Diether Hopf

Pädagogische Diagnostik

first published in:
Die Psychologie des 20. Jahrhunderts: Bd. 2: Konsequenzen für die Pädagogik; Entwicklungsstörungen und therapeutische Modelle / hrsg. von Walter Spiel. - Zürich : Kindler, 1980, S. 896-919

Postprint published at the Institutional Repository of Potsdam University:
In: Postprints der Universität Potsdam
Humanwissenschaftliche Reihe ; 102
http://opus.kobv.de/ubp/volltexte/2009/3639/
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-36393

Postprints der Universität Potsdam
Humanwissenschaftliche Reihe ; 102

PÄDAGOGISCHE DIAGNOSTIK

von Diether Hopf

EINFÜHRUNG

Lehren und Lernen in der Schule wären undenkbar, wenn Schülern und Lehrern keine Informationen über die Vorkenntnisse, die Lernfortschritte und die Lernergebnisse des Schülers sowie die Bedingungen, unter denen eine Leistung zustande gekommen ist, zur Verfügung stünden. Wie unvollkommen solche Informationen auch sein mögen, sie steuern das Verhalten von Lehrern und Schülern im Unterricht und bilden darüber hinaus die Grundlage für wichtige Entscheidungen über die Schullaufbahn des Schülers. Schulleistungen zu beurteilen, gehört zum Alltagsgeschäft des Lehrers. Die meisten Urteile veralten rasch, weil die Schüler sich verändern, weil wechselnde Inhalte ständig neue Anforderungen stellen und unterschiedliche Vorkenntnisse abrufen. Jeder Schultag ist daher für jeden einzelnen Schüler durchsetzt von zahlreichen Urteilen über seine Leistungen, die meist vom Lehrer ausgehen. Zählt man nur die Klassenarbeiten und Urteile über mündliche Leistungen zusammen, so werden in den öffentlichen Schulen der deutschsprachigen Länder täglich über eine Million Leistungsurteile abgegeben. Diese Zahl würde sich wesentlich erhöhen, wenn man die unausgesprochenen Leistungsurteile hinzurechnete, die sich ergeben, wenn ein Lehrer zum Beispiel mit der verbreiteten Methode des Frage-Antwort-Unterrichts in engen Schritten einen neuen Sachverhalt einführt, indem er auf jede Frage nur diejenigen Antworten akzeptiert, die seinen Vorstellungen von der Entwicklung des Gedankens entsprechen, die übrigen aber kommentarlos ablehnt.

Pädagogische Diagnostik beschäftigt sich mit den Vorgängen, Verfahren und Maßnahmen, die der Messung und Beurteilung des Lernens und der Lernumwelt dienen. Sie richtet sich nicht nur auf die Beschreibung individueller Merkmale und Zustände und auf die Erfassung individueller Unterschiede in der Schulleistung, wie es Ulich u. Mertens (1973, 9) in Analogie zur psychologischen Diagnostik formuliert haben, sondern auch auf die Identifikation ihrer Ursachen, seien sie in der schulischen Institution, in der Familie des Schülers, in den Einflüssen der Gleichaltrigen oder in der Qualität des Unterrichts zu suchen. Der Begriff Pädagogische Diagnostik wird hier deshalb weit gefaßt, weil Leistungsurteile sich nicht auf ein Phänomen beziehen, welches einen eindeutig faßbaren Gegenstand darstellt, der unmittelbar am Verhalten eines Schülers erkennbar wäre und über den unterschiedliche Beurteiler stets zum gleichen Ergebnis kommen würden, sondern weil sie ein Konstrukt interpretieren: »Schulleistung« ist eine gedankliche Größe, die in einem noch weithin ungeklärten Verhältnis zu den Indikatoren steht, aus welchen man auf sie schließt. Schulnoten zum Beispiel, Testergebnisse oder Einschätzungen von Lehrern sind solche Indikatoren für Schulleistung. Nun läßt sich ein Konstrukt nicht ein für allemal und endgül-

tig bestimmen, sondern nur durch darauf bezogene Forschung zunehmend präzisieren und inhaltlich füllen. Je konsistenter es sich am Verhalten festmachen läßt und je weniger es abhängig ist von den Methoden, die zu seiner Messung verwendet werden, als desto besser gesichert kann es gelten. »Intelligenz« beispielsweise ist eines der am besten gesicherten psychologischen Konstrukte, weil unterschiedliche Tests relativ ähnliche Intelligenzquotienten auf relativ zuverlässige Weise liefern und weil zahlreiche empirische Studien ihr Verhältnis zu anderen Variablen und individuellen Merkmalen in einer Weise beleuchteten, die das Konstrukt bereichert und im Detail modifiziert, im Kern aber unberührt gelassen haben als die »zusammengesetzte oder globale Fähigkeit des Individuums, zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinanderzusetzen«, wie Wechsler (1958) es formuliert hat. Die Interpretation der Intelligenztestergebnisse und ihre Nutzung für Entscheidungen unterliegt freilich in den letzten Jahrzehnten dramatischen Schwankungen, je nachdem, ob die Beurteiler sie als »Substanz« (und dann gewöhnlich als genetisch fixiert) verstehen oder ob sie sich ihres psychologischen Konstruktcharakters bewußt sind (und damit meist ihre Genese und Lernbarkeit stärker berücksichtigen).

Indikatoren für Schulleistung sind vor allen Dingen die Zensuren, die von Lehrern für schriftliche oder mündliche Leistungen erteilt werden, Gutachten, informelle Einschät-

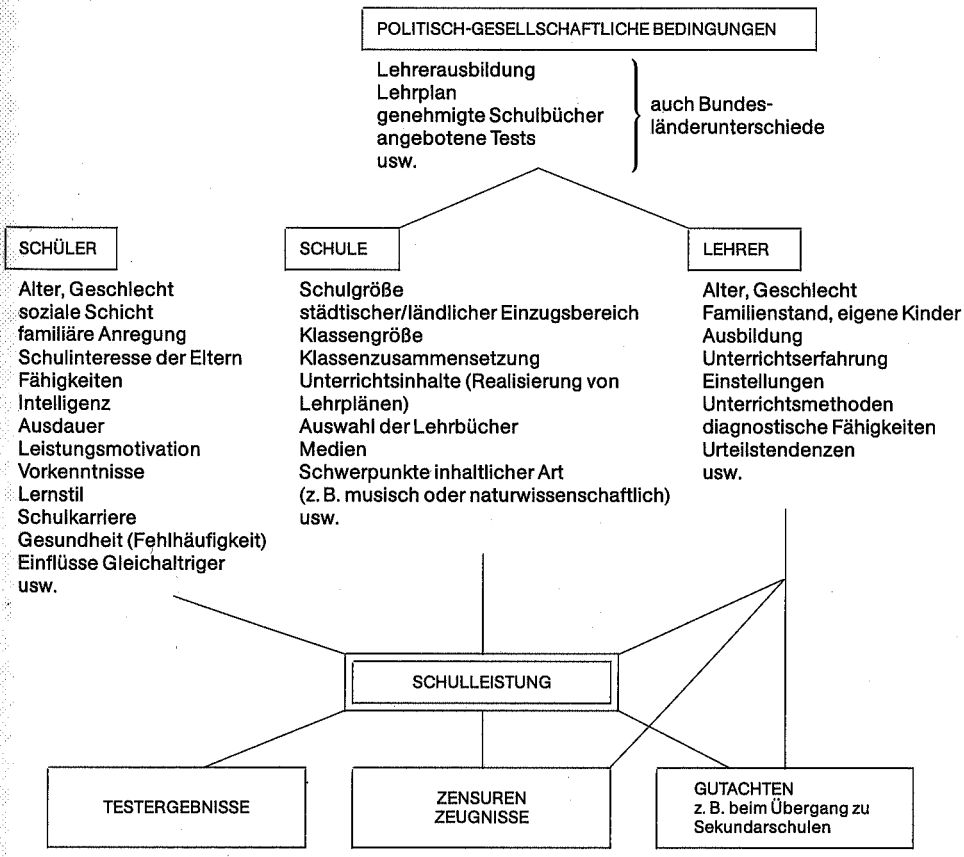

Abb. 1

zungen oder die Ergebnisse von Schulleistungstests. Das Verhältnis zwischen dem Konstrukt Schulleistung und seinen Indikatoren ist insbesondere hinsichtlich der Schulnoten noch bei weitem nicht voll aufgeklärt. Soviel aber ist gewiß, daß zwischen Schulleistung und Zensur kein direktes Verhältnis besteht, daß also Schulleistungen sich nicht unmittelbar, vollständig und ohne Verzerrung in den Schulnoten abbilden. Die Ursache hierfür liegt in der Komplexität und vielfältigen Bestimmtheit des Sachverhalts, den wir mit Schulleistung bezeichnen. Denn am Zustandekommen der Schulleistung ist mittelbar oder unmittelbar nicht nur der Schüler, sondern auch der Lehrer, der Unterricht, die Schule und die Gesellschaft beteiligt, und auch das, was der Schüler in die Leistung einbringt, stammt nicht nur von ihm als Individuum. Abb. 1 mag dies illustrieren.

Die Darstellung deutet nur an, von wie vielen Faktoren Schulleistungen beeinflußt werden. Denn neben den direkten und teilweise unmittelbar plausiblen Zusammenhängen muß man mit einer großen Zahl von Wechselwirkungen rechnen, über deren Stärke und Richtung erst spärliche Untersuchungen vorliegen. So gibt es beispielsweise Zusammenhänge zwischen Vorkenntnissen und Lernstilen der Schüler einerseits und Unterrichtsmethoden des Lehrers andererseits, weil nicht jede Methode in jeder Lage auch für jeden Schüler die richtige ist (vgl. u. a. Flammer 1975). Eine große Zahl von Schülern erreicht aufgrund der nicht optimalen Passung der Unterrichtsmethoden ihres Lehrers zu den ihnen angemessenen Formen des Lernens bei weitem nicht die Leistung, die sie bei anderer Konstellation hätten erreichen können. Läßt sich »ihre« Schulleistung dann so beurteilen und interpretieren, als seien sie allein für ihr Zustandekommen verantwortlich? Derartige Bedingungen gehen im übrigen nicht nur in die Zensuren, sondern auch in die Ergebnisse von Schulleistungstests ein.

An Hand der oben wiedergegebenen schematischen Darstellung sei hier abschließend nochmals die weite Fassung des Begriffs Pädagogische Diagnostik unterstrichen, nach der diese sich beispielsweise auch auf die Beschreibung von Lehrerverhalten sowie die Evaluation von Curricula, kurz auf die Messung und Beurteilung von »Lernumwelten« bezieht, weil nicht nur individuelle, sondern auch situative Merkmale und Unterschiede erfaßt sein müssen, ehe diagnostische Befunde sinnvoll interpretierbar werden. Dieser Gesichtspunkt wird in den folgenden Abschnitten, in welchen das Verhältnis von Pädagogischer Diagnostik und Lernen sowie die Möglichkeiten und Grenzen der Pädagogischen Diagnostik bei der Schülerauslese und Prognose zukünftiger Leistung im Vordergrund stehen, mehrfach wiederkehren.

PÄDAGOGISCHE DIAGNOSTIK UND SCHULISCHES LERNEN

Beurteilung und Bewertung stellen zentrale Elemente des schulischen Lernens dar und üben großen Einfluß auf Lernprozeß, Lernergebnis und auf das weitere Lernen aus. Insbesondere beeinträchtigen wiederholte Mißerfolge, die sich in schlechten Zensuren ausdrükken, nachhaltig die Motivation zu weiterem Lernen. 75 Prozent aller schlechten Noten aber betreffen nach einer groben Schätzung immer wieder dieselben 25 Prozent der Schüler, so daß man mit einem gewissen Recht von dem »schlechten Schüler« wie von einem Typus gesprochen und die negativen Begleit- und Folgewirkungen dieser Einstufung beschrieben hat (vgl. z. B. Höhn 1967). Umgekehrt können gute Zensuren als Erfolg erlebt werden und zu einer Steigerung der Anstrengungsbereitschaft und Ausdauer führen, was wiederum die Wahrscheinlichkeit erhöht, auch weiterhin gute Noten zu erhalten. Aber nicht nur nach Abschluß einer Lernsequenz üben Leistungsurteile große Wirkung auf das weitere Lernen aus, sondern sie können auch während des Lernprozesses als Orientierungsmarken dienen und so den Lernverlauf nachhaltig positiv oder negativ beeinflussen.

Die Zensuren in der Schule beziehen sich, wenn man von den »Kopfnoten« im Zeugnis

beispielsweise über Betragen, Mitarbeit oder Fleiß absieht, auf die Resultate »kognitiven« Lernens. Daß daneben in der Schule und im Unterricht auch ganz andere Dinge gelernt werden, wie zum Beispiel Selbstvertrauen, Angst, Kooperation, Unterordnung, daß also auch »soziales« und »affektives« Lernen hier seinen Platz hat – und sei es auch unbeabsichtigt –, steht außer Frage (vgl. z. B. Jackson 1973); viele Betrachter der Vorgänge in der Schule halten den damit angesprochenen Bereich des sogenannten hidden curriculum für eher noch wichtiger und wirksamer als den Erwerb von Kenntnissen und Fertigkeiten, der auf den ersten Blick im Vordergrund des Schulunterrichts zu stehen scheint. Es würde allerdings zu einer künstlichen Unterscheidung eng miteinander zusammenhängender Lernbereiche führen, würde man sie getrennt voneinander behandeln. Denn es zeigt sich beispielsweise, daß fundamentale kognitive Fortschritte oft gerade in Situationen erzielt werden, in welchen »soziale« Probleme der Interaktion mit subjektiv wichtigen Personen gelöst werden müssen, es also nicht um intentionales kognitives Lernen geht (vgl. z. B. Krappmann 1973). Der Leser wird deshalb im folgenden auf Beispiele aus allen drei genannten, traditionell unterschiedenen Lernbereichen treffen.

Zur Ordnung des komplexen Gefüges, als welches sich das Lernen in der Schule darstellt, sei ein von Carroll (1963) vorgelegtes »Modell schulischen Lernens« eingeführt, mit dessen Hilfe sich die Aufgaben der Pädagogischen Diagnostik gut beschreiben lassen.

Carroll unterscheidet fünf Variablen, deren Kenntnis es ermöglicht, schulisches Lernen sowie den Erwerb von Fertigkeiten zu erklären. Die Variablen lassen sich in Carrolls Lernmodell weitgehend in Zeiteinheiten ausdrücken. Insbesondere gilt dies für die zunächst genannten drei Variablen: Fähigkeit, Ausdauer und Gelegenheit zum Lernen.

Die *Fähigkeit* eines Schülers wird ausgedrückt in der Zeit, die er braucht, um unter optimalen Lernbedingungen bei einer bestimmten Aufgabe ein festgelegtes Kriterium zu erreichen. Unter diesen Voraussetzungen, d. h. also, wenn sowohl die Qualität des Unterrichts als auch die Motivation des Schülers, das Kriterium zu erreichen, hoch sind, gibt die Lerngeschwindigkeit Auskunft über die Lernfähigkeit des Schülers.

Die *Ausdauer* erweist sich an der Zeitdauer, die ein Schüler mit Lernen zu verbringen bereit ist. Der Lernerfolg liegt nach den Annahmen des Modells um so höher, je mehr Zeit auf Lernen angewendet wird, welches auch immer die Ursachen für die Ausdauer des Schülers sein mögen.

Die *Gelegenheit zum Lernen* ist als die Zeit definiert, die zum Lernen einer bestimmten Aufgabe tatsächlich zur Verfügung steht. Selbst bei hoher Motivation wird der Lernerfolg ausbleiben, wenn die Zeit zu kurz bemessen ist. Dies dürfte eine insbesondere für die langsam lernenden Schüler ziemlich normale Situation im Klassenunterricht darstellen, bei dem der Lehrer sein Unterrichtstempo nur auf eine bestimmte, beispielsweise die mittlere Leistungsgruppe zuschneidet (vgl. Dahllöf 1973).

Mit *Qualität des Unterrichts* ist der Grad der Strukturiertheit der Unterrichtsinhalte und die Angemessenheit der Unterrichtsmethode für den Schüler gemeint. Hierzu gehören beispielsweise die richtige und verständliche Erklärung der Inhalte, die Wahl adäquater Methoden bei der Einführung eines neuen Sachverhalts, optimale Formen von Wiederholung, Übung und Feedback sowie die Feststellung und Berücksichtigung besonderer Schwierigkeiten, die bei einzelnen Lernschritten auftauchen.

Die fünfte und letzte Variable in dem Modell schulischen Lernens besteht in der *Fähigkeit, den Unterricht zu verstehen.* Nicht optimale Qualität des Unterrichts erfordert von den Schülern zusätzliche Anstrengungen, um dem Unterricht zu folgen. Fällt dies dem Schüler schwer, kann sich der Lernerfolg erheblich verringern, ja er kann sogar ganz ausbleiben. Untersuchungsbefunde deuten darauf hin, daß sich Unterschiede in der Art und Qualität des Unterrichts auf Schüler mit hoher verbaler Intelligenz nur gering auswirken, daß sie jedoch von erheblicher Bedeutung sind für Schüler mit einer gering ausgeprägten Fähigkeit,

den Unterricht zu verstehen, auch wenn sie über ein hinreichendes Maß an Ausdauer und sonstigen Fähigkeiten verfügen und genügend Gelegenheit zum Lernen gegeben ist.

In Carrolls Modell werden die genannten Variablen in differenzierter Weise zueinander in Beziehung gesetzt. Es kann hier nicht die Aufgabe sein, die zahlreichen Anregungen für die Gestaltung des schulischen Lernens, die dort erörtert werden, im einzelnen zu diskutieren. Die theoretische wie praktische Fruchtbarkeit des Modells hat sich jedenfalls mehrfach bestätigt; so hat beispielsweise Bloom (1971) aus dem Carrollschen Modell das Konzept des »zielerreichenden Lernens« (mastery learning) abgeleitet, dessen wichtigstes Merkmal darin besteht, durch Variation der Variable Gelegenheit zum Lernen dafür zu sorgen, daß nicht mehr nur etwa die Hälfte, sondern 80 bis 90 Prozent einer Lerngruppe ein bestimmtes Kriterium erreichen (vgl. z. B. Block 1971). Nach Carrolls Modell des schulischen Lernens bestimmt sich der Grad des Lernerfolgs, d. h. der Grad, in dem ein Schüler bei einer bestimmten Aufgabe ein bestimmtes Kriterium erreicht, als eine Funktion der Zeit, die der Schüler tatsächlich mit dem Lernen der Aufgabe verbringt, im Verhältnis zu der Zeit, die er angesichts seiner Fähigkeit, der Qualität des Unterrichts und der Fähigkeit, den Unterricht zu verstehen, benötigen würde, um das Kriterium zu erreichen.

An dem Modell des Lernens in der Schule läßt sich erkennen, auf wie vielfache Weise die Leistung eines Schülers determiniert ist, selbst wenn man nur den Mikrokosmos des Unterrichtsgeschehens und einige individuelle Merkmale betrachtet. Ein Urteil über die Leistungen eines Schülers wäre jedenfalls wenig zuverlässig, genau und aussagekräftig, wenn nicht die fünf im Modell genannten Variablen bei seinem Zustandekommen mitbedacht würden. Daß selbst dies nicht ausreicht, um insbesondere den Ursachen für Mißerfolg auf die Spur zu kommen und um gezielte pädagogische Maßnahmen zu ihrer Überwindung zu ergreifen, zeigt ein Blick auf die oben gegebene schematische Darstellung (Abb. 1), in der die von Carroll genannten Variablen nur einen Teil der uns bekannten Determinanten von Schulleistung umschreiben. Immerhin handelt es sich bei ihnen um die wichtigsten direkten und unmittelbaren Einflußgrößen, so daß die Erörterung ihrer Implikationen für pädagogisch-diagnostische Fragen von besonderer Bedeutung ist.

Im folgenden sollen die einzelnen Variablen genauer erläutert werden. Zur Vorbereitung und Durchführung seines Unterrichts, insbesondere aber zur Planung individueller Lernhilfen muß der Lehrer die Fähigkeiten seiner Schüler kennen. Nach dem Modell Carrolls lassen sich diese bestimmen als die Zeitdauer, die ein Schüler braucht, um unter sonst optimalen Bedingungen eine Aufgabe zu lösen. Der Zeitbedarf verschiedener Schüler zur Bewältigung einer bestimmten Aufgabe variiert, wie jeder Lehrer weiß, in einer Klasse beträchtlich, und zwar nicht nur in heterogenen Schulklassen, wie man sie beispielsweise in der Grundschule oder in der Gesamtschule antrifft, sondern auch in vermeintlich homogenen Lerngruppen wie den Schulklassen im Gymnasium oder in der Hauptschule. Wenn der Lehrer diese Unterschiede unberücksichtigt läßt, wird er bei den einen Schülern Langeweile mit nachfolgender Unlust und Desinteresse erzeugen oder aber seine Schüler überfordern, so daß sie nach kurzer Zeit den Versuch aufgeben, die gestellten Aufgaben zu bearbeiten. Wenn man gerade den langsam lernenden Schülern Mißerfolge ersparen will, wird man ihnen daher mehr Gelegenheit zum Lernen geben müssen. Welche erstaunlichen Erfolge erzielt werden können, wenn der Zeitfaktor den Lernfähigkeiten der Schüler entsprechend variiert wird, zeigen die bei Block (1971) wiedergegebenen Berichte über Erfahrungen mit zielerreichendem Lernen, bei denen es nur noch wenige Prozent der Schüler sind, welche die jeweiligen Unterrichtsziele und -inhalte nicht bewältigen.

Eine große Zahl von Forschungsbefunden deuten nun allerdings darauf hin, daß man kaum von einer »allgemeinen Lernfähigkeit« sprechen kann, sondern daß die Lernfähigkeit stärker aufgabenspezifisch ist, als man gemeinhin angenommen hat (vgl. z. B. Carroll 1970; Lindvall, Bolvin 1967; Roeder 1974). Unterschiedliche Lerninhalte bedingen offenbar un-

terschiedliche Lernverläufe. Der Zuwachs an Kenntnissen in einer bestimmten Zeiteinheit bei einem Schüler wird also unterschiedliche pädagogische Konsequenzen haben müssen, je nachdem ob es sich beispielsweise um ein Gebiet handelt, bei dem man mit linear ansteigenden Lernkurven rechnen kann, oder um einen Aufgabenbereich, bei dem der Lernzuwachs im Laufe der Zeit immer geringer wird. Genaugenommen müßte man demnach einerseits wissen, wie die Lernverläufe bei dem jeweils bearbeiteten Gebiet aussehen, andererseits aber auch genau diagnostizieren können, an welchem Punkt im Verlauf des Lernprozesses ein Schüler steht, um seine Lernfortschritte interpretieren zu können. Ein wichtiges Nebenprodukt dieser Überlegungen besteht darin, daß Intelligenztests als ungeeignete Verfahren für die genannten Zwecke angesehen werden müssen, weil sie die Aufgabenspezifität der Lernfähigkeit nicht berücksichtigen. Langfristige Schulversuche haben, ganz entsprechend diesen Überlegungen, bestätigt, daß es keinen Zusammenhang zwischen der Geschwindigkeit des Lernfortschritts und der Intelligenz gibt, sobald man die Aufgabenspezifität der Lernfähigkeit berücksichtigt und dementsprechende Gelegenheit zum Lernen bietet – ein Befund, dessen Bedeutung angesichts der im Kontext des traditionellen Unterrichts immer wieder bestätigten Korrelation zwischen Intelligenz und Schulleistung kaum überschätzt werden kann (vgl. hierzu Michel 1960, Schwarzer 1972, Lohnes 1973). Demgegenüber helfen Informationen über die verbale Intelligenz von Schülern bei der Einschätzung ihrer Fähigkeit, dem Unterricht zu folgen. Viel häufiger nämlich, als Lehrer gemeinhin annehmen, sind die Ursachen für schulischen Mißerfolg darin zu suchen, daß der Lehrer bei der Einführung eines neuen Sachverhalts die Schüler nicht durch die Schwierigkeit des Gegenstandes, sondern durch die vermeidbare Kompliziertheit seiner Erklärungen überfordert.

Wenn ein Schüler nicht über die nötige Ausdauer oder die Motivation verfügt, so lange zu lernen, bis er das vorgegebene oder selbst gesetzte Ziel erreicht hat, wird sein Lernerfolg auch bei hinreichenden Fähigkeiten und ausreichender Gelegenheit zum Lernen hinter seinen Möglichkeiten zurückbleiben. Die Ursachen für Motivationsdefizite sind freilich vielfältig (vgl. z. B. Heckhausen 1972 sowie Knörzer in diesem Bd.), und es wäre für den Lehrer eine unschätzbare Hilfe, die potentielle Ausdauer, d. h. also die maximalen Werte, die ein Schüler ereichen könnte, zu kennen sowie die Gründe zu wissen, warum ein Schüler seine Möglichkeiten nicht ausschöpft, da es viele Wege gibt, um Schüler zu motivieren, sei es durch die Veränderung der Unterrichtsmethode oder durch Präsentation besonders attraktiver Materialien oder die Wahl anderer Lernformen. Auch hier gibt es beträchtliche individuelle Unterschiede bei den Schülern, da diese sich nicht nur nicht für jedes Fach gleichermaßen interessieren, sondern gewöhnlich auch nur an bestimmten, ausgewählten Themen innerhalb eines Faches wirkliches Interesse gewinnen und dann auch ausdauernd lernen werden. Zudem sind Maßnahmen des Lehrers zur Verbesserung der Lernmotivation von Schülern nicht bei allen Schülern gleichermaßen wirksam, und auch hier wäre dem Lehrer und den Schülern sehr damit geholfen, wenn eine differenzierte Diagnose möglich wäre. Freilich stehen derzeit diagnostische Verfahren, die diesen Zwecken dienen könnten und sich ohne übermäßigen Aufwand im Unterricht verwenden ließen, nicht zur Verfügung (vgl. Veroff 1973), so daß der Lehrer hier auf seine Erfahrung und die gründliche Kenntnis seiner Schüler angewiesen ist. Aus diesen Überlegungen folgt im übrigen – dies sei hier in Parenthese vermerkt, gilt aber in entsprechender Weise für die übrigen Variablen –, daß unsere Schulzensuren höchst komplexe Urteile darstellen, an deren Zustandekommen der Lehrer unter Umständen mehr Anteil hat als der Schüler.

Wie oben bereits erwähnt wurde, wirken sich Mängel in der Qualität des Unterrichts stärker auf solche Schüler aus, deren Fähigkeit, dem Unterricht zu folgen, unterentwickelt ist. Aus diesem und aus vielen anderen Gründen dürfte die Qualität des Unterrichts eine besonders wichtige Ursache für unbefriedigende Lernerfolge darstellen. Guter Unterricht

setzt darüber hinaus, wie jeder Lehrer weiß, erhebliche Kräfte bei den Schülern frei, sich den Anstrengungen des Lernens zu unterziehen. Im Unterschied zu der großen Zahl hochentwickelter Meßinstrumente für die Erfassung individueller Merkmale gibt es aber kaum brauchbare Verfahren, um die Qualität des Unterrichts eines Lehrers einzuschätzen (vgl. z. B. Rosenshine 1970, Gage 1972). Obwohl sowohl die Alltagserfahrung wie auch Carrolls Modell schulischen Lernens auf die Bedeutsamkeit dieser Variable hinweisen, ist es bisher nur in Ausnahmefällen gelungen, Merkmale des Unterrichts zu Lernerfolgen der Schüler in Beziehung zu setzen. Aber in diesem Bereich hat die Forschung kaum erst begonnen, was vermutlich seine Ursache darin hat, daß man seit jeher gewohnt ist, Mißerfolge und Fehler in den Leistungen der Schüler nicht auf Defizite in der Fähigkeit und auf das Verhalten des Lehrers zurückzuführen, sondern den Schülern anzulasten. Eine weitere Ursache für den Mangel an diagnostischen Instrumenten zur Erfassung der Qualität des Unterrichts dürfte in dem Umstand zu suchen sein, daß die Identifikation voneinander klar unterscheidbarer Unterrichtsmethoden auf Schwierigkeiten stößt, weil der in den Schulen konkret vorfindbare Unterricht eine schwer entwirrbare Mischung von Elementen theoretisch unterscheidbarer Unterrichtsmethoden darzustellen scheint (vgl. Hopf 1980).

Pädagogische Interventionen werden um so wirksamer sein, je genauere Diagnosen über die Ausprägung der genannten fünf Variablen dem Lehrer zur Verfügung stehen, weil sie ihm die Möglichkeit geben, die Genese einer schulischen Leistung besser zu verstehen und den Gründen für Erfolg und Mißerfolg auf die Spur zu kommen. An dem geschilderten Modell läßt sich erkennen, daß unsere Schulzensuren ein so komplexes, intern und extern determiniertes Phänomen wie die Schulleistung nicht in einer Weise abbilden können, die es erlaubt, jedem Schüler die ihm angemessene Förderung zuteil werden zu lassen, sondern daß der Lehrer gegenwärtig nur mit globalen, undifferenzierten pädagogischen Maßnahmen arbeiten kann, aus denen jeder einzelne Schüler sich selbst das ihm Zugängliche und Angemessene heraussuchen muß. Prüfungen, Zensuren, Zeugnisse, Gutachten, standardisierte oder informelle Tests richten sich nämlich auf keine der genannten Variablen direkt, sondern auf ein von diesen auf undurchsichtige Weise bedingtes Produkt.

Die traditionelle Leistungsbeurteilung versucht, die Frage zu beantworten, ob und in welchem Grade ein Schüler eine gestellte Aufgabe gelöst oder ein Lernziel erreicht hat. Hierfür ist eine wichtige, gleichwohl durchaus nicht selbstverständliche Voraussetzung die Reflexion über die jeweilige Lernaufgabe, die oft in einer präzisen Beschreibung ihrer Komponenten sowie der Lernziele in einer überprüfbaren Form bestehen wird. Denn ohne die genaue Passung des Meßinstruments – worunter hier alle Verfahren wie Klassenarbeiten, mündliche Prüfungen oder auch standardisierte Tests verstanden werden – zu den Zielen und Inhalten des Unterrichts besteht nicht einmal die Möglichkeit, über die Produkte des Lernens, wie immer sie zustande gekommen sein mögen, gültige Auskunft zu erhalten. Zu wie unterschiedlichen Ergebnissen eine Messung führen kann, mag ein bei Mager (1965) erwähntes Beispiel illustrieren: »Ein Lehrer äußert gelegentlich, daß er es als ein wichtiges Lernziel ansieht, das Musikverständnis seiner Schüler zu entwickeln. Was aber bedeutet das? Wie sollen sich Schüler verhalten, wenn sie das Lernziel erreicht haben?

Einige mögliche Antworten wären:

a) der Lernende seufzt ekstatisch, wenn er Bach hört;

b) der Lernende kauft eine Hi-Fi-Einrichtung und Schallplatten im Wert von 1500 DM;

c) der Lernende beantwortet 95 Auswahl-Antwort-Fragen zur Musikgeschichte richtig;

d) der Lernende schreibt einen flüssigen Aufsatz über die Bedeutung von sieben Opern;

e) der Lernende sagt: ›Mann, glaub mir, ich bin Fachmann. Es ist einfach großartig!‹«

Dieses vielleicht etwas überzogene Beispiel macht gleichwohl deutlich, wie unterschiedlich Lernerfolg bei gleichlautenden Zielvorstellungen aussehen kann, wenn auf die Präzisierung der Unterrichtsabsichten zuwenig Mühe verwendet wurde. Man muß freilich da-

von ausgehen, daß zwischen Lehrern in verschiedenen Schulen, ja sogar innerhalb derselben Schule erhebliche Unterschiede bezüglich der Zielvorstellungen im selben Fach bestehen werden, so daß Zensuren über den Lernerfolg eines Schülers sehr Verschiedenes bedeuten können. Das Beispiel zeigt darüber hinaus, daß Schüler mit unterschiedlichen Interessen in demselben Fach je nach den Zielvorstellungen des Lehrers sich angesprochen oder abgestoßen und gelangweilt fühlen werden. Denn der Lehrer wird im Normalfall diejenigen Leistungen hoch bewerten, die seinen Zielvorstellungen des Unterrichts entsprechen.

Wir können hier festhalten, daß zur Beurteilung von Schulleistungen im Kontext des traditionellen Unterrichts ein großes Arsenal diagnostischer Verfahren notwendig wäre, wenn die Urteile die Grundlage für pädagogische Maßnahmen bilden sollen, die der individuellen Förderung des Schülers dienen. Zugleich zeigt sich schon auf den ersten Blick, daß die traditionellen Leistungsurteile für den gezielten Ausgleich von Lerndefiziten und die Steuerung des Lernprozesses entsprechend der spezifischen Lage eines Schülers keine zureichende Grundlage darstellen können. Hierfür müßten vielmehr differenzierte Verfahren zur Messung des aktuellen, individuellen Zustandes bezüglich der Lernfähigkeit, den Lernvoraussetzungen, der Ausdauer oder der Fähigkeit, dem Unterricht zu folgen, zur Verfügung stehen, darüber hinaus aber auch Informationen über »Lernumwelten«, beispielsweise also die Unterrichtsmethoden und Unterrichtsgestaltung durch den Lehrer, die lernwirksamen Gruppenprozesse in der Klasse sowie die Lernbarrieren, die außerhalb des Unterrichts liegen und Ursache für Mißerfolg beim schulischen Lernen sein können. Ferner haben wir gesehen, daß Schulnoten von Schulklasse zu Schulklasse, aber auch von Schüler zu Schüler Unterschiedliches bedeuten, weil einerseits innerhalb desselben Faches und derselben Schulform Ziele und Inhalte des Unterrichts je nach Lehrer variieren und andererseits dasselbe Verhalten eines Lehrers auf den einen Schüler fördernd, auf den anderen hemmend wirken kann. Hier liegen die Hauptursachen für die vielfach beschriebene Fragwürdigkeit der Zensurengebung, die insbesondere von Ingenkamp (z. B. 1966, ^{6}1976) ausführlich dargestellt worden ist.

PÄDAGOGISCHE DIAGNOSTIK UND SCHULISCHE AUSLESE

Wir haben gesehen, daß Pädagogische Diagnostik für das Lernen in der Schule von außerordentlich großer Bedeutung ist, daß die gegenwärtig verwendeten Beurteilungsformen und -verfahren jedoch in unzureichendem Maße der Förderung des Lernens dienen, vielmehr häufig zu einer mit der Zeit wachsenden Entmutigung zahlreicher Schüler führen. Dies wird sich auch so lange nicht ändern, wie Zensuren auf Lernprodukte (und nicht Lernprozesse) gerichtet sind und somit keine Hinweise für das weitere Lernen geben, und wie Lehrer und Schulverwaltung an der Vorstellung festhalten, Zensuren müßten sich annähernd normal verteilen. Wie immer der absolute Kenntnisstand einer Klasse aussehen mag, bei Normalverteilung der Zensuren liegt zwangsläufig etwa die Hälfte der Schüler unter dem Durchschnitt, und etwa jeder dritte Schüler erzielt nach den klassenimmanenten Standards nichtausreichende Ergebnisse. Die naheliegende Überlegung, daß die Förderung des Lernens Hauptaufgabe der Schule sei, Unterricht und Leistungsbewertung aber der bestmöglichen Erfüllung dieses Zieles dienen und somit in einer Form realisiert werden sollten, die sich an den Erfordernissen des übergeordneten Zieles orientiert, scheint, was die Pädagogische Diagnostik betrifft, weithin aus dem Blickfeld geraten zu sein; andernfalls wäre kaum verständlich, warum täglich zahllose Kinder vermeidbar entmutigt werden, und warum Unterrichtsmodelle, in denen Mißerfolgserlebnisse minimiert und zugleich insgesamt bessere Lernergebnisse erzielt werden, weitgehend unbeachtet bleiben (vgl. z. B. Lindvall, Bolvin 1967; Bloom 1973).

Vom Standpunkt der Schul- und Unterrichtsziele her betrachtet, dürfte die Förderung

des Lernens als die wichtigste Aufgabe der Pädagogischen Diagnostik bezeichnet werden. De facto scheint aber, jedenfalls was die individuellen Folgen betrifft, von gleichrangiger, wenn nicht sogar größerer Bedeutung ihre Verwendung zu Zwecken der Auslese und der Prognose zukünftiger Schulleistung zu sein.

Ausleseentscheidungen begleiten den Schüler durch seine gesamte Schulzeit. So wird zunächst in Schulreifeuntersuchungen überprüft, ob ein Kind den Anforderungen der Schule entspricht; bei nicht gegebener Schulreife wird es zurückgewiesen und dabei entweder einem Schulkindergarten zugeführt, in welchem Maßnahmen spezifischer Förderung und Vorbereitung auf die Schule vorgesehen sind, oder aber, in der Mehrzahl der Fälle, zurück in seine Familie verwiesen, wo die Chancen des Ausgleichs der bei ihm festgestellten Defizite meist gering sind. Um Auslese handelt es sich ferner, wenn jährlich Entscheidungen über Versetzung oder Sitzenbleiben der Schüler getroffen werden. Ein weiterer markanter Punkt in der Schullaufbahn jedes Schülers besteht in der Übergangsauslese für die Sekundarschulen am Ende des vierten oder sechsten Grundschuljahres. Hier werden die Schüler, sofern nicht im Einzugsbereich eine Gesamtschule zur Verfügung steht, auf die drei Sekundarschultypen Gymnasium, Realschule oder Hauptschule verteilt, die sich nicht nur in Anspruch und Niveau, sondern auch im Curriculum voneinander unterscheiden. Innerhalb der Gesamtschulen erfolgt in der Regel nach kurzer Zeit eine Verteilung der Schüler in den sogenannten Hauptfächern auf Fachleistungskurse unterschiedlichen Niveaus, deren Besuch für die Art des Schulabschlußzeugnisses von entscheidender Bedeutung ist. Auch die Auswahl der für geeignet gehaltenen Abiturienten bei den Numerus-clausus-Fächern an der Universität ist ein Beispiel schulischer Selektion.

Cronbach u. Gleser (1965) haben die im Rahmen des Auslesevorgangs und seiner Folgen für die betroffenen Schüler wichtige Unterscheidung zwischen Selektion und Klassifikation eingeführt. In beiden Fällen sollen diejenigen Schüler, die ein bestimmtes Merkmal in einer bestimmten Ausprägung besitzen, ausfindig gemacht werden. Bei der Klassifikation freilich macht man sich nicht nur Gedanken darüber, was mit den Ausgewählten, sondern auch was mit den Zurückgewiesenen geschehen soll. Beide Gruppen werden hier einer ihren spezifischen Bedürfnissen entsprechenden Betreuung zugeführt. Im Unterschied dazu ist bei der Selektion das Interesse nur auf diejenigen gerichtet, die ausgewählt werden; die Zurückgewiesenen bleiben ihrem Schicksal überlassen. Nun braucht hier nicht begründet zu werden, daß dem pädagogischen Auftrag der Schule Selektion in der hier gegebenen Definition nicht entsprechen dürfte, sondern daß vielmehr um der Förderung des Lernens willen jeder Schüler den genau für ihn geeigneten Unterricht bekommen sollte. Die angeführten Beispiele zeigen demgegenüber, daß in unserem Schulwesen Selektionsvorgänge im Vordergrund stehen. So werden beispielsweise die für schulunreif befundenen Kinder in der Mehrzahl der Fälle gerade nicht spezifisch pädagogisch betreut, sondern – wie bereits erwähnt – in ihre in der Regel wenig anregende familiäre Umwelt zurückgeschickt. Ähnliches gilt für die Abiturienten, welchen es nicht gelingt, das Fach ihrer Wahl zu studieren; sie bleiben auf der Suche nach für sie geeigneten Alternativen sich selbst überlassen. Aber auch bei der Auswahl der Schüler für die Sekundarschultypen spielt der Gesichtspunkt, eine Zuordnung zu treffen von Schülermerkmalen einerseits und darauf speziell abgestimmten pädagogischen Maßnahmen andererseits, eine untergeordnete Rolle. Vielmehr geht es überwiegend darum, nach Maßgabe einer aus den Zensuren der Hauptfächer oder entsprechenden Übergangsprüfungen vermeintlich ermittelten allgemeinen Schulleistungsfähigkeit darüber zu entscheiden, ob ein Schüler zum Gymnasium oder zur Realschule zugelassen wird oder ob man ihn der Hauptschule zuführt, weil er die festgesetzten Voraussetzungen der Realschule oder des Gymnasiums nicht erfüllt. Hier handelt es sich also nicht um den Versuch, eine Passung von Schülermerkmalen und pädagogischen Maßnahmen herzustellen, sondern es geht um Zulassung oder Zurückweisung. Denn zwischen den Lernanforderungen der ver-

schiedenen Sekundarschultypen bestehen nicht zuletzt qualitative Unterschiede in den Anforderungen, so daß eine Korrelation der Leistungen beim Übergang mit den Leistungen in Hauptschule oder Gymnasium zwei verschiedene empirische Bedeutungen besitzt. Damit läßt sich also im Einzelfall aufgrund derselben Ausleseprüfung gar nicht entscheiden, ob ein Schüler sich für die Hauptschule oder für das Gymnasium eignet.[1] Auch bei den Entscheidungen über das Sitzenbleiben eines Schülers aufgrund seiner Zensuren in zwei oder drei Fächern handelt es sich bei näherer Betrachtung eher um Selektion als um Klassifikation. Denn die Wiederholung einer Klassenstufe bedeutet für den Schüler nicht, daß er eine pädagogische Förderung erfährt, die auf seine spezifischen Leistungsausfälle zugeschnitten wäre. Vielmehr bleiben seine Defizite im einzelnen unberücksichtigt, und er sieht sich einem Unterricht ausgesetzt, der ihn zu Beginn so stark unterfordert, daß Interessen und Anstrengungsbereitschaft rasch und oft endgültig zum Erliegen kommen.

Hinter der schulischen Selektion oder Klassifikation steht die Annahme, daß man künftige Leistungen eines Schülers prognostizieren könne. Leistungsurteile stellen zwar zunächst eine Analyse vorgefundenen Schülerverhaltens oder, wenn sie die Erfahrungen einer längeren Zeit summieren, etwa in einer Jahreszensur, das kumulative Ergebnis mehrerer Analysen dar. Zugleich enthält jede solche Analyse aber auch ein prognostisches Element. So verleiht beispielsweise ein Zeugnis die Berechtigung, in die nächsthöhere Klasse aufzusteigen, womit die Voraussage gemacht wird, daß der Schüler auch in der höheren Klassenstufe mit hoher Wahrscheinlichkeit hinreichende Leistungen erzielen wird. Noch deutlicher, geradezu dramatisch in ihrer individuellen Bedeutung, wird die prognostische Komponente der Leistungsurteile, wenn beispielsweise auf der Grundlage von Abschlußzeugnissen der Grundschule oder von Übergangsausleseprüfungen nach der vierten Grundschulklasse entschieden wird, welchem Sekundarschultyp ein Schüler zugeführt wird; oder wenn Abiturnoten die Grundlage für die Zulassung zum Studium bilden, also angenommen wird, daß Schüler mit sehr guten Schulleistungen beispielsweise auch sehr gute Ärzte sein werden. Auch hier geht man davon aus, daß zwischen früheren und späteren Leistungen ein enger Zusammenhang besteht.

Die Prognose künftigen Verhaltens eines Schülers setzt voraus, daß das in Frage stehende Merkmal präzis gemessen werden kann und daß es über die Zeit, auf welche sich die Prognose bezieht, stabil bleibt. In die Selektion gehen also Annahmen über die Konstanz der schulischen Leistungsfähigkeit und über die Möglichkeit ihrer genauen Feststellung ein. Sofern für ein schulfernes Kriterium, beispielsweise für den Arztberuf, selegiert wird wie beim Numerus-clausus-Verfahren für das Medizinstudium, impliziert die Auslese darüber hinaus, daß das beurteilte Merkmal, hier also die Schulleistung, mit der Kriteriumsleistung, in unserem Beispiel der Tätigkeit als Arzt, in Zusammenhang steht.

Betrachten wir zunächst die Frage der Meßgenauigkeit und der Zuverlässigkeit (Reliabilität) im Rahmen der schulischen Leistungsbeurteilung. Für eine Prognose und eine darauf gegründete Auslese wäre es untragbar, wenn beispielsweise ein Schüler bei der Wiederholung eines gleichartigen Probediktats für den Übergang von der Grundschule zum Gymnasium in dem einen Fall eine 2, in dem anderen eine 4 erhielte. Wenn man davon ausgeht, daß das gemessene Merkmal, hier also die Fähigkeit, orthographisch korrekt zu schreiben, sich nicht in kurzer Zeit verändert, muß die Ursache für die Schwankung entweder in der Qualität des Meßverfahrens liegen oder beim Schüler gesucht werden. Wenn sich dann zeigt, daß solche Schwankungen in der Bewertung auch in zahlreichen anderen Fällen vorkommen, liegt die Annahme nahe, daß man es mit einem unzuverlässigen Meßinstrument zu tun hat.

Diese Frage ist nun speziell an Hand vorliegender Schulleistungs- oder Intelligenztests seit Jahrzehnten mit besonderer Aufmerksamkeit verfolgt worden. Dabei hat sich gezeigt, daß auch sehr sorgfältig konstruierte Tests niemals vollkommen zuverlässig sind, sondern

ihre Ergebnisse bei Wiederholungsmessungen schwanken. Dies trifft auch dann zu, wenn alle erdenkliche Vorsorge dafür getroffen ist, daß der Test unter streng standardisierten Bedingungen ausgefüllt sowie seine Auswertung in genau vorgeschriebener Weise vorgenommen wird. Die vorkommenden Schwankungen lassen sich daher vor allem mit Veränderungen in den äußeren Bedingungen (z. B. Wetter) oder in der Disposition der Prüflinge (z. B. Angst), auf die das Meßinstrument im Grunde nicht ansprechen dürfte, erklären.

Die Zuverlässigkeit eines Meßinstruments wird nach psychometrischem Brauch durch die Korrelation zwischen den Ergebnissen zweier Messungen, die nur kurze Zeit auseinanderliegen, bei denselben Probanden bestimmt (zu den komplizierten Detailfragen der Reliabilitätsbestimmung vgl. u. a. Lienert 1967). Die Reliabilitätskoeffizienten bei Schulleistungstests liegen mit ungefähr 0,90 relativ nahe am Maximalwert einer Korrelation von 1,00, der die vollkommene Übereinstimmung der beiden korrelierten Datensätze zeigt, während die Zuverlässigkeit von Schulzensuren mit einem durchschnittlichen Korrelationskoeffizienten von 0,50 erheblich niedriger liegt. Die Implikationen einer unvollkommenen Reliabilität eines Instruments lassen sich am greifbarsten an Hand des sogenannten Meßfehlers illustrieren (s. Bd. V dieser Enzyklopädie). So beträgt beispielsweise der Meßfehler eines gebräuchlichen Schulleistungstests, der einen Mittelwert von 100 Punkten hat, ungefähr ± 5 Punkte. Dies bedeutet, daß es unzulässig wäre, einen Schüler mit einem Testwert von 106 für besser zu halten als einen Schüler mit einem Testwert von 98, wie Abb. 2 zeigt.

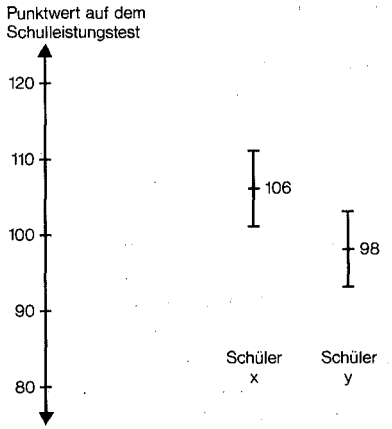

Abb. 2

Wie das Schaubild demonstriert, muß man davon ausgehen, daß bei einer Wiederholungsmessung der Punktwert dieser Schüler (mit der Wahrscheinlichkeit von 2:1) irgendwo in dem Bereich von ± 5 Punkten um den bei der ersten Messung erhaltenen Wert schwankt. Schüler Y könnte dann bei der zweiten Messung ohne weiteres eine höhere Leistung auf dem Test erzielen als Schüler X.

Dieses Beispiel wurde ausgewählt, um zu zeigen, daß selbst bei Meßverfahren, die mit besonderer Sorgfalt in bezug auf ihre Zuverlässigkeit konstruiert zu werden pflegen, große Vorsicht bei der Interpretation der Ergebnisse geboten ist. Der Meßfehler anderer Beurteilungsverfahren ist im allgemeinen noch beträchtlich größer; so beträgt der in unserem Beispiel gewählte Meßfehler, der sogenannte Standardmeßfehler, bei den üblichen Schulzensuren bei einer Skala von nur 6 Punkten (Note 1 bis 6) ungefähr plus oder minus eine Zensur. Auf die Gründe für eine derart gravierende Meßungenauigkeit wird später einzugehen sein.

Die zweite Voraussetzung für eine Prognose ist die Stabilität des gemessenen Merkmals.

Jede Schülerauslese und Verteilung auf unterschiedliche Schultypen geht davon aus, daß der Grad der Lernfähigkeit eines Schülers, wie er sich in einem gegebenen Moment in den Zensuren ausdrückt, über lange Jahre stabil bleiben wird. Nun wird man in der Tat bei der Zuteilung von Berufs- und Lebenschancen nur relativ selten Fehlentscheidungen treffen, wenn man sie auf der Grundlage eines stabilen Merkmals vornehmen und wenn darüber hinaus dies Merkmal mit hinreichender Genauigkeit festgestellt werden kann. Auf diese Annahme gründet sich auch beispielsweise die Zuweisung der Schüler in die drei Sekundar-Schultypen: schulische Leistung, Leistungsfähigkeit oder allgemeine Begabung werden als Merkmale betrachtet, bei denen man im Laufe der Schulzeit mit so geringer Schwankung rechnen muß, daß sich eine schulorganisatorische Vorsorge für Veränderungen, die nach dem zehnten Lebensjahr eintreten, weitgehend erübrigt.

Nun steht hinter der Aufteilung der etwa zehnjährigen Schüler in die Sekundarschultypen eine lange Tradition, und sie scheint sich, oberflächlich betrachtet, ja auch zu bewähren. Unterstützung von wissenschaftlicher Seite ist ihr vor allem durch das berühmt gewordene Buch Benjamin Blooms (1964) zuteil geworden, in welchem der Autor auf die überragende Bedeutung der frühen Jahre für die Entwicklung der Intelligenz und der Schulleistung hinweist. Bloom kommt in seinem Buch zu dem Ergebnis, daß sich bis zum sechsten Lebensjahr ein Drittel und bis zum dreizehnten Lebensjahr drei Viertel der allgemeinen Schulleistung, soweit man sie aus den Leistungstestergebnissen ablesen könne, entwickelt habe (a. a. O., 110). Er folgert daraus, daß vorschulisches Lernen auf die Lernformen und Lernergebnisse der Kinder weitreichende Einflüsse habe und daß die ersten drei Schuljahre der Grundschule vermutlich die entscheidene Periode seien, in der sich allgemeine Lernformen entwickelten; alles nachfolgende Lernen in der Schule sei großenteils determiniert durch das, was das Kind im Alter vor neun Jahren gelernt habe. Noch einschneidender ist die Interpretation, die Bloom jahrzehntelangen Einzelbefunden aus der Intelligenzforschung zuteil werden läßt. Nach seinen Berechnungen sollen 50 Prozent der Intelligenz bereits bis zum vierten Lebensjahr und 80 Prozent bis zum achten Jahr entwickelt sein. Sensible Phasen beschleunigten Wachstums lägen danach sogar vor der Schulzeit eines Kindes. Aber abgesehen davon, daß Blooms Aussagen auf nicht haltbaren psychometrischen und inhaltlichen Voraussetzungen beruhen (vgl. Hopf 1971; Krapp, Schiefele 1976), stehen sie in Widerspruch zu einer großen Zahl von Untersuchungsergebnissen, aus denen hervorgeht, daß sowohl hinsichtlich der Intelligenzentwicklung als auch der Schulleistungsentwicklung erhebliche Veränderungen positiver oder negativer Art auch in höherem Alter häufig beobachtet werden und insbesondere dann eintreten, wenn Kinder oder Jugendliche einer bezüglich ihres Anregungspotentials andersartigen Umwelt ausgesetzt werden. Außerdem muß man davon ausgehen, daß große Unterschiede in den Entwicklungsverläufen der Komponenten der globalen Merkmale Intelligenz oder Schulleistung bestehen. So entwickeln sich beispielsweise die Intelligenzfaktoren »sprachliches Verständnis« oder »Schnelligkeit der Wortfindung« wesentlich langsamer als andere Intelligenzfaktoren wie »Schnelligkeit der Wahrnehmung«; entsprechend wird man bei dem globalen Merkmal »allgemeine Schulleistung« differenzieren müssen zwischen Entwicklungsverläufen für einzelne Fächer und innerhalb der Fächer für die unterschiedlichen Fähigkeits- und Kenntniskomponenten, aus denen sich eine Fachleistung zusammensetzt (vgl. Roeder 1974, Treumann 1974). Wenn man sich vor Augen hält, wie wechselnd die Anregungen und Gelegenheiten zum Lernen für ein Kind im Laufe seiner Schulgeschichte aussehen, wird man die Instabilität des Merkmals Schulleistung für nicht verwunderlich ansehen.

Das angesprochene Problem wird innerhalb der klassischen Testtheorie unter dem Stichwort prognostische Validität von Messungen abgehandelt. Die prognostische Validität oder Gültigkeit läßt sich empirisch dadurch bestimmen, daß man den Zusammenhang zwischen der aktuellen Leistung und der Leistung zu einem späteren Zeitpunkt, der sogenann-

ten Kriteriumsleistung, auf die eine Prognose hin gestellt werden soll, bestimmt. Je höher die Korrelationskoeffizienten zwischen den beiden Datensätzen ausfallen, desto sicherer kann man das Kriterium aus der aktuellen Leistung vorhersagen. In zahlreichen empirischen Untersuchungen, die zum Problem der Voraussage von Schulerfolg durchgeführt worden sind, hat sich nun aber nicht nur gezeigt, daß eine perfekte Prognose unmöglich ist, sondern daß man sogar mit einem außerordentlich hohen »Schätzfehler« zu rechnen hat. Aus dem Schätzfehler läßt sich ableiten, wie häufig man aufgrund bestimmter Testresultate oder sonstiger Leistungsurteile Fehlentscheidungen in Selektionssituationen trifft.

Bei allen Ausleseentscheidungen werden zwei Gruppen geschaffen: die Ausgewählten und die Zurückgewiesenen. Schon nach unserer Betrachtung der Meßfehlerproblematik und der relativen Instabilität des Merkmals Schulleistung wird man nicht die Hoffnung haben können, stets die Richtigen auszuwählen beziehungsweise zurückzuweisen. Um wieder auf das Beispiel der Übergangsauslese für das Gymnasium zurückzukommen: Einerseits wird man also einige Kinder aufnehmen, die nach den Vorstellungen des Gymnasiums nicht hätten aufgenommen werden dürfen (Fehlertyp A), andererseits wird man aber auch potentielle Gymnasiasten zu Unrecht abweisen (Fehlertyp B). Die folgende Abbildung soll den Vorgang der Schülerauslese illustrieren; die in der Skizze angegebenen Prozentzahlen geben die beim Übergang zum Gymnasium herrschenden Verhältnisse ungefähr wieder.

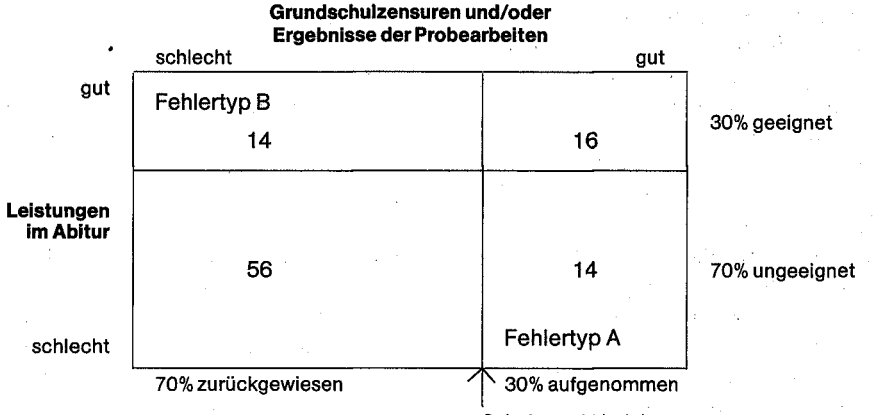

Abb. 3

Die beiden Fehlerarten lassen sich nun mit Hilfe unterschiedlicher Vorgehensweisen verringern. Möchte man den Fehler vom Typ A reduzieren, also möglichst wenige ins Gymnasium aufnehmen, die nach Meinung der Gymnasiallehrer dort nicht hingehören, müßte man die Aufnahmequoten drastisch reduzieren (etwa auf 5 Prozent); möchte man andererseits unbedingt vermeiden, daß potentielle Gymnasiasten zu Unrecht abgewiesen werden, muß man die Aufnahmequote erheblich vergrößern (etwa auf 70 Prozent). Mit der Verringerung des einen Fehlertyps ist notwendigerweise eine Vergrößerung des anderen verbunden. Nach allem, was man über die prognostische Gültigkeit der bei uns gebräuchlichen Aufnahmeverfahren weiß (vgl. zum Beispiel Schultze 1964), muß man in der Bundesrepublik Deutschland schätzungsweise mit einer Fehlerquote (Fehler vom Typ A und B) von etwa 30 Prozent rechnen. Obwohl dies seit langem bekannt ist, gibt es noch Bundesländer, in denen die Auslese mit ähnlichen Verfahren betrieben wird, die Vernon schon 1957 und Schultze 1964 untersucht und für unzureichend befunden haben.

Angesichts dieser Ergebnisse wird die Notwendigkeit bildungspolitischer Vorentscheidungen deutlich. Während man bei der Auslese von Flugzeugpiloten vermutlich ungern Fehler vom Typ A in Kauf nehmen wird, dürfte es mit den in Präambeln von Schulgesetzen wiederkehrenden Zielformulierungen von der optimalen Förderung jedes einzelnen Kindes nur schwer zu vereinbaren sein, eine aufgrund der mangelhaften Genauigkeit der Auslese-verfahren große Gruppe von Kindern jährlich von denjenigen Schulformen fernzuhalten, in denen sie eine für sie geeignete, anregende Lernumwelt vorfinden würden, zumal es vielerlei Möglichkeiten gibt, der eventuell daraus folgenden Überforderung der fälschlich Aufgenommenen pädagogisch zu begegnen. Eine perfekte Prognose der geistigen Entwicklung von Kindern und Jugendlichen wird es aller Voraussicht nach nie geben. Man wird daher nach schulorganisatorischen Modellen Ausschau halten müssen, bei denen die Folgen falschgetroffener Ausleseentscheidungen revidierbar bleiben oder bei denen von vornherein auf Auslese verzichtet wird.

Auf ein besonderes Problem verweist die Diskussion über die Selektion von Bewerbern um Studienplätze an den Hochschulen. Während im Fall der Übergangsauslese von der Grundschule zu den Sekundarschulen noch eine gewisse Ähnlichkeit zwischen der Entscheidungsgrundlage bei der Auslese und dem prognostizierten Kriterium zu bestehen scheint, da es sich in beiden Fällen um Leistungen in der Institution Schule handelt, wird im Numerus-clausus-Verfahren eine Prognose von den Schulleistungen in der Oberstufe des Gymnasiums auf Endleistungen im Studium, im Grunde sogar auf Berufsbewährung gestellt. Das Kriterium unterscheidet sich also hier schon auf den ersten Blick inhaltlich erheblich von dem Prädiktor. So ist es auch nicht weiter erstaunlich, daß die bisher untersuchten Zusammenhänge zwischen Zensuren in der Schule und Zensuren im Hochschulstudium sehr schwach sind und daß es keinerlei Zusammenhänge zwischen Schulzensuren und Berufserfolg gibt, ja nicht einmal Studienerfolg und Berufserfolg miteinander korrelieren. Auch diese Forschungsbefunde sind freilich nicht besonders überraschend, da universitäre und berufliche Anforderungen stark voneinander abweichen können. Auch läßt sich schwer entscheiden, welchen Indikator man für Berufserfolg wählen sollte: Läßt er sich beim Mediziner oder beim Juristen etwa am Einkommen ablesen oder an der Berufszufriedenheit oder an der Einschätzung durch Kollegen?

Eine weitere Schwierigkeit besteht darin, daß Berufsgruppen aus recht heterogenen Untergruppen bestehen können, bei denen sehr verschiedenartige Kenntnisse und Fähigkeiten abgerufen werden. So wird beispielsweise die erfolgreiche Arbeit eines Psychiaters und Psychotherapeuten von ganz anderen Fähigkeiten abhängen als die eines Pathologen oder Augenarztes. Es wäre verwunderlich, wenn die Zensuren in den Schulfächern in gleicher Weise voraussagen könnten, wer in dem einen oder anderen Fall ein guter Arzt sein wird.

Am Beispiel der Auslese im Numerus-clausus-Verfahren wird das Problem der inhaltlichen Beziehung zwischen Prädiktor und Kriterium besonders deutlich. Entgegen dem ersten Anschein gilt dies aber auch für das Verhältnis von Schulleistungen in der Grundschule und vom Abschneiden in Probediktaten oder -rechenarbeiten einerseits zu den Schulleistungen in der Oberstufe und der Abiturprüfung andererseits; auch hier werden qualitativ jeweils andere Fähigkeiten angesprochen, so daß sich der schwache Zusammenhang mit daraus erklärt, daß recht heterogene Dinge zueinander in Beziehung gesetzt werden, deren Unterschiedlichkeit durch das Etikett Schulleistung nur verschleiert wird.

Wenn Schulzensuren, Prüfungen oder Schulleistungstests zu Auslesezwecken verwendet werden, lassen sich also eine große Zahl von Fehlentscheidungen nicht vermeiden. Die uns zur Verfügung stehenden Verfahren der Pädagogischen Diagnostik lassen allerdings nicht nur die Erwartungen unerfüllt, die hinsichtlich Genauigkeit und prognostischer Gültigkeit an sie gestellt werden, sondern sie haben darüber hinaus noch unerwünschte Nebeneffekte,

die oft ein beträchtliches Ausmaß annehmen, gewöhnlich aber nicht auf den ersten Blick erkennbar sind. Wenn Leistungsurteile Verteilerfunktion erhalten, wirken sie nämlich auf die Bildungsinstitutionen zurück. So kann man beispielsweise in denjenigen Regionen der Bundesrepublik, in denen noch Probearbeiten für die Übergangsauslese zu den Sekundarschulen geschrieben werden, beobachten, daß diese Prüfungsverfahren massive Einflüsse auf die Auswahl der Lehrinhalte und auf die Gestaltung des Unterrichts zumindest in den beiden Schuljahren, die der Prüfung vorausgehen, ausüben. Die Schüler werden hier nämlich oft schwerpunktmäßig auf jene Prüfung vorbereitet – mit dem Erfolg, daß die sogenannten Nebenfächer, insbesondere die musischen Fächer, oft völlig in den Hintergrund treten. Darüber hinaus üben auch die Eltern auf ihre Kinder und nicht zuletzt auch auf die Lehrer einen erheblichen Leistungsdruck aus, der in Regionen ohne Übergangsausleseprüfung nicht zu beobachten ist. Durch die Ausleseprüfung tritt also eine Verarmung des Unterrichts ein, und das Interesse der Lehrer, mit neuen Unterrichtsformen und Unterrichtsinhalten, die ihnen besonders interessant und erfolgversprechend zu sein scheinen, zu experimentieren, erlahmt schon bald nach dem Beginn der Arbeit in der Schule, weil es nur gegen den Widerstand der Eltern durchzusetzen wäre. Nicht unerwartet häufen sich auch in Regionen, wo die schulische Auslese großes Gewicht besitzt, die Berichte über erhebliche psychosomatische Störungen bei den Schülern, weil schulischer Leistungsdruck ausgeglichene Entwicklungsprozesse behindert.

ZUR GEGENWÄRTIGEN ZENSURENGEBUNG[2]

Aus unseren Überlegungen zur Bedeutung der Pädagogischen Diagnostik für schulisches Lernen und die individuelle Förderung des einzelnen Schülers wurde bereits deutlich, daß die herkömmliche Zensurengebung den hieraus erwachsenden Ansprüchen nicht gerecht werden kann. Nun sind Zensuren zu diesem Zweck gewöhnlich auch nur nebenbei verwendet worden; hauptsächlich und überwiegend dienten und dienen sie vielmehr der Auslese und der Erteilung von Berechtigungen. Vor allem in dieser Funktion sind sie daher auch Gegenstand der Kritik geworden; diese Kritik richtet sich nicht prinzipiell gegen jegliche Beurteilung von Schülerleistung und schon gar nicht gegen die Lehrer, sondern gegen die notorische Überinterpretation und Überforderung der Tragfähigkeit schulischer Leistungsurteile.

In Frage gestellt wurden im Zuge dieser Auseinandersetzung vor allem die Präzision der Zensuren sowie ihre Vergleichbarkeit zwischen verschiedenen Schulen. So weiß man bereits seit vielen Jahrzehnten, (vgl. z. B. Lietzmann 1927, 46 ff), daß die Prämisse nicht zutrifft, unter der Zensuren und Zeugnisse gelesen, interpretiert und zur Grundlage wichtiger Entscheidungen gemacht werden, gleiche Zensuren in demselben Fach würden gleiche Bedeutung besitzen.

Beispielsweise ist nachgewiesen worden, daß bei einem Vergleich der Leistungen von Schülern aus unterschiedlichen Klassen mit Hilfe eines gemeinsamen, gültigen Maßstabes dieselben Leistungen in der einen Klasse zu einer guten, in einer anderen dagegen zu einer mangelhaften Zensur führten; und dies war der Fall, obwohl die untersuchten Schulklassen sich in enger räumlicher Nachbarschaft befanden. Auch unterscheiden sich die Durchschnittsleistungen von Schulklassen desselben Schultyps erheblich (vgl. Schultze 1975). Natürlich erklären sich solche Beobachtungen leicht daraus, daß die Lehrer gar nicht über die Möglichkeit verfügen, andere als klassenimmanente Maßstäbe zu benutzen. Darüber hinaus bedingt die Vergleichbarkeit von Leistungsurteilen unter anderem auch ein gleiches Unterrichtsangebot, gleiche Unterrichtsmethoden sowie gleiche Vorerfahrungen und Vorkenntnisse der Schüler. Keine dieser notwendigen Voraussetzungen für die Sicherung der Vergleichbarkeit ist jedoch in unserem derzeitigen Beurteilungssystem erfüllt. Der In-

formationswert von Zensuren ist daher über den Rahmen der Schulklasse hinaus, in welcher sie erteilt worden sind, außerordentlich gering. Ingenkamp (⁶1976, 200) kommt daher mit Recht zu dem Schluß, »daß für unser gesamtes schulisches Berechtigungswesen keine sachliche Rechtfertigung besteht«.

Auch die geringe Präzision herkömmlicher Schulzensuren läßt sich aus den Befunden zahlreicher empirischer Unterrichtsuntersuchungen belegen. So hat sich beispielsweise gezeigt, daß dieselbe Leistung, sei es eine Rechenarbeit oder ein Deutschaufsatz, von verschiedenen Lehrern ganz unterschiedlich beurteilt wird: Dieselbe Klassenarbeit kann von dem einen Lehrer mit mangelhaft, von dem anderen mit gut oder gar sehr gut bewertet werden. Dies ist nicht verwunderlich, weil die Lehrer unterschiedliche Beurteilungskriterien heranziehen oder den Kriterien (z. B. Orthographie, Sprache, Gedankenreichtum bei der Beurteilung eines Deutschaufsatzes) ein unterschiedliches Gewicht beimessen. Daraus erklärt sich auch die oft berichtete Erfahrung, daß erhebliche Verschiebungen im Zensurenbild auftreten, wenn während des Schuljahres ein Lehrerwechsel stattfindet. Aus dem Mangel an Übereinstimmung zwischen den Lehrern folgt also eine große Interpretationsunsicherheit; kein Dritter kann ohne breite Zusatzinformationen genau wissen, was ein Leistungsurteil bedeutet.[3] Ferner lassen sich bei Lehrern interindividuell variierende, intraindividuell jedoch recht konstante Urteilstendenzen nachweisen, welche eine gewisse Ähnlichkeit mit den subjektiv konstanten Wahrnehmungsverzerrungen bei der Beobachtung von Sternen in der Astronomie aufweisen, die man dort als »persönliche Gleichung« bezeichnet (vgl. Lietzmann 1927). Manche Lehrer tendieren beispielsweise zu einer Bevorzugung der Normalverteilungskurve beim Zensieren und vergeben eine gleiche Quote schlechter und guter Zensuren in ihrer Klasse. Andere Lehrer wieder bevorzugen eine schiefe Verteilung und geben entweder mehr gute beziehungsweise mehr schlechte Zensuren. Ferner variieren die Streuungen von Zensurenverteilungen innerhalb der Klassen, indem manche Lehrer nur sehr selten, andere dagegen großzügig von den extremen Noten 1 und 6 Gebrauch machen. Schließlich werden Zensuren dadurch beeinflußt, daß der Lehrer über Informationen verfügt, die über das beurteilte Leistungsprodukt hinausgehen. Das Verhalten eines Schülers im Unterricht, seine soziale Herkunft, Vorurteile des Lehrers, Sympathie oder Antipathie sowie viele andere subjektive Faktoren beeinflussen, wie man aus streng kontrollierten Experimenten weiß, die Zensurengebung. Darüber hinaus hat sich gezeigt, daß sogar derselbe Lehrer in der Regel dieselben Klassenarbeiten nach einem Zeitabstand von einigen Wochen anders bewertet als beim ersten Mal.

Es wäre freilich ein Mißverständnis, die Tatsache, daß in Leistungsurteile auch subjektive Einflüsse eingehen, insgesamt als Beweis für die Untauglichkeit der Leistungsbeurteilung für jeglichen Zweck anzusehen. Insbesondere die erwähnte, auf der Verschiedenheit von Beurteilungskriterien beruhende Unvergleichbarkeit der Zensuren zeigt auch, daß sich die traditionelle Zensurengebung ganz zu Recht am Unterricht des individuellen Lehrers orientiert, da sie seine jeweiligen (und gerade nicht klassenübergreifenden) Ziele und Inhalte aufnimmt und verstärkt. Da weder Unterrichtsziele noch Unterrichtsmethoden oder Unterrichtsinhalte im einzelnen festgelegt sind und von Lehrer zu Lehrer variieren, kann man die Unvergleichbarkeit der Zensuren als einen Beleg für die Abstimmung der Leistungsbewertung auf den jeweiligen Unterricht eines Lehrers betrachten. In dieser Hinsicht genügen deshalb die Zensuren in ihrer derzeitigen Form bis zu einem gewissen Grade einer zentral wichtigen Forderung, die man an jegliche Pädagogische Diagnostik richten muß: dem Unterricht zu *folgen*, nicht ihn zu *bestimmen*. Insofern sind alle Versuche, Präzision und Vergleichbarkeit der schulischen Leistungsbeurteilung zu erhöhen, stets auch unter dem Gesichtspunkt zu beurteilen, ob sie nicht möglicherweise zu einer Einbuße an inhaltlicher Validität oder aber zu einer problematischen Normierung des Unterrichtsangebots führen. Andererseits kann aber auch kein Zweifel daran bestehen, daß auf der Grundlage unserer

Zensuren mit ihrem geringen Maß an Präzision und Vergleichbarkeit keine verläßliche Einteilung von Schülern – zum Beispiel in Gymnasiasten, Realschüler und Hauptschüler oder auch in Sitzenbleiber und Versetzte – vorgenommen werden kann, sondern daß es dabei immer und notwendigerweise eine unvertretbar hohe Anzahl von Fehlentscheidungen geben wird. Innerhalb der einzelnen Schulklasse ist das Leistungsurteil des Lehrers gewiß bis zu einem gewissen Grad brauchbar, wenn es darum geht, den Leistungsstand eines Schülers im Verhältnis zu seinen Klassenkameraden zu bestimmen, gelegentlich vielleicht auch als Anreiz für den Schüler, intensiver zu lernen, oder auch als Information über den Unterrichtserfolg für den Lehrer. Es wird jedoch in gefährlicher Weise überinterpretiert, wenn auf der Grundlage von Zensuren Berechtigungen erteilt werden oder gar große Gruppen von Schülern von besonders anregenden Lernbedingungen, wie sie etwa im Gymnasium herrschen, ausgeschlossen werden.

Zensuren stellen nicht nur die Basis für Ausleseentscheidungen und für die Vergabe von Berechtigungen dar, sondern sie sollen auch die Schüler motivieren und Impulse zu verstärktem Arbeiten geben. Wie weit dies Ziel erreichbar ist, hängt nicht zuletzt davon ab, ob der einzelne Schüler auf die Dauer die Anforderungen des Unterrichts erfüllt und Erfolge verzeichnet oder ob er sich einer Kette von Mißerfolgserlebnissen ausgesetzt sieht. Denn man kann nicht erwarten, daß ein Schüler, der monatelang oder gar jahrelang schlechte Zensuren erhält, diese als Lernanreiz empfindet; auf die Dauer wird nur der Erfolg weiteres Lernen begünstigen. Schulischer Mißerfolg freilich ist immer dann für eine große Gruppe von Schülern unvermeidlich, wenn die Zensuren in einer Klasse, wie es derzeit üblich ist, jeweils das ganze Spektrum der Notenskala umfassen, etwa in der Form einer Normalverteilung. Solange auf diese Weise innerhalb der Schulklasse Rangreihen aufgestellt werden, können die Zensuren für zahlreiche Schüler nicht den Anreiz zu weiterem Lernen darstellen, sondern sie werden im Gegenteil Entmutigung und Desinteresse hervorrufen. Vergleichende Beurteilung, und sei es auch nur innerhalb einer Klasse, läßt sich daher nicht mit der Intention der Motivationssteigerung durch Leistungsurteile vereinbaren. Vielmehr bedarf es hier der Verwendung intrasubjektiver Normen, bei denen sich die Leistungsurteile an der Ausgangslage und den Möglichkeiten des individuellen Schülers, nicht aber an den Leistungen seiner Mitschüler orientieren. Von dieser Möglichkeit der Notenverwendung wird in unseren Schulen nur sehr selten, beispielsweise im Sportunterricht, Gebrauch gemacht, obgleich sich ihre Wirksamkeit insbesondere in Schulversuchen, die mit Formen der programmierten Instruktion arbeiten, zweifelsfrei erwiesen hat (vgl. Lindvall, Bolvin 1967).

SCHULTESTS

Die herkömmliche Zensurengebung vermag, wie seit langem bekannt ist, kaum eine der besprochenen Aufgaben Pädagogischer Diagnostik befriedigend zu erfüllen. So entstanden aus dem Wunsch, objektivere und zuverlässigere Beurteilungsverfahren zu entwickeln, insbesondere in den angelsächsischen Ländern eine kaum mehr überschaubare Zahl von Schulleistungstests sowie einzelne Testverfahren, die auf andere Merkmale als Schulleistung – zum Beispiel Intelligenz, Fremdspracheneignung, Konzentration, Interessen – gerichtet sind (s. auch den Beitrag von B. Schuch in Bd. V dieser Enzyklopädie). Die überwiegende Mehrzahl dieser Tests sind sogenannte normorientierte oder vergleichsorientierte Verfahren, die an einer bestimmten, für die künftigen Benutzer möglichst repräsentativen Zielgruppe geeicht wurden und erlauben, jedes individuelle Testergebnis mit den durch die Eichung gewonnenen Resultaten (Normen) zu vergleichen.[4] Im Unterschied hierzu stellen die sogenannten kriterienorientierten Tests, auf die sich in jüngster Zeit zunehmend das Interesse der wissenschaftlichen Diskussion richtet (vgl. z. B. Klauer u. a. 1972), eine Stich-

probe aus dem Kriterium (Lernziel, Lerninhalt) dar, über welches eine Aussage getroffen werden soll; das Testresultat des einzelnen Schülers gibt an, wie weit er das gewünschte Kriterienverhalten aufweist beziehungsweise das entsprechende Unterrichtsziel erreicht hat; ein Vergleich zu Mitschülern oder zu Angehörigen der Eichstichprobe kann entfallen. Dadurch entsteht eher die Möglichkeit für den Schüler, die oben erwähnten intrasubjektiven Normen aufzubauen, das Testergebnis zur Feststellung des Lernfortschritts zu verwenden und als Erfolgsbestätigung anzusehen.

Objektivität, Zuverlässigkeit und Gültigkeit sind in der klassischen Testtheorie die Kriterien, nach denen sich die Güte eines Beurteilungsverfahrens bestimmt (s. Bd. V dieser Enzyklopädie). Insbesondere hinsichtlich der Objektivität und Zuverlässigkeit gelten Tests als der traditionellen Leistungsbeurteilung überlegen. Für objektiver wird das Resultat eines Tests vor allem deshalb gehalten, weil hier eine Verhaltensstichprobe unter genau bestimmten Umständen beobachtet, registriert und möglichst auch interpretiert wird. Hierzu wird der Versuch unternommen, die Testsituation für alle Schüler möglichst gleich zu gestalten. Denn die individuellen Unterschiede in der Schulleistung sollen nur durch die Verschiedenheit zwischen den Schülern, nicht durch die Verschiedenheit zwischen Situationen zustande kommen. Um die Testsituation zu standardisieren, verwendet man beispielsweise eine fest vorgeschriebene Instruktion, die dafür sorgen soll, daß bei den Schülern die gleiche Motivationslage entsteht; ferner werden einheitliche Zeitgrenzen vorgeschrieben, an alle Schüler das gleiche Testmaterial verteilt, dem Lehrer im einzelnen vorgeschrieben, wie er sich bei Rückfragen der Schüler zu verhalten hat und vieles andere mehr. Darüber hinaus werden Anstrengungen unternommen, um bei der Auswertung subjektive Elemente auszuscheiden, indem genaue Vorschriften bestehen, wie man ein Testresultat ermittelt. Auch hinsichtlich der Interpretation der Testergebnisse versucht man, soweit möglich, den subjektiven Einflüssen des Auswerters Grenzen zu setzen.

Über die Zuverlässigkeit von Beurteilungsverfahren war oben bereits im Zusammenhang mit der Diskussion des Meßfehlers die Rede. Ein Test gilt als zuverlässig, wenn seine Resultate bei denselben Personen relativ stabil sind. Erst wenn eine hinreichende Übereinstimmung wiederholter Messungen empirisch gesichert ist, kann man davon ausgehen, mit einer gewissen Wahrscheinlichkeit auch bei einer einmaligen Messung eine gute Schätzung des »wahren« Wertes erhalten zu haben.

Auch die Gültigkeit, insbesondere die prognostische Gültigkeit von Verfahren der Leistungsbeurteilung hatten wir bereits erörtert. Wo es um die Diskussion von Gütekriterien für Schulleistungstests geht, steht jedoch meist eine andere Form der Validität im Vordergrund, nämlich die sogenannte Inhaltsgültigkeit. Ein Test ist danach in dem Grade valide, in dem er mißt, was er messen soll. Der Test hat dabei die Aufgabe, von einem begrenzten Verhaltensausschnitt her Aussagen zu machen über das entsprechende breite Verhaltensrepertoire, welches nicht in ganzem Umfang zutage tritt. So kann beispielsweise ein muttersprachlicher Wortschatztest, der aus dreißig Wörtern besteht, erst dann als valide gelten, wenn sich zeigen läßt, daß zwischen der Testleistung und der gesamten Wortkenntnis des Schülers eine enge Beziehung besteht. Dabei muß der Test auf die Unterrichtsinhalte und Unterrichtsziele genau zugeschnitten sein und beispielsweise nach einer Unterrichtseinheit, die sich auf ein bestimmtes Lehrbuchkapitel stützt, auch auf die darin enthaltenen Inhalte und Ziele bezogen sein.

Das Kriterium der Inhaltsgültigkeit verweist, wenn man einmal von den bereits beschriebenen Problemen der Selektion absieht, auf die Hauptschwierigkeit der Testkonstruktion und Testverwendung. Die größere Objektivität von Tests und ihre Zuverlässigkeit im Vergleich zu Zensuren als Leistungsurteilen verleiten leicht zu der Annahme, daß die Verwendung von Tests zugleich Unbestechlichkeit und Gerechtigkeit des Urteils garantieren. Man vergißt dabei, daß die Einflüsse subjektiver Willkür beim Test lediglich an anderer Stelle

auftreten als bei den Zensuren und daß sie weniger leicht zu erkennen sind. Zwar kommen verschiedene Beurteiler mit Hilfe von Tests zu stärker übereinstimmenden Urteilen, subjektive Momente spielen jedoch in der Phase der Testkonstruktion eine wesentliche Rolle. Schulleistungstests, die für den Unterricht verschiedener Lehrer valide sind, lassen sich nämlich nicht schulklassenübergreifend konstruieren, sondern die Inhalte der Testitems sind das Ergebnis eines langen Entscheidungsprozesses, der durchaus nicht frei von subjektiver, wenn auch kontrollierter Willkür ist. Ein gutes Beispiel für die erstaunliche Heterogenität von Unterrichtsinhalten und Unterrichtszielen auf einer Klassenstufe geben Edelstein u. a. (1968) sowie Zeiher (1972) in ihren Untersuchungen der Stoffverteilung in den siebten Klassen der Gymnasien der Bundesrepublik Deutschland. Wie immer man sich bei der Testkonstruktion entscheidet, stets wird es eine Gruppe von Schulklassen geben, die durch den Test weniger angemessen behandelt werden als andere. Selbst wenn man die Tests auf den gemeinsamen Kern der Unterrichtsstoffe aller Klassen beschränkt, wird dieser Kern in der einen Klasse einen großen Teil, in einer anderen nur einen geringen Bruchteil der Unterrichtszeit abdecken, und der Test wird damit nicht in gleicher Weise »fair« beziehungsweise »unfair« sein können. Im Vergleich hierzu decken die herkömmlichen Klassenarbeiten sehr viel genauer Unterrichtsstoffe und -ziele ab, die der Lehrer in der jeweiligen Klasse anstrebt. Um es etwas überspitzt auszudrücken: Entweder gewinnen wir subjektive Urteile bei validen Prüfungsaufgaben oder objektive Urteile bei invaliden Testitems. So viel ist jedenfalls sicher, daß normierte, standardisierte Schulleistungstests schwerlich der individuellen Lernförderung dienen, weil sie nur in Ausnahmefällen auf die spezielle Lage des einzelnen Schülers im Detail zugeschnitten sein können.

Man hat als Ausweg aus diesem Dilemma die Konstruktion sogenannter informeller Tests empfohlen. In der Tat stellen informelle Tests einen akzeptablen Kompromiß dar, insofern der Lehrer hier die Testitems selbst konstruiert und sie damit auf seinen Unterricht genau zuschneiden kann und zugleich die Objektivität und Zuverlässigkeit durch die Verwendung entsprechender Aufgabenformate sowie durch die Bestimmung der Eigenschaften jedes Testitems bei wiederholter Anwendung erhöht. Obwohl es inzwischen mehrere brauchbare Anleitungen zur Konstruktion informeller Testverfahren für die Hand des Lehrers gibt, liegen in der Praxis die Grenzen dieses Verfahrens hauptsächlich in der dadurch entstehenden Zusatzbelastung des Lehrers, der auf die Konstruktion solcher Verfahren in seiner Ausbildung in der Regel nicht vorbereitet wurde, sowie auch in der weithin feststellbaren Beschränkung informeller Tests auf leicht erfaßbare, wenig komplexe Kenntnisse und Fertigkeiten.

Ein anderer Ansatz zur Lösung der mit der Forderung nach inhaltlicher Gültigkeit von Testverfahren verbundenen Schwierigkeiten, zumal im Kontext zunehmend individualisierten Lernens, ist mit dem Vorschlag der Konstruktion einer sogenannten Itembank gegeben. Im Unterschied zu den in sich geschlossenen, aus einer größeren Zahl von Items bestehenden standardisierten Schulleistungstests bietet die Itembank die beliebige Kombination unterschiedlicher Items an, die sich der Lehrer aus einer großen Aufgabensammlung entsprechend den Inhalten und Zielen seines Unterrichts aussuchen kann und die im Prinzip sogar für die Bedürfnisse jedes einzelnen Schülers in jedem Stadium eines Lernprozesses zusammengestellt werden können. So lassen sich mit Hilfe einer Itembank spezifische, kriterienbezogene Meßinstrumente ad hoc konstruieren, mit deren Hilfe Vorkenntnisse, lernzielspezifische Stärken oder Schwächen des Schülers oder auch der Unterrichtserfolg des Lehrers festgestellt werden können. Wenn man sich vor Augen hält, ein wie komplexes Produkt aus individuellen und situativen Komponenten Schulleistung in jedem Schulfach zu jedem Zeitpunkt darstellt und wie rasch es sich in Reaktion auf Einflüsse von außen und aufgrund der Wechselwirkung zahlreicher Faktoren verändert (vgl. Flammer 1975), muß der von Wood u. Skurnik (1969) entwickelte Vorschlag zur Konstruktion einer Itembank we-

gen ihrer großen Flexibilität als einer der wichtigsten Vorschläge zur Verbesserung der Pädagogischen Diagnostik in den letzten Jahrzehnten betrachtet werden.

Die derzeit verfügbaren Schultests hingegen sind weit davon entfernt, der eigentlichen Aufgabe der Schule, nämlich das Lernen der Schüler zu unterstützen, zu dienen. Denn sie stellen wiederum interindividuelle Normen für die Interpretation von Testresultaten bereit, deren Fragwürdigkeit bereits aufgewiesen wurde. Auch Schultests werden daher die langsam lernenden Schüler vor dauerhaftem Mißerfolg nicht bewahren, sondern sie im Gegenteil durch ihre dem ersten Anschein nach größere Dignität noch stärker als Zensuren entmutigen.

Schultests können auf der anderen Seite eine Orientierungsfunktion erfüllen, sofern der Lehrer sie dazu benutzt, den Stand seiner Klasse mit dem anderer Schulklassen zu vergleichen. Hierdurch ließe sich eine gewisse Einheitlichkeit von Lernzielen und Lerninhalten im Schulwesen erzielen. Jedoch ist es fraglich, ob gerade dazu Anlaß besteht, wenn es Rahmenrichtlinien gibt, die einen verbindlichen Kern von Stoffen und Zielen enthalten.

Der wichtigste Vorzug von Schultests gegenüber den traditionellen Verfahren der Leistungsbeurteilung dürfte darin bestehen, daß bei der Testkonstruktion, der Testverwendung und der Interpretation der Ergebnisse durch den Lehrer aufgrund der vom Testkonstrukteur vorgegebenen Hinweise die Grenzen der Leistungsbeurteilung sehr viel klarer ins Auge springen als beim üblichen Zensieren. So wird bei der Beschäftigung mit der Zuverlässigkeit eines Tests und mit dem Meßfehler bei der Interpretation eines Testpunktwertes unübersehbar deutlich, daß ein Ergebnis nie den »wahren« Wert darstellt, aufgrund dessen man eindeutig über die Leistung eines Schülers Auskunft erhalten hat, sondern daß es sich stets um einen Schätzwert handelt, in dessen Umgebung sich der wahre Wert mit angebbarer Wahrscheinlichkeit befindet. Neuere Tests, die den mit jedem Testergebnis verbundenen Meßfehler ernst nehmen, geben deshalb das Ergebnis nicht als Punkt, sondern als ein relativ breites Band an, innerhalb dessen der wahre Wert zu suchen ist. Durch eine gründliche Beschäftigung mit Schultests dürfte daher vielen Lehrern, Schülern oder Eltern erstmals klar werden, mit welcher Vorsicht Leistungsurteile interpretiert werden müssen: Die Rangfolge innerhalb einer Klasse kann sich bei der nächsten Messung stark verändern; die Zensur in der einen Klasse wird bei gleicher Leistung in einer parallelen anderen Klasse völlig anders aussehen; das Ergebnis von Schulleistungstests hängt stark davon ab, in welchem Grade die durch den Test abgedeckten Ziele und Inhalte von dem Lehrer im vorhergehenden Unterricht behandelt wurden; Urteile über die zukünftige Entwicklung eines Schülers sind nur mit großer Irrtumswahrscheinlichkeit zu treffen. Aus der wachsenden Erkenntnis der Grenzen jeder Leistungsbeurteilung können dann neue Impulse dafür entstehen, sich weniger der Perfektionierung der pädagogisch-diagnostischen Verfahren zu widmen, als vielmehr die Anstrengungen auf die gezielte Förderung der Schüler und den Ausgleich von Lernschwächen einzelner Schüler zu richten.

SCHLUSSBEMERKUNG

Die in unseren Schulen gebräuchlichen Formen der Leistungsbewertung sind weit davon entfernt, Aussagen über einen Schüler zu erlauben, die zuverlässig und gültig genug wären, um auf ihrer Grundlage Entscheidungen über seine Schullaufbahn zu fällen. Subjektive Elemente gehen in so hohem Maße in die Leistungsurteile ein, daß ihre Genauigkeit und Vergleichbarkeit mehr als fraglich genannt werden muß. Aber auch Schultests in ihrer heutigen Form stellen weder eine befriedigende Grundlage für Selektionsentscheidungen und für die Erteilung von Berechtigungen dar, noch eignen sie sich zur individuellen Förderung des Lernens.

Noten wie Testresultate interpretieren schulische Mißerfolge von Kindern darüber hin-

aus einseitig zu Lasten der Schüler; sie müßten sich im Prinzip ebenso auf den Unterricht beziehungsweise die Schule richten. Wie läßt sich beispielsweise ohne Rekurs auf den Einfluß von Schule und Unterricht der Befund erklären, daß gerade in den sogenannten Hauptfächern, in welchen den Lehrern besonders viele Wochenstunden für die Förderung auch der schwachen Schüler zur Verfügung stehen, die höchste Versagerquote auftritt und die im Durchschnitt niedrigsten Zensuren erteilt werden? Hier tritt offenkundig die Förderungsfunktion der Schule hinter der von ihr ebenfalls erwarteten Verteilerfunktion in den Hintergrund. Daß die Verteilerfunktion freilich bereits in der Grundschule durchschlagen muß mit der Folge, daß einer sehr großen Zahl von Kindern die Chance zu differenziertem Lernen und zur Bewährung vorenthalten wird und die Freude am Lernen frühzeitig verkümmert, steht nicht im Einklang mit den öffentlich vertretenen Zielen in Schule und Unterricht.

Die Kritik an der gegenwärtigen Pädagogischen Diagnostik läßt sich in wenigen Punkten zusammenfassen:

– Zensuren und sonstige Leistungsurteile werden überinterpretiert und zu Aufgaben herangezogen, denen sie von vornherein nicht genügen können.

– Als Folge davon werden zahlreiche Schüler frühzeitig durch wiederholte schlechte Zensuren entmutigt und von maximal fördernden Lernbedingungen ausgeschlossen.

– Die Leistungsurteile führen nicht zu pädagogischen Maßnahmen, die eine abgestimmte Antwort auf die festgestellten Mängel darstellen; Diagnose und »Therapie« sind nicht aufeinander bezogen.

Diese Kritik an der gegenwärtigen Praxis der Leistungsbeurteilung besagt nicht weniger, als daß in unseren Schulen Jahr für Jahr ein großer Teil unserer Kinder bei weitem nicht in einer Weise gefördert und betreut werden, die ihnen erlauben würde, ihre Lernmöglichkeiten voll zu entfalten und ihre Fähigkeiten optimal zu entwickeln. Darüber hinaus verursachen Prüfungsarbeiten, Zensuren und Zeugnisse bei zahllosen Schülern Ängste und führen zu familiären Konflikten, die vielleicht vermieden würden, wenn Beurteiler und Betroffene die seit Jahrzehnten bekannte und nicht widerlegte Kritik an der traditionellen Leistungsbeurteilung beherzigen und lernen würden, der geläufigen Überinterpretation von Zensuren oder auch Testergebnissen skeptisch zu begegnen und Widerstand entgegenzusetzen. Beispielsweise könnten neue Beurteilungssysteme erprobt werden, die sich nicht mehr einer breiten Skala von 1 bis 6, 1 bis 15 oder gar 1 bis 100 bedienen und somit einen uneinlösbaren Anspruch auf Präzision vorspiegeln, sondern die nur konstatieren, ob ein bestimmtes Lernziel erreicht wurde oder nicht (vgl. Viebahn 1977); oder man könnte sich die Erfahrungen zunutze machen, die an den Jena-Plan-Schulen oder an den Freien Waldorfschulen gemacht worden sind. So hat Petersen ([55]1974) anstelle der herkömmlichen Zeugnisse zwischen einem objektiven und einem subjektiven Bericht des Lehrers unterschieden. Im ersteren wurden die Eltern in Form einer inhaltlichen Beschreibung (ohne Ziffernzensuren) über den Leistungsstand ihres Kindes aufgeklärt; den subjektiven Bericht richtete der Lehrer an den Schüler, informierte ihn darin in kindgemäßer Form über seine Stärken und Schwächen und gab Hinweise auf die weiteren Arbeitschritte. Auch in den Rudolf-Steiner-Schulen verzichtet man auf Zensuren sowie auf das Sitzenbleiben. Statt dessen begleiten den Schüler ausführliche Charakterisierungen, die sich nicht nur auf seine Leistungen in den üblichen Schulfächern beziehen, sondern die soziale und emotionale Entwicklung sowie die Fähigkeit im musischen Bereich mit bedenken (vgl. z. B. Rauer 1973 sowie Wehr in diesem Bd.).

Langfristig wird man freilich vor allem die Pädagogische Diagnostik ihrem ursprünglichen Zweck, schulisches Lernen zu unterstützen, wieder zuführen müssen. Hierzu gehört zunächst die Ablösung des vorherrschenden interindividuellen durch den intraindividuellen Normenbezug, also die Ermutigung des einzelnen Kindes durch den Hinweis auf die

Lernfortschritte, die es gegenüber seiner eigenen Ausgangslage erzielt hat. Hierzu gehört ferner, diejenigen Merkmale der »Lernumwelt« der Kinder zu identifizieren, welche anregende Wirkungen ausüben und Freiraum für die Entfaltung individueller Interessen bieten, um dann entsprechende Verbesserungen vornehmen zu können. Denn mehr als andere Determinanten der Schulleistung ist die Lernumwelt gezielten Veränderungen zugänglich. Wichtiger als die Verfeinerung des pädagogisch-diagnostischen Instrumentariums ist also die Erforschung und Verbesserung der Bedingungen des Bildungsprozesses, in welchem Pädagogische Diagnostik nur einen Teilaspekt darstellt, der den Überlegungen zu den Zielen, Inhalten, Methoden sowie der Organisation des Unterrichts nachgeordnet sein muß. Die Lösung der wichtigsten Probleme, welche die derzeitige Pädagogische Diagnostik aufwirft, wird nicht darin liegen, ein Arsenal pädagogisch-diagnostischer Verfahren zu entwikkeln, mit dessen Hilfe sich in jedem Moment feststellen läßt, an welcher Stelle sich einzelne Schüler in einer bestimmten Unterrichtsstunde in einem bestimmten Schulfach genau befinden, um darauf mit der Anweisung geeigneter, weiterführender Aufgaben reagieren zu können. Vielmehr folgt aus der Einsicht in die vielfache Determiniertheit jedes Leistungsprozesses und jedes Leistungsproduktes durch individuelle Faktoren sowie durch die Eigenschaften der Lernumwelt zwangsläufig entweder Resignation angesichts der Fülle der zu bewältigenden diagnostischen Aufgaben oder die Suche nach anderen Unterrichtsmodellen, in welchen man ohne ein solches Arsenal an Verfahren auskommt. Ein Beispiel hierfür bietet der »Offene Unterricht«, wie man ihn an zahlreichen englischen, amerikanischen und kanadischen Grundschulen und an einigen Sekundarschulen beobachten kann (vgl. Calliess 1976). In der hier vorherrschenden flexiblen Unterrichtsorganisation, welche der Selbststeuerung des Lernens durch die Schüler einen zentralen Platz einräumt, sind die Lehrer nicht darauf angewiesen, mit Hilfe einer großen Zahl der herkömmlichen Beurteilungsverfahren den Leistungsstand der Schüler festzustellen, sondern die heutzutage üblichen Verfahren treten völlig in den Hintergrund zugunsten informeller Evaluationsvorgänge, die darin bestehen, daß einerseits die Lehrer eher beiläufig aus ihrer Kenntnis des einzelnen Schülers seine Lernprozesse zu steuern helfen, daß andererseits aber vor allem die Schüler lernen, weitgehend selbst darüber zu befinden, wo sie stehen, welche Aufgaben ihnen entsprechen und was sie weiterbringt (vgl. Hopf 1975).

Freilich werden sich derartige Unterrichtsmodelle im deutschsprachigen Raum kaum verwirklichen lassen, solange die Auslesefunktion der Schule weiterhin im Vordergrund steht und ihren entstellenden Einfluß auf Ziele, Inhalte und Formen des Unterrichts ausübt. So wie gegenwärtig die Proportion der für schulunreif gehaltenen Kinder bei der Einschulung in erstaunlichem Maße sinkt und die Schule sich offenbar den Kindern stärker als vorher anzupassen bereit ist, wird vielleicht in wenigen Jahren, wenn die schwachen Schülerjahrgänge auf die Sekundarschulen übergehen und die selektiven Schultypen um jeden Schüler »werben« müssen, um ihren Bestand nicht zu gefährden, das pädagogische Klima entstehen, welches für eine freiere Entfaltung kindlicher Fähigkeiten und Interessen Spielraum gewährt und in welchem Pädagogische Diagnostik vor allem der Förderung des Lernens dient.

ANMERKUNGEN

1

Zum hiermit angesprochenen Problem der »differentiellen Validität« vgl. u. a. Cronbach u. Suppes: »Research for tomorrow's schools« (1969, 82 ff). Differentiell valide sind Leistungsurteile übrigens nicht nur für verschiedenartige Lernumwelten, sondern auch für verschiedene Gruppen der Bevölkerung wie Frauen und Männer, Schwarze und Weiße, soziale Unterschicht und Mittelschicht usw., vgl. u. a. Shulman: »Reconstruction of educational research« (1970, 380) und die breite Literatur über die sogenannten kulturfreien Tests.

2

Belege für die meisten der in diesem Abschnitt aufgeführten Argumente zur Fragwürdigkeit der Notengebung finden sich vor allem in dem Buch Ingenkamps »Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte« (⁶1976), auf das der Leser hier ausdrücklich hingewiesen sei.

3

Zur prinzipiellen Unlösbarkeit des Problems objektiver Textbeurteilung vgl. Zeiher u. Zeiher (1977).

4

Ein Überblick über die deutschsprachigen Schultests findet sich bei Wulf (1974, 668–678).

LITERATUR

Barkey, P., Langfeldt, H.-P., Neumann, G.: Pädagogisch-psychologische Diagnostik am Beispiel von Lernschwierigkeiten. Bern 1976

Block, J. H. (Ed.): Mastery learning. New York 1971

Bloom, B. S.: Stability and change in human characteristics. New York 1964

Individual differences in school achievement: A vanishing point? New York 1971. Deutsch: Individuelle Unterschiede in der Schulleistung: ein überholtes Problem? In: W. Edelstein, D. Hopf (Hg.): Bedingungen des Bildungsprozesses. Psychologische und pädagogische Forschungen zum Lehren und Lernen in der Schule. Stuttgart 1973

Calliess, E.: Open education – die radikale Alternative? Was man von der englischen Primarschule lernen kann. Gesamtschule, 2, 1976, 12–15

Carroll, J. B.: A model of school learning. Teachers College Record, 64, (8), 1963. Deutsch: Ein Modell schulischen Lernens. In: W. Edelstein, D. Hopf (Hg.): Bedingungen des Bildungsprozesses. Psychologische und pädagogische Forschungen zum Lehren und Lernen in der Schule. Stuttgart 1973

Problems of measurement related to the concept of learning for mastery. Educational Horizons, 1970, 71–80

Cole, N. S.: Bias in selection. ACT research report No. 51, May 1972. The American College Testing Program, Iowa City

Cronbach, L., Gleser, G. C.: Psychological tests and personnel decisions. Urbana/Ill. 1965

Cronbach, L., Snow, R.: Aptitudes and instructional methods. A handbook of research on interactions. New York 1977

Cronbach, L., Supples, P. (Ed.): Research for tomorrow's schools. Disciplined inquiry for education. London 1969

Dahllöf, U.: Rahmenfaktoren und zielerreichendes Lernen. In: W. Edelstein, D. Hopf (Hg.): Bedingungen des Bildungsprozesses. Stuttgart 1973, 271–284

Döscher, D., Kuhr, H. J., Ziegenspeck, J.: Pädagogische Diagnostik. Annotierte Bibliografie (Aufsätze 1969–1976), BiB-Report. Duisburg 1977

Edelstein, W., Sang, F., Stegelmann, W.: Unterrichtsstoffe und ihre Verwendung in der BRD (Teil I). Studien und Berichte, Bd. 12. Berlin 1968

Flammer, A.: Individuelle Unterschiede im Lernen. Weinheim 1975

Floud, J., Halsey, A. H.: Social class, intelligence tests, and selection for secondary schools. In: A. H. Halsey, J. Floud, C. A. Anderson (Eds): Education, economy, and society. New York 1963

Gage, N. L.: Teacher effectiveness and teacher education. The search for a scientific basis. Palo Alto 1972

Garten, H.-K. (Hg.): Diagnose von Lernprozessen. Braunschweig 1977

Guthke, J.: Zur Diagnostik der intellektuellen Lernfähigkeit. Stuttgart 1977

Heckhausen, H.: Die Interaktion der Sozialisationsvariablen in der Genese des Leistungsmotivs. In: C. F. Graumann, Sozialpsychologie. Handbuch der Psychology, Bd. 7/2. Göttingen 1972, 955–1019

Höhn, E.: Der schlechte Schüler. Sozialpsychologische Untersuchungen über das Bild des Schulversagers. München 1967

Hörmann, H.: Aussagemöglichkeit psychologischer Diagnostik. Göttingen 1964

Hopf, D.: Übergangsauslese und Leistungsdifferenzierung. Eine Untersuchung am Beispiel der Grammar und Comprehensive Schools in England. Frankfurt/M. 1970

Entwicklung der Intelligenz und Reform des Bildungswesens. Bemerkungen zu B. S. Bloom, Stability and change in human characteristics. Neue Sammlung, 11, 1971, 33–51

Das Numerus-clausus-Verfahren. Möglichkeiten, Grenzen, Folgewirkungen. Neue Sammlung, 14, 1974, 180–189

Flexible Unterrichtsorganisation: Möglichkeiten und Grenzen. Neue Sammlung, 6, 1975, 520–537

Mathematikunterricht. Eine empirische Untersuchung zur Didaktik und Unterrichtsmethode in der 7. Klasse des Gymnasiums. Stuttgart 1980

Hoyt, D. P.: The relationship between college grades and adult achievement, a review of the literature. ACTP Res. Rep. No. 7, Iowa City 1965

Ingenkamp, K. H. (Hg.): Möglichkeiten und Grenzen des Lehrerurteils und des Schultests. In: H. Roth (Hg.): Begabung und Lernen. Stuttgart 1966, 407–432, ⁹1973

Pädagogische Diagnostik. Ein Forschungsbericht über Schülerbeurteilung in Europa. Trendbericht im Auftrag des Europarats in Straßburg. Weinheim 1975

(Hg.): Die Fragwürdigkeiten der Zensurengebung. Texte und Untersuchungsberichte. Weinheim ⁶1976

(Hg.): Schüler- und Lehrerbeurteilung. Empirische Untersuchungen zur pädagogischen Diagnostik. Weinheim 1977

Jackson, P. W.: Life in classrooms. New York 1968

Die Welt des Schülers. In: W. Edelstein, D. Hopf (Hg.): Bedingungen des Bildungsprozesses. Psychologische und pädagogische Forschungen zum Lehren und Lernen in der Schule. Stuttgart 1973

Klauer, K. J., Fricke, R., Herbig, M.: Lernzielorientierte Tests. Beiträge zur Theorie, Konstruktion und Anwendung. Düsseldorf 1972

Krapp, A., Schiefele, H.: Lebensalter und Intelligenzentwicklung. Eine Analyse des Entwicklungsmodells von B. S. Bloom. München 1976

Krappmann, L.: Spiele und soziale Lernziele. In: B. Daublebsky u. a. (Hg.): Spielen in der Schule. Stuttgart 1973

Kutscher, J. (Hg.): Beurteilen oder verurteilen. München 1977

Lienert, G. A.: Testaufbau und Testanalyse. Weinheim 1967

Lietzmann, W.: Über die Beurteilung der Leistungen in der Schule. Leipzig 1927

Lindvall, C. M., Bolvin, J. O.: Programed instruction in the schools. An application of programing principles in »individually prescribed instruction«. In: Programed instruction. 66th yearbook, II, NSSE, Chicago 1967

Lohnes, P. R.: Evaluating the schooling of intelligence. Educ. Researcher, 2, 1973, 6–11

MAGER, R. G.: Lernziele und programmierter Unterricht. Weinheim 1965

MCCLELLAND, D. C.: Testing for competence rather than for »intelligence«. In: American Psychol., Jan. 1973, 1–14

MICHEL, W. E.: Some lessons from high school physics. In: Proceedings, 1959 Invitational Conference on Testing Problems, ETS, Princeton, 1960, S. 17–26

PETERSEN, P.: Der kleine Jena-Plan. Weinheim [55]1974

RAUER, W.: Zur Beurteilung von Schülerleistungen in den Freien Waldorfschulen. Lebendige Schule, 28, 1973, 75–78

RAUH, H.: Schulleistungen und Übertrittsempfehlungen am Ende des 4. Schuljahres in ihrer Beziehung zur Entwicklung während der Grundschulzeit. Empirische Längsschnittanalyse eines komplexen Begabungsurteils. In: K. H. Ingenkamp (Hg.): Schüler- und Lehrerbeurteilung. Empirische Untersuchungen zur pädagogischen Diagnostik. Weinheim 1977, 15–64

ROEDER, P. M.: Dimensionen der Schulleistung. Modelle der Differenzierung in Abhängigkeit von Leistungsdimensionen einzelner Fächer. Gutachten und Studien der Bildungskommission des Deutschen Bildungsrates. Band 21, Teil 1. Stuttgart 1974

ROSENSHINE, B.: Teaching behaviors and student achievement. Stanford 1970

SCHREINER, G.: Sinn und Unsinn der schulischen Leistungsbeurteilung. Die Deutsche Schule, 62, 1970, 226–237

SCHULTE, D. (Hg.): Diagnostik in der Verhaltenstherapie. München 1974

SCHULTZE, W. (Hg.): Über den Voraussagewert der Auslesekriterien für den Schulerfolg am Gymnasium. Forschungsberichte der Max-Traeger-Stiftung, Heft 1, Frankfurt/M. 1964

Die Leistungen im Englischunterricht in der Bundesrepublik im internationalen Vergleich. Mitteilungen und Nachrichten des Deutschen Instituts für Internationale Pädagogische Forschung. Frankfurt/M. 1975

SCHWARZER, C., SCHWARZER, R. (Hg.): Diagnostik im Schulwesen. Studientexte zur pädagogischen Diagnose, Beratung und Entscheidung. Braunschweig 1977

SCHWARZER, R.: Mastery Learning durch programmierte Instruktion? Eine Untersuchung der Beziehungen zwischen Lernerfolg, Intelligenz und Arbeitszeit beim programmierten Unterricht. Phil. Diss. Kiel 1972

SHULMAN, L. S.: Reconstruction of educational research. Review of Educational Research, 40, 3, 1970, 371–396

TENT, L., FINGERHUT, W., LANGFELDT, H.-P.: Quellen des Lehrerurteils. Untersuchungen zur Aufklärung der Varianz von Schulnoten. Weinheim 1976

TREUMANN, K.: Dimensionen der Schulleistung. Leistungsdimensionen im Mathematikunterricht. Gutachten und Studien der Bildungskommission des Deutschen Bildungsrates. Bd. 21, Teil 2. Stuttgart 1974

ULICH, D., MERTENS, W.: Urteile über Schüler. Zur Sozialpsychologie pädagogischer Diagnostik. Weinheim 1973

VERNON, P. E. (Hg.): Secondary school selection. A British Psychology Society Inquiry. London 1957

VEROFF, J.: Wie allgemein ist das Leistungsmotiv? In: W. Edelstein, D. Hopf (Hg.): Bedingungen des Bildungsprozesses. Psychologische und pädagogische Forschungen zum Lehren und Lernen in der Schule. Stuttgart 1973

VIEBAHN P.: »Bestanden – nicht bestanden« als Bewertungskategorien in Prüfungen. Überblick über empirische Befunde zum Vergleich von stark und schwach gegliederten Beurteilungssystemen. Psychologie in Erziehung und Unterricht, 24, 1977, 231–240

WECHSLER, D.: The measurement and appraisal of adult intelligence. Baltimore 1958. Deutsch: Die Messung der Intelligenz Erwachsener. Bern 1964

WOOD, R., SKURNIK, L. S.: Item banking. National Foundation of Educ. Research, London 1969

WULF, D., u. a. (Hg.): Wörterbuch der Erziehung. München 1974

ZEIHER, H., ZEIHER, H. J.: Objektive Textbeurteilung – ein unerreichbares Ziel und seine didaktischen Folgen. Zeitschrift für Pädagogik, 23, 1977, 183–194

ZEIHER, H. J.: Unterrichtsstoffe und ihre Verwendung in der 7. Klasse des Gymnasiums in der BRD (Teil II): Deutschunterricht. Studien und Berichte, Bd. 24. Berlin 1972