Bayerische Schule

Heute unter anderem:

Herausgeber: Bayerischer Lehrer- und Lehrerinnenverband e. V., 8000 München 2, Bavariaring 37, Telefon 08 11 / 77 61 03 und 76 67 80
Redaktionsleitung: Korbinian Huber
Verantwortlich für
Aktuelles: Dieter Krause, 8000 München 15, Bavariaring 37, Telefon 08 11 / 77 61 03 und 76 67 80
Wissenschaft und Praxis: Dr. Lothar F. Katzenberger, 8700 Würzburg, Tilsiter Straße 6, Telefon 09 31 / 7 85 10
Aus dem Verband, letzte Seite: Korbinian Huber, 8226 Altenmarkt, Marktplatz 3, Telefon 0 86 21 / 72 69
Leserbriefe: Direkt an die BLLV-Geschäftsstelle 8000 München 2, Bavariaring 37, oder Redaktionsleitung.
Funk – TV: Ute Moeller-Andresen, 8000 München 23, Dunantstraße 5, Telefon 08 11 / 33 36 86
Bücherschau: Dr. Hans-Peter Winkel, 8000 München 50, Siegmund-Schacky-Straße 4, Telefon 08 11 / 1 41 18 30
Gestaltung und Schlußredaktion: Korbinian Huber
Anzeigen: Erwin Huber, 8223 Trostberg, Gabelsbergerstraße 4–6, Telefon 0 86 21 / 30 64
Verlag und Druck: Druck- und Verlagshaus A. Erdl KG, 8223 Trostberg, Gabelsbergerstraße 4–6, Telefon 0 86 21/30 64, Fernschreiber 56 3183
Titelseite: „die agentour", München
Die „Bayerische Schule" erscheint zweimal monatlich. Sie wird allen Mitgliedern des BLLV geliefert. Der Bezugspreis ist im Mitgliedsbeitrag enthalten. Nichtmitglieder können bei der Post bestellen. Bezugspreis DM 15,– jährlich (auf Postscheckkonto München Nr. 40 677/806).

Für unverlangt eingesandte Manuskripte übernehmen wir keine Haftung. Falls kein Rückporto beiliegt, schicken wir sie auch nicht an den Autor zurück.
Namentlich gekennzeichnete Beiträge stellen die Meinung der Verfasser, nicht unbedingt die der Redaktion und des BLLV dar.

Redaktionsschluß für alle Beiträge dieser Ausgabe: Freitag, 29. Juni
Redaktionsschluß der nächsten Ausgabe: Freitag, 13. Juli
Redaktionsschluß der übernächsten Ausgabe: Montag, 30. Juli

WISSENSCHAFT UND PRAXIS

Die Diskussion hält an:

Über die fragwürdigen Gütekriterien von Schulleistungstests

Eine Kritik herkömmlicher Diagnostik am Beispiel des „Schulleistungs-Prüfsystems auf Testbasis" – Von Lutz Mauermann

Lutz Mauermann, Lehrer und cand. phil., Jahrgang 1943, verheiratet. 1962 Abitur an der Oberrealschule Erding, 1962–65 Studium an der PH München-Pasing. Während dieser Zeit: Vorsitzender der BLLV*-Studentengruppe, AStA-Mitglied, 1. Vorsitzender der Landesstudentengruppe des* BLLV*, Lehrer an verschiedenen Schulen, 1968 2. LAP, Beginn eines nebenberuflichen Zweitstudiums in den Fächern Pädagogik, Psychologie und Soziologie an der Universität München, 1971 Abordnung vom Schuldienst zur Dienstleistung bei der »Projektgruppe Schullaufbahnberatung« am Pädagogischen Institut II der Universität München (Prof. Dr. Hans Schiefele) und am Institut für Unterrichtsforschung der Universität Augsburg (Prof. Dr. Rolf Oerter); seitdem Mitarbeiter dieser Forschungsgruppe und mit Problemen pädagogischer Diagnostik betraut.*

Anschrift des Verfassers: 8058 Klettham, Fuggerstr. 1.

In der Bayerischen Schule vom 10.4.73 erschien ein Beitrag zur Diskussion um die Übertrittsverfahren für weiterführende Schulen, in dem der Autor, Professor Rüdiger, das Problem der Auslese und speziell dessen bayerische Lösung darstellt und das von ihm herausgegebene „Schulleistungs-Prüfsystem auf Testbasis" (Verlag Wolf, Regensburg) als Mittel zur Verbesserung der Übertrittsauslese bezeichnet. In einer Replik geht Rüdiger auch auf meinen kurzen Artikel ein, der vor gut einem Jahr – leider etwas verkürzt und ohne Literaturverzeichnis – in der Bayerischen Schule veröffentlicht worden ist. Im folgenden will ich versuchen, meine nach wie vor bestehende Skepsis an herkömmlichen Schulleistungstests am Beispiel des Regensburger „Prüfsystems" zu begründen. Dabei stütze ich mich im wesentlichen auf Forschungsergebnisse der Projektgruppe Schullaufbahnberatung am Institut für Pädagogik II der Universität München und am Institut für Unterrichtsforschung der Universität Augsburg.

Rüdiger betont in der Anleitung zum Prüfsystem, daß hiermit „in Form einer Nahlösung ... gerechter und verläßlicher" ausgelesen werden könne (1972, S. 1). Der Beweis von gesteigerter Gerechtigkeit wird mit Hilfe von Test-Gütekriterien angetreten.

Objektivität

Im testtheoretischen Sinn ist Objektivität nur die Unabhängigkeit des Ergebnisses vom Untersucher, die abgetrennt von inhaltlicher Fragestellung betrachtet wird. Es handelt sich dabei lediglich um einen Formalismus: Das natürliche und vielschichtige Lernverhalten, wie es der Schüler z. B. im Unterricht zeigt, wird auf Situationen beschränkt, die leicht beobachtbar und für andere mitteilbar sind. Beim Prüfsystem sieht das dann so aus, daß nur bestimmte Wörter (Lückentextdiktat) beachtet werden, zwei Zahlen über den Begriff der Zahlenreihe Auskunft geben und eindeutig lösbare Zahlenrechnungen Fertigkeiten nachweisen sollen. Sobald jedoch komplexere Lernziele in Angriff genommen werden (Aufsatz, Textrechnen), gelten nur noch Auswertungs-„Empfehlungen"[1]) (Rüdiger 1972, S. 3) bzw. verschwinden als Punktwert in einer anderen Kategorie. Gerade die fehlerhafte Beurteilung von komplexen Schulleistungen wirft man jedoch den Lehrern vor. Wegen Diktat- oder überschaubarer Mathematikkorrekturen hat es noch selten Proteste seitens der Öffentlichkeit gegeben. Einheitliche Probearbeiten, die von regionalen Arbeitsgruppen für das Übertrittsverfahren bislang erstellt wurden und noch werden, genügen mindestens genauso gut dem Kriterium der Testobjektivität wie das Prüfsystem.

Was das Prüfsystem angeblich mehr hat, ist die Bereitstellung von Urteilsnormen (Prozentränge nach repräsentativen Stichproben). Daß dies jedoch wenig mit Objektivität zu tun hat, sondern eher eine politische Entscheidung darstellt, geht aus der Vorbemerkung zur Notentabelle hervor: Da bisherige Übertrittsquoten soundso groß waren, dürfen eben nur soundso viel Schüler gute Noten bekommen und die Hürde zur weiterführenden Schule nehmen. Regional oder schulortsbezogen wird es sicher zu ungerechten Normierungen kommen, da keinerlei Rücksicht

[1]) Ich bezweifle, ob eine Aufsatzbeurteilung objektiver allein dadurch wird, daß beispielsweise für die Rubrik „differenziertere Satzkonstruktionen, Hauptsatz-Nebensatz-Verbindungen" eine 12-Punkte-Skala zur Verfügung steht. Ich persönlich traute mir nicht zu, anhand eines erfahrungsgemäß durchschnittlich 15 Sätze umfassenden Aufsätzchens eines Mädchen oder Jungen der 4. Jahrgangsstufe auf einer solchen differenzierenden Skala ein „gerechtes" Urteil abzugeben.

auf die jeweilige Situation (Ausstattung mit Lehrern, Erkrankungen, Schulbusschwierigkeiten, Lehr- und Lernmittelausstattung u. ä.) und die mit ihr verknüpften Auswirkungen auf das Lernverhalten der Schüler genommen wird. „Gerechter" ist das Prüfsystem keinesfalls als diejenigen Probearbeiten, die von Lehrerteams in den Landkreisen erstellt werden und regionale Voraussetzungen berücksichtigen können. Wie sieht es nun mit der Verläßlichkeit aus?

Reliabilität

Zunächst einmal beruhen die Verläßlichkeitsschätzungen für das Prüfsystem auf der Ermittlung des Konsistenzkoeffizienten (Testhalbierungsmethode), der „stets eine ‚optimistischere' Auskunft über die Reliabilität gibt als dies etwa Retest- oder Paralleltestkoeffizienten geben, die mehr die empirischen Gegebenheiten der Praxis berücksichtigen" (Lienert 1969, S. 214).

Der Reliabilitätskoeffizient soll angeben, inwieweit ein Testergebnis meßgenau ist. Er wird in die Berechnung des Standardmeßfehlers (s_e) einbezogen

$$s_e = s_x\sqrt{1 - r_{tt}},$$

wobei s_x die Standardabweichung (= Maß für die Streuung der Testwerte) und r_{tt} der Reliabilitätskennwert ist. Da im diesbezüglichen Abschnitt bei Rüdigers Artikel die tabellarisch aufgeführten Daten fehlen – ich vermute, daß diese aus technischen Gründen von der Redaktion der Bayerischen Schule weggelassen wurden –, greife ich zur Veranschaulichung der Zusammenhänge auf Ergebnisse bei der Erprobung des Prüfsystems zurück, die an anderer Stelle veröffentlicht worden sind (Rüdiger 1971, S. 135). Dort wird für die Aufgabengruppe „Zahlenreihen" im 4. Schülerjahrgang aufgeführt:

$$r_{tt} = 0{,}79 \text{ und } s_x = 3{,}60.$$

Das ergibt einen Meßfehler von

$$s_e = 3{,}60\sqrt{1 - 0{,}79} = 3{,}60\sqrt{0{,}21} \approx \underline{1{,}65}$$

Dies gilt als Unsicherheitsmaß für individuelle Testpunktwerte. Zu jedem von den Schülern erzielten Punktwert kann nun ein Vertrauensintervall (CL_X) berechnet werden, d. h. der Bereich auf der Rohwertskala, innerhalb dessen der „wahre" Testwert des Schülers bei Berücksichtigung des Meßfehlers liegen kann.[2]) Wird die Wahrscheinlichkeit eines Irrtums mit 5 Prozent in Kauf genommen, dann ist das Konfidenzintervall (vgl. Lienert 1969, S. 453):

$$CL_X = X \pm 1{,}96 \cdot s_e$$

Als Vertrauensbereich ergibt sich dann bei einem Meßfehler von 1,65:

$$\begin{aligned} CL_X &= X \pm 1{,}96 \cdot 1{,}65 \\ &= X \pm 3{,}234 \\ &= X \pm \sim 3 \end{aligned}$$

Hat also ein Schüler beispielsweise 11 (= X) von den 15 Zahlenreihen richtig ergänzt, so ist seine „wahre" Leistung in 95 Prozent aller Fälle in einem Bereich zwischen 8 und 14 Testpunkten zu suchen. Soll gar ein Irrtum von nur 1 Prozent zugelassen sein – was bei der Bedeutung einer Ausleseentscheidung nicht so abwegig erscheint –, dann erhöht sich der Faktor, mit dem der Meßfehler multipliziert werden muß, auf 2,58[3]):

$$\begin{aligned} CL_X &= X \pm 2{,}58 \cdot 1{,}65 \\ &= X \pm 4{,}257 \\ &= X \pm \sim 4 \end{aligned}$$

Der „wahre" Wert eines Schülers mit einer Leistung von 11 Punkten läge dann bei 99-prozentiger Wahrscheinlichkeit in einem Bereich von 7 bis 15 Testpunkten. Übertragen auf Noten der dem Prüfsystem beigefügten Normentabelle wäre die Meßgenauigkeit in dem angenommenen Beispiel auf dem Bereich zwischen Note 4 und Note 1 beschränkt.

Dieses Rechenexempel läßt verständlich erscheinen, weshalb Testautoren viel daran liegt, die Verläßlichkeit ihrer Produkte in permanenter Revision möglichst hochzuschrauben. Deshalb werden testtheoretische Tricks angewandt: Immer stärkere Ausschaltung von sogenannten Störfaktoren, Trennschärfenerhöhung der Einzelaufgaben (Items), Annäherung der Lösungswahrscheinlichkeit der Aufgaben an das Idealmaß von 50 Prozent, Vermehrung der Einzelaufgaben. *Dabei wird übersehen, daß eine konsequente Entfremdung von der Schulwirklichkeit einherläuft.* „Man erhält hierdurch zwar eine hohe Reproduzierbarkeit der Ergebnisse, schränkt dabei aber gleichzeitig die Übertragbarkeit der Testresultate auf natürliche Situationen hoffnungslos ein." (Frenz, Krüger & Tröger 1973, S. 25) Hier wird bereits das Problem der Gültigkeit von Tests angesprochen. Wie steht es damit beim Prüfsystem?

Validität

Rüdiger schreibt, daß „die inhaltliche Validität nach dem Grad der Lehrplanübereinstimmung des Verfahrens als gegeben erachtet werden muß" (Rüdiger 1973, S. 177). Damit mag er insofern recht haben, als alle Testaufgaben dem Stoffverteilungsplan entnommen sein könnten. Inhaltliche Gültigkeit verlangt jedoch auch eine repräsentative Auswahl aus dem fraglichen Aufgabenbereich. Für den Bereich Deutsch mag selbst das noch gelten (obwohl bemerkt werden muß, daß bereits das Kultusministerium eine Ausklammerung von Bereichen wie „Umgang mit Texten" und „Mündliche Sprachgestaltung" vorgenommen hat). Prekär wird die Situation für die drei Aufgabengruppen „Rechnen". Ausgehend von dem Gewicht, das den Gruppen „Zahlenreihen", „Zahlenrechnen" und „Textrechnen" beigemessen wird, müßte der Rechenunterricht in allen Volksschulklassen Bayerns zur Hälfte aus dem Training von „logischem Denken" mittels Zahlenreihen bestehen, während – entsprechend der maximal erreichbaren Punktzahl im 4. Schuljahr – 19 Prozent auf formales Rechnen und 31 Prozent auf Textrechnen fielen. Eine so geartete Gewichtung der Stoffbereiche ist aber weder aus gängigen Lehrbüchern noch aus den Lehrplänen abzuleiten. Hinzu kommt, daß gerade bei standardisierten Tests und ihren formalisierten Auswerteverfahren auch rein technische Eigenarten von entscheidendem Einfluß auf die Lösungsmöglichkeit eines Schülers oder einer Gruppe sein können. Von inhaltlicher Validität des Prüfsystem-Rechnens kann somit nicht die Rede sein.[4])

Nun läßt sich sagen, die Zahlenreihen stellten eigentlich nicht einen Rechentest, sondern einen „gleichzeitigen Begabungstest" (Rüdiger 1973, S. 178) dar. Sie seien klassische Intelligenztestaufgaben, die den Schulerfolg „besser vorhersagen als das testfreie Lehrerurteil" (a. a. O., S. 177). Folglich: Das Zahlenreihen-Lösen kann als Prediktor für den Erfolg am Gymnasium verwendet werden. Wenn Rüdiger behauptet, die Psychologen orientierten sich schon seit langem nicht mehr an einem starren Begabungsbegriff, so beweist sein Plädoyer für herkömmliche Intelligenztestaufgaben, deren Sinn und Fragwürdigkeit z. B. Liungman (1973) in seinem Buch aufzuzeigen versucht, bezüglich der Rigidität das Gegenteil. Die Aussage: die Fähigkeit, Zahlenreihen zu lösen, ist ein Kriterium, das für den erfolgreichen Weg durch das Gymnasium vorhanden sein muß, entspricht genau einem statischen Prognosemodell (vgl. Frenz, Krüger & Tröger 1973, S. 34). Ich kann dieser Aussage nicht zustimmen, da sie meinem pädagogischen Tun als Lehrer das Motiv raubt, indem sie mir weismachen will: Nur Begabte schaffen den Zugang zu privilegierenden Ausbildungsgängen, und mit den übrigen sollten schlichtere Tätigkeiten geübt werden. *Die neuere Sozialisationsforschung hat gezeigt, daß unterschiedliche Schulleistungen im wesentlichen auf unterschiedliche, gesellschaftlich bedingte Voraussetzungen zurückzuführen sind.* Die Grenzen der individuellen Lernfähigkeit sind beim gegenwärtigen Stand der Forschung vorerst unbestimmt (vgl. Gottschalch, Neumann-Schönwetter & Soukup 1971, S. 43 f.).

---

[2]) Dabei wird angenommen – was selten geprüft und wohl nie zutrifft –, daß für jeden Bereich der Skala die Verhältnisse gleich sind. Untersuchungen und Vernunftgründe sprechen dafür, daß der „Meßfehler" im unteren Skalenbereich z. B. größer ist als im oberen und auf andere Faktoren zurückzuführen ist.

[3]) Es würde an dieser Stelle zu weit führen, den Ableitungszusammenhang für diesen Faktor darzustellen. Er ist bei Lienert (1969) nachzulesen.

[4]) Zu denken gibt auch, daß sich bislang noch kein Didaktiker des Mathematikunterrichts dazu hergegeben hat, sein Plazet unter das Prüfsystem zu setzen.

Schulversuche mit einer Strategie des „zielerreichenden Lernens" (mastery learning) deuten an, wie optimistisch einerseits die Lernfähigkeit gesehen werden kann und wie fragwürdig andererseits langfristige Prognosen über schulischen Erfolg werden (s. Bloom 1970, S. 21 ff.).

Zum Nachweis der kriterienbezogenen Validität bemüht der Herausgeber des Prüfsystems die Noten der Schüler im nächsten Jahr an den weiterführenden Schulen und Ergebnisse bei den etablierten Tests PSB[5]), KLI[6]), AST[7]). Korrelationen zwischen dem Prüfsystem und den Kriterien werden zur Ermittlung des Bewährungskoeffizienten gerechnet.

Was bedeutet nun z. B. ein Korrelationskoeffizient (r) von 0,54 zwischen PSB (logisches Denken) und den Zahlenreihen des Prüfsystems? Ein r gibt den Zusammenhang an zwischen den Meßwerten zweier Merkmale bei ein und derselben Stichprobe (etwa bei Schülern). Wenn r den theoretisch möglichen Maximalwert von +1,00 annimmt, so würde das bedeuten, daß zwischen den Meßwerten des Merkmals X (PSB) und denen des Merkmals Y (Prüfsystem) ein positiver, streng linearer Zusammenhang besteht: Jedem hohen X-Wert in dem einen Test entspricht ein hoher Y-Wert im anderen, jedem niedrigen X-Wert entspricht ein niedriger Y-Wert. Ein r von 0,54 darf nicht als Prozentwert gedeutet werden (also daß in 54 von 100 Fällen die Werte von Prüfsystem und PSB übereinstimmen). Vielmehr muß als Bestimmtheitsmaß r^2 verwendet werden. Dieser Wert gibt an, in welchem Ausmaß die Varianz der Meßwerte bei einer Variablen durch die Varianz der Meßwerte der anderen Variablen bestimmt werden kann (und das gilt nur für die gegebene Stichprobe!).

So führt ein $r = 0{,}54$ auf $r^2 = 0{,}2916$, und dieser Wert läßt die Aussage zu, daß 29 Prozent (!) der Varianz beider Meßwertreihen übereinstimmt, wobei nichts darüber ausgesagt ist, ob dies für den ganzen Bereich der Skala gilt (vgl. Fußnote 2).

Der Korrelationskoeffizient r kann dazu verwendet werden, die Vorhersage-Effizienz (E) nach der Formel

$$E = 1 - \sqrt{1 - r^2}$$

zu berechnen (vgl. Cronbach & Gleser 1965, S. 31). Der Effizienzwert gibt an, um wieviel die Vorhersage der Varianz des Kriteriumswertes durch den Test gegenüber dem reinen Zufall verbessert wird. Im vorliegenden Fall:

$$E = 1 - \sqrt{1 - 0{,}2916}$$
$$= 1 - \sqrt{0{,}7084}$$
$$\approx 0{,}16 \text{ oder } 16\ \%$$

Die beiden Rechenbeispiele sollten deutlich gemacht haben, welche Bedeutung solchen Bewährungskoeffizienten zugemessen werden darf.

Die Gültigkeit der herbeizitierten „qualifizierten Übertrittstests" PSB, KLI und AST sollte nicht unbedingt als gegeben hingenommen werden. Eine Analyse der entsprechenden Handanweisungen zu den Tests zeigt,

- ☐ daß die Validierung des Prüfsystems für Schul- und Bildungsberatung (PSB) von Horn recht undurchsichtig an mehreren unvergleichbaren Stichproben, teilweise an der Notengebung der Lehrer, teilweise durch Verweis auf ähnliche, bereits bewährte Subtests, nachgewiesen wird (vgl. Horn 1969, S. 17 ff.);
- ☐ daß der Kombinierte Lern- und Intelligenztest (KLI) von Schröder seine Gültigkeit an den Notendurchschnitten des 1. und 2. Gymnasialschuljahres in Mathematik, Deutsch, 1. Fremdsprache und Geografie nachweist (vgl. Schröder 1968, S. 12 ff.);
- ☐ daß dem Allgemeinen Schulleistungstest für 4. Klassen (AST) von Fippinger lediglich „logische" Validität zugrundeliegt, indem die Aufgaben von Lehrern in Augenschein genommen und zu „schultypischen Leistungsanforderungen der Hauptfächer in vierten Klassen" erklärt wurden (Fippinger 1967, S. 15).

Auch für das Prüfsystem müssen die zur Genüge angefeindeten Noten (vgl. Ingenkamp 1971, Schröter 1971) als „Bewährungshelfer" herhalten. *Was ist davon zu halten, wenn die aufgrund offensichtlich minder qualifizierter Methoden der Lehrer entstandenen Noten (vgl. Rüdiger 1973, S. 177) einerseits in Frage gestellt und andererseits zur Bestimmung der Validität der eigenen Aussagen herangezogen werden?*

Noch zwei Bemerkungen zu den Validitätskoeffizienten für das Prüfsystem:

a) Es entspricht nicht den empirisch-wissenschaftlichen Gepflogenheiten, Koeffizienten, die aus verschiedenen Stichproben gewonnen wurden (Prüfsystem – Noten – Vergleich basiert auf N = 286, Übertrittsgutachten – Noten – Vergleich auf N = 126), gegeneinander auszuspielen (Rüdiger 1963, S. 178), ohne dabei die Problematik der Stichprobe in der pädagogisch-psychologischen Forschung zu erwähnen. Wenn schon die Statistik in eine Argumentationskette einbezogen wird, für Leser bestimmt, die wahrscheinlich wenig mit diesem Gebiet vertraut sind, dann sollten auch die wichtigsten Implikationen angedeutet werden (hierzu: Becker-Freyseng, Krüger & Rietbrock 1973, S. 4).

b) Sollten die Ergebnisse aus dem Prüfsystem wirklich das „Eignungsurteil abgestützt und u. U. auch verbessert haben" (Rüdiger 1973, S. 178), so wäre interessant gewesen, welchen Wert eine Korrelation zwischen dem „testfreien Lehrerurteil" und den Kriterien der aufnehmenden Schule annimmt. Ich möchte behaupten, daß dieser Wert kaum schlechter als 0,48 sein wird. Dann allerdings wäre in Anbetracht des Aufwandes zur Durchführung des Prüfsystems dessen Einsatz kaum gerechtfertigt.

Test und Schulwirklichkeit

Unter dem Druck der klassischen Testtheorie mit ihrer Forderung nach Zuverlässigkeit und Gültigkeit haben viele Psychologen die Frage vernachlässigt, *was* und *wozu* gemessen werden soll. Um signifikante experimentelle Befunde zu erhalten, werden die zu beobachtenden Verhaltensweisen immer mehr zersplittert, eingeschränkt und durch Ausschluß von Störfaktoren labilisiert. Damit entfernt sich jedoch die experimentelle immer mehr von der praktischen Realität, und es stellt sich folglich das Problem der Bedeutung experimenteller Ergebnisse für die Praxis. Jeder Test greift einen mehr oder minder eingeschränkten Ausschnitt aus dem Universum aller möglichen Verhaltensweisen heraus. Meist handelt es sich dabei um recht einfache und leicht nachprüfbare Kenntnisse und Fertigkeiten (s. o.). Holzkamp schreibt dazu in seinen Überlegungen „Zum Problem der Relevanz psychologischer Forschung für die Praxis":

„Wenn der Praktiker nun einen solchen Test benutzen will, so hätte er zunächst zu prüfen, ob die Situationen, an denen der Test validiert worden ist, tatsächlich hinreichende Strukturähnlichkeiten mit den praktischen Situationen hat, über die er mit Hilfe des Tests Aussagen machen will, da nur so die Relevanz jedes möglichen Testbefundes gesichert wäre." (Holzkamp 1972, S. 26)

Da der Praktiker in der Regel kaum Mittel zur Hand hat, „Strukturähnlichkeiten" nachzuweisen, sieht er sich gezwungen, die Befunde als irgendwie aussagekräftig hinzunehmen. Die Konsequenz daraus ist abzusehen: *Es wird versucht, die Wirklichkeit an das Experiment anzupassen, anstatt eine wirklichkeitsgetreue Testsituation zu schaffen.* Welche Folgen eine derartig mißverstandene Testpraxis hat, geht aus einem Sammelreferat von Kirkland (1971) hervor. Darin werden über 200 Untersuchungen aufgeführt, die sich mit den Auswirkungen des Testens in den USA befassen, wo jährlich zwischen 3 und 5 standardisierte Tests pro Schüler durchgeführt werden. Im Falle des Prüfsystems wäre es durchaus denkbar, daß ein verantwortungsbewußter Lehrer weit mehr Zeit für das

[5]) PSB: Prüfsystem für Schul- und Bildungsberatung von W. Horn.

[6]) KLI 4+: Kombinierter Lern- und Intelligenztest von H. Schröder.

[7]) AST 4: Allgemeiner Schulleistungstest für 4. Klassen von F. Fippinger.

Üben mit Zahlenreihen verwendete, um seinen Schülern ein möglichst gutes Abschneiden zu ermöglichen.[8])

Der Autor eines herkömmlichen Tests darf es nicht hinnehmen, wenn 50 % der Schülerschaft in Hinsicht auf ein bestimmtes Merkmal eine „sehr gute" Leistung erbringt. Seine zugrundegelegten Meßmodelle verlangen nämlich eine Normalverteilung der Daten. *Die Gaußsche Glockenkurve scheint wie ein unabwendbares Fatum über unserem Beurteilungssystem zu liegen.*

Dem Lehrer wird nahegelegt, seine Notengebung so einzurichten, daß 5 % Note 1, 15 % Note 2, je 30 % Note 3 oder 4, 15 % Note 5 und 5 % Note 6 erhalten (vgl. Feigel & Keitel 1967, S. 77). Auch hier soll sich die Wirklichkeit an ein von außen herangetragenes Modell anpassen. Daß sich in einem so dynamischen Bereich wie dem menschlichen Lernen eine so starre Festlegung verbietet, erweist unter anderem der bereits erwähnte Bericht von Bloom (1970). Der Glaube versetzt Berge, sagt der Volksmund. Wieviel Wahrheit hinter diesem Sprichwort steckt, beweist ein Experiment von Jacobson & Rosenthal (1970): Ein Team von Wissenschaftlern ließ die Kinder einer Schule einen Begabungstest machen. Anschließend wurden per Zufall die Namen von zehn Kindern jeder Klasse herausgeschrieben und den betreffenden Lehrern gesagt, daß diese Kinder besonders gut abgeschnitten haben – was nicht den Tatsachen entsprach –, und von ihnen in den nächsten Monaten gute Leistungen zu erwarten seien. Am Ende des Schuljahres konnte der Effekt dieser „selbsterfüllenden Prophezeiung" (self-fulfilling prophecy) statistisch nachgewiesen werden: Die fraglichen Kinder waren im Vergleich zu ihren anderen Mitschülern tatsächlich besser geworden.

So unbegründet erscheint jetzt die Aussage des Direktors eines Forschungsprogramms der Universität von Kalifornien (UCLA) gar nicht mehr: *„... jeder Lehrer, der eine Normalverteilung der Leistung seiner Schüler erwartet, sollte auf der Stelle entlassen werden"* (Sheldon 1970, S. 3).

Folgerungen

Eine nüchterne Betrachtung der Erkenntnisse aus den vorausgegangenen Überlegungen läßt, zusammen mit den wahrhaftig nicht wenigen Kritikpunkten, die Rüdiger (1973, S. 178) selbst aufzählt, nur den Schluß zu:

a) *Das Prüfsystem ist weder gerecht noch verläßlich.* Daß es dem Lehrerurteil oder „fragwürdigen handgestrickten Schultests" überlegen ist, konnte bislang nicht nachgewiesen werden. Die Errechnung von Kennwerten der Güte trägt in nur sehr beschränktem Maße zur Aussagekraft bei. Ihre nachträgliche Ermittlung ist sinnlos, wenn zum Zeitpunkt der Bekanntgabe das Gutachten bereits erstellt ist.[9])

b) Infolge seiner mangelnden inhaltlichen Gültigkeit vollzieht das Prüfsystem eine unstatthafte Trennung zwischen Unterricht und Schülerbeurteilung. *Diese Trennung macht ein Lehrer nicht. Er hat bedeutend mehr Daten über den einzelnen Schüler gespeichert,* als dies jegliche Testbatterie erbringen könnte: Er weiß,

☐ wie die eine oder andere Unterrichtseinheit abgelaufen ist,

☐ wie die Schüler sich an diesem Unterricht beteiligt haben,

☐ welche Störungen durch Unterrichtsausfälle, Erkrankungen u. ä. während des Lernprozesses auftraten,

☐ welche Unterrichtsmittel eingesetzt oder nicht verwendet werden konnten,

☐ welche Lerngeschichte der einzelne Schüler aufzuweisen hat,

☐ welchem familiären Milieu der Schüler entstammt.

Vor diesem Hintergrund kann er das beobachtete Lernverhalten in Beziehung zu Lernvoraussetzungen setzen und Erklärungen dafür suchen, und er könnte, so es die organisatorischen Gegebenheiten zuließen, differenzierende pädagogische Maßnahmen vorschlagen oder selbst treffen.

c) Die Nachteile (z. B. Verhinderung curricularer Innovation) des Prüfsystems überwiegen bei weitem dessen angebliche Vorteile. Unverständlich ist – bei allem guten Willen, der ihm unterstellt werden darf –, weshalb die akzeptierte Kritik am System des Übertrittsverfahrens den Herausgeber des Prüfsystems nicht bewog, die selbstgewählte Rolle des Erfüllungsgehilfen einer „ministeriellen Interimsanordnung" (Rüdiger 1973, S. 177) abzulegen und mit Nachdruck dafür zu sorgen, „daß sich ein ‚Schulleistungstest' bei einer sinnvollen, weitsichtigen und entschiedenen Neuregelung von selbst erübrigt" (a. a. O., S. 178). Die Chancen für einen ehrenhaften Rückzug aus dieser Affäre stehen gut: Neueste Presseverlautbarungen verheißen eine Verbannung des Namens „Schulleistungstest" für die regional einheitlichen Probearbeiten. Außerdem soll für eine unauffälligere Integration in den normalen Unterrichtsbetrieb gesorgt werden. Jetzt dem Lehrer zu helfen, seine allseits angezweifelten Beurteilungstechniken zu verfeinern, ist m. E. eine weitaus lohnendere Aufgabe einer Erziehungswissenschaftlichen Fakultät. *Dann wäre der Weg frei für eine Pädagogisierung der Diagnostik: von einem durch das Selektionsprimat geleiteten zu einem vom Gedanken des optimalen Förderns aller Schüler getragenen Unterricht* (vgl. Mauermann, Schulte & Seisenberger 1973). (Hervorhebungen durch Red.) ■

Literaturverzeichnis

BECKER-FREYSENG, W., KRÜGER, K., & RIETBROCK, G.: Materialien zur Kritik und Neuorientierung der Testdiagnostik in der Schule. Arbeiten zur Pädagogischen Diagnostik, Nr. 7. München: Zentrum für Bildungsforschung 1973 (erscheint demnächst als Beitrag in: Diagnostik in der Schule, Reihe PPF, Oldenbourg-Verlag München)

BLOOM, B. S.: Alle Schüler schaffen es. betrifft: erziehung, 1970, 3, Heft 4, 15–27

CRONBACH, L. J., & GLESER, G. C.: Psychological Tests and Personnel Decisions. Urbana: University of Illinois Press 1965

FIPPINGER, F.: Allgemeiner Schulleistungstest für 4. Klassen (AST 4). Beiheft mit Anleitung und Normentabellen. Weinheim: Beltz 1967

FEIGEL, K., & KEITEL, E.: Bayerische Landesvolksschulordnung. München: Beck 1967 (4. Aufl.)

FRENZ, H.-G., KRÜGER, K., & TRÖGER, H.: Die Unangemessenheit der herkömmlichen Testdiagnostik für schulische Entscheidungen. Arbeiten zur Pädagogischen Diagnostik, Nr. 6. München: Zentrum für Bildungsforschung 1973 (erscheint demnächst als Beitrag in: Diagnostik in der Schule, Reihe PPF, Oldenbourg-Verlag München)

HOLZKAMP, K.: Kritische Psychologie. Frankfurt/Main: Fischer 1972

HORN, W.: Prüfsystem für Schul- und Bildungsberatung (PSB). Handanweisung für die Durchführung, Auswertung und Interpretation. Göttingen: Hogrefe 1969

INGENKAMP, K. (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Weinheim: Beltz 1971

JACOBSON, L., & ROSENTHAL, R.: Schüler leisten, was ihre Lehrer von ihnen erwarten. betrifft: erziehung, 1970, 3, Heft 12, 21–25

KIRKLAND, M. C.: The effects of tests on students and schools. Review of Educational Research, 1971, 41, 303–350

LIENERT, G.-A.: Testaufbau und Testanalyse. Weinheim: Beltz 1969 (3. Aufl.)

LIUNGMAN, C. G.: Der Intelligenzkult. Hamburg: Rowohlt 1973 (rororo-Sachbuch 6792)

MAUERMANN, L.: Brauchen wir Tests? Bayerische Schule 1972, 25, 106

MAUERMANN, L., SCHULTE, H.-J., & SEISENBERGER, G.: Anmerkungen zu einer pädagogischen Orientierung der Schülerbeurteilung. Arbeiten zur Pädagogischen Diagnostik, Nr. 5. München: Zentrum für Bildungsforschung 1973 (erscheint demnächst als Beitrag in: Diagnostik in der Schule, Reihe PPF, Oldenbourg-Verlag München)

RÜDIGER, D.: Ein „Übertritts-Prüfsystem" im Ad-hoc-Einsatz. Heimatliche Schule, 1971, 132–137

RÜDIGER, D. (Hrsg.): Schulleistungs-Prüfsystem für den Übergang an weiterführende Schulen. Anleitung. Regensburg: Wolf 1972 (Dezember)

RÜDIGER, D.: „Schulleistungstests" – ein Weg zur Verbesserung der Übertrittsauslese. Bayerische Schule, 1973, 26, 175–178

SCHRÖDER, H.: Kombinierter Lern- und Intelligenztest (KLI 4+). Beiheft mit Anleitung und Normentabellen. Weinheim: Beltz 1968

SCHRÖTER, G.: Die ungerechte Aufsatzzensur. Bochum: Kamp 1971

SHELDON, M. S.: Entrance and placement testing for the Junior College. Junior College Research Review, 1970, 5, Heft 4

[8]) Diese Annahme legt ein Studium der Normentabellen zum Prüfsystem von 12/1971 und 12/1972 nahe: 1971 lösten 2,6 % von 1366 Schülern keine einzige der 16 Zahlenreihen, 1972 waren es 6,9 % von 1234. Das deutet auf eine Erhöhung der Schwierigkeit hin. Der Mittelwert lag jedoch 1971 bei 7,2, 1972 bei 8,2 Punkten. Das bessere Abschneiden trotz größerer Schwierigkeit könnte u. U. mit dem Übungseffekt in den Klassen von 1972 erklärt werden.

[9]) Welche Konsequenzen werden gezogen, wenn sich herausstellen sollte, daß die Gütekriterien – etwa bei der Durchführung im April 1973 – aus unerklärlichen Gründen **unter** die Kennwerte vergleichbarer qualifizierter Übertrittstests gesunken sind?