SEEQ-DE: Konstruktion und Überprüfung einer deutschsprachigen Adaption des Instruments "Student Evaluation of Educational Quality" (SEEQ; Marsh, 1982, 2007)

Martin Daumiller, Robert Grassinger, Tobias Engelschalk, Markus Dresel

Angaben zur Veröffentlichung / Publication details:

Daumiller, Martin, Robert Grassinger, Tobias Engelschalk, and Markus Dresel. 2021. "SEEQ-DE: Konstruktion und Überprüfung einer deutschsprachigen Adaption des Instruments 'Student Evaluation of Educational Quality' (SEEQ; Marsh, 1982, 2007)." *Diagnostica* 67 (4): 176–88. https://doi.org/10.1026/0012-1924/a000274.

SEEQ-DE

Konstruktion und Überprüfung einer deutschsprachigen Adaption des Instruments „Student Evaluation of Educational Quality“ (SEEQ; Marsh, 1982, 2007)

Martin Daumiller[1], Robert Grassinger[2], Tobias Engelschalk[1] und Markus Dresel[1]

[1] Lehrstuhl für Psychologie, Philosophisch-Sozialwissenschaftliche Fakultät, Universität Augsburg

[2] Pädagogische Psychologie, Fakultät I, Pädagogische Hochschule Weingarten

Zusammenfassung: Berichtet wird die Konstruktion und Überprüfung einer deutschsprachigen Adaption des Fragebogens „Student Evaluation of Educational Quality“ (SEEQ) von Marsh (1982, 2007), ein umfassend überprüftes und international etabliertes Instrument zur Erfassung von Studierendenurteilen der Lehrqualität. Es wurde übersetzt, geringfügig erweitert und anhand einer Stichprobe von 76 687 Studierendenurteilen zu 3 660 Lehrveranstaltungen überprüft. Interne Konsistenzen und Intraklassenkorrelationen indizierten eine hohe Messgenauigkeit. Faktorenanalysen bestätigten die im SEEQ unterschiedenen Dimensionen. Neben den Produktfaktoren (Lernzuwachs, Gesamtbeurteilung) ließen sich als Faktoren des Lehrhandelns Engagement, Stoffstrukturierung und -präsentation, Aktivierung der Studierenden, Sozialklima, Stoffbreite, Leistungsbewertung und Aufgaben wie im Originalinstrument unterscheiden. Mit der Adaption wird studentische Beiträge als optionaler Faktor vorgeschlagen. Das Instrument erwies sich als messinvariant über verschiedene Veranstaltungsformen hinweg. Insgesamt legen die Ergebnisse nahe, dass mit der deutschsprachigen Adaption des SEEQ die Qualität hochschulischer Lehre international anschlussfähig und mit hoher Güte erfasst werden kann.

Schlüsselwörter: Lehrqualität, Lehrerfolg, Studierendenevaluationen, SEEQ

Construction and Confirmation of a German Adaption of the Student Evaluation of Educational Quality Questionnaire (SEEQ)

Abstract: The Student Evaluation of Educational Quality Questionnaire (SEEQ) by Marsh (1982, 2007) was adapted to German and tested using assessments from 76,687 students in 3,660 university courses. Internal consistencies and intraclass correlations indicted a high reliability. Two-level CFAs and ESEM analyses confirmed the separability of all original SEEQ dimensions: learning and overall as product factors, and enthusiasm, organization / clarity, group interaction, individual rapport, breadth, examinations / grading, and assignments / readings as factors on the level of instructional behaviors. In this adaption, we additionally proposed student contributions as an optional factor (the extent to which contributions of fellow students are considered helpful, and whether they are effectively controlled by the instructor), especially for contexts – such as those found in Germany – where student-directed teaching methods are prevalent. Additionally, we expanded the overall course rating by adding two items using a grade scale. These two adaptions are optional, and the scale worked equally well without them. We could confirm measurement invariance across different types of courses. Taken together, our findings indicate that the German adaption of the SEEQ measures teaching quality in accordance with established testing standards.

Keywords: educational evaluation, students’ evaluations of teaching (SET), teaching effectiveness, SEEQ

Die Evaluation der Qualität von Lehrveranstaltungen durch Studierende ist ein zentrales Element der Qualitätssicherung im Hochschulbereich und an den meisten Hochschulen nicht mehr wegzudenken, insbesondere da sie ein wichtiges Feedback zur Verbesserung der Hochschullehre bietet (Spinath et al., 2018; Spinath & Stehle, 2011). Lehrqualität ist multifaktoriell und umfasst sowohl Verhaltensweisen der Lehrenden als auch Ergebnisse auf Studierendenseite, wobei zu einem umfassenden Verständnis beide Merkmalsgruppen – also Prozess- und Produktmerkmale der Lehrqualität – berücksichtigt werden sollten (Abrami, d’Apollonia & Rosenfield, 2007; Marsh, 2007; Rindermann, 2009). Hierbei wird angenommen, dass bestimmte Dimensionen des Verhaltens der Lehrenden für die Prozessqualität der Lehre bedeutsam sind (Döring, 2005; Feldman, 1976). Auf der Produktebene wird argumentiert, dass gute Lehre zu einem Zuwachs von Wissen und Fertigkeiten und langfristig zu Berufserfolg und einer erfolgreichen Lebensbewältigung führt (Helmke, 1996). Entsprechend sollte sich qualitätsvolle Lehre in günstigen Bewertungen sowohl von Prozess- als auch Produktmerkmalen widerspiegeln, wobei diese Aspekte in einem positiven Zusammenhang stehen. Umfassend belegt ist, dass Studierende Bewertun-

gen von Prozess- und Produktmerkmalen der Lehrqualität valide vornehmen können (Stehle & Spinath, 2011).

Studentische Lehrveranstaltungsevaluationen weisen international eine lange Tradition auf. Seit den 1990er Jahren werden verstärkt auch im deutschsprachigen Raum multifaktorielle Fragebogeninstrumente zur Erfassung von Lehrqualität herangezogen (ESM 1). Diese Instrumente erfüllen oft hohe Standards der Messgüte und gelten als im Praxiseinsatz bewährt. Dazu zählen häufig auch standortspezifische Fragebögen, die auf Grundlage generellerer Fragebögen adaptiert und psychometrisch überprüft wurden (Sengewald, 2016). Nicht zuletzt, um die Qualität dieser standortspezifischen Instrumente zu untersuchen (z. B. bei Untersuchungen der konvergenten Validität) und die Anschlussfähigkeit der erhaltenen Befunde an einschlägige internationale Arbeiten zu wahren, sind auch deutschsprachige Versionen international weit verbreiteter Evaluationsbögen notwendig.

Ein international besonders häufig verwendetes Instrument zur Erfassung von Studierendenurteilen der Lehrqualität ist der Fragebogen „Student Evaluation of Educational Quality" (SEEQ; Marsh, 1982, 2007). Die Messgüte dieses Instruments gilt durch eine große Zahl publizierter Studien als umfassend belegt – auch über die Hauptgütekriterien hinausgehend, etwa in Bezug auf Ökonomie und Testfairness (Marsh, 2007). Der SEEQ ist damit ein international etabliertes und akzeptiertes Instrument in der Praxis der Evaluation von Lehrqualität im Hochschulbereich (Marsh & Roche, 1997). Daneben wird der SEEQ in einer Vielzahl internationaler Forschungsarbeiten eingesetzt, beispielsweise um Lehrqualität in Abhängigkeit von Merkmalen der Lehrenden (z. B. Motivation) oder des Kontexts (z. B. Institutionsformen) zu analysieren, Veränderungen der Lehrqualität zu erforschen oder die Wirkung von Interventionen zu untersuchen (Marsh, 2007). Somit ist dieses Instrument nicht nur für die Evaluationspraxis, sondern auch für internationale Forschung zur Lehrqualität bedeutsam.

Im SEEQ unterscheidet Marsh (2007) insgesamt neun Dimensionen. Als Indikatoren für den Lehrerfolg auf Produktebene sind (1) der *Lernzuwachs* der Studierenden (learning / value) und (2) die *Gesamtbeurteilung* (overall) der Studierenden enthalten. Als Faktoren des Lehrhandelns auf Prozessebene werden (3) *Engagement der Lehrperson* (instructor enthusiasm), (4) *Stoffstrukturierung und -präsentation* (organization / clarity), (5) *Aktivierung der Studierenden* (group interaction), (6) *Sozialklima* (individual rapport), (7) *Stoffbreite* (breadth of coverage), (8) *Leistungsbewertung* (examinations / grading) und (9) *Aufgaben* (assignments / readings) erfasst. Marsh und Dunkin (1992) evaluierten den Inhalt des SEEQ hinsichtlich der von Feldman (1976) und Fincher (1985) berichteten, übergreifenden Prinzipien des Lehrens und Lernens in tertiären Einrichtungen und fanden, dass die erfassten Dimensionen der Lehrqualität diese Prinzipien adäquat adressieren. Zwar gibt es in der Fachliteratur eine Diskussion über die Anzahl und Qualität der Dimensionen von Lehrqualität im Allgemeinen (Spooren, Brockx & Mortelmans, 2013), sodass auch andere Faktoren als jene des SEEQ herangezogenen werden könnten – in Hinblick auf den SEEQ hat eine Vielzahl an Studien jedoch mittels Faktorenanalysen belegt, dass die genannten neun Dimensionen des SEEQ von Studierenden voneinander separierbar wahrgenommen und beurteilt werden (Marsh & Hocevar, 1991; vgl. aber auch Toland & De Ayala, 2005) und über verschiedene sprachliche sowie kulturelle Kontexte hinweg gültig sind (Marsh, 2007; Watkins, 1994). Auch wenn Dozierende die Qualität ihrer eigenen Lehrveranstaltungen beurteilen, unterscheiden sie zwischen diesen Dimensionen (Marsh, Overall & Kesler, 1979). Eine multidimensionale und damit differenzierte Auffassung erscheint auch mit Blick auf die Rückmeldung spezifischer veränderbarer Aspekte der Lehre notwendig. Zahlreiche Studien erbrachten, dass jede der unterschiedenen Prozessdimensionen der Lehrqualität für die Effektivität und Verbesserung der Lehre relevant sein kann (Marsh, 2007).

Neben diesen Aspekten wird im SEEQ eine Reihe an Hintergrundvariablen erfasst. Darunter fallen der Aufwand / Anspruch der Veranstaltung (workload / difficulty) sowie Besuchsgrund, Vorleistungen, Erfolgserwartung, Semester und Geschlecht der Studierenden. Von diesen Aspekten wird angenommen, dass sie nicht unmittelbar für die Lehrqualität relevant sind, sich jedoch auf die Studierendeneinschätzungen der Lehrqualität auswirken können (Dresel & Rindermann, 2011). Zuletzt werden spezifische Stärken und Schwächen der evaluierten Veranstaltungen im SEEQ mit einer offenen Frage erfasst.

Der SEEQ wurde für unterschiedliche Länder wie Australien, China, Griechenland, Großbritannien, Indien, Iran, Malaysia, Nigeria, Spanien, Thailand und die Vereinigten Staaten sprachlich adaptiert und überprüft (ESM 1). Es gab bislang jedoch keine überprüfte deutschsprachige Adaption, die sich eng am Originalinstrument orientiert – möglicherweise, da für den deutschsprachigen Raum wie dargestellt andere etablierte Instrumente vorliegen, die ebenfalls auf einem multidimensionalen Verständnis von Lehrqualität fußen (Rindermann, 2009). Um in der Evaluationspraxis die solide theoretische und methodische Basis des SEEQ nutzen zu können und in der Hochschulforschung ein international direkt vergleichbares Instrument bereitzustellen, wird eine deutsche Adaption benötigt, die nah am Original konstruiert ist. Die vorliegende Arbeit berichtet die Konstruktion und Überprüfung einer solchen Adaption. Der Überprüfung liegt die Frage zugrunde, ob die mit dem Instrument erhobenen Werte valide interpretiert werden können. Für die intendierte

Nutzung als Evaluations- und Forschungsinstrument sind Objektivität und Reliabilität zentrale Voraussetzungen. Während für erstere eine standardisierte Durchführung und Datenerfassung entscheidend sind, gilt es hinsichtlich der Reliabilität insbesondere zwei Aspekte abzusichern (Marsh, 2007): die internen Konsistenzen der Lehrqualitätsdimensionen sowie die Übereinstimmung der Studierendenangaben bei der Beurteilung derselben Lehrveranstaltung. Erst durch ausreichend hohe Reliabilität gemäß dieser beiden Aspekte ist es gerechtfertigt, die einzelnen Items bzw. die einzelnen Studierendenangaben in Form von Mittelwerten zusammenzufassen, die in der Praxis wiederum als Evaluationsergebnis rückgemeldet werden oder als Grundlage für aufbauende Analysen in der Forschungspraxis dienen. Ohne ausreichend hohe Reliabilität wäre somit keine Nutzung mit dem Instrument gewonnener, aggregierter Werte sinnvoll. Für die valide Interpretierbarkeit der Messergebnisse im Rahmen der angestrebten Nutzung liefern darauf aufbauend folgende Punkte Hinweise:

1. substanzielle Unterschiede zwischen Lehrveranstaltungen, da Aspekte erfasst werden sollen, hinsichtlich derer sich Lehrveranstaltungen oder Lehrpersonen unterscheiden; sollten sich entsprechende Unterschiede nicht nachweisen lassen, wäre ein Einsatz des Instruments zum Vergleich von Lehrveranstaltungen oder Lehrpersonen nicht angezeigt
2. die interne Struktur der Qualitätsdimensionen, die sich wie theoretisch angenommen und wie im Originalinstrument voneinander abgrenzen lassen sollte; ohne entsprechende Abgrenzbarkeit wäre es nicht gerechtfertigt, von unterschiedlichen Teilaspekten der Lehrqualität zu sprechen
3. die Messinvarianz dieser Dimensionen über verschiedene Veranstaltungsformen (Vorlesungen, Seminare, Übungen), da die Items für diese gleichermaßen gut funktionieren sollten; ohne diesen Nachweis wäre es nicht sinnvoll, die Mittelwerte zwischen verschiedenen Veranstaltungsformen zu vergleichen
4. eine Replikation bekannter mit dem Originalinstrument gewonnener empirischer Befunde etwa in Form von Unterschieden in der Bewertung verschiedener Veranstaltungsformen (Vorlesungen werden durchschnittlich schlechter bewertet als Seminare; Marsh, 2007) sowie positiven Zusammenhängen zwischen Produkt- und Prozessfaktoren (vgl. Marsh, 2007); sollte ein Nachweis entsprechender Befunde ausbleiben, könnte das implizieren, dass der SEEQ-DE anders funktioniert als das Originalinstrument und somit nicht adäquat als deutschsprachige Adaption zu verwenden wäre

Ein Nachweis dieser Aspekte liefert begründete Hinweise für die valide Interpretation der Testergebnisse gemäß der intendieren Verwendung des Verfahrens als Evaluations- und Forschungsinstrument.

Bei der Überprüfung dieser Aspekte ist zu beachten, dass die Studierendenangaben, die unterschiedlichen Veranstaltungen und Dozierenden zuzuordnen sind, auf verschiedenen Ebenen analysiert werden können (Toland & De Ayala, 2005; Sengewald & Vetterlein, 2015). Im Anwendungskontext ist häufig die Veranstaltungsebene naheliegend, da sich Aussagen zur Lehrqualität unmittelbar auf Lehrveranstaltungen bzw. das Handeln der Lehrpersonen darin beziehen (Marsh, 2007). Gleichwohl ist es ebenfalls wichtig, die Ebene der einzelnen Studierendenangaben zu berücksichtigen. Unabhängig von ihrer Gruppierung in Veranstaltungen, sollten sich Studierendenurteile der Lehrqualität in die unterschiedlichen Faktoren der Lehrqualität differenzieren lassen und die entsprechenden Items sollten diese reliabel abbilden (Marsh, Hau, Chung & Siu, 1998). Reliabilitätsnachweise auf Ebene der Studierenden sind zudem gerade für Forschungszwecke (in denen häufig mehrebenenanalytisch Studierendenurteile als abhängige Variablen betrachtet werden; z. B. Daumiller et al., 2021) eine wichtige Voraussetzung. Für die Überprüfung von faktorieller Struktur und Messinvarianz kann eine simultane Berücksichtigung beider Ebenen mit konfirmatorischen Zwei-Ebenen-Faktorenanalysen (Zwei-Ebenen-CFAs) als besonders geeignet erachtet werden, da diese die hierarchische Datenstruktur berücksichtigen und zudem einen hypothesenprüfenden Ansatz ermöglichen (Sengewald & Vetterlein, 2015). Nachweise der faktoriellen Struktur des SEEQ über diesen mehrebenenanalytischen Zugang stehen trotz einer Vielzahl an Analysen auf einer Ebene noch aus (Toland & De Ayala, 2005). Die jüngste Strukturprüfung führte Marsh et al. (2009) mit Hilfe von Exploratory Structural Equation Modeling (ESEM) Analysen auf Veranstaltungsebene durch (d. h. mit aggregierten Werten). Diese Methode ebenfalls anzuwenden, ermöglicht uns zu prüfen, ob die deutsche Adaption ähnliche Ergebnisse liefert wie von Marsh et al. (2009) berichtet. Zusätzliche Zwei-Ebenen-CFA erlauben darüber hinaus den Forschungsstand zum SEEQ zu erweitern und erhärtende Evidenz bezüglich der faktoriellen Struktur zu liefern.

Adaption des SEEQ

Die Original-Items des SEEQ wurden unter Anwendung einer konsensualen Strategie ins Deutsche übersetzt. In einem mehrstufigen Prozess übersetzten drei der Autoren die Items zunächst unabhängig voneinander und führten die Übersetzungen anschließend zusammen. Zur Absicherung inhaltlicher Nuancen übersetzte sie ein englischsprachiger Muttersprachler zurück ins Englische. Abwei-

chungen zwischen Originalitems und rückübersetzten Items wurden schließlich für nochmalige Anpassungen genutzt. Etwaige Unwägbarkeiten der sprachlichen und kulturellen Anpassung berücksichtigend wurden in der Konstruktionsversion des deutschsprachigen Fragebogens einige Faktoren um zusätzliche Items verlängert und in einer vorangehenden Pilotstudie mit Studierendenangaben zu 777 Lehrveranstaltungen getestet. Aufbauend darauf konnten einige der zusätzlichen Items für die endgültige Version der deutschsprachigen Adaption (SEEQ-DE) wieder getilgt werden. Die finalen Items sind in ESM 2 aufgeführt.

Die Subskalen *Lernzuwachs* sowie *Stoffstrukturierung und -präsentation* wurden um jeweils ein Item erweitert. Die Erweiterung der Subskala zum Lernzuwachs liegt darin begründet, dass ein Originalitem zwei inhaltliche Aspekte umfasste (Verständnis und Lernen), die sich im Deutschen nicht adäquat in einem einzigen Item abbilden ließen – sie wurden daher auf zwei Items aufgeteilt (Item 3 und 4). *Stoffstrukturierung und -präsentation* wurde um das Item „Die Inhalte werden gut strukturiert" erweitert, da durch die deutschsprachige Übersetzung sonst nicht hinreichend sichergestellt wäre, dass das gesamte inhaltliche Spektrum dieses Faktors abgebildet ist.

Das Instrument wurde zudem um zwei inhaltliche Aspekte ergänzt: Zum einen wurde ein weiterer Faktor hinzugenommen, *Studentische Beiträge*, der das Ausmaß misst, zu dem Beiträge von Mitstudierenden ertragreich für das eigene Lernen sind und effektiv durch die Lehrperson gesteuert (d. h. ausgewählt, moderiert und ergänzt) werden. Dies erschien sinnvoll, um einen wesentlichen – aber im SEEQ bislang nicht fokussierten – Aspekt von Lehrformen einzubeziehen, bei denen die Erarbeitung von Inhalten vermehrt interaktiv oder durch Studierendenbeiträge erfolgt, was insbesondere im deutschsprachigen Raum verbreitet ist. Dies mag in Seminaren mit umfangreichen Studierenden-Inputs beispielsweise sehr relevant sein, kann aber auch in interaktiveren Vorlesungen, in denen etwa Studierendenbeiträge aufgegriffen oder kleinere Diskussionen geführt werden, einen informativen Evaluationsaspekt darstellen. Zum anderen betrifft dies eine Erweiterung der *Gesamtbeurteilung* von Lehrveranstaltung und Lehrperson durch zwei Notenurteile. Obwohl davon ausgegangen wird, dass die Gesamtbeurteilung hinreichend messgenau durch die beiden Originalitems erfasst werden kann, schlagen wir die Erweiterung durch Notenurteile vor, da Notenurteile in der Evaluationspraxis von allen Akteursgruppen vielfach als besonders anschaulich eingeschätzt werden (Rindermann, 2009).

Während die zusätzlichen Items zum Lernzuwachs und zur Stoffstrukturierung und -präsentation obligatorisch sind, sind die gegenüber dem SEEQ neuen inhaltlichen Aspekte der Qualität studentischer Beiträge und der Notenurteile zur Gesamtbeurteilung optionale Fragebogenteile. Sie können je nach Evaluationskontext genutzt werden.

Insgesamt umfasst die deutschsprachige Adaption des SEEQ somit zehn Subskalen zur Lehrqualität (vgl. Tabelle 1 für eine Übersicht sowie ESM 2 für alle Items und Antwortformate). Lernzuwachs, Engagement der Lehrperson, Stoffstrukturierung und -präsentation, Aktivierung der Studierenden, Sozialklima, Stoffbreite, Leistungsbewertung und Aufgaben werden mit 2–5 Items auf einer Skala von 1 (*stimme gar nicht zu*) bis 5 (*stimme voll zu*) erfasst. Zusätzlich dazu gibt es die Antwortmöglichkeit *nicht beurteilbar*, die später als fehlender Wert behandelt wird. Dies erlaubt eine einfache Darstellung in der Rückmeldepraxis (z. B. in Form von Mittelwerten) und eine Schätzung des entsprechenden latenten Merkmals in den statistischen Analysen (Holman & Glas, 2005). Die Gesamtbeurteilung wurde mit vier Items auf einer Zensurskala von 1 (*mangelhaft*) bis 5 (*sehr gut*) erfasst. Als Hintergrundvariablen werden wie im Originalinstrument Aufwand bzw. Anspruch der Veranstaltung, Erfolgserwartung, Interesse vor der Veranstaltung, Arbeitsaufwand in Stunden, Vorleistungen, Besuchsgrund, Semester und Geschlecht der Studierenden erfasst. Zudem wird am Ende des Fragebogens eine offene Frage zu spezifischen Stärken und Schwächen der Veranstaltung gestellt.

Ziel der vorliegenden Untersuchung

Das Ziel der vorliegenden Arbeit ist, die Messgüte der deutschsprachigen Adaption des SEEQ zu untersuchen. Die Analysen prüfen die zuvor dargestellten Aspekte der Reliabilität und Validität, um damit die intendierte Verwendung des Verfahrens als Evaluations- und Forschungsinstrument abzusichern. Im Speziellen wurden zunächst grundlegende Item- und Skaleneigenschaften analysiert. Im Anschluss wurden interne Konsistenzen und die Übereinstimmung der studentischen Urteile in Lehrveranstaltungen bestimmt. Zudem wurden Unterschiede zwischen Lehrveranstaltungen betrachtet sowie die Faktorenstruktur und die Messinvarianz mit ESEM und konfirmatorischen Zwei-Ebenen-Analysen überprüft. Schließlich wurden Unterschiede zwischen verschiedenen Veranstaltungsformen sowie Zusammenhänge zwischen Produkt- und Prozessmerkmalen der Lehrqualität analysiert.

Tabelle 1. Deskriptive Statistiken

		Studierendenebene (*N* = 76687)										Veranstaltungsebene (*N* = 3660)							
	Items[a]	*M*	*SD*	Range	Schiefe	Kurtosis	r_{it}	ω_h	α	ICC1	ICC2	*M*	*SD*	Range	Schiefe	Kurtosis	r_{it}	ω_h	α
Lehrqualität																			
Lernzuwachs	4/5	3.9	0.8	1.0 – 5.0	−0.6	0.2	.47 – .71	.78	.84	.40	.90	4.0	0.4	2.2 – 5.0	−0.3	−0.1	.54 – .86	.88	.91
Gesamtbeurteilung	2/4	4.0	0.8	1.0 – 5.0	−0.9	0.7	.80 – .83	.86	.92	.47	.92	4.2	0.5	2.1 – 5.0	−0.8	0.7	.92 – .93	.93	.97
Engagement der Lehrperson	4/4	3.9	0.9	1.0 – 5.0	−0.9	0.5	.71 – .71	.85	.88	.50	.93	4.1	0.5	1.7 – 5.0	−0.7	0.4	.82 – .91	.92	.93
Stoffstrukturierung und -präsentation	4/5	4.1	0.7	1.0 – 5.0	−0.9	1.2	.52 – .70	.77	.82	.24	.84	4.2	0.3	2.5 – 5.0	−0.7	1.2	.53 – .84	.84	.87
Aktivierung der Studierenden	4/4	4.2	0.8	1.0 – 5.0	−1.1	1.2	.77 – .91	.87	.92	.34	.91	4.4	0.4	2.0 – 5.0	−1.0	1.3	.84 – .93	.89	.95
Sozialklima	4/4	4.3	0.7	1.0 – 5.0	−1.2	1.8	.59 – .81	.85	.87	.22	.85	4.4	0.5	2.6 – 5.0	−0.9	1.4	.68 – .90	.89	.90
Stoffbreite	4/4	3.8	0.8	1.0 – 5.0	−0.6	0.3	.63 – .73	.83	.85	.26	.86	4.0	0.4	2.1 – 5.0	−0.4	0.2	.67 – .82	.89	.88
Leistungsbewertung	3/3	4.1	0.8	1.0 – 5.0	−1.0	1.1	.65 – .74	.85	.85	.19	.79	4.2	0.4	2.3 – 5.0	−0.6	0.7	.63 – .76	.84	.84
Aufgaben	2/2	4.0	0.9	1.0 – 5.0	−0.9	0.8	.71 – .71	.83	.83	.15	.70	4.1	0.4	2.0 – 5.0	−0.5	1.0	.67 – .67	.80	.80
Studentische Beiträge	–/4	3.9	0.8	1.0 – 5.0	−1.0	1.4	.57 – .68	.76	.81	.30	.85	4.0	0.4	2.0 – 5.0	−0.7	0.7	.71 – .81	.82	.82
Hintergrundvariablen[b]																			
Aufwand/Anspruch der Veranstaltung	3/3	3.5	0.7	1.0 – 5.0	−0.2	0.1	.57 – .67	.79	.78	.45	.90	3.5	0.4	1.9 – 5.0	−0.1	−0.1	.73 – .82	.88	.86
Erfolgserwartung der Studierenden	1/1	3.6	0.8	1.0 – 5.0	−0.4	0.3				.19		3.7	0.4	2.3 – 5.0	−0.1	0.5			
Interesse vor Veranstaltung	1/1	3.4	1.1	1.0 – 5.0	−0.4	−0.4				.20		3.5	0.5	1.5 – 5.0	−0.3	0.1			
Arbeitsaufwand in Stunden	1/1	3.2	1.3	1.0 – 6.0	0.5	−0.5				.42		3.3	1.0	1.0 – 6.0	0.8	0.3			
Vorleistungen der Studierenden	1/1	2.3	0.6	1.0 – 5.0	0.4	0.7				.16		2.2	0.3	1.0 – 4.0	0.2	0.5			
Semester der Studierenden	1/1	2.1	1.1	1.0 – 6.0	0.9	0.5				.59		2.2	0.9	1.0 – 5.6	0.4	−0.5			
Geschlecht der Studierenden[c]	1/1	0.6		0.0 – 1.0						.17		0.7	0.2	0.0 – 1.0	−0.4	−0.6			

Anmerkungen: [a] Itemanzahl im Original SEEQ / in der deutschen Adaption. Reliabilitäten bezogen auf alle Items. [b] Der „Besuchsgrund der Studierenden" (polytomes Item auf Nominalskalenniveau): 40 % Pflicht, 27 % Wahlpflicht, 1 % Termin, 17 % Interesse, 14 % Prüfungsvorbereitung. [c] 0 = männlich, 1 = weiblich.

Methode

Stichprobe

Zur Überprüfung des SEEQ-DE wurde eine Stichprobe von 3660 Lehrveranstaltungen genutzt. Diese fanden zwischen Sommersemester 2013 und Wintersemester 2017/18 an einer mittelgroßen Universität im süddeutschen Raum im gesamten Fachspektrum statt (davon 2 497 Seminare, 797 Übungen, 105 Vorlesungen und 261 sonstige Veranstaltungen). Insgesamt umfasst die Stichprobe 76687 Studierendenurteile (M = 21.0 Urteile pro Veranstaltung, SD = 22.5; Seminare: M = 16.6, SD = 9.9; Übungen: M = 31.0, SD = 33.5; Vorlesungen: M = 51.9, SD = 59.2; sonstige Veranstaltungen: M = 18.9, SD = 21.5). Die beurteilenden Studierenden waren zu 64% weiblich und durchschnittlich in der Mitte des 3. Semesters (SD = 2.1).

Prozedur und Einsatz des SEEQ-DE

In der vorliegenden Stichprobe wurde die Lehrveranstaltungsevaluation durchgeführt, indem zunächst Professor_innen auf freiwilliger Basis und in Abstimmung mit ihren Mitarbeiter_innen für den von ihnen verantworteten Bereich die zu evaluierenden Veranstaltungen auswählten. Die Lehrpersonen dieser Veranstaltungen erhielten anschließend eine schriftliche Durchführungsanweisung sowie die Evaluationsbögen (Muster: ESM 3 und ESM 4). Die ausgefüllten Bögen wurden durch Studierende eingesammelt und in einem versiegelten Umschlag an die Qualitätssicherungsstelle der Universität weitergegeben, wo die Daten scannergestützt erfasst wurden. Ergänzt durch Interpretationshilfen (Definition der Dimensionen, Vergleichswerte) wurden die Ergebnisse veranstaltungsspezifisch an die einzelnen Dozierenden rückgemeldet.

Die deutschsprachige Adaption des SEEQ wurde in der oben beschriebenen Form eingesetzt, einschließlich der optionalen Items. Die Professor_innen entschieden je nach Konzeption der Vorlesungen in ihrem Bereich, ob darin die zusätzliche Skala *studentische Beiträge* mit eingesetzt wurde oder nicht (in 53% der Vorlesungen wurde diese eingesetzt). Der Bogen wurde – wie in der Durchführungsanweisung empfohlen – überwiegend zu Beginn einer Sitzung etwa in der Mitte des Semesters eingesetzt. Die dafür erforderliche Zeit betrug etwa 5–10 Minuten.

Ergebnisse

Grundlegende Skaleneigenschaften

Alle Subskalen der Lehrqualität nahmen auf Studierendenebene das gesamte Spektrum möglicher Werte an (vgl. Tabelle 1). Zudem waren bei allen Subskalen sowohl auf Studierenden- als auch auf Veranstaltungsebene nennenswerte Varianzen bzw. Standardabweichungen beobachtbar, was auf substanzielle Unterschiede zwischen Urteilenden und Veranstaltungen schließen lässt. Die Unterschiede zwischen den Studierenden waren dabei deskriptiv etwas größer als die Unterschiede auf Veranstaltungsebene (was auch zu erwarten ist, da die Varianz auf Studierendenebene neben Veranstaltungsunterschieden auch Nicht-Übereinstimmungen zwischen Studierenden derselben Veranstaltungen reflektiert). Obwohl durchgehend Mittelwerte oberhalb der theoretischen Skalenmitte von 3.0 beobachtet wurden, indizierten die Werteverteilungen sowie Schiefe und Kurtosis (die stets innerhalb der gebräuchlichen Grenzen von +/– 2 lagen; Gravetter & Wallnau, 2014) keine Verletzung der Normalverteilung auf beiden Ebenen, was eine wichtige Voraussetzung für die nachfolgenden Analysen ist.

Interne Konsistenzen

Zur Bestimmung der internen Konsistenzen wurde McDonald's Omega (ω_h) berechnet (Dunn, Baguley & Brunsden, 2014). Zur Vergleichbarkeit mit älteren Studien zum SEEQ berichten wir zudem Cronbach's α. Alle internen Konsistenzen der Subskalen lagen in einem guten bis sehr guten Bereich (vgl. Tabelle 1). Dies ist eine entscheidende Bedingung für eine messgenaue Einschätzung einzelner Lehrveranstaltungen. Auch die Subskala zur Gesamtbeurteilung ohne die zwei zusätzlichen Items lieferte gute Kennwerte (ω_h = .82 auf Studierenden- und ω_h = .92 auf Veranstaltungsebene).

Urteilerübereinstimmung

Für eine reliable Erfassung von Veranstaltungsmerkmalen mit Hilfe studentischer Urteile ist es ferner nötig, dass die Studierenden innerhalb von Lehrveranstaltungen in hinreichendem Maße in ihren Einschätzungen übereinstimmen. Hierbei ist zu berücksichtigen, dass kein einheitlicher Effekt des Lehrhandelns von Dozierenden auf die einzelnen Studierenden erwartet werden kann (z. B. differenzielle Lerneffekte aufgrund heterogener Voraussetzungen), sodass perfekte Urteilerübereinstimmung

nicht realistisch ist. Die Übereinstimmung wird durch die Intraklassenkorrelation ICC2 quantifiziert, die wie herkömmliche Reliabilitätskoeffizienten interpretiert werden kann (Lüdtke, Robitzsch, Trautwein & Kunter, 2009). Werte nahe 1 geben an, dass Studierende in denselben Veranstaltungen sehr ähnliche Einschätzungen abgeben. Die ICC2 Werte (vgl. Tabelle 1) lagen bei allen Subskalen in einem zufriedenstellenden bis sehr guten Bereich.

Darauf aufbauend kann bestimmt werden, wie viele Studierendenurteile nötig sind, um auch für jede Einzelveranstaltung eine hinreichende Reliabilität der Einschätzungen zu erhalten (Lüdtke et al., 2009). Wird dafür ICC2 ≥ .60 angesetzt, erhält man für die Subskala mit dem geringsten Varianzanteil auf Veranstaltungsebene (Aufgaben, ICC1 = .15), dass dafür mindestens 8.5 Urteile vorliegen sollten. Daraus folgt, dass eine Lehrveranstaltung von mindestens 9 Studierenden beurteilt werden sollte, um bei allen Subskalen eine hinreichende Zuverlässigkeit der auf die einzelne Lehrveranstaltung bezogenen Ergebnisse zu gewährleisten.

Unterschiede zwischen Lehrveranstaltungen

Eine basale Voraussetzung für eine valide Interpretation der erfassten Aspekte der Lehrqualität ist, dass sich Lehrveranstaltungen bzw. Lehrpersonen bezüglich dieser Aspekte unterscheiden. Dementsprechend sollte eine hinreichende Varianz zwischen Veranstaltungen vorliegen. Ohne eine solche wäre eine Messung aus Gründen der Ökonomie sowie aus der damit implizierten nicht gegebenen Veränderbarkeit auch wenig sinnvoll.

Der Anteil der Zwischenveranstaltungsvarianz an der Gesamtvarianz wird durch die Intraklassenkorrelation ICC1 quantifiziert, wobei Werte von ICC1 = .05 als kleine und Werte von ICC1 = .20 als große Unterschiede zwischen Kursen interpretiert werden können (Snijders & Bosker, 2020). Bei allen Subskalen des SEEQ-DE waren moderate bis sehr große Unterschiede zwischen Lehrveranstaltungen zu verzeichnen, vor allem beim Lernzuwachs, bei der Gesamtbeurteilung und beim Engagement der Lehrperson (vgl. Tabelle 1).

Faktorielle Struktur

Zur Untersuchung der internen Struktur verwendeten wir zunächst ESEM-Analysen mit Geomin Rotation auf Veranstaltungsebene analog zu Marsh et al. (2009), um zu prüfen, ob der SEEQ-DE ähnliche Ergebnisse liefert. Zudem führten wir konfirmatorische Zwei-Ebenen-Analysen durch, die im vorliegenden Fall besonders angemessenen sind (Sengewald & Vetterlein, 2015).[1] Wie bei Marsh et al. (2009) ließen wir Ladungen der beiden Items zur Gesamtbewertung der Lehrveranstaltung auf die Dimension *Lernzuwachs* sowie Ladungen der beiden Items zur Gesamtbewertung der Lehrperson auf die Dimension *Engagement der Lehrperson* zu. Die Analysen wurden mit Mplus 8.1 (Muthén & Muthén, 2017) durchgeführt. Fehlende Werte wurden modellbasiert mittels FIML-Schätzer und EM-Algorithmus geschätzt (Peugh & Enders, 2004). Zur Evaluation der Modellfits wurden den Vorschlägen von Kline (2015) folgend χ^2, CFI, TLI, RMSEA und SRMR als Fit-Indizes herangezogen. Hierbei ist zu berücksichtigen, dass gerade bei umfangreichen und komplexen Messmodellen (d. h. mit vielen Items und Faktoren) eher liberale Cut-Off-Werte verwendet werden sollten, da es hierbei unwahrscheinlich ist, strenge Kriterien zu erreichen (Cheung & Rensvold, 2002; Fan & Sivo, 2007). Als Indikatoren für einen akzeptablen Modellfit verwendeten wir daher CFI ≥ .90 und TLI ≥ .90 (Byrne, 2013) sowie RMSEA ≤ .08 und SRMR ≤ .10 (Schermelleh-Engel, Moosbrugger & Müller, 2003). Zudem ist zu berücksichtigen, dass gegeben der Neuheit der ESEM-Analysen zu diesem Zeitpunkt keine abweichenden Empfehlungen zur Beurteilung der Güte der damit geschätzten Modelle vorliegen.

Die ESEM-Ergebnisse (ESM 5) waren sehr ähnlich zu den von Marsh et al. (2009) für den Original-SEEQ berichteten und indizierten einen zufriedenstellenden Fit zu den Daten. Die Faktorladungen (vgl. ESM 6) bestätigten die 10-faktorielle Struktur und waren vergleichbar mit den Ladungsmustern bei Marsh et al. (2009) sowie den für explorative Faktorenanalysen typischerweise berichteten Ergebnissen (Marsh, 2007). Bemerkenswert ist, dass die höheren Querladungen der Items zur Gesamtbeurteilung genau mit den Ergebnissen von Marsh et al. (2009) korrespondieren (auf die Lehrveranstaltung bezogene Items laden auch auf den Faktor *Lernzuwachs*; auf die Lehrperson bezogene Items laden indes stärker auf den Faktor *Engagement*).

Die Zwei-Ebenen-CFA (ESM 5) erbrachte ebenfalls einen akzeptablen Modellfit. Angesichts der Modellkomplexität und der teils substanziellen Querladungen in den ESEM-Analysen, die für diese Analyse auf null gesetzt wurden, ist es nicht verwunderlich, dass die erreichten Fit-Indizes nicht besser ausfielen.

Bei beiden Analysen zeigte sich, dass die Zusatzitems zur Gesamtbeurteilung auf dem gleichen Faktor wie die

[1] Zusätzlich führten wir konfirmatorische Faktorenanalysen auch ausschließlich auf Studierendenebene durch. Die Resultate waren sehr ähnlich zu den hier berichteten Ergebnissen.

beiden etablierten Items luden und dass die optionale Dimension *studentische Beiträge* auch empirisch einen separaten Faktor darstellte.[2] Da die Subskala *studentische Beiträge* nur in etwa der Hälfte der evaluierten Vorlesungen eingesetzt wurde und um die Vergleichbarkeit zu den Analysen des Originalinstruments zu erhöhen, führten wir zusätzlich analoge Analysen ohne diese Subskala durch. Die Ergebnisse unterschieden sich nicht substanziell von den hier berichteten Resultaten und werden im Detail in ESM 7 und ESM 8 berichtet.

Zusammenfassend sprechen die faktorenanalytischen Befunde dafür, dass mit dem adaptierten Instrument die multifaktorielle Lehrqualität gemäß der international etablierten und für die deutsche Adaption postulierten Struktur erfasst werden kann.

Messinvarianz für unterschiedliche Veranstaltungsformen

Darauf aufbauend ist es ferner wichtig, dass diese Faktorenstruktur über verschiedene Veranstaltungsformen hinweg besteht und die Items der einzelnen Faktoren dasselbe messen (Fondel, Lischetzke, Weis & Gollwitzer, 2015). Deshalb betrachten wir die Messinvarianz des Instruments über Übungen, Seminare und Vorlesungen hinweg. Diese Analysen führten wir basierend auf den Zwei-Ebenen-CFAs durch, da in diesen die genestete Datenstruktur am passendsten berücksichtigt wird.[3] Dazu wurden mehrere zunehmend restriktive Multi-Gruppen-Modelle geschätzt (Muthén & Muthén, 2017). Zunächst wurde eine äquivalente Struktur der Faktorladungsmatrix über die drei Gruppen hinweg modelliert (*konfigurale Messinvarianz*). Anschließend wurden zusätzlich die Faktorladungen zwischen den Gruppen gleichgesetzt (*metrische Messinvarianz*). Schließlich wurden noch die Intercepts der Indikatoren zwischen den Gruppen restringiert (*skalare Messinvarianz*). Unterschiede im Modellfit evaluierten wir durch Betrachtung der Unterschiede in den CFI- und RMSEA-Werten (unter Verwendung der empfohlenen Cut-Off Werte von: ΔCFI = .01, ΔRMSEA = .015, ΔSRMR = .03 für metrische und ΔSRMR = .01 für skalare Messinvarianz; vgl. Chen, 2007). Angesichts der großen Stichprobe sind diese Werte χ^2-Tests vorzuziehen und dabei wiederum vor allem die ΔCFI Werte relevant, da ΔRMSEA und ΔSRMR negativ mit der Güte des Gesamtmodellfits assoziiert sind, d. h. bei schlechterem Gesamtmodellfit automatisch größer werden, sodass weniger restriktive Cut-Off-Werte angelegt werden müssten (Cheung & Rensvold, 2002).

Die Ergebnisse (vgl. ESM 5) zeigten, dass die restriktiveren Modelle die Daten nur unwesentlich schlechter beschrieben als die vorangehenden Modelle (insbesondere hinsichtlich CFI, aber auch hinsichtlich RMSEA und SRMR). Dies indiziert skalare Messinvarianz für die verschiedenen Veranstaltungsformen und erlaubt es somit, die manifesten und latenten Mittelwerte der Lehrqualitätsdimensionen zwischen Veranstaltungen verschiedenen Typus zu vergleichen.

Unterschiede in der Beurteilung verschiedener Veranstaltungsformen

Basierend darauf wurde überprüft, ob das Instrument in der Lage ist, die gut dokumentierten Niveauunterschiede bei der Beurteilung von unterschiedlichen Formen von Veranstaltungen abzubilden, insbesondere jene zwischen Vorlesungen und Seminaren (Rindermann, 2009). Dazu führten wir auf Veranstaltungsebene eine MANOVA mit nachfolgenden Bonferroni-Post-Hoc Vergleichen durch.[4] Darin betrachteten wir die Lehrqualitätsdimensionen als abhängige Variablen und die Veranstaltungsform als dreistufigen Faktor. Wir fanden große, statistisch signifikante Unterschiede zwischen den drei Veranstaltungsformen (Wilks λ = .684; *multivariates* $F(20, 6706) = 70.04$; $p < .001$; $\eta^2 = .17$), die in ESM 8 dargestellt sind. Vorlesungen wurden hinsichtlich aller Faktoren durchschnittlich signifikant schlechter eingestuft als Seminare. Zusätzliche Validitätshinweise lieferte das Profil der Mittelwertdifferenzen, das erwartungsgemäß bei der Aktivierung von Studierenden und studentischen Beiträgen (eher schwer in Vorlesungen umfassend zu realisierende Aspekte) besonders große und beim Engagement der Lehrperson, der Stoffstrukturierung und -präsentation sowie der Aufgaben (veranstaltungsformunabhängig relevante Aspekte) eher kleinere Unterschiede aufwies. Ferner wurden Übungen im Durchschnitt schlechter bewertet als Seminare, was möglicherweise auf die geringere

[2] Interessanterweise fielen die Ladungen der Items 1 (passende Themen) und 4 (sinnvolle Ergänzungen), die stärker das Verhalten der Lehrperson als jenes der Kommiliton_innen beurteilen, niedriger aus als die anderen beiden Items. Eine zusätzlich durchgeführte explorative Faktorenanalyse lieferte jedoch keine Hinweise darauf, dass die 4 Items mehr als einen Faktor bilden.

[3] Zur Absicherung führten wir die Analysen ebenfalls auf Basis des ESEM-Modells durch. Diese Analysen erbrachten zur Messinvarianz gleiche Resultate wie die Analysen auf Basis der Zwei-Ebenen-CFAs.

[4] Zur Absicherung dieser Resultate setzten wir zudem im Modell der skalaren Messinvarianz die Mittelwerte der zehn Lehrqualitätsdimensionen gleich und fanden im Einklang zu den manifesten Analysen eine statistisch signifikante Verschlechterung des Modellfits.

didaktische Erfahrung der Tutor_innen, die in der Regel selbst Studierende sind, zurückgeführt werden kann.

Beziehungen der Prozessmerkmale des Lehrhandelns mit Lernzuwachs und Gesamtbeurteilung

Zur weiteren Prüfung bekannter empirische Befunde zum Originalinstrument wurden auf Grundlage des 10-faktoriellen Modells Korrelationen zwischen den beiden Produktfaktoren des Lehrerfolgs (*Lernzuwachs*, *Gesamtbeurteilung*) mit den acht Prozessfaktoren des Lehrhandelns bestimmt. Dem lag die Annahme zugrunde, dass ein gültiges Instrument zur Beurteilung der Lehrqualität Lehrdimensionen erfassen muss, die für den Lernzuwachs der Studierenden und deren Gesamtbeurteilung der Lehrveranstaltungen von Relevanz sind. Im Einklang mit Befunden zum Originalinstrument (vgl. Marsh, 2007) erwarteten wir positive Zusammenhänge zwischen den Prozess- und den Produktmerkmalen der Lehrqualität. Dies war durchgehend der Fall: Für alle acht im SEEQ-DE einbezogenen Prozessfaktoren des Lehrhandelns waren auf Veranstaltungsebene substanzielle Zusammenhänge mit beiden Lehrerfolgsindikatoren zu beobachten, die indizierten, dass eine stärkere Ausprägung des jeweiligen Lehrhandlungsfaktors mit Bewertungen eines insgesamt größeren Lehrerfolgs einherging (Korrelationen auf Grundlage des ESEM Modells für Engagement: $\rho = .37$ mit dem Lernzuwachs / $\rho = .58$ mit der Gesamtbeurteilung; Stoffstrukturierung und -präsentation: $\rho = .32 / .40$; Aktivierung der Studierenden: $\rho = .25 / .33$; Sozialklima: $\rho = .23 / .48$; Stoffbreite: $\rho = .32 / .33$; Leistungsbewertung: $\rho = .32 / .05$; Aufgaben: $\rho = .38 / .35$; studentische Beiträge: $\rho = .41 / .18$). Die Höhe dieser Korrelationen war deskriptiv sehr ähnlich zu den bei Marsh et al. (2009) berichteten (alle $|\Delta\rho| \leq .13$).[5] Wurde die Gesamtbeurteilung ohne die beiden optionalen Items berechnet, fanden sich nahezu identische Korrelationen. Regressionsanalysen verwiesen auf einen inkrementellen Erklärungswert durch den zusätzlichen Faktor *studentische Beiträge*, der auch unter Kontrolle der anderen Prozessfaktoren einen statistisch signifikanten Effekt auf den Lernzuwachs ($\beta = .27$, $p < .001$) sowie die Gesamtbeurteilung ($\beta = .05$, $p = .001$) hatte.

Weiterführende Analysen erbrachten, dass diese Zusammenhänge in allen Veranstaltungsformen vorlagen – für die Korrelationen zeigten sich zwischen den drei Veranstaltungsformen keine nennenswerten Unterschiede. Dies ist ein Hinweis darauf, dass die einbezogenen Prozessfaktoren in Übereinstimmung mit der Forschungsliteratur (Marsh, 2007) von globaler Relevanz für alle Veranstaltungsformen und nicht nur für bestimmte Formen von Lehrveranstaltungen bedeutsam sind.

Diskussion

Ziel der vorliegenden Untersuchung war die Konstruktion und Überprüfung einer deutschsprachigen Adaption des Fragebogens „Student Evaluation of Educational Quality" von Marsh (1982, 2007). Dazu wurde das Inventar im Rahmen einer separaten Pilotierungsstudie übersetzt und geringfügig erweitert sowie anhand eines umfangreichen Datensatzes umfassend untersucht. Da der SEEQ zwar für viele Sprachen, aber noch nicht für den deutschsprachigen Raum, adaptiert wurde, wird mit der vorliegenden Arbeit eine Lücke geschlossen und ein Instrument zur Verfügung gestellt, das auf äußerst umfangreichen Validierungsbelegen fußt, die auf internationaler Ebene gewonnenen wurden. Es ergänzt die im deutschsprachigen Raum weit verbreiteten multifaktoriellen Instrumente zur Erfassung der Lehrqualität, die oft hohe Standards der Messgüte erfüllen und im Praxiseinsatz als bewährt gelten. Der SEEQ-DE ist neben der Evaluation einzelner Lehrveranstaltungen insbesondere auch auf die Hochschulforschung ausgelegt, um die Anschlussfähigkeit an die internationale Forschungsliteratur, insbesondere jener zum SEEQ, zu gewährleisten. Um begründete Schlussfolgerungen zur Validität der Interpretation der Testergebnisse gemäß der intendieren Verwendung als Evaluations- und Forschungsinstrument zu erlangen, wurden zentrale Aspekte der Messgüte aufgegriffen und empirisch geprüft (interne Konsistenz, Urteilerübereinstimmung, Unterschiede zwischen Lehrveranstaltungen, interne Struktur der Lehrqualitätsdimensionen und deren Messinvarianz, Replikation bekannter Befunde). Die Stärken der berichteten Analysen liegen in der umfassenden Datengrundlage sowie den Berechnungen unter Berücksichtigung der Studierenden- und Veranstaltungsebene. Zusammenfassend erbrachten die durchgeführten Analysen klare Hinweise auf eine zulässige Nutzung des Messinstruments als Evaluations- und Forschungsinstrument.

In Bezug auf die Objektivität kann zunächst festgehalten werden, dass Durchführungs- und Auswertungsobjektivität aufgrund der standardisierten Testdurchführung (schriftliche Instruktion, exakte Durchführungsvor-

[5] Da in den ESEM Analysen bei Marsh et al. (2009) keine Ergebnisse zur Gesamtbeurteilung berichtet wurden, wurden nur die Korrelationen mit dem Lernzuwachs verglichen.

gaben) und Datenerfassung (im Workflow der universitären Qualitätssicherung typischerweise maschinell) als sehr hoch angenommen werden können. Die Interpretationsobjektivität kann durch Vergleichswerte aus anderen Veranstaltungen unterstützt werden (wobei Normwerte auf Basis einer repräsentativen Veranstaltungsstichprobe bislang allerdings nicht zur Verfügung stehen). In Anbetracht der aufgezeigten Unterschiede zwischen unterschiedlichen Veranstaltungsformen sollten hierfür spezifische Vergleichswerte genutzt werden (Vetterlein & Sengewald, 2015). Zur Erweiterung des Lernpotentials von Evaluationsrückmeldungen ist ferner eine persönliche Rückmeldung denkbar, mit der die einzelnen Dozierenden den Verlauf ihrer Lehrqualität über die Zeit betrachten können (Marsh, 2007). Unabhängig vom Verwendungszweck und der angelegten Bezugsnorm erlaubt das Verfahren eine Korrektur der Evaluationsergebnisse für Bias- und Unfairnesseffekte mit Hilfe der erfassten Hintergrundvariablen, um die Angemessenheit der Ergebnisse und deren sachgemäße Interpretation zusätzlich zu verbessern (Staufenbiel, Seppelfricke & Rickers, 2016). Eine solche Korrektur – beispielsweise in Form der Berücksichtigung dieser Variablen als Kontrollvariablen – ist insbesondere in Forschungskontexten oftmals unumgänglich, um zu validen Aussagen über inter- und intraindividuelle Unterschiede in der Lehrqualität zu gelangen (z. B. Dresel & Rindermann, 2011). In der Evaluationspraxis kann schließlich die Akzeptanz der Evaluationsrückmeldungen durch eine inhaltliche Beschreibung der erfassten Lehrqualitätsdimensionen unterstützt werden (ESM 9).

Die Analyse der grundlegenden Skaleneigenschaften erbrachte, dass ein großes Wertespektrum angenommen wurde und deutliche Unterschiede insbesondere zwischen Veranstaltungen bestanden, aber auch innerhalb von Veranstaltungen zwischen den Urteilen der einzelnen Studierenden. Dabei fand sich das für Lehrqualitätseinschätzungen durch Studierende typische und auch für den SEEQ bekannte Muster, dass die Bewertungen im Schnitt deutlich über dem theoretischen Skalenmittelwert lagen (Marsh, 1982).

Im Hinblick auf die Reliabilität erbrachten die Analysen durchgehend sehr zufriedenstellende Kennwerte. Erstens zeigte sich, dass die verschiedenen Items die jeweiligen Faktoren hinreichend intern konsistent messen und die entsprechenden Reliabilitätskennwerte ähnlich zu jenen sind, die für den Original-SEEQ berichtet wurden (vgl. Marsh, 2007). Dass die auf Veranstaltungsebene guten bis sehr guten internen Konsistenzen auf Studierendenebene tendenziell etwas weniger hoch ausfielen (aber mit ω_h = .76–.87 immer noch sehr zufriedenstellend), spricht gegen Antworttendenzen oder undifferenziertes Antwortverhalten. Eine Sichtkontrolle einer Teilstichprobe von 150 Fragebögen bestärkt dies, da in diesen nur sehr selten konsistente Antwortmuster vorzufinden waren. Zweitens verwiesen die hohen Intraklassenkorrelationen ICC2 darauf, dass Studierende derselben Veranstaltungen zu hinreichend ähnlichen Einschätzungen kommen, so dass messgenau auf die einzelnen Lehrqualitätsdimensionen als Veranstaltungsmerkmale geschlossen werden kann. Zu betonen ist, dass aus den Ergebnissen selbst für ein in dieser Hinsicht reliables Verfahren für die Evaluationspraxis folgt, dass in jeder zu bewertenden Veranstaltung eine relativ große Mindestanzahl von Studierendenurteilen (hier: neun) vorliegen muss, um hinreichend präzise Einzelergebnisse zu erhalten. Mit weniger Studierenden sollte eine Lehrveranstaltungsevaluation – auch aus Gründen der Anonymität der Studierenden – eher nicht durchgeführt werden. Zudem sollte berücksichtigt werden, dass Unterschiede zwischen den Urteilen der Studierenden auch ein wichtiges Feedback für die Dozierenden darstellen, da diese eine differenziell wirksame Lehre ausdrücken können. Daher bietet sich zusätzlich zu den Mittelwerten eine Rückmeldung des gesamten Wertespektrums der Studierendenurteile an.

Für die valide Interpretation der mit dem SEEQ-DE erfassten Lehrqualität spricht in Analogie zum Originalinstrument zunächst, dass sich die inkludierten Faktoren aus theoretischen Annahmen zur Effektivität von Instruktion ableiten (u. a. Berücksichtigung von Prozess- wie Produktmerkmalen; Döring, 2005; Rindermann, 2009) sowie in Korrespondenz zu übergreifenden Prinzipien des Lehrens und Lernens in Hochschulen stehen (z. B. Feldman, 1976; Fincher, 1985). Darüber hinaus belegt eine große Zahl an empirischen Studien in unterschiedlichen instruktionalen und kulturellen Kontexten die Relevanz dieser Faktoren für den Lernertrag der Studierenden (Perry & Smart, 2007; Spooren et al., 2013). Insbesondere werden die Items des Originalinstruments in der Forschung zur Effektivität von Hochschullehre seit vielen Jahren als inhaltsgültig für die jeweiligen Faktoren der Lehrqualität betrachtet (Marsh, 2007). Beide Aspekte – solide theoretische Fundierung und umfassende empirische Evidenz – verweisen auf eine umfassende inhaltliche Abdeckung des zu messenden Konstrukts, die entsprechend auch in der deutschsprachigen Adaption anzunehmen ist (aufgrund der Inhaltsgleichheit und der prinzipiellen Vergleichbarkeit der Qualitätsaspekte universitärer Lehre in unterschiedlichen Hochschulkontexten und Ländern).

Die weiterführenden Analysen der deutschsprachigen Adaption bestätigten die Unterscheidung in die verschiedenen Prozess- und Produktfaktoren der Lehrqualität: Auf Ebene der Lehrveranstaltungen und unter Berücksichtigung der geschachtelten Datenstruktur ließen sich in Faktorenanalysen die zehn erfassten Faktoren klar voneinander trennen. Dies untermauert die Annahme einer multifaktoriellen Lehrqualität, die die postulierten Pro-

zessfaktoren des Lehrhandelns und Produktfaktoren des Lehrerfolgs umfasst (Marsh & Roche, 1997). Dies gilt ebenso für den im SEEQ-DE zusätzlich aufgenommenen Faktor *studentische Beiträge*. Zu betonen ist, dass zwei unterschiedliche Analyseverfahren zur Prüfung der faktoriellen Struktur verwendet wurden. Die zum Vergleich mit den aktuellsten zur Struktur des SEEQ vorgelegten Befunden durchgeführten ESEM-Analysen erbrachten ähnliche Ergebnisse wie für das Originalinstrument (Marsh et al., 2009). Die Zwei-Ebenen-CFAs erlaubten es, das hierarchische Design von Lehrveranstaltungsevaluationen abzubilden (Sengewald & Vetterlein, 2015), und lieferten durch Replikation der faktoriellen Abgrenzbarkeit der verschiedenen Lehrqualitätsaspekte unter simultaner Berücksichtigung der Studierenden- und Dozierendenebene erhärtete Evidenz zu ihrer faktoriellen Struktur. Die unseres Wissens erstmalige konfirmatorische Bestätigung der Faktorenstruktur des SEEQ durch Zwei-Ebenen-Analysen erweitert den Forschungsstand zu dessen Struktur auch methodisch.

Einen weiteren Beleg für die valide Interpretierbarkeit der mit dem SEEQ-DE erfassten Aspekte der Lehrqualität besteht darin, dass die theoretisch erwartete Faktorenstruktur in Übungen, Seminaren und Vorlesungen gleichermaßen gegeben ist. Dies impliziert die Generalisierbarkeit dieser Faktoren über unterschiedliche Veranstaltungsformen hinweg und ist eine wichtige Voraussetzung für die Ergebnisinterpretation: Auf Basis der gezeigten skalaren Messinvarianz können die Ergebnisse unterschiedlicher Veranstaltungen sinnvoll miteinander verglichen werden.

Zwei weitere Validitätsbelege lieferte die Replikation von Befunden, die mit dem Originalinstrument gewonnen wurden. Die teils erheblichen Unterschiede zwischen den durchschnittlichen Lehrqualitätsbeurteilungen von Vorlesungen, Seminaren und Übungen fielen analog zu den bisherigen Ergebnissen in der Literatur aus (z. B. Rindermann, 2009). Zudem fanden wir theoriekonforme und in der Höhe vom Original-SEEQ bekannte Interkorrelationen zwischen den erfassten Prozessmerkmalen der Lehre sowie dem Lernzuwachs der Studierenden und deren Gesamtbeurteilung der Veranstaltungen (Marsh, 2007). Trotz dieser Belege ist als Einschränkung der vorliegenden Arbeit anzuführen, dass keine externen Merkmale wie Leistungs- oder Beobachtungsdaten berücksichtigt werden konnten. Gerade die Analyse von Zusammenhängen mit objektiven Leistungsdaten ist eine wichtige Perspektive für zukünftige Forschungsarbeiten (Stehle, Spinath & Kadmon, 2012). Darüber hinaus läge ein zukünftiger Validierungsschritt auch in der parallelen Darbietung des SEEQ-DE und des englischsprachigen Originals in einer Stichprobe von Studierenden, die beide Sprachen ausreichend beherrschen.

Der SEEQ ist ein ökonomisches und transparentes Verfahren, das als besonders fair gilt; es ist von Studierenden wie Dozierenden akzeptiert und wird von ihnen vielfach als hilfreich eingestuft (Marsh, 2007). Die in der deutschen Adaption des SEEQ zusätzlich aufgenommenen Aspekte (Gesamtbeurteilung in Form von Noten, Subskala zu studentischen Beiträgen), die anlassbezogen optional eingesetzt werden können, tragen zusätzlich zur Nützlichkeit des Verfahrens bei.

Abseits der Konstruktion und Überprüfung des SEEQ-DE liefern die präsentierten Resultate einen zusätzlichen Beitrag zur Generalisierbarkeit der Dimensionen von Lehrqualität über verschiedene Kontexte hinweg und fügen sich damit in eine Reihe internationaler Arbeiten dazu ein (vgl. ESM 1). Darüber hinaus stehen die Befunde im Einklang mit der Forderung, dass unterschiedliche Dimensionen der Lehrqualität bei Studierendenevaluationen der Lehre berücksichtigt werden sollten (Spooren et al., 2013), und sie bestätigen, dass diese über verschiedene Veranstaltungsformen hinweg ähnlich strukturiert sind (Marsh et al., 2009). Von hohem praktischem Wert ist ferner, dass sich die im Messinstrument enthaltenen Dimensionen des Lehrhandelns unabhängig von der Veranstaltungsform als relevant für den Lehrerfolg erwiesen. Dies spricht dafür, dass die unterschiedenen Lehrqualitätsdimensionen für alle Veranstaltungsformen von praktischer Relevanz sind und für die Verbesserung der Lehre berücksichtigt werden sollten.

Trotz der genannten Stärken der vorliegenden Untersuchung sind bei der Interpretation der Ergebnisse wichtige Limitationen zu berücksichtigen. Zunächst ist zu bedenken, dass die vorliegenden Befunde auf einer papierbasierten Durchführung der Lehrveranstaltungsevaluationen beruhen (siehe Janke et al., 2020, für eine elektronische Umsetzung). Zwar ist von einer ähnlichen Güte bei einer Online-Befragung auszugehen, jedoch könnte es Unterschiede vor allem in der Teilnahmebereitschaft, aber in gewissem Umfang auch in den Antwortprozessen geben (Dresel & Tinsner, 2008; Spooren et al., 2013). Des Weiteren ist die vorliegende Datengrundlage dadurch limitiert, dass die Teilnahme freiwillig war (sowohl auf Studierenden- als auch auf Dozierendenseite). Es ist nicht anzunehmen, dass sich dies auf die Güte des Fragebogens auswirkt, jedoch sind die gefundenen Ausprägungen der Lehrqualität dadurch wohl positiv verzerrt. Ferner ist zu berücksichtigen, dass die Daten über einen längeren Zeitraum erhoben wurden und dass Studierende mehrmals in der Stichprobe enthalten sind und mehrere Veranstaltungen evaluierten. Da die Studierendeneinschätzungen der Lehrqualität zum Teil auch auf die Merkmale der evaluierenden Studierenden selbst zurückgeführt werden können (Feistauer & Richter, 2017), mag eine Überschätzung der Korrelationen auf Studierenden-

ebene resultiert (für die Veranstaltungsebene trifft diese Limitation angesichts der reliablen Messung auf dieser Ebene nicht zu) und die ML-Schätzungen beeinflusst haben.

Resümierend kann trotz der Limitationen festgehalten werden, dass mit der vorgelegten deutschsprachigen Adaption des SEEQ ein Instrument zur Verfügung steht, mit dem die multifaktorielle Qualität von Lehrveranstaltungen im Hochschulbereich international anschlussfähig, transparent und ökonomisch erfasst werden kann und für das umfangreiche Hinweise für eine valide Interpretation der Testergebnisse gemäß der intendieren Verwendung als Evaluations- und Forschungsinstrument vorliegen.

Elektronische Supplemente (ESM)

Die elektronischen Supplemente sind mit der Online-Version dieses Artikels verfügbar unter https://doi.org/10.1026/0012-1924/a000274

ESM 1. Beispiele für weitere Evaluationsinstrumente sowie internationale Übersetzungen des SEEQ

ESM 2. Items des SEEQ-DE

ESM3. Muster für den Einsatz des SEEQ-DE mittels EvaSys (Electric Paper, 2015)

ESM 4. Durchführungsinstruktion für Testleiter bzw. Dozierende

ESM 5. Ergebnisse der Faktorenanalysen

ESM 6. Ergebnisse der ESEM-Analyse zur Faktorstruktur auf Veranstaltungsebene

ESM 7. Ergebnisse der Faktorenanalysen ohne Faktor „Studentische Beiträge“

ESM 8. Mittelwerte und 95%-Konfidenzintervalle aller Subskalen

ESM 9. Beschreibung der Evaluationsergebnisse

Literatur

Abrami, P. C., d'Apollonia, S. & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry and J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385 – 456). Dordrecht, Netherlands: Springer.

Byrne, B. (2013). *Structural equation modeling with Mplus.* New York, NY: Routledge.

Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464 – 504. https://doi.org/10.1080/10705510701301834

Cheung, G. & Rensvold, R. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233 – 255. https://doi.org/10.1207/S15328007SEM0902_5

Daumiller, M., Janke, S., Hein, J., Rinas, R., Dickhäuser, O. & Dresel, M. (2021). Do teachers' achievement goals and self-efficacy beliefs matter for students' learning experiences? Evidence from two studies on perceived teaching quality and emotional experiences. *Learning and Instruction.* [Advanced online publication]. https://doi.org/10.1016/j.learninstruc.2021.101458

Döring, N. (2005). Für Evaluation und gegen Evaluitis. In B. Berendt, H.-P. Voss & J. Wildt (Hrsg.), *Neues Handbuch Hochschullehre.* (S. 2 – 22). Berlin: Raabe.

Dresel, M. & Rindermann, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness. *Research in Higher Education*, *52*, 717 – 737. https://doi.org/10.1007/s11162-011-9214-7

Dresel, M. & Tinsner, K. (2008). Onlineevaluation von Lehrveranstaltungen. *Zeitschrift für Evaluation, 7*, 183 – 211.

Dunn, T., Baguley, T. & Brunsden, V. (2014). From alpha to omega. *British Journal of Psychology, 105*, 399 – 412. https://doi.org/10.1111/bjop.12046

Fan, X. & Sivo, S. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*, 509 – 529. https://doi.org/10.1080/00273170701382864

Feistauer, D. & Richter, T. (2017). How reliable are students' evaluations of teaching quality? *Assessment & Evaluation in Higher Education, 42*, 1263 – 1279. https://doi.org/10.1080/02602938.2016.1261083

Feldman, K. (1976). The superior college teacher from the students' view. *Research in Higher Education, 5*, 243 – 288. https://doi.org/10.1007/BF00991967

Fincher, C. (1985). Learning theory and research. In J. Smart (Ed.), *Higher education* (pp. 63 – 96). New York, NY: Agathon.

Fondel, E., Lischetzke, T., Weis, S. & Gollwitzer, M. (2015). Zur Validität von studentischen Lehrveranstaltungsevaluationen. *Diagnostica*, *61*, 124 – 135. https://doi.org/10.1026/0012-1924/a000141

Gravetter, F. & Wallnau, L. (2014). *Essentials of statistics for the behavioral sciences* (8th ed.). Belmont, CA: Wadsworth.

Helmke, A. (1996). Studentische Evaluation der Lehre. *Zeitschrift für Pädagogische Psychologie, 10*(3/4), 181 – 186.

Holman, R. & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*(1), 1 – 17. https://doi.org/10.1348/000711005x47168

Janke, S. et al. (2020). Open Access Evaluation: Lehr-Evalution Online (LEO) als Instrument zu studentischen Lehrveranstaltungsevalution. *Qualität in der Wissenschaft*, *14*(4), 120 – 125.

Kline, R. (2015). *Principles and practice of structural equation modeling.* New York: Guilford.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, *34*(2), 120 – 131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Marsh, H. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*(1), 77 – 95. https://doi.org/10.1111/j.2044-8279.1982.tb02505.x

Marsh, H. (2007). Students' evaluations of university teaching. In R. Perry & J. Smart (Eds.), *The scholarship of teaching and learning in higher education* (pp. 319 – 383). Dordrecht, Netherlands: Springer.

Marsh, H. & Dunkin, M. (1992). Students' evaluations of university teaching. *International Journal of Educational Research*, *11*, 253 – 388. https://doi.org/10.1016/0883-0355(87)90001-2

Marsh, H., Hau, K., Chung, C. & Siu, T. (1998). Confirmatory factor analyses of Chinese students' evaluations of university tea-

ching. *Structural Equation Modeling*, *5*, 143 – 164. https://doi.org/10.1080/10705519809540097
Marsh, H. & Hocevar, D. (1991). Students' evaluations of teaching effectiveness. *Teaching and Teacher Education*, *7*, 303 – 314. https://doi.org/10.1016/0742-051X(91)90001-6
Marsh, H., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. et al. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 439 – 476. https://doi.org/10.1080/10705510903008220
Marsh, H., Overall, J. & Kesler, S. (1979). Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal*, *16*(1), 57 – 70. https://doi.org/10.3102/00028312016001057
Marsh, H. & Roche, L. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist*, *52*, 1187 – 1197. https://doi.apa.org/doi/10.1037/0003-066X.52.11.1187
Muthén, L. & Muthén, B. (2017). *Mplus User's guide.* Los Angeles, CA: Muthén & Muthén.
Perry, R. & Smart, J. (Eds.). (2007). *The scholarship of teaching and learning in higher education.* Dordrecht, Netherlands: Springer.
Peugh, J. & Enders, C. (2004). Missing data in educational research. *Review of Educational Research, 74,* 525 – 556. https://doi.org/10.3102/00346543074004525
Rindermann, H. (2009). *Lehrevaluation* (2. Aufl.). Landau: VEP.
Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models. *Methods of Psychological Research Online, 8,* 23 – 74.
Sengewald, E. (2016). *Überprüfung und Anwendung von Multilevel-Messmodellen für Fragebögen zur Lehrveranstaltungsevaluation* (Doktorarbeit). Friedrich-Schiller-Universität Jena, Jena.
Sengewald, E. & Vetterlein, A. (2015). Multilevel Faktorenanalyse für Fragebögen zur Lehrveranstaltungsevaluation. *Diagnostica, 61*, 116 – 123. https://doi.org/10.1026/0012-1924/a000140
Snijders, T. A. B. & Bosker, R. J. (2020). *Multilevel analysis* (2nd ed.). London, UK: Sage.
Spinath, B., Antoni, C., Bühner, M., Elsner, B., Erdfelder, E., Fydrich, T. et al. (2018). Empfehlungen zur Qualitätssicherung in Studium und Lehre. *Psychologische Rundschau, 69*, 183 – 192. https://doi.org/10.1026/0033-3042/a000408
Spinath, B. & Stehle, S. (2011). Evaluation von Hochschullehre. In L. Hornke & M. Amelang (Hrsg.), *Enzyklopädie der Psychologie* (S. 617 – 667). Göttingen: Hogrefe.
Spooren, P., Brockx, B. & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The State of the Art. *Review of Educational Research*, *83*, 598 – 642. https://doi.org/10.3102/0034654313496870
Staufenbiel, T., Seppelfricke, T. & Rickers, J. (2016). Prädiktoren studentischer Lehrveranstaltungsevaluationen. *Diagnostica, 62,* 44 – 59. https://doi.org/10.1026/0012-1924/a000142
Stehle, S., Spinath, B. & Kadmon, M. (2012). Measuring teaching effectiveness: Correspondence between students' evaluations of teaching and different measures of student learning. *Research in Higher Education, 53*, 888 – 904. https://doi.org/10.1007/s11162-012-9260-9
Stehle, S. & Spinath, B. (2011). Zur Validität studentischer Lehrbeurteilungen. In M. Krämer, S. Preiser & K. Brusdeylins (Hrsg.), *Psychologiedidaktik und Evaluation VII.* (S. 347 – 356). Aachen: Shaker.
Toland, M. & De Ayala, R. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65*, 272 – 296. https://doi.org/10.1177/0013164404268667
Vetterlein, A. & Sengewald, E. (2015). Ergebnisdarstellung in der Lehrveranstaltungsevaluation. *Diagnostica, 61,* 153 – 162. https://doi.org/10.1026/0012-1924/a000128
Watkins, D. (1994). Student evaluations of university teaching: A cross-cultural perspective. *Research in Higher Education*, *35*, 251 – 266. https://doi.org/10.1007/BF02496704

Danksagung
Wir danken Christian Eibl und Susanne Reeß für ihre Unterstützung bei der Planung und Durchführung der Lehrveranstaltungsevaluationen.

ORCID
Martin Daumiller
https://orcid.org/0000-0003-0261-6143

Dr. Martin Daumiller
Lehrstuhl für Psychologie
Universität Augsburg
86179 Augsburg
martin.daumiller@phil.uni-augsburg.de