erwartungen an eben diese Makrostruktur textsortenbedingt sind.
3. Die korrespondierenden sprachlichen Kohäsionsmittel werden unter ***linking language*** bei den sprachlichen Kriterien bewertet.

5. Ausblick auf Praxis

⇨ Erarbeitung jedes einzelnen Kriteriums anhand schon eingestufter *Benchmark*-Texte (Merkmale herausarbeiten und Einstufungen nachvollziehen)
⇨ Übungstexte selbst analysieren und einstufen (Konzentration auf je ein Kriterium)
⇨ Sukzessives Einüben des gesamten Bewertungsprozesses

Bibliographie:

J. Charles Alderson: “Bands and Scores”. In: J. Charles Alderson & Brian North (eds): *Language Testing in the 1990s: The Communicative Legacy*. London: Macmillian 1991, 71-86.

Geoff Brindley: “Describing Language Development? Rating Scales and SLA”. In: Lyle F. Bachman & Andrew D. Cohen (eds): *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: CUP 1998, 112-140.

Andrew D. Cohen: *Assessing Language Abilities in the Classroom*. Boston: Heinle & Heinle, 1994².

Council of Europe: *A Common European Framework of Reference for Language Learning and Teaching*. Strasbourg: 1996.

Council of Europe: *A Common European Framework of Reference for Language Learning and Teaching. User's Guide for Examiners*. Strasbourg: 1996.

Alister Cumming: “Expertise in evaluating second language compositions”. In: *Language Testing* 7, 1 (1990), 31-51.

Peter Elbow: “Writing Assessment: Do it better, do it less“. In: Edward White, William Lutz u. Sandra Kamusikiri (eds): *Assessment of writing: Politics, policies, practices*. NY: MLAA 1996, 120-134.

Europarat: *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheidt 2001.

Arthur Hughes: *Testing for Language Teachers*. CUP 1986. Insb. Chapter 9: “Testing Writing“, 75-100.

Walter Kintsch & Teun A. van Dijk: “Towards a Model of Text Comprehension and Production”. In: *Psychological Review* 85 (5), September 1978, 363-394.

Hans P. Krings: „Schreiben in der Fremdsprache – Prozessanalysen zum ‘vierten skill’.” In: Antons G. u. Krings Hans P.: *Textproduktion – ein interdisziplinärer Forschungsüberblick*. Tübingen: Niemeyer 1989, 377-436.

Rainer H. Lehmann: „Aufsatzbeurteilung – Forschungsstand und empirische Daten.“ In: Karlheinz Ingenkamp u. Reinhold S. Jäger: *Tests und Trends* 8, Weinheim: Beltz 1990, 64-94.

Tom Lumley: “Assessment criteria in a large-scale writing test: what do they really mean to the raters?“ In: *Language Testing* 2002 19 (3), 246-276.

Michael Milanovic, Nick Saville & Shen Shuhong: “A study of the decision-making behaviour of composition markers”. In: Michael Milanovic, Nick Saville (eds): *Language Testing 3 – Performance, Testing, Cognition and Assessment*. Cambridge: CUP 1996, 92-114.

Alastair Pollitt: “Giving Students a Sporting Chance: Assessment by Counting and by Judging”. In: J. Charles Alderson, Brian North (eds): *Language Testing in the 1990s: The Communicative Legacy*. London: Macmillian 1991, 46-59.

Alastair Pollitt & Neil L. Murray: “What raters really pay attention to”. In: Michael Milanovic, Nick Saville (eds): *Language Testing 3 – Performance, Testing, Cognition and Assessment*. Cambridge: CUP 1996, 74-91.

Doug Shale: “Essay Reliability: Form and Meaning”. In: In: Edward White, William Lutz u. Sandra Kamusikiri (eds): *Assessment of writing: Politics, policies, practices*. NY: MLAA 1996, 76-96.

Günther Schneider & Brian North: *Fremdsprachen können – was heißt das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit*. Zürich: Ruegger 2000.

Elana Shohamy, Claire M. Gordon & Roberta Kraemer: “The Effect of Raters’ Background and Training on the Reliability of Direct Writing Tests”. In: *The Modern Language Journal*, Vol. 76 (1) 1992, 27-33.

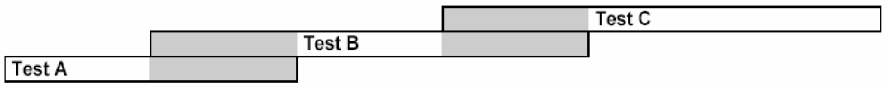
Glossar

Administration Die Administration eines Tests bezeichnet den Einsatz des Tests in einer Probandengruppe; siehe auch *Testdurchführung*.

Faktorenanalyse „Die Faktorenanalyse ist ein statistisches Verfahren zur Datenreduktion. Es wird häufig in der sozialwissenschaftlichen und psychologischen Forschung eingesetzt.
Ein Anwendungsbeispiel sind Persönlichkeitstests: Die Probanden füllen zunächst einen Fragebogen mit z. B. 60 skalierten Fragen zur Persönlichkeit aus. Aus diesen 60 Einzelwerten lässt sich jedoch kein schlüssiges Persönlichkeitsbild vermitteln.
Die Faktorenanalyse sucht jetzt zunächst über die Gesamtstichprobe (z. B. 1000 Personen) nach den dahinterliegenden Dimensionen der Einzelvariablen. Das können dann bei 60 Ursprungsvariablen zum Beispiel 8, 10, 12 oder auch mehr Dimensionen sein. (...) So lässt sich bei Persönlichkeitstests zum Beispiel in der Regel ein Faktor Extraversion/Introversion feststellen.
Grundlage für die Berechnung Faktorenanalyse ist eine Korrelationsmatrix.
Die häufigste Variante der Faktorenanalyse ist die Hauptkomponentenanalyse (oder auch principal components analysis genannt, abgekürzt PCA). Bei dieser Methode geht man davon aus, dass die Faktoren untereinander nicht korrelieren. Wie die Ursprungsvariablen zu der errechneten Dimension beitragen, wird dabei aus den Faktorladungen deutlich: Eine Ladung von 1 bedeutet, die Variable ist mit dem Faktor identisch, eine Ladung von 0 bedeutet, die Variable ist von dem Faktor vollkommen unabhängig.
Nach der Berechnung der Faktorenanalyse gibt der Faktorwert für jeden einzelnen Probanden seine Ausprägung auf den einzelnen Faktoren an.“[1]

Gleichung Siehe *Vergleich von Tests oder Prüfungen*.

Item-Response Theorie Siehe auch *probabilistische Testtheorie*.
„IRT stellt eine Gruppe von Mess- oder Skalierungsmodellen zur Verfügung. Das direkteste und stabilste ist das *Rasch*-Modell, benannt nach dem dänischen Mathematiker Georg Rasch. Die IRT ist eine Weiterentwicklung, basierend auf der Probabilitätstheorie, und wird vor allem dazu benutzt, den Schwierigkeitsgrad einzelner Testaufgaben in einer Itembank zu bestimmten. Fortgeschrittene Lernende haben hohe Chancen, eine elementare Frage richtig zu beantworten, Anfänger haben sehr geringe Chancen, eine anspruchsvolle Aufgabe zu lösen. Diese einfache Tatsache ist beim Rasch-Modell zu einer Skalierungsmethode entwickelt worden, die man benutzen kann, um Items auf der gleichen Skala zu kalibrieren. Eine Weiterentwicklung dieses Ansatzes kann sowohl zur Skalierung von Deskriptoren der Kommunikationsfähigkeit als auch zur Skalierung von Testitems benutzt werden. Bei einer Rasch-Analyse können verschiedenen Tests oder Fragebögen zu einer überlappenden Kette zusammengefügt werden, indem man 'Ankeritems' benutzt, die den aneinander grenzenden Elementen gemeinsam sind. Im folgenden Diagramm sind die Ankeritems grau schattiert. Auf diese Weise können die Test- oder Fragebögen auf bestimmte Gruppen abgestimmt werden, sie bleiben aber mit einer gemeinsamen Skala verknüpft. Allerdings muss man bei diesem Prozess sehr sorgfältig vorgehen, weil das Rasch-Modell die jeweils besten und niedrigsten Ergebnisse bei jedem Test verzerrt.



Der Vorteil einer Rasch-Analyse ist, dass sie ein stichproben- und skalenunabhängiges Maß liefern kann, d. h. eine Skalierung, die unabhängig ist von den Stichproben und den Tests/Fragebögen, die bei der Analyse benutzt wurden. Sie liefert Skalenwerte, die bei zukünftigen Gruppen konstant bleiben, vorausgesetzt, die zukünftigen Probanden können als neue Gruppen innerhalb der gleichen statistischen Population gelten. Systematische

[1] Quelle der Definition: http://de.wikipedia.org/wiki/Faktorenanalyse, Zugriff letztmalig am 13.09.2005.

Veränderungen in den Werten im Verlauf der Zeit (z. B. aufgrund curricularer Veränderungen oder von Prüfertraining) können quantifiziert und in Anpassungen berücksichtigt werden. Ebenso kann systematische Variation zwischen Lernertypen bzw. Typen von Beurteilenden quantifiziert und ausgeglichen werden (Wright & Masters 1982; Linacre 1989)."[2]

Kalibrierung Siehe *Vergleich von Tests oder Prüfungen*.

Klassische Testtheorie „Die klassische Testtheorie ist die meistverbreitete psychometrische Testtheorie. Der Schwerpunkt des Modells der klassischen Testtheorie liegt auf der Genauigkeit einer Messung bzw. auf der Größe des jeweiligen Messfehlers. Daher wird sie oft auch als Messfehlertheorie bezeichnet.
Die klassische Testtheorie versucht zu klären, wie, ausgehend von einem Testwert einer Versuchsperson, auf die wahre Ausprägung des zu messenden Persönlichkeitsmerkmals geschlossen werden kann."[3] Siehe auch *Testtheorie*.

Korrelation „Die Korrelation ist eine Beziehung zwischen zwei oder mehr quantitativen statistischen Variablen. Es gibt positive und negative Korrelationen. Ein Beispiel für eine positive Korrelation (je mehr, desto mehr) ist: Je mehr Futter, desto dickere Kühe. Ein Beispiel für eine negative Korrelation (je mehr, desto weniger) ist: Je mehr Verkauf von Regenschirmen, desto weniger Verkauf von Sonnencreme.
Die Korrelation beschreibt nicht unbedingt eine Ursache-Wirkungs-Beziehung in die eine oder andere Richtung. So darf man über die Tatsache, dass man Feuerwehren oft bei Bränden findet, nicht folgern, dass sie sie legt. Die direkte Kausalität kann auch gänzlich fehlen. So kann es durchaus eine Korrelation zwischen dem Rückgang der Störche im Burgenland und einem Rückgang der Anzahl Neugeborener geben, aber diese Ereignisse haben natürlich direkt nichts miteinander zu tun (weder bringen Störche Kinder noch umgekehrt), das heißt, sie haben kausal allenfalls über eine dritte Größe etwas miteinander zu tun, etwa über die Verstädterung, die Nistplätze vernichtet.
Im Gegensatz zur Proportionalität ist die Korrelation nur ein stochastischer Zusammenhang, das heißt, es kann nur eine ungefähre Zu- oder Abnahme prognostiziert werden. Zum Beispiel kann eine 200-prozentige Steigerung der Futtermenge mal eine Gewichtszunahme der Kühe von 10%, mal von 20% bewirken, wohingegen eine Verdoppelung der Masse eines Hammers bei gleicher Beschleunigung immer eine Verdoppelung der Kraft bewirkt, da hier ein proportionaler Zusammenhang besteht."[4]

Kriterienorientiertes Testen Das kriterienorientierte Testen nutzt, anders als das *normorientierte Testen*, ein Kriterium (sei es ein konkretes Lernziel oder ein Standard) als Bezugspunkt: Man erhält Aussagen darüber, wie gut ein Proband ein bestimmtes Kriterium beherrscht respektive ob ein Proband einen bestimmten Standard erreicht.

Leistungsdimensionen siehe *Testgegenstand*.

Modell-basierter Testansatz Bei diesem Ansatz ist das *Testkonstrukt* in theoretischen Modellen verortet, welche den *Testgegenstand* beschreiben.

Normorientiertes Testen Im Gegensatz zum *kriterienorientierten Testen* wird beim normorientierten Testen die Probandengruppe als Bezugsnorm genutzt: Man erhält Aussagen darüber, wo das Individuum in Bezug auf die Probandengruppe steht.

Operationalisierung „Die Operationalisierung beschreibt die Art und Weise, wie ein theoretisches Konstrukt [siehe auch *Testkonstrukt*, Anm. d. V.] gemessen werden soll. Die Operationalisierung hat in allen Wissenschaften eine große Bedeutung, da sie die Grundlage dafür sind, dass Experimente wiederholt werden können, denn nur dadurch, dass ein Experiment wiederholt wird und dabei

[2] Quelle der Definition: GER (2001: 204f).
[3] Quelle der Definition: http://de.wikipedia.org/wiki/Klassische_Testtheorie, Zugriff letztmalig am 13.09.2005.
[4] Quelle der Definition: http://de.wikipedia.org/wiki/Korrelation, Zugriff letztmalig am 13.09.2005.

die gleichen Ergebnisse erreicht werden, kann eine Hypothese zuverlässig geprüft werden.
Die Verarbeitungsgeschwindigkeit des menschlichen Gehirns zum Beispiel kann mit Hilfe der Reaktionsgeschwindigkeit operationalisiert werden.
Neben der Messgröße, wie hier zum Beispiel der Reaktionsgeschwindigkeit, muss für die Operationalisierung aber auch noch die Erhebungsmethode, das Erhebungsinstrument und dabei insbesondere die Teile, mit denen die empirische Informationen gewonnen werden sollen, beschrieben werden. Schließlich ist noch das Verfahren zu erläutern, wie die Informationen für die eigentliche Analyse aufbereitet werden.
Auf das oben genannte Beispiel bezogen hieße also das theoretische Konstrukt "Verarbeitungsgeschwindigkeit des menschlichen Gehirns". Die Operationalisierung könnte ein Experiment sein, in dem ein Proband vor ein Gerät gesetzt wird, das aus einer einzelnen Lampe und einem Druckschalter besteht. Lampe und Druckschalter sind an eine Computer-Stoppuhr angeschlossen, die gestartet wird, wenn die Lampe aufleuchtet und stoppt, wenn der Proband den Schalter drückt. Das ist die Beschreibung des Erhebungsinstruments. Der Schalter ist dabei so konstruiert, dass er die Stoppuhr in dem Moment anhält, in dem der Finger des Probanden den Schalter berührt. Der Proband muss demnach keinen Widerstand überwinden, um den Schalter auszulösen (Beschreibung der Teile des Instruments, die zur Gewinnung der Information benutzt werden). Der Proband erhält die Aufgabe, den Schalter zu drücken, sobald die Lampe aufleuchtet. Soweit die Beschreibung der Erhebungsmethode.“ [5]

Probabilistische Testtheorie

„Die probabilistische Testtheorie (*Item-Response Theorie*, s. a. dort) beschreibt, wie man aus Ergebnissen standardisierter psychometrischer Tests auf Persönlichkeitseigenschaften zurückschließen kann. Testtheorie wird sowohl auf Multiple Choice-Tests wie auch auf Tests mit offeneren Antwortformaten angewandt.“[6] Siehe auch *Testtheorie*.
Dabei werden Methoden der Psychometrie auf Basis probabilistischer mathematischer Modelle (z. B. *Rasch-Modell*, s. a. dort) eingesetzt.

Qualitative/quantitative Verfahren

Siehe *Testauswertung*.

Rasch-Modell

Das Rasch-Modell ist ein mathematisches Skalierungsmodell der probabilistischen Testtheorie, siehe *Item-Response Theorie*.

Regressionsanalysen

Regressionsanalysen werden genutzt, um den Einflusses mehrerer (unabhängiger) Variablen auf eine von ihnen abhängige Variable zu untersuchen mittels Bestimmung eines Regressionsgewichts, das der Einflussgröße auf die abhängige Variable entspricht – Testschwierigkeiten oder Testperformanzen können beispielsweise durch verschiedene Variablen und deren mittels Regression bestimmtem Gewicht erklärt werden.
„Die Regressionsanalyse ist ein statistisches Verfahren zur Analyse von Daten und geht von der Aufgabenstellung aus, sog. "einseitige" statistische Abhängigkeiten (d.h. statistische Ursache-Wirkung-Beziehungen) durch so genannte "Regressionsfunktionen" zu beschreiben. Dazu verwendet man oft lineare Funktionen, aber auch quadratische Funktionen und Exponentialfunktionen.
Es wird eine metrische Variable y betrachtet, die von einer oder mehreren metrischen unabhängigen Variablen bestimmt wird. Ein Beispiel wäre die Abhängigkeit der Arbeitslosenzahl von den Exporten und dem Inlandskonsum. Mit Hilfe der Regressionsanalyse wird die Struktur der Abhängigkeit zwischen y und den unabhängigen Variablen untersucht. Die interessierende Variable y wird abhängige Variable oder Zielvariable und die erklärenden Variablen x werden unabhängige Variablen oder Regressoren genannt.
Ein spezielles Verfahren der Regressionsanalyse ist die lineare Regression, bei der angenommen wird, dass ein interessierendes Merkmal y gut durch eine lineare Kombination anderer Merkmale x erklärt werden kann. Die Gewichtung der Einflüsse der erklärenden Merkmale wird dabei aus Daten geschätzt.
Betrachtet man den Fall mit nur einer unabhängigen Variablen, so spricht man von linearer

[5] Quelle der Definition: http://de.wikipedia.org/wiki/Operationalisierung, Zugriff letztmalig am 13.09.2005.
[6] Quelle der Definition: http://de.wikipedia.org/wiki/Probabilistische_Testtheorie, Zugriff letztmalig am 13.09.2005.

Einfachregression, den Fall mit 2 oder mehr unabhängigen Variablen bezeichnet man als multiple lineare Regression.“ [7]

Statistische Prüfung Siehe *Vergleich von Tests oder Prüfungen*.

Stimulus Unter *Stimulus* wird in dieser Arbeit jener Teil einer offenen (Test-)Aufgabenstellung verstanden, der in der Regel aus Text oder Bildern besteht und die Probanden zur Bearbeitung der Aufgabe anregen soll. Neben den Stimulus tritt die Arbeitsanweisung, welche Situation, Kommunikationsanlass und -Partner sowie die eigentliche Handlungsanweisung enthält.

Streuung „Unter Streuung fasst man in der Statistik verschiedene Maßzahlen zusammen, die der Einschätzung der Streubreite von Stichprobenwerten um ihren Mittelwert dienen.“ [8]

Task Unter *Task* wird in der vorliegenden Arbeit eine handlungsorientierte Aufgabenstellung verstanden, die sich in der Regel aus *Stimulus* und Arbeitsanweisung zusammensetzt.

Testauswertung Bei der Testauswertung, der Bewertung der Testitems oder Performanzen nach der *Testdurchführung*, kann man zwischen quantitativen und qualitativen Verfahren unterscheiden:
Quantitative Zählverfahren werden bei geschlossenen Formaten angewandt: Jede richtige Amtwort wird beispielsweise als ein Punkt gezählt.
Qualitative Verfahren kommen in der Regel bei offenen Formaten zum Einsatz; dabei wird die Qualität der Performanz beurteilt. Siehe auch *Testformat*.

Testdurchführung Bezeichnet den eigentlichen Testeinsatz, den Testlauf; siehe auch *Administration*.

Testformate Tests haben unterschiedliche Formate; die Klassifizierung der Formate kann unter verschiedenen Gesichtspunkten erfolgen: Bezogen auf die Antwortmöglichkeiten werden offene von geschlossenen Formaten unterschieden; bezogen auf die Art, wie sprachliche Aspekte in einem Test erfasst werden können, werden direkte von indirekten Formaten unterschieden; und bezogen auf die Dimensionalität des zu erfassenden *Testgegenstands* werden *discrete-point tests* von integrativen Formaten unterschieden:
Offene Formate geben den Probanden die Möglichkeit, ihre Antwort frei zu formulieren; diese Formate elizitieren Performanzbeispiele. Bei geschlossenen Formaten hingegen gibt es nur eine korrekte Antwortmöglichkeit; diese Formate lassen in der Regel Auswahl- oder Ankreuzmöglichkeiten zu. Das bekannteste geschlossene Format ist das *Multiple-Choice* Format.
Direkte Formate erfassen die zu testende sprachliche Handlung oder Produktion direkt, wobei es sich dabei in der Regel um Performanzen im produktiven und interaktiven Bereich handelt, wohingegen indirekte Formate die Probandenfähigkeit über Indikatoren erfassen. Beispielsweise sind rezeptive Fertigkeiten nicht direkt beobachtbar, weshalb sie indirekt erfasst werden, beim Leseverstehen etwa über Fragen zum Textverständnis.
Discrete-point tests erfassen isolierte sprachliche Elemente auf Basis der Annahme, dass sich verschiedene sprachliche Dimensionen diskret voneinander unterscheiden lassen, während integrative Formate sich dem Sprachvermögen ganzheitlich zu nähern versuchen.

Testgegenstand Das, was ein Test erfassen soll, wird als Gegenstand oder Leistungsdimension im *Testkonstrukt* beschrieben.

Testitem Testaufgaben, die das *Testkonstrukt operationalisieren*, werden Testitems genannt.
Testitems können verschiedene Formate haben, siehe *Testformate*.

Testkonstrukt Das Testkonstrukt beschreibt das theoretische Verständnis der Wissensbestände, Kompetenzen, Fertigkeiten etc., die ein Test erfassen soll. Dieses theoretische Verständnis muss in konkrete *Testitems* umgesetzt werden. Dieser Prozess wird *Operationalisierung* genannt.

[7] Quelle der Definition: http://de.wikipedia.org/wiki/Regressionsanalyse, Zugriff letztmalig am 13.09.2005.
[8] Quelle der Definition: http://de.wikipedia.org/wiki/Streuung_%28Statistik%29, Zugriff letztmalig am 13.09.2005.

Testtheorie

„In der Psychometrie beruht eine Testtheorie auf einem mathematischen Modell, das bestimmte statistische Zusammenhänge zwischen Persönlichkeitsmerkmalen und empirischen Testwerten erwarten lässt. Nach einer Testdurchführung schließt man mit Hilfe der Testtheorie von den Testergebnissen auf die Persönlichkeitsmerkmale zurück. Die Testtheorie liefert ferner Gütekriterien, anhand derer die Signifikanz der Ergebnisse und damit die Qualität des Tests beurteilt werden können.
Man unterscheidet zwischen *klassischer* und *probabilistischer* Testtheorie; letztere heißt auch Item-Response Theorie."[9] Siehe auch *klassische, probabilistische, Item-Response Theorie*.

Variablen

„Eine Variable ist eine Größe, die verschiedene Werte annehmen kann. Sie ist also in ihrer Größe veränderlich oder variabel. Variablen werden auch Platzhalter oder Unbekannte genannt. Sie kommen in Formeln und Termen vor."[10]
Bei der Testauswertung oder Testanalyse werden diejenigen Größen als Variablen bezeichnet, die untersucht werden sollen. Dabei gibt es unabhängige Variablen und solche, die von anderen Variablen abhängig sind. Zusammenhänge zwischen Variablen können über *Korrelations-, Regressions-* oder *Faktorenanalysen* untersucht werden.

Varianz

„Die Varianz ist in der Statistik ein Streuungsmaß, d.h. ein Maß für die Abweichung einer Zufallsvariable X von ihrem Erwartungswert $E(X)$."[11] Siehe auch *Streuung*.

Vergleich von Tests oder Prüfungen

Um verschiedene Tests und Prüfungen zu vergleichen, gibt es traditionell drei Verfahren:[12]
1. *Gleichung*: Dabei werden alternative Versionen des gleichen Tests in verschiedenen Testpopulationen produziert. Die Testresultate können dann problemlos verglichen werden.
2. *Kalibrierung*: Dies ist ein psychometrisches Skalierungsverfahren, bei dem Resultate aus verschiedenen Tests auf eine gemeinsame Skala kalibiriert oder geeicht werden, also verschiedene Testresultate gemeinsam skaliert werden.
3. *Statistische Prüfung*: Dabei werden Testresultate unter Zuhilfenahme statistischer Rechenverfahren bereinigt, um den unterschiedlichen Schwierigkeitsgraden der zu vergleichenden Testaufgaben und/oder der unterschiedlichen Strenge der Bewerter einer offenen Testaufgabe gerecht zu werden und die Tests auf diese Weise vergleichbar zu machen.

[9] Quelle der Definition: http://de.wikipedia.org/wiki/Testtheorie, Zugriff letztmalig am 13.09.2005.
[10] Quelle der Definition: http://de.wikipedia.org/wiki/Variablen, Zugriff letztmalig am 13.09.2005.
[11] Quelle der Definition: http://de.wikipedia.org/wiki/Varianz, Zugriff letztmalig am 13.09.2005.
[12] Nach GER 2001: 176f.

Bibliographie

Aguado, Karin (Hrsg.) (2000): *Zur Methodologie in der empirischen Fremdsprachenforschung*. Hohengehren: Schneider.

Ahrendt, Manfred (1991): „Die vier Arten der Einsprachigkeit." In: *Praxis des neusprachlichen Unterrichts* 38/2, 115-122.

Alderson, J. Charles (1991a): "Dis-sporting life. Response to Alastair Pollitt's paper: 'Giving Students a Sporting Chance: Assessment by Counting and by Judging'." In: Alderson & North (eds): *Language Testing in the 1990s*, 60-70.

Alderson, J. Charles (1991b): "Bands and Scores." In: Alderson & North (eds): *Language Testing in the 1990s*, 71-86.

Alderson, J. Charles (ed.) (2002): *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Case Studies*. Strasbourg: Council of Europe.

Alderson, J. Charles & Brian North (eds) (1991): *Language Testing in the 1990s: The Communicative Legacy*. London: Macmillan.

Alderson, J. Charles, Caroline Clapham & Dianne Wall (1995): *Language Test Construction and Evaluation*. Cambridge: University Press.

Alderson, J. Charles, Neus Figueras, Henk Kuijper, Günter Nold, Sauli Takala and Claire Tardieu (2004): *The Development of Specifications for Item Development and Classification within the CEF: Reading and Listening. Final Report of The Dutch CEF Construct Project*. Amsterdam: Unpublished Document.

Allen, Harold B. & Russel N. Campbell (eds) (1972): *Teaching English as a Second Language*. New York: McGraw-Hill.

Amor, Stuart (1999): "A Project on Photography: The Dynamics of Authenticity." In: *Der fremdsprachliche Unterricht Englisch* 4, 10-16.

Ansorge, Rainer (Hrsg.) (1994): *Schlaglichter der Forschung*. Hamburg: Reimer.

Antons, Gerd & Hans P. Krings (Hrsg.) (1989): *Textproduktion. Ein interdisziplinärer Forschungsüberblick*. Tübingen: Niemeyer.

Bachman, Lyle F. (^{2}1991a): *Fundamental Considerations in Language Testing*. Oxford: University Press.

Bachman, Lyle F. (1991b): "What does language testing have to offer?" In: *TESOL Quarterly* 25, 671-704.

Bachmann, Lyle F. & Adrian S. Palmer (1987): "The Construct Validation of Some Components of Communicative Proficiency." In: Grotjahn et al. (eds), 91-110.

Bachmann, Lyle F. & Adrian S. Palmer (1996): *Language Testing in Practice*. Oxford: University Press.

Bachman, Lyle F. & Andrew D. Cohen (eds) (1998): *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: University Press.

Baker, Colin (^{3}2001): *Foundations of Bilingual Education and Bilingualism*. Clevedon: Multilingual Matters.

Barkowski, Hans (2003): „*Skalierte Vagheit* – der europäische Referenzrahmen für Sprachen und sein Versuch, die sprachliche Kommunikationskompetenz des Menschen für Anliegen des Fremdsprachenunterrichts niveaugerecht zu portionieren.“ In: Bausch et al. (Hrsg): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 22-28.

Barkowski, Hans & Armin Wolff (Hrsg.) (1999): *Materialien Deutsch als Fremdsprache 52*. Regensburg: Fachverband Deutsch als Fremdsprache.

Bausch, Karl-Richard, Herbert Christ, Frank G. Königs & Hans-Jürgen Krumm (Hrsg.) (^{3}1995): *Handbuch Fremdsprachenunterricht*. Tübingen: Franke.

Bausch, Karl-Richard, Herbert Christ, Frank G. Königs & Hans-Jürgen Krumm (Hrsg.) (2003): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*. Tübingen: Narr.

Bausch, Karl-Richard, Herbert Christ & Hans-Jürgen Krumm (Hrsg.) (^{4}2003): *Handbuch Fremdsprachenunterricht*. Tübingen: Franke.

Beck, Bärbel & Eckhard Klieme (Hrsg.) (2005): *Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie*. Weinheim: Beltz (in Vorbereitung).

Becker-Mrotzek, Michael (1997): *Schreibentwicklung und Textproduktion*. Opladen: Westdeutscher Verlag GmbH.

Becker, Gerold (Hrsg.) (2001): *Qualität entwickeln: Evaluieren*. Friedrich Jahresheft XIX.

Berchem, Theodor (1992): „Europa: Kulturelle Identität und sprachliche Vielfalt. Gedanken zu einer möglichen europäischen Sprachenpolitik.“ In: Gnutzmann et al. (Hrsg), 48-63.

Bereiter, Carl (1980): “Development in Writing.“ In: Gregg & Steinberg (eds.), 73-93.

Bleyhl, Werner (1996): „Der Fallstrick des traditionellen Lehrens und Lernens fremder Sprachen. Vom Unterschied zwischen linearem und nicht-linearem Fremdsprachenunterricht.“ In: *Praxis des neusprachlichen Unterrichts* 43/4, 339-347.

Bleyhl, Werner (2003): „Die sprachliche Leistungsbeurteilung und die Chance zur Verbesserung des Fremdsprachenunterrichts dank des *backwash*-Effekts.“ In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 36-44.

Bleyhl, Werner & Johannes P. Timm (1998): „Wortschatz und Grammatik im Kontext.“ In: Timm (Hrsg.), 259-271.

Börner, Wolfgang (1989): „Didaktik schriftlicher Textproduktion in der Fremdsprache.“ In: Antons et al. (Hrsg.), 348-376.

Bolten, Jürgen (1994): „Im Spiel der Lebenswelten. Zur theoretischen Grundlegung interkulturellen Kommunikationstrainings.” In: Bungarten (Hrsg.), 17-34.

Bolton, Sybille (Hrsg.) (2000): *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests*. Köln: Gilde-Verlag.

Brindley, Geoff (1998): “Describing Language Development? Rating Scales and SLA.” In: Bachman et al. (eds), 112-140.

Brinker, Klaus (^{2}1988): *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden*. Berlin: Schmidt Verlag.

Broadbent, John & Leonardo Oriolo (1991): ”Language Education Across Europe: Towards an Intercultural Perspective.“ In: Buttjes & Byram (eds), 306-322.

Brown, Roger (1973): *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Brown, Douglas (^{3}1994): *Principles of Language Learning and Teaching*. New Jersey: Prentice Hall.

Buhren, Claus G., Dagmar Killus & Sabine Müller (^{5}2002): *Wege und Methoden der Selbstevaluation. Ein praktischer Leitfaden für Schulen*. Dortmund: IFS.

Bungarten, Theo (Hrsg.) (1994): *Kommunikationstraining im wirtschaftlichen Umfeld*. Tostedt: Attikon.

Burkard, Christoph (1997): *Lernfall externe Evaluation*. Bönen: Kettler.

Buttjes, Dieter (1991): ”Mediating Languages and Cultures: The Social and Intercultural Dimension Restored.“ In: Buttjes & Byram (eds), 3-16.

Buttjes, Dieter (^{3}1995): „Landeskunde-Didaktik und landeskundliches Curriculum.“ In: Bausch et al. (Hrsg.): *Handbuch Fremdsprachenunterricht*, 146.

Buttjes, Dieter & Michael Byram (eds) (1991): *Mediating Languages and Cultures: Towards an Intercultural Theory of Foreign Language Education*. Clevedon: Multilingual Matters.

Bybee, Joan L. (1976): *An Introduction to Natural Generative Phonology*. New York: Academic Press.

Bybee, Joan L. (1991): “Natural Morphology: The Organization of Paradigms and Language Acquisition.” In: Huebner et al. (eds), 67-92.

Byram, Michael (1991): ”Teaching Culture and Language: Towards an Integrated Model.“ In: Buttjes & Byram (eds), 17-32.

Camp, Roberta (1996): “New Views on Measurement and New Models for Writing Assessment.” In: White et al. (eds), 135-147.

Canale, Michael (1983): “On some Dimensions of Language Proficiency.“ In: Oller (ed.), 333-342.

Canale, Michael & Merill Swain (1980): “Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing.” In: *Applied Linguistics* 1/1, 1-47.

Canale, Michael & Merill Swain (1981): “A Theoretical Framework for Communicative Competence.” In: Palmer et al. (eds), 31-36.

Carroll, Brendan J. (1980): *Testing Communicative Performance. An Interim Study*. Oxford: Pergamon.

Carroll, John B. (1972): “Fundamental Considerations in Testing for English Language Proficiency of Foreign Students.” In: Allen et al. (eds), 313-321.

Carroll, John B. (1983): “Psychometric Theory and Language Testing.” In: Oller (ed.), 80-107.

Cattliff, Roslyn and Sydney Thorne (1988): *English in the Classroom. Englisch – wie es im Unterricht klingen soll*. Frankfurt/Main: Diesterweg.

Centre for Educational Research and Innovation (Hrsg.) (2001): *Bildung auf einen Blick. OECD-Indikatoren. Ausbildung und Kompetenzen*. Paris: OECD.

Chaudron, Craig (1988): *Second language classroom*. Cambridge: University Press.

Chomsky, Noam (1959): ”Review of B.F. Skinner (1957) Verbal behavior.“ In: *Language* 35, 26-58.

Chomsky, Noam (1965): *Aspects of the Theory of Syntax*. Cambridge, MA: The MIT Press.

Chomsky, Noam (1975): *Reflections on Language*. London: Temple Smith.

Chomsky, Noam (1980): *Rules and Representations*. Oxford: Blackwell.

Christ, Herbert (1991): *Fremdsprachenunterricht für das Jahr 2000. Sprachenpolitische Betrachtungen zum Lehren und Lernen fremder Sprachen*. Tübingen: Narr.

Christ, Herbert (2003): „Was leistet der ‘Gemeinsame europäische Referenzrahmen für Sprachen: lernen, lehren, beurteilen’?“ In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 57-66.

Christ, Ingeborg (2003): „Auf dem Wege zu einer neuen Evaluationskultur im Fremdsprachenunterricht.“ In: *Neusprachliche Mitteilungen aus Wissenschaft und Praxis* 3/56, 157-169.

Coffman, William E. (1971): “On the Reliability of Ratings of Essay Examination in English.” In: *Research in the Testing of English* 5/1, 24-36.

Cohen, Andrew D. (²1994): *Assessing Language Abilities in the Classroom*. Boston: Heinle & Heinle.

Cole, Peter & Jerry L. Morgan (eds) (1975): *Syntax and Semantics. Vol. 3: Speech Acts*. New York: Academic Press.

Connor-Linton, Jeff (1996): “Looking behind the Curtain: What do L2 Composition Ratings Really Mean?” In: *TESOL Quarterly* 30, 762-765.

Cooper, Robert L. (1972): “Testing.” In: Allen et al. (eds), 330-345.

Council of Europe (1996a): *A Common European Framework of Reference for Language Learning and Teaching*. Strasbourg. Online: http://culture2.coe.int/portfolio/documents/0521803136txt.pdf, Zugriff letztmalig am 13.09.2005.

Council of Europe (1996b): *A Common European Framework of Reference for Language Learning and Teaching. User's Guide for Examiners*. Strasbourg. Online: http://www.coe.int, Zugriff am 15.01.2002.

Revised Version:

Council of Europe (²2002): *Common European Framework of Reference for Language Learning and Teaching. Language Examination and Test Development*. Strasbourg. Online: http://culture2.coe.int/portfolio/documents/Guide%20October%202002%20revised%20version1.doc, Zugriff letztmalig am 13.09.2005.

Council of Europe (2003a): *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). Manual. Preliminary Pilot Version*. Strasbourg: September 2003, DGIV/EDU/LANG (2003) 5 rev. 1. 10. Online: http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual, Zugriff letztmalig am 13.09.2005.

Council of Europe (2003b): *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). Manual. Overwiev of Preliminary Pilot Version*. Strasbourg: September 2003, DGIV/EDU/LANG (2003) 10. Online: http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual, Zugriff letztmalig am 13.09.2005.

Council of Europe (2004): *Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg: December 2004, DGIV/EDU/LANG (2004) 13. Online:
http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/language_Policy/Manual/-CEF%20reference%20supplement%20version%203.pdf?L=E, Zugriff am 23.03.2005.
Jetzt zugänglich über:
http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/language_Policy/Manual, Zugriff letztmalig am 13.09.2005.

Cumming, Alister (1998): "Theoretical Perspectives on Writing." In: Grabe (ed.), 61-78.

De Beaugrande, Robert (1985): *Writing Step by Step. Easy Strategies for Writing and Revising*. San Diego, CA: Harcourt Brace Jovanovich.

De Beaugrande, Robert & Wolfgang U. Dressler (1981): *Einführung in die Textlinguistik*. Tübingen: Niemeyer.

de Jong, John (1988): "Rating Scales and Listening Comprehension." In: *Australian Review of Applied Linguistics* 11/2, 73-87.

de Jong, John (2004): *The Role of the Common European Framework*. Paper presented at the Inaugural Conference of EALTA, Kranjska Gora, Slovenia, 14th – 16th May 2004. Online: http://www.ealta.eu.org/conference/ppt/EALTA2004_John.ppt, Zugriff am 03.02.2005.

Deutsches PISA-Konsortium (Hrsg.) (2001): *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Doyé, Peter (^{3}1995): „Lehr- und Lernziele." In: Bausch et al. (Hrsg.): *Handbuch Fremdsprachenunterricht*, 161-166.

Dulay, Heidi C. & Marina K. Burt (1974): "Natural Sequences in Child Second Language Acquisition." In: *Language Learning* 24, 37-53.

Edmondson, Willis (1999): „Die fremdsprachliche Ausbildung kann nicht den Schulen überlassen werden!" In: *Praxis des neusprachlichen Unterrichts* 46/6, 115-123.

Edmondson, Willis (2003): „Bildunsgpolitik und Referenzrahmen." In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 67-74.

Edmondson, Willis & Juliane House (1993): *Einführung in die Sprachlehrforschung*. Tübingen: UTB.

Ekholm, Mats (1996): *Wirksamkeit und Zukunft der Lehrerfortbildung in NRW. Abschlussbericht der Evaluationskommission*. Düsseldorf: Concept.

Elbow, Peter (1996): "Writing Assessment: Do it better, do it less." In: White et al. (eds), 120-134.

Ellis, Rod (1990): *Instructed Second Language Acquisition: Learning in the Classroom*. Oxford: Blackwell.

Ellis, Rod (1994): *The Study of Second Language Acquisition*. Oxford: University Press.

Europäische Union (1995): *Weißbuch Lehren und Lernen. Auf dem Wege zur kognitiven Gesellschaft*. Brüssel, Luxemburg: Amt für amtliche Veröffentlichungen der europäischen Gemeinschaften.

Europarat (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheidt.

Færch, Claus & Gabriele Kasper (1983): *Strategies in Interlanguage Communication*. London: Longman.

Farhady, Hossein (1979): ”The Disjunctive Fallacy between Discrete-point and Integrative Tests.“ In: *TESOL Quarterly* 13/3, 347-357.

Feilke, Helmut (2005): *Kommentar zur Konzeption der DESI-Module „Schreibfähigkeit”*. Unveröffentlichtes Manuskript im Rahmen der DESI-Fachtagung. Frankfurt/Main: September 2004.

Fetscher, Doris & Volker Hinnenkamp (1994): „Interkulturelles Kommunikationstraining und das Managen der interkulturellen Situation.” In: *Sprache und Literatur in Wissenschaft und Unterricht*, 67-87.

Finkenstaedt, Thomas & Konrad Schröder (Hrsg.) (1989): *Zwischen Empirie und Machbarkeit: Erstes Symposium zum Bundeswettbewerb Fremdsprachen*. Augsburger I & I Schriften, Bd. 50.

Frey, Evelyn (2002): *Prototypenorientierte Untersuchungen zur Pluralbildung der Substantive und ihre didaktischen Folgen*. Frankfurt/Main: Lang.

Gaile, Dorothee (1999): „Wie im richtigen Leben... .” In: *Praxis des neusprachlichen Unterrichts* 46/4, 356-362.

Gass, Susan & Carolin Madden (eds) (1985): *Input in Second Language Acquisition*. Cambridge, MA: Newbury House.

Gnutzmann, Claus & Frank G. Königs (1992): „Methodische und politische Dimensionen des Fremdsprachenunterrichts zu Beginn eines neuen Jahrzehnts.“ In: Gnutzmann et al. (Hrsg.), 9-47.

Gnutzmann, Claus, Frank G. Königs & Waldemar Pfeiffer (Hrsg.) (1992): *Fremdsprachenunterricht im internationalen Vergleich: Perspektive 2000*. Frankfurt/Main: Diesterweg.

Gogolin, Ingrid (1994): *Der monolinguale Habitus der multilingualen Schule*. Münster: Waxmann.

Gogolin, Ingrid (2003): „Der Gemeinsame europäische Referenzrahmen.“ In: Bausch et al. (Hrsg): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 85-94.

Grabe, William (ed.) (1998): *Foundations of Second Language Teaching. Annual Review of Applied Linguistics* Vol. 18. Cambridge: University Press.

Grabe, William & Robert B. Kaplan (1996): *Theory and Practice of Writing: An Applied Linguistics Perspective*. New York: Longman.

Gregg, Lee W. & Erwin R. Steinberg (eds) (1980): *Cognitive Processes in Writing*. Hillsdale, NJ: Erlbaum.

Grice, Herbert P. (1975): “Logic and Conversation.” In: Cole et al. (eds.), 41-58.

Grotjahn, Rüdiger (2000): „Determinanten der Schwierigkeit von Leseverstehensaufgaben.“ In: Bolton (Hrsg.), S. 7-56.

Grotjahn, Rüdiger, Christine Klein-Braley & Douglas K. Stevenson (eds) (1987): *Taking their Measure: The Validity and Validation of Language Tests*. Bochum: Brockmeyer.

Gudykunst, William S., Ruth M. Guzley & Mitchell R. Hammer (1996): ”Designing Intercultural Training.” In: Landis et al. (eds), 61-80.

Gumpertz, John J. (1982): *Discourse Strategies*. Cambridge: University Press.

Gumpertz, John J. (1984): *Communicative Competence Revisited*. Berkeley: University of California Press.

Gumpertz, John J. & Dell H. Hymes (eds) (1972): *Directions in Sociolinguistics: The Ethnography of Communication*. New York: Holt, Rinehart & Winston.

Halliday, Michael A. K. (1976): *Explorations in the Functions of Language*. London: Arnold.

Halliday, Michael A. K. & Ruqaiya Hasan (1976): *Cohesion in English*. London: Longman.

Halliday, Michael A. K. & Ruqaiya Hasan (1989): *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford: University Press.

Hamp-Lyons, Liz (ed.) (1991): *Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex.

Hamp-Lyons, Liz (1996a): "Rating Nonnative Writing: The Trouble with Holistic Scoring." In: *TESOL Quarterly* 30, 759-762.

Hamp-Lyons, Liz (1996b): "The Challenges of Second-Language Writing Assessment." In: White et al. (eds), 226-240.

Hamp-Lyons, Liz & Barbara Kroll (1996): "Issues in ESL Writing Assessment: An Overview." In: *College ESL* 6/1, 52-72.

Hansen, Georg (2000a): *Vielfalt lernen – die europäische Perspektive. Ein multimedialer Dateikurs*. Hagen: Fernuniversität. CD-Rom-Nr. 3806.

Hansen, Georg (2000b): *Vielfalt lernen – die europäische Perspektive. Ein multimedialer Dateikurs*. Hagen: Fernuniversität. CD-Rom-Nr. 3808.

Harris, David P. (1969): *Testing English as a Second Language*. New York: McGraw-Hill.

Harsch, Claudia (2004): "Writing Assessment: How the Common European Framework of Reference can be used for developing assessment scales." In: British Council Germany (ed.): *Standards in Language Learning and the Common European Framework*. Berlin, 5-6 March 2004, Conference Report.

Harsch, Claudia & Konrad Schröder (2005a): „Schule zwischen Selbst- und Fremdbestimmung: Internationale Sprachtests, PISA, DESI und die ‚neue Evaluationskultur'." In: Maisch (Hrsg.), 22-36.

Harsch, Claudia & Konrad Schröder (2005b): „Kompetenzmodelle und Kompetenzniveaus im Bereich des Englischen: Textrekonstruktion C-Test." In: Beck & Klieme (Hrsg.), 238-252.

Harsch, Claudia, Rainer H. Lehmann, Astrid Neumann & Konrad Schröder (2005): „Übergreifende Konzeptualisierung sprachlicher Kompetenzen: Schreibfähigkeit." In: Beck & Klieme (Hrsg.), 50-72.

Hartig, Johannes (2005): „Messung sprachlicher Kompetenzen: Skalierung und Kompetenzniveaus." In: Beck & Klieme (Hrsg.), 95-112.

Hartig, Johannes & Eckhard Klieme (2005): „Kompetenz und Kompetenzdiagnostik." In: Schweizer, Karl: *Leistung und Leistungsdiagnostik*. (In Vorbereitung).

Hayes, John & Linda Flower (1980): "Identifying the Organization of Writing Process." In: Gregg et al. (eds), 3-30.

Henning, Grant (1992): "Dimensionality and Construct Validity of Language Tests." In: *Language Testing* 9, 1-11.

Henrici, Gert (1993): „Fremdsprachenerwerb durch Interaktion?" In: *Fremdsprachen Lehren und Lernen* 22, 215-237.

Henrici, Gert (1995): *Spracherwerb durch Interaktion? Eine Einführung in die fremdsprachenerwerbsspezifische Diskursanalyse*. Hohengehren: Schneider.

Hertel, Elke (1994): *Der Schüler der Sekundarstufe I im Bundeswettbewerb Fremdsprachen*. Augsburger I & I Schriften, Bd. 70.

Hopes, Clive (1998): *Beurteilung, Evaluation und Sicherung der Qualität an Schulen in der Europäischen Union*. Frankfurt/Main: GFPF und DIPF.

House, Juliane (2003): „Der Gemeinsame europäische Referenzrahmen für Sprachen – Anspruch und Realität." In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 95-104.

Huebner, Tom & Charles A. Fergusonn (eds) (1991): *Crosscurrents in Second Language Acquisition and Linguistic Theories*. Amsterdam: Benjamins.

Hughes, Arthur (1986): *Testing for Language Teachers*. Cambridge: University Press.

Huxley, Renira & Elisabeth Ingram (eds) (1971): *Language Acquisition, Models and Methods*. New York: Academic Press.

Hymes, Dell (1971): "Competence and Performance in Linguistic Theory." In: Huxley et al. (eds), 3-28.

Hymes, Dell (1972a): "Models of Interaction of Language and Social Life." In: Gumpertz et al. (eds), 35-71.

Hymes, Dell (1972b): "On Communicative Competence." In: Pride & Holmes (eds), 269-293.

Ingenkamp, Karlheinz & Reinhold S. Jäger (Hrsg.) (1990): *Tests und Trends 8*. Weinheim: Beltz.

Ingram, Elisabeth (1978): "The Psycholinguistic Basis." In: Spolsky (ed.), 1-14.

Jost, Axel & Uwe Multhaup (1996): "Prozessorientierte Interpretation eines Telekommunikationsprojektes." In: *Der fremdsprachliche Unterricht Englisch* 1, 31-36.

Jude, Nina & Eckhardt Klieme (2005): „Definitionen sprachlicher Kompetenz – ein Differenzierungsansatz." In: Beck & Klieme (Hrsg.), 12-27.

Kaftandjieva, Felianka, Norman Verhelst & Sauli Takala (1999): *DIALANG: A Manual for Standard Setting Procedure*. (Unpublished Document).

Kahl, Peter W. (1977): „Leistungsmessung." In: Schröder et al. (Hrsg.), 133-137.

Kast, Bernd (1999): *Fertigkeit Schreiben*. Berlin: Langenscheidt.

Kelly, George A. (1955): *The Psychology of Personal Constructs.* Vols. I and II. New York, NY: Norton.

Kennedy, Alan & Alan Wilkes (eds) (1975): *Studies in Long-Term Memory*. London: John Wiley & Sons.

Kintsch, Walter & Teun A. van Dijk (1978): "Towards a Model of Text Comprehension and Production." In: *Psychological Review* 85/5, 363-394.

Klemm, Klaus (2001): „Ein fatales Verständnis von Qualität." In: Frankfurter Rundschau, 15.11.2001.

Kleppin, Karin (2003): „Der Gemeinsame europäische Referenzrahmen für Sprachen: Ärgernis oder Fortschritt?" In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 105-112.

Klieme, Eckhard (2001): *Erfassung sprachlicher Leistungen in der Sekundarstufe. Konstrukte, Messmodelle und Befunde aus Sicht der empirisch-pädagogischen Forschung*. (Unveröffentlichtes internes Arbeitspapier des DIPF zum DESI-Projekt, Stand 20.06.2001).

Klieme, Eckhard (2004): *Einleitender Vortrag auf der DESI-Fachtagung*. Frankfurt/Main: September 2004. (Unveröffentlichter Vortrag).

Klieme, Eckhard, Hermann Avenarius, Werner Blum, Peter Döbrich, Hans Gruber, Manfred Prenzel, Kristina Reiss, Kurt Riquarts, Jürgen Rost, Heinz-Elmar Tenorth & Helmut J. Vollmer (Hrsg.) (2003): *Zur Entwicklung nationaler Bildungsstandards – Eine Expertise*. Frankfurt/Main: Deutsches Institut für Internationale Pädagogische Forschung (DIPF).

Klieme, Eckhard, Wolfgang Eichler, Andreas Helmke, Rainer H. Lehmann, Günter Nold, Hans-Günter Rolff, Konrad Schröder, Günther Thomé & Heiner Willenberg (2003): *DESI-Bericht über die Entwicklung und Erprobung der Erhebungsinstrumente*. Vertrauliches Manuskript für die Kultusministerkonferenz. Frankfurt/Main: Deutsches Institut für Internationale Pädagogische Forschung (DIPF).

Königs, Frank G. (^{3}1995): „Die Dichotomie Lernen/Erwerben." In: Bausch et al. (Hrsg.): *Handbuch Fremdsprachenunterricht*, 428-431.

Königs, Frank G. (2003): „(K)Eine Referenz für den Referenzrahmen? Überlegungen zum ‚Gemeinsamen europäischen Referenzrahmen für Sprachen'." In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 113-119.

Köpcke, Klaus-Michael (1988): "Schemas in German Plural Formation." In: *Lingua* 74/4, 303-335.

Köpcke, Klaus-Michael (1993): *Schemata bei der Pluralbildung im Deutschen. Versuch einer kognitiven Morphologie*. Tübingen: Narr.

Köpcke, Klaus-Michael (1995): „Die Klassifikation der schwachen Maskulina in der deutschen Gegenwartssprache. Ein Beispiel für die Leistungsfähigkeit der Prototypentheorie." In: *Zeitschrift für Sprachwissenschaft* 14, 159-180.

Kultusministerkonferenz (2003): *Vereinbarung über Bildungsstandards für den mittleren Abschluss (Jahrgangsstufe 10)*. Beschluss der Kultusministerkonferenz vom 04.12.2003. Online: http://www.kmk.org/schul/Bildungsstandards/Rahmenvereinbarung_MSA_BS_04-12-2003.pdf, http://www.kmk.org/schul/Bildungsstandards/Deutsch_MSA_BS_04-12-03.pdf, http://www.kmk.org/schul/Bildungsstandards/1.Fremdsprache_MSA_BS_04-12-2003.pdf, Zugriff letztmalig am 13.09.2005.

Krashen, Stephen D. (1982): *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon.

Krashen, Stephen D. (1985): *The Input Hypothesis*. London: Longman.

Krings, Hans P. (1989): „Schreiben in der Fremdsprache – Prozessanalysen zum 'vierten skill'." In: Antons et al. (Hrsg.), 377-436.

Kroll, Barbara (1998): "Assessing Writing Abilities." In: Grabe (ed.), 219-240.

Krumm, Hans-Jürgen (2003): „Der Gemeinsame europäische Referenzrahmen – ein Kuckucksei für den Fremdsprachenunterricht." In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 120-126.

Krumm, Hans-Jürgen (^{4}2003): „Lehr- und Lernziele." In: Bausch et al. (Hrsg.): *Handbuch Fremdsprachenunterricht*, 116-121.

Lado, Robert (1961): *Language Testing*. New York: McGraw-Hill.

Landesinstitut für Schule und Weiterbildung (1999): *Qualitätssicherung und Qualitätsentwicklung im fremdsprachlichen Bereich – Maßnahmen zur Intensivierung des Sprachenlernens: Europäisches Portfolio der Sprachen, Europäischer Referenzrahmen*. Soest.

Landis, Dan & Rabi S. Bhagat (eds) ([2]1996): *Handbook of Intercultural Training*. Thousand Oaks u. a.: Sage.

Larson, Richard L. (1996): "Portfolios in the Assessment of Writing: A Political Perspective." In: White et al. (eds), 271-283.

Lee, Tony (1996): "Taking a Multifaceted View of the Unidimensional Measurement from Rasch Analysis in Language Tests." In: Milanovic & Saville (eds), 266-275.

Lehmann, Rainer H. (1990): „Aufsatzbeurteilung – Forschungsstand und empirische Daten." In: Ingenkamp & Jäger (Hrsg.), 64-94.

Lehmann, Rainer H. (1994): "Research on National and International Writing Assessments: Contributions from the Hamburg Study of Achievement in Written Composition." In: Ansorge (Hrsg.), 173-184.

Lehmann, Rainer H., Rüdiger Gänsfuß & Rainer Peek (1999): *Aspekte der Lernausgangslage und der Lernentwicklung – Klassenstufe 7*. Hamburg: Amt für Schule.

Lehmann, Rainer H., Rainer Peek, Rüdiger Gänsfuß & Vera Husfeldt (2000): *Aspekte der Lernausgangslage und der Lernentwicklung – Klassenstufe 9*. Hamburg: Amt für Schule.

Lehmann, Rainer H., Rainer Peek, Rüdiger Gänsfuß & Vera Husfeldt (2002): *Aspekte der Lernausgangslage und der Lernentwicklung – Klassenstufe 9. Ergebnisse einer Längsschnittstudie in Hamburg*. Hamburg: Behörde für Bildung und Sport.

Lienert, Gustav A. & Ulrich Raatz ([5]1994): *Testaufbau und Testanalyse*. Weinheim: Beltz.

Linacre, John M. (1989): *Multi-faceted Measurement*. Chicago: MESA Press.

Long, Michael H. (1983): "Native Speaker/Non-Native Speaker Conversation and the Negotiation of Comprehensible Input." In: *Applied Linguistics* 4, 126-141.

Lumley, Tom (2002): "Assessment Criteria in a Large-Scale Writing Test: What do They Really Mean to the Raters?" In: *Language Testing* 19/3, 246-276.

Macht, Konrad (1982): *Leistungsaspekte des Englischlernens*. Frankfurt/Main: Diesterweg.

Maisch, Josef (Hrsg.) (2005): *Evaluation und Analyse in der Schulentwicklung. Ansätze, Methoden und Beispiele für die Schulpraxis*. Donauwörth: Auer.

Martin, Jean-Pol & Rudolf Kelchner (1998): "Lernen durch Lehren." In: Timm (Hrsg.), 211-219.

McNamara, Tim (1996): *Measuring Second Language Performance*. London: Longman.

Milanovic, Michael & Nick Saville (eds) (1996): *Language Testing 3 – Performance, Testing, Cognition and Assessment*. Cambridge: University Press.

Milanovic, Michael, Nick Saville & Shen Shuhong (1996): "A Study of the Decision-Making Behaviour of Composition Markers." In: Milanovic & Saville (eds), 92-114.

Ministerium für Bildung, Wissenschaft und Kultur Mecklenburg-Vorpommern (2001): *Rahmenplan. Orientierungsstufe und Jahrgangsstufen 5 und 6 der integrierten Gesamtschule*. Schwerin.

Ministerium für Schule und Weiterbildung NRW (Hrsg) (1997): *‚...und sie bewegt sich doch!' Entwicklungskonzept ‚Stärkung der Schule'*. Frechen: Ritterbach (Schriftenreihe Schule in NRW 9014).

Ministerium für Schule und Weiterbildung NRW (Hrsg) (1998a): *Schulprogramm – eine Handreichung*. Frechen: Ritterbach (Schriftenreihe Schule in NRW 9027).

Ministerium für Schule und Weiterbildung NRW (Hrsg.) (1998b): ‚*Qualität als gemeinsame Aufgabe.' Rahmenkonzept ‚Qualitätsentwicklung und Qualitätssicherung schulischer Arbeit'*. Frechen: Ritterbach (Schriftenreihe Schule in NRW 9029).

Ministerium für Schule und Weiterbildung NRW (Hrsg.) (1998c): *Qualitätsentwicklung und Qualitätssicherung. Aufgabenbeispiele Klasse 10: Englisch.* Frechen: Ritterbach.

Mosenthal, Peter B. (1996): "Understanding the Strategies of Document Literacy and Their Conditions of Use." In: *Journal of Educational Psychology* 88/2, 314-332.

Murphy, Sandra & Barbara Grant (1996): "Portfolio Approaches to Assessment: Breakthrough or More of the Same?" In: White et al. (eds), 284-300.

Neuner, Gerhard (2003): „Der Gemeinsame europäische Referenzrahmen für Sprachen (RR) – neue Impulse für die Weiterentwicklung der Fremdsprachendidaktik und die Sprachlehrforschung." In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 140-144.

Neville, Helen & Daphne Bavelier (1998): "Neural Organization and Plasticity of Language." In: *Current Opinion in Neurobiology* 8, 254-258.

Nold, Günter (2000): „Ist schulisches Fremdsprachenlernen prognostizierbar? Überlegungen zum empirischen Forschungsprojekt ‚Lernstrategien zur Förderung sprachlicher Verstehensstrukturen in Englisch als Fremdsprache'." In: Aguado (Hrsg.), 75-92.

North, Brian (1995): "The Development of a Common Framework Scale of Descriptors of Language Proficiency based on a Theory of Measurement." In: *System* 23, 446-465.

North, Brian (2000): *The Development of a Common Framework Scale of Language Proficiency*. PhD Thesis, Thames Valley University 1996. Reprinted 2000, New York u. a.: Lang.

North, Brian & Günther Schneider (1998): "Scaling Descriptors for Language Proficiency Scales." In: *Language Testing* 15/2, 217-263.

Oller, John W. (1976): "Evidence for a General Language Proficiency Factor: An Expectancy Grammar." In: *Die Neueren Sprachen* 75, 165-174.

Oller, John W. (1979): *Language Tests at School. A Pragmatic Approach*. London: Longman.

Oller, John W. (ed.) (1983): *Issues in Language Testing Research*. Rowley, MA: Newbury House.

Palmer, Adrian S., Peter J. Groot & George A. Trosper (eds) (1981): *The Construct Validation of Tests of Communicative Competence*. Washington, DC: TESOL.

Paris, Scott G., T. A. Lawton, J. C. Turner & J. L. Roth (1991): "A Developmental Perspective on Standardized Achievement Testing." In: *Educational Researcher* 20/5, 12-20.

Picht, Robert (^{3}1995): „Kultur- und Landeswissenschaften." In: Bausch et al. (Hrsg): *Handbuch Fremdsprachenunterricht*, 66-73.

Pienemann, Manfred (1989): "Is Language Teachable?" In: *Applied Linguistics* 10, 52-79.

Piepho, Hans-Eberhard (1974): *Kommunikative Kompetenz als übergeordnetes Lernziel im Englischunterricht*. Dornburg-Frickhofen: Frankonius.

Pollitt, Alastair (1991a): "Giving Students a Sporting Chance: Assessment by Counting and by Judging." In: Alderson & North (eds): *Language Testing in the 1990s*, 46-59.

Pollitt, Alastair (1991b): “Response to Charles Alderson’s Paper: ‘Bands and Scores’.” In: Alderson & North (eds): *Language Testing in the 1990s*, 87-94.

Pollitt, Alastair & Neil L. Murray (1996): “What Raters Really Pay Attention to.” In: Milanovic & Saville (eds), 74-91.

Pommerin, Gabriele (1999): „Deutsch als Zweitsprache und Konzepte interkulturellen Lernens.“ In: Barkowski et al. (Hrsg.), 528-551.

Portmann, Paul (1991): *Schreiben und lernen*. Tübingen: Niemeyer.

Pride, John B. & Janet Holmes (eds) (1972): *Sociolinguistics*. Harmandsworth: Penguin.

Quetz, Jürgen (2003): „Der Gemeinsame europäische Referenzrahmen: Ein Schatzkästchen mit Perlen, aber auch mit Kreuzen und Ketten.“ In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 145-155.

Rost, Jürgen (1996): *Lehrbuch Testtheorie Testkonstruktion*. Bern: Huber.

Saunders, Malcolm (1982): *Multicultural Teaching. A Guide for the Classroom*. London: McGraw-Hill.

Savignon, Sandra J. (1983): *Communicative Competence: Theory and Classroom Practice*. Reading, MA: Addison-Wesley.

Schaeffer, Benson (1975): “Skill Integration during Cognitive Development.” In: Kennedy et al. (eds), 165-180.

Schiffrin, Deborah (1994): *Approaches to discourse*. Oxford: Blackwell.

Schneider, Günther & Brian North (2000): *Fremdsprachen können – was heißt das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit*. Zürich: Ruegger.

Schröder, Konrad (1971): „Englisch als Schulfach.” In: *Neusprachliche Mitteilungen aus Wissenschaft und Praxis* 24, 144-152.

Schröder, Konrad (1993): “Languages.“ In: Shelly et al. (eds), 13-64.

Schröder, Konrad (1999): „Den Englischunterricht europatauglich machen.“ In: *Mitteilungsblatt des FMF Hessen/Thüringen* 14. Online:
http://www.schule.bremen.de/schulen/IEF/FMF-HB/Schroe.pdf, Zugriff am 7.10.2003.

Schröder, Konrad (2004): *The Teaching of English in a European Context*. Unveröffentlichter Vortrag auf der ExpoLingua. Prag: November 2004.

Schröder, Konrad (2005): „Kommt nach dem ‚PISA-Schock’ der ‚DESI-Schock’? Sprachenzertifikate, PISA, DESI, die Bildungsstandards und die ‚neue Evaluationskultur’ an unseren Schulen.“ In: *Neusprachliche Mitteilungen aus Wissenschaft und Praxis* 3, 36-46.

Schröder, Konrad & Thomas Finkenstaedt (Hrsg.) (1977): *Reallexikon der englischen Fachdidaktik*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Schwarz, Paul (2001): „Entwickeln statt vermessen." In: *Frankfurter Rundschau*, 22.11.2001.

Schwerdtfeger, Inge (2003): „Der europäische Referenzrahmen – oder: Das Ende der Erforschung des Sprachenlernens?“ In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 173-179.

Selinker, Larry (1972): “Interlanguage.“ In: *International Review of Applied Linguistics in Language Teaching* 10, 209-231.

Selinker, Larry (1992): *Interlanguage Revisited*. London: Longman.

Shale, Doug (1996): “Essay Reliability: Form and Meaning.” In: White et al. (eds), 76-96.

Shelly, Monica & Margaret Winck (eds) (1993): *What is Europe? Volume 2: Aspects of Cultural Diversity*. London: Routledge.

Shohamy, Elana, Claire M. Gordon & Roberta Kraemer (1992): “The Effect of Raters’ Background and Training on the Reliability of Direct Writing Tests.” In: *The Modern Language Journal* 76/1, 27-33.

Sinclair, John & Malcolm R. Coulthard (1975): *Towards an Analysis of Discourse: The English used by Teachers and Pupils*. Oxford: University Press.

Solmecke, Gert (2000): „Faktoren der Schwierigkeit von Hörtests.” In: Bolton (Hrsg.), S. 57-76.

Spolsky, Bernhard (ed.) (1978a): *Approaches to Language Testing. Advances in Language Testing Series 2*. Arlington, VA: Center for Applied Linguistics.

Spolsky, Bernhard (1987b): “Linguists and Language Testers.” In: Spolsky (ed.) (1978a), v-x (*Introduction*).

Statistisches Bundesamt (Hrsg.) (2002): *Bildung im Zahlenspiegel*. Stuttgart: Metzler-Poeschel.

Swain, Merill (1985): “Communicative Competence: Some Roles of Comprehensible Input and Comprehensible Output in its Development.“ In: Gass et al. (eds), 235-253.

Takala, Sauli (2004): *Manual for relating examination to the Common European Framework: aims and procedures*. Paper presented at the Inaugural Conference of EALTA. Kranjska Gora, Slovenia: 14th – 16th May 2004. Online: http://www.ealta.eu.org/conference/ppt/takala14may.ppt, Zugriff letztmalig am 13.09.2005.

Tarone, Elaine E. (1981): “Some Thoughts on the Notion of Communicative Strategy.“ In: *TESOL Quarterly* 15, 285-295.

Thiele, Burkard (2000): *Die Bildungspolitik der Europäischen Gemeinschaft*. Münster: LIT.

Thomas, Alexander (1996): „Analyse der Handlungswirksamkeit von Kulturstandards.” In: Alexander Thomas (Hrsg.): *Psychologie interkulturellen Handelns*. Göttingen: Hogrefe, 107-135.

Thomas, Alexander & Karl Heinz Wagner (1999): „Von Fremdheitserfahrung zu interkulturellem Verstehen.” In: *Praxis des neusprachlichen Unterrichts* 46/3, 227-236.

Thürmann, Eike (2000): „Impulse aus der Praxis der Curriculumentwicklung für die Weiterentwicklung in der Fremdsprachendidaktik.“ In: *Fremdsprachen Lehren und Lernen* 29, 124-145.

Thurstone, Leon L. (1959): *The Measurement of Values*. Chicago: The University of Chicago Press.

Timm, Johannes-P. (Hrsg.) (1998): *Englisch lernen und lehren*. *Didaktik des Englischunterrichts*. Berlin: Cornelsen.

Torrance, Harry (1998): “Learning from Research in Assessment.” In: *Assessing Writing* 5/1, 31-37.

Upshur, John A. & Carolyn E. Turner (1995): “Constructing Rating Scales for Second Language Tests.” In: *ELT Journal* 49/1, 3-12.

van Ek, Jan A. (1975): *The Threshold Level in a European Unit/Credit System for Modern Language Learning by Adults*. Strasbourg: Council of Europe.

van Ek, Jan A. (1980): *Threshold Level English*. London: Pergamon Press.

van Ek, Jan A. & John L. M. Trim (1990): *Threshold Level 1990*. Strasbourg: Council of Europe.

van Ek, Jan A. & John L. M. Trim (1991): *Waystage 1990*. Cambridge: University Press.

van Ek, Jan A. & John L. M. Trim (1997): *Vantage Level*. Strasbourg: Council of Europe.

van Dijk, Teun A. (1977): *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. London: Longman.

Vollmer, Helmut J. (2003): „Ein gemeinsamer europäischer Referenzrahmen für Sprachen: Nicht mehr, nicht weniger." In: Bausch et al. (Hrsg.): *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion*, 192-206.

Wagner, Johannes (1992): „Perspektiven für den Ausbau von Fremdsprachenunterricht: Interkulturelle Kommunikation als Unterrichtsgegenstand." In: Gnutzmann et al. (Hrsg.), 138-152.

Weigle, Sara C. (1998): "Using FACETS to Model Rater Training Effects." In: *Language Testing* 15, 263-287.

White, Edward, William Lutz & Sandra Kamusikiri (eds) (1996): *Assessment of writing: Politics, Policies, Practices*. New York: MLAA.

Wiater, Werner (2005): „Evaluation in Schule und Unterricht." In: Maisch (Hrsg.), 8-19.

Wright, Benjamin D. & Geoff N. Masters (1982): *Rating Scale Analysis*. Chicago: MESA Press.

Bildnachweise:

Die Bilder in dieser Arbeit unterscheiden sich teils von denen, die in den Untersuchungen eingesetzt wurden, da bei einigen Bildern die Urheberrechte nicht eingeholt werden konnten.
Bei manchen der hier verwendeten Bilder war die Autorenschaft nicht eindeutig festzustellen.
Sollte es Probleme mit dem Copyright geben, kontaktieren Sie mich bitte und ich werde die Bilder unverzüglich ersetzen oder entfernen.

S. 265:
oben links: www.espn.go.com/boxing/columns/kellerman_max/1504893.html, Zugriff am 01.06.2006.
oben rechts: www.painetworks.com/pages2rf/ce/ce0225.html, Zugriff am 01.06.2006.
unten links: www.dietmarwanko.com, Zugriff am 01.06.2006.
unten rechts: www.spiegel.de/bildergalerie, Zugriff am 08.05.2002.

S. 333:
oben links: www.morgenpost-berlin.de, Zugriff am 07.01.2002.
oben rechts: www.painetworks.com/pages2rf/ce/ce0225.html, Zugriff am 01.06.2006.
unten links: www.dietmarwanko.com, Zugriff am 01.06.2006.
unten rechts: www.drehscheibe-deutschland.de/Mode/Miss_World/miss_world.html, Zugriff am 01.06.2006.

S. 334:
oben links: www.spiegel.de/bildergalerie, Zugriff am 08.05.2002.
oben rechts: www.painetworks.com/pages2rf/ce/ce0225.html, Zugriff am 01.06.2006.
unten links: www.dietmarwanko.com, Zugriff am 01.06.2006.
unten rechts: www.spiegel.de/bildergalerie, Zugriff am 08.05.2002.

Internetseiten:

ALTE: http://www.alte.org, Zugriff letztmalig am 13.09.2005.

ALTE CEFR Grid for the Analysis of Writing Tasks:
http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual/,
Zugriff letztmalig am 13.09.2005.

Cambridge Exams: http://www.cambridgeesol.org/exams, Zugriff letztmalig am 13.09.2005.

CEF: http://culture2.coe.int/portfolio/documents/0521803136txt.pdf, Zugriff letztmalig am 13.09.2005.

DESI-Projekt: http://www.dipf.de/desi, Zugriff letztmalig am 13.09.2005.

DIALANG-Projekt: http://www.dialang.org, Zugriff letztmalig am 13.09.2005.

Dutch Grid: http://www.ling.lancs.ac.uk/cefgrid, Zugriff letztmalig am 13.09.2005,
http://www.ealta.eu.org/dutch/grid.htm, Zugriff am 27.01.2005,
http://bowland-files.lancs.ac.uk/cefgrid/, Zugriff letztmalig am 13.09.2005.

EALTA: http://www.ealta.eu.org/conference, Zugriff am 03.02.2005.

EALTA Präsentation Harsch:
http://www.ealta.eu.org/conference/2006/docs/Harsch_ealta2006.ppt, Zugriff am 06.07.2006.

EU-Sprachenportfolio: http://www.coe.int/portfolio, Zugriff letztmalig am 13.09.2005.

GER-Deskriptoren-Datenbank:
http://www.unifr.ch/ids/Portfolio/descriptors.htm, Zugriff letztmalig am 13.09.2005.

International English Language Testing System*:* http://www.ielts.org, Zugriff letztmalig am 13.09.2005.

KMK-Bildungsstandards, Zugriff letztmalig am 13.09.2005:
http://www.kmk.org/schul/Bildungsstandards/Rahmenvereinbarung_MSA_BS_04-12-2003.pdf
http://www.kmk.org/schul/Bildungsstandards/Deutsch_MSA_BS_04-12-03.pdf
http://www.kmk.org/schul/Bildungsstandards/1.Fremdsprache_MSA_BS_04-12-2003.pdf,

Manual for Relating Examinations to the CEF:
http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual,
Zugriff letztmalig am 13.09.2005.

Manual for Relating Examinations to the CEF: Reference Supplements:
http://www.coe.int/T/E/Cultural_Co-peration/education/Languages/Language_Policy/
Manual/CEF%20reference%20supplement%20version%203.pdf?L=E, Zugriff am 23.03.2005.
http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Manual,
Zugriff letztmalig am 13.09.2005.

Max-Planck-Institut: http://www.mpipf-muenchen.mpg.de/MPIPF/forsch_g.htm, Zugriff letztmalig am 13.09.2005.

Nano: http://www.3sat.de/nano/cstuecke/42778/index.html, Zugriff letztmalig am 13.09.2005.

PISA: http://www.pisa.oecd.org/pisa/outcome.htm, Zugriff am 02.12.2003,
oder http://www.mpib-berlin.mpg.de/pisa, Zugriff letztmalig am 13.09.2005.

User's Guide for Examiners (*Language Examining and Test Development*):
http://culture2.coe.int/portfolio/documents/Guide%20October%202002%20revised%20version1.doc,
Zugriff am 15.01.2002 auf Originalversion, letztmalig am 13.09.2005 auf *revised version*.

Claudia Harsch

Persönliche Angaben

Geburtsdatum:	14.11.1969
Geburtsort:	Augsburg
Staatsangehörigkeit:	deutsch

Ausbildung

1996 – 2001	**Universität Augsburg**
	Magisterstudium Deutsch als Fremdsprache, Didaktik des Englischen, Anglistik
November 2001	Abschluss: Magisterprüfung
2002 – 2005	**Universität Augsburg**
	Promotionsstudium Didaktik des Englischen
09/2005	Dissertation: „Der Gemeinsame europäische Referenzrahmen für Sprachen: Leistung und Grenzen“
07/2006	Disputatio

Berufserfahrung

09/1999 - 05/2000	**Pädagogischer Austauschdienst**
	Assistant Teacher in Huddersfield/ England am Technical College
07/2000 - 08/2000	**DiD-Sprachferien Augsburg**
	Lehrtätigkeit im Bereich Deutsch als Fremdsprache
01/2002 - jetzt	**Universität Augsburg**
	Wissenschaftliche Mitarbeiterin, Lehrstuhl Didaktik des Englischen:
01/2002 - 04/2006	DESI-Projekt: Schulleistungsstudie im Auftrag der KMK
09/2005 - jetzt	SEE1-Projekt: Evaluierung und Implementierung der Bildungsstandards in Zusammenarbeit mit dem Institut zur Qualitätssicherung im Bildungswesen Berlin
	Lehraufträge/Lehrtätigkeit für Deutsch als Fremdsprache und Didaktik des Englischen