Psychologie in Erziehung und Unterricht

21. Jahrgang 1974

ZEITSCHRIFT FÜR FORSCHUNG UND PRAXIS

Herausgegeben von:
Prof. Dr. Heinz-Rolf Lückert, München
Prof. Dr. Horst Nickel, Düsseldorf
Dr. Anne-Marie Tausch, Hamburg

FORTSETZUNG VON „SCHULE UND PSYCHOLOGIE"

Ernst Reinhardt Verlag München/Basel

INHALT

(Ziffern in Klammer hinter den Seitenzahlen verweisen auf die Heft-Nummer des 21. Jg.)

Originalarbeiten

Übersichtsartikel

Sammelreferate

Zeitschriftenreferate

Forschung und Praxis – aktuell

Ernst Reinhardt Verlag, München/Basel
Psychol. in Erzieh. u. Unterr., 21. Jg., Seite 105–124 (1974)

Übersichtsartikel

Kurt Heller

Zur Problematik der Leistungsbeurteilung in der Schule*

Problems of comprehending and judging achievement at school

Summary: This contribution brings a review of the actual and often discussed problems in comprehending and judging achievement at school. The author deals with theoretical and practical questions in like manner. After a short discussion of terms, an operational determination of the thing generally called "achievement at school" (Schulleistung) follows. Then the intellectual and non-intellectual determinants of achievement at school are shown and their importance for the advancement of achievement at school is emphasized. The discussion of the objective and subjective methods for judging achievement at school is the main theme. Classical and modern testtheoretical tendencies are critically analysed concerning their relevance in judging achievements of pupils. Special accent is put to the description of objective methods in judging achievements. Intensive discussion is dedicated to subjective methods in judging achievements of pupils by teachers. New varianceanalytical findings are discussed to the problem of giving marks as well as experimental results about different influences of judging essays. The last paragraph shows ways to unify the marks-giving to school-essays.

Zusammenfassung: Der Beitrag vermittelt einen Überblick über das aktuelle und vieldiskutierte Problem der Erfassung und Beurteilung schulischer Leistungen. Dabei wird auf theoretische und praxisrelevante Fragen gleichermaßen eingegangen. Nach einer kurzen Begriffsdiskussion wird zunächst eine operationale Bestimmung von Schulleistungen versucht. Danach werden die intellektuellen und außerintellektuellen Determinanten der Schulleistung aufgewiesen und ihre Bedeutung für die Schulleistungsförderung unterstrichen. Im Mittelpunkt steht die Erörterung von objektiven und subjektiven Methoden der Schulleistungsbeurteilung Dabei werden klassische und moderne testtheoretische Ansätze auf ihre Relevanz für die Beurteilung von Schülerleistungen kritisch untersucht. Besonderes Gewicht wird auf die Beschreibung der objektiven Verfahren der Leistungsbeurteilung gelegt, aber auch die subjektiven Verfahren der Schülerbeurteilung durch Lehrer werden eingehend erörtert. Hier kommen u. a. neue varianzanalytische Befunde zur Notengebung sowie umfangreiche experimentelle Untersuchungsergebnisse über die verschiedenen Einflüsse der Aufsatzbeurteilung zur Sprache. Abschließend werden Möglichkeiten zur Vereinheitlichung der Zensierung von Schüleraufsätzen aufgezeigt.

1. Gegenstand schulischer Leistungsbeurteilungen

1.1. Zur Problematik des Leistungsbegriffs in der Pädagogik

Mit dem Begriff der *Schulleistung* sei hier das gesamte Leistungsverhalten, soweit es im Kontext schulischer Bildungsbemühungen virulent wird, angesprochen. Dabei verdienen der dynamische Aspekt (Lernprozeß) und der statische Aspekt (Lernprodukt) gleichermaßen Beachtung – entgegen manchen Begriffsvorstellungen, die nur das Ergebnis der Schülerleistung und nicht auch deren Bedingungsgefüge im

* Gekürzte und teilweise überarbeitete Wiedergabe des Einleitungs- und Übersichtsreferats in *K. Heller* (Hrsg.): Leistungsbeurteilung in der Schule. Quelle & Meyer, Heidelberg 1974. – Der Abdruck erfolgt mit freundlicher Genehmigung des Verlags Quelle & Meyer.

Auge haben. Diesen mehr oder weniger operational definierbaren Kategorien können schließlich ethische (Leistungspflicht) und ökonomische Überlegungen (Leistungsnotwendigkeit) – als Sinndimensionen menschlicher Leistungen überhaupt – hinzugefügt werden.

Daraus resultiert nach *Furck* (1964, S. 118 ff.) eine doppelte Aufgabenfunktion der Schulleistung, nämlich die Persönlichkeitsbildung des Schülers und die Sicherung des volkswirtschaftlichen Leistungspotentials. „Das Problem der Leistung in der Schule spitzt sich so zu der Frage nach dem rechten Verhältnis von individueller Bildsamkeit und ihr angemessener Anforderung zu" (loc. cit.). Damit ist im weiteren Sinne das pädagogisch-didaktische Anliegen der Leistungsförderung angesprochen. Sicherlich lassen sich noch andere Aspekte und möglicherweise ganz neue Perspektiven zum Schulleistungsproblem aufweisen, die bislang artikulierten, realisierbaren Ansätze (z. B. von *Furck, v. Hentig* oder *Gaude & Teschner* u. a.) kulminieren letzten Endes doch immer wieder um die hier skizzierten Problembereiche.

Unsere Definitionsformel „Schulleistung meint das gesamte Leistungsverhalten im Kontext der Schule" schließt prinzipiell das Schüler- (bzw. Eltern-) *und* Lehrerverhalten ein, wenngleich die Vokabel „Schulleistung" häufig nur die Leistung des Schülers indiziert. Ein solchermaßen eingeengter Leistungsbegriff ist jedoch nur akzentuierend, d. h. im Hinblick auf jeweils thematisierte Fragestellungen, angebracht. Die Gefahr, hierbei den Gesamtkomplex interdependenter Zusammenhänge aus dem Auge zu verlieren, ist freilich nicht von der Hand zu weisen. Wir müssen davon ausgehen, daß Schulleistungen mehrdimensional bedingt sind und in hohem Maße von persönlichkeits- und sozialpsychologischen Faktoren abhängige Variablen(bündel) darstellen.

1.2. Hauptdimensionen der Schulleistung

Ausgehend von den üblichen Indikatoren der Schulleistung versuchten *Landfeldt & Fingerhut* (1974) eine empirische Beschreibung dieses Phänomens. Die zu diesem Zweck faktorenanalysierten Schulzeugnisse teils eigener, teils fremder Provenienz deuten übereinstimmend an, „daß die Schulleistung – so wie sie durch Schulnoten erfaßt wird – bei weitem nicht so differenziert ist, wie es Zeugnisse mit zehn und mehr Einzelnoten nahelegen" (S. 40). Für die Schülerleistungen auf der Sekundarstufe konnten lediglich drei Faktoren nachgewiesen werden: ein Fremdsprachenfaktor, ein mathematisch-naturwissenschaftlicher Faktor und ein sachkundlicher Faktor (sensu *Denig & Weis* 1970). Für die Primarstufe (Grundschüler und lernbehinderte Sonderschüler) ergaben sich sogar nur zwei Faktoren: ein allgemeiner Schulleistungsfaktor (vor allem auf die Fachzensuren in Deutsch und Rechnen bezogen) und ein Faktor der schulischen Disziplin (mit entsprechenden Ladungen auf den sog. Kopfnoten). Die Ergebnisse von *Funke* (1972) und *Zimmermann* (1968) bestätigen – unabhängig voneinander – diese gegenüber dem wesentlich differenzierteren Zeugnisbild stark reduzierte Faktorenstruktur der tatsächlichen Schülerleistung. Im Hinblick auf die Schulleistungsbeurteilung herkömmlicher Art (Zensurierung) ergäbe sich hieraus die Konsequenz, künftig nicht mehr als zwei oder drei Schulnoten zu vergeben. Zumindest unter dem Gesichtspunkt streng leistungsbezogener Beurteilung sind mehr als drei Einzelnoten irreführend und überflüssig zugleich.

Sofern man neben den Lehrerurteilen in Form von Zensuren noch weitere Daten, z. B. Schulleistungstests, Intelligenztests, biographische Informationen usw., in die Faktorenanalyse miteinbezieht, gewinnt man ein höher strukturiertes Modell der Schulleistung. Wie *Fingerhut & Langfeldt* (1971) unter Bezug auf ein entsprechendes Datenmaterial von 1756 Viertklaßkindern nachweisen konnten, ergeben sich hierbei folgende vier Faktoren: 1. „Schulische Leistungsfähigkeit" mit Ladungen auf den Zensuren der Kernfächer (Deutsch, Rechnen, Heimatkunde), sämtlichen Skalen des AST 4 (Allgemeiner Schulleistungstest von *Fippinger)* und den Untertests 3 + 4 des LPS (Leistungsprüfsystem von *W. Horn);* 2. „Genereller Notenfaktor" mit Ladungen auf den Fachleistungs- und Kopfnoten sowie den Variablen Alter und Geschlecht; 3. „Test-Intelligenz" mit Ladungen auf sämtlichen Skalen des LPS; 4. „Rechenleistungen" mit Ladungen auf der Rechennote, den Rechenskalen der AST 4 und den LPS-Subtests 3 + 4, 14 und 15 (Arbeitsprobe).

Diese Ergebnisse lassen einen gewissen Zusammenhang zwischen einigen Intelligenzfaktoren (Reasoning, Accuracy, Number) und der (Grund-)Schulleistung, besonders hinsichtlich des allgemeinen Leistungsstandards und der Rechenleistung, erkennen. Die gerade in der jüngeren Literatur häufig vertretene Auffassung, wonach zwischen Intelligenz und Schulleistung keine oder nur schwache Zusammenhänge bestehen, muß demnach korrigiert bzw. folgendermaßen präzisiert werden. „Es kann ... nur gesagt werden, daß Schul*noten* und Intelligenz relativ unabhängig von einander sind", nicht aber Schul*leistung* und Intelligenz. Da Schulnoten allein nach den Befunden von *Langfeldt & Fingerhut* kein adäquates Abbild der Schulleistung geben, können diese „nur in mäßigem Umfange etwas über die *tatsächlichen* Bedingungen der Schulleistung aussagen" (1974, S. 42). Dieses Ergebnis erhellt indirekt die Bedeutung kognitiver Faktoren im Hinblick auf die Konstituierung von Schulleistungen, ohne hiermit den Einfluß nicht-kognitiver Determinanten als gering einzuschätzen.

Eine modifizierte Faktorenanalyse der Daten von *Fingerhut & Langfeldt* ergab schließlich fünf abgrenzbare Cluster: 1. Intelligenzleistungen (LPS), 2. Schulleistungen (AST 4), 3. Fachleistungsnoten (Deutsch, Rechnen, Heimatkunde), 4. Kopfnoten (Betragen, Fleiß, Aufmerksamkeit, Ordnung), 5. Merkmale biographischer Art (Alter, Geschlecht, soziale Herkunft). Aus der Distanz der Cluster zueinander postulieren die Autoren jetzt ein zweidimensionales Konstrukt „Schulleistung", demzufolge sich Schulleistungen im gegenwärtigen Bildungssystem als Funktion von „tatsächlicher Leistung aufgrund von Begabung" (Faktor I) und „Anpassung an dieses System" (Faktor II) erweisen. Zur Stützung ihres Interpretationsmodells, das die Autoren als – vorläufigen – Diskussionsbeitrag gewertet wissen wollen, können sie auf faktorenanalytische Untersuchungsbefunde von *Seitz* und *Löser* aus den Jahren 1969 bis 1971 verweisen.

1.3. Bedingungskomponenten der Schulleistung

Wenn im folgenden von *Determinanten* die Rede ist, so gilt es zu beachten, daß sich dieser Begriff auf *korrelative* Beziehungen beschränkt, d. h. nicht ohne weiteres etwas über Ursache-Wirkungs-Zusammenhänge aussagt. Strenggenommen kann bei

Beziehungen dieser Art nicht exakt „zwischen ‚verursachenden' Determinanten und ‚enthaltenen' Faktoren" *(Gaedike)* unterschieden werden. Während der zweite Aspekt bereits Gegenstand der vorhergehenden Betrachtungen war, sollen nun in Anlehnung an den *Gaedike*schen Beitrag (1974) vorrangig Bedingungskomponenten (erster Bedeutungsaspekte von ‚Determinante') der Schulleistung untersucht werden. Für die Interpretation der hier einschlägigen Befunde gelten freilich die angeführten (methodologischen) Kautelen ungemindert.

Zunächst werden die *kognitiven* Determinanten der Schulleistung in den Blick genommen, also die Frage, „ob ‚gute' Schüler auch zugleich ‚intelligente' Schüler sind und umgekehrt". Die meisten der von *Gaedike* gesammelten Untersuchungsergebnisse zur Korrelation zwischen (allgemeinen) Intelligenztest- und Schulleistungswerten liegen im Bereich von $r = 0.40$ und $r = 0.70$, gelegentlich auch darüber oder darunter. Dabei korrelieren „schulnahe" Intelligenztests (Verbaltests) im allgemeinen mit der Schulleistung höher als „schulferne" (Handlungstests). Ferner korrelieren Intelligenztests höher mit Schulleistungs*tests* als mit anderen Indikatoren, z. B. Schulzensuren. Entsprechende Zusammenhänge fallen bei jüngeren Schülern (Grundschule) deutlicher aus als bei älteren Schülern (Sekundarstufe) und dort wiederum deutlicher bei Hauptschülern als bei Realschülern und Gymnasiasten, was analog auch – durchgängig – für den Vergleich von Mädchen und Jungen gilt.

Allerdings tragen nicht alle kognitiven Faktoren in gleichem Maße zum Schulerfolg bei. So bestimmen vor allem verbale Fähigkeiten, Faktoren des logischen und schlußfolgernden Denkens (Reasoning) sowie Zahlenverständnis, teilweise auch technische Fähigkeiten (Space) und Wahrnehmungsfaktoren (Perceptual Speed, Accuracy) den Bildungserfolg, gemessen an den traditionellen Formen der Schulleistung (siehe noch *Heller* 1970, S. 127 ff.). „Es kommt also nicht (nur) darauf an, intelligent oder kreativ zu sein, sondern (von den Bildungsinstitutionen) bevorzugt werden Schüler, die in ganz bestimmter Weise intelligent sind" *(Gaedike* 1974, S. 60).

Ausführlicher werden von derselben Autorin die *nicht-kognitiven* Determinanten der Schulleistung erörtert. Hierbei gewinnen die Lern- und Leistungsmotivation sowie Faktoren der Arbeitshaltung des Schülers, Persönlichkeitsvariablen wie Ängstlichkeit, Selbstachtung, Extraversion vs. Introversion, aber auch Merkmale des Lehrerverhaltens (z. B. Setzung „sachfremder" Leistungsmotivation, Entwicklung und Pflege kognitiver Stile, Unterrichtsatmosphäre, direktives Verhalten, Werthaltungen und Einstellungen) sowie äußere und innere Bedingungen des soziokulturellen Hintergrundes (Familie, Peergroups) mehr oder weniger starken Einfluß auf das Leistungsverhalten des Schülers. Die aufgezählten Determinanten bilden ein Bezugsgeflecht vielseitiger und mehrschichtiger Einflußgrößen. Dabei scheinen die Bezugsmuster einem Variationsspielraum ausgesetzt zu sein, der größer ist als der, dem die Phänomene der Schulleistung selbst unterliegen. Somit liegt nach dem resümierenden Urteil der Autorin „noch ein weites Feld für entsprechende Forschungen vor uns", bis alle Beziehungen aufgedeckt und deren Kenntnis für eine optimale Schulleistungsförderung eingebracht werden können.

2. Aufgaben und Ziele schulischer Leistungsbeurteilung

2.1. Leistungsbeurteilung im Dienste der Unterrichtsorganisation und Bildungsreform

Diese Funktion der Leistungsbeurteilung steht in unmittelbarem Zusammenhang mit schulpädagogischen bzw. unterrichtsdidaktischen Problemen. „Nicht die Beurteilung und Benotung eines Schülers, die fast schon zum Selbstzweck geworden ist, sondern die Bewertung des Unterrichtserfolges und die Erfassung von Lernschwierigkeiten ist u. E. die vornehmste Aufgabe pädagogisch-psychologischer Prüfmethoden, ausgerichtet auf das Ziel, rechtzeitiges und wirkungsvolles pädagogisches Handeln zu ermöglichen und zu initiieren" *(Zielinski* 1973, S. 58). Unter dieser Prämisse dienen schulische Leistungskontrollen vorab zur Überprüfung aufgestellter Lernziele und zur Diagnose des Lehr-Lernprozesses. Die so gewonnenen Informationen bieten unverzichtbare Hilfen für die Curriculumentwicklung und -revision, die Lernzielbestimmung, die didaktische Planung und Analyse des Unterrichtsverlaufs u. dgl. m. Die Leistungsbeurteilung ist somit notwendiger Bestandteil einer optimalen Unterrichtsgestaltung.

Unter dem Aspekt der (inneren) Bildungsreform wäre hier die Wechselwirkung von Curriculumentwicklung und Leistungsbeurteilung hervorzuheben (vgl. Symposium des Europarates in Berlin 1971). So erfordert die Überprüfung festgelegter Curricula, z. B. im Hinblick auf schulpädagogische und lernpsychologische Möglichkeiten, jeweils bestimmte Methoden ihrer Bewertung. Theoretisch bedeutet dies, daß *alle* kognitiven und nicht-kognitiven Determinanten der konkret geforderten Schülerleistung in einem solchen Bewertungssystem Berücksichtigung finden müssen. Praktisch ist man dieser Idealforderung bislang allenfalls bezüglich der Kontrolle der kognitiven Variablen nahegekommen, wohingegen nicht-kognitive Variablen nur gelegentlich Beachtung fanden. Andererseits bewirkt die Entwicklung bestimmter Beurteilungsmethoden oft auch Modifikationen und Ergänzungen schulischer Curricula, deren Beschränkung auf kognitive Inhalte keine Notwendigkeit darstellt. Daraus läßt sich die Forderung ableiten, die Diskussion über schulmethodische und curriculare Neuerungen nicht isoliert, sondern in wechselseitiger Abstimmung voranzutreiben. Nur so kann man der Gefahr unerwünschter Stabilisierungseffekte wirksam begegnen.

2.2. Leistungsbeurteilung als Funktion individueller Beratung

Dieser Aspekt der Leistungsbeurteilung zielt auf die „Auto-Evaluation", die Orientierung des Schülers (bzw. seiner Eltern) über seine eigenen Schulleistungen, seine Lernfortschritte (intraindividueller Vergleich) und seine Position innerhalb der Klassen- oder Altersgruppe (interindividueller Vergleich). Dazu ist zweierlei notwendig: erstens die genaue Kenntnis der Lernanforderungen, was eindeutig definierte Lernziele erfordert, und zweitens die Transparenz des Beurteilungssystems, dessen unerläßliche Bestandteile objektive, zuverlässige und gültige Bewertungskriterien bilden. Ferner müssen dem Schüler Notwendigkeit, Zweck und Formen der Leistungsbeurteilung einsichtig gemacht werden.

Während im vorigen Abschnitt unterrichtsdidaktische und somit lehrerzentrierte bzw. gruppenimmanente Interessen (z. B. der Schulklasse bzw. Lerngruppe) im Vordergrund der Betrachtung standen, kommt unter dem Gesichtspunkt schülerzentrierter Aufgabenfunktionen der Individualdiagnose besondere Bedeutung zu. Hier sind diagnostizierte Begabungs-Leistungsdiskrepanzen (z. B. Phänomene des sog. Underachievement versus Overachievement, d. h. unter- versus übererwartungsgemäße Schulleistungen), vorübergehende oder andauernde Lern- und Leistungshemmungen (z. B. Konzentrationsstörungen, sozio-kulturelle oder sensorische Deprivationserscheinungen wie Sprachbarrieren oder Hör- und Sehfehler), aber auch partielle versus allgemeinere Leistungsschwächen (z. B. Legasthenie, Rechenschwäche usw. versus Lernbehinderung[en] im Sinne der Sonderschulbedürftigkeit) sehr oft unerläßliche Voraussetzung für die Bestimmung angemessener (schul)pädagogischer Maßnahmen bzw. therapeutischer Behandlungsansätze. Darüber hinaus will – und sollte – jeder Schüler, auch der ‚unauffällige', konstinuierlich über den persönlichen Stand und Fortschritt und erst recht über eventuelle Rückschritte im Lernprozeß informiert werden. Ob diese Information durch Tests, sog. Diagnosebögen oder Zensuren erfolgt, ist prinzipiell von zweitrangiger Bedeutung, sofern die verwendeten Methoden der Leistungsbeurteilung hinreichend objektiv, verläßlich und gültig sind. Daß diese Anforderungskriterien in der Praxis der Schülerbeurteilung – besonders bei den subjektiven Verfahren (z. B. der Notengebung) – häufig nicht erfüllt sind, betrifft eine Reihe von Problemen, die im folgenden Kapitel (Kap. 3) zur Sprache kommen.

2.3. Leistungsbeurteilung als Funktion der Schullaufbahn- bzw. Systemberatung

Die bisher erörterten Funktionen schulischer Leistungsbeurteilung sind nicht unabhängig von der Struktur des jeweiligen Schul- und Bildungssystems, wie umgekehrt eine gewisse feed-back-Wirkung auf die Bildungsinnovation zu erwarten ist. Von unterschiedlichen Aufgabenschwerpunkten und entsprechend modifiziert eingesetzten Beurteilungstechniken (etwa im dreigliedrigen versus Gesamtschulsystem) abgesehen, kulminieren sämtliche Beurteilungs- und Beratungsansätze letztlich um die Differenzierungsproblematik. An dieser Frage und ihrer Lösung führt m. E. kein Weg vorbei, wenigstens ist bislang kein funktionierendes Schulsystems, das diese Problematik ohne Schaden ausklammern könnte, bekannt.

Die Schullaufbahnberatung im traditionellen (vertikal gegliederten) Schulsystem orientiert sich vornehmlich an den Anforderungskriterien der einzelnen Schularten (Hauptschule, Realschule, Gymnasium), für die – über kürzere oder längere Perioden – in etwa konsistente Lern-Leistungsbedingungen bzw. invariante Schülermerkmale postuliert werden. Der Gefahr systemstabilisierender Tendenzen versucht man hier durch regelmäßige Leistungskontrollen mit entsprechenden Übergangsmöglichkeiten (Prinzip der Durchlässigkeit) zu begegnen. Eine solche Laufbahnberatung müßte sich allerdings weniger am Selektions- als vielmehr am Klassifikationsmodell orientieren (vgl. *Heller* 1970).

In der Gesamtschule verlagern sich analoge Wahl- und Entscheidungsprozesse mehr nach „innen", d. h. in den Unterricht und das didaktische Handeln. Hier werden dann Kriterien benötigt, die angemessene Gruppierungen im Niveauunter-

richt, Wechsel zwischen Fachleistungskursen u. ä. erlauben bzw. sinnvolle Wahlpflichtkombinationen im Hinblick auf die Berechtigungsfunktion der Schulabschlußqualifikation garantieren. Dazu bedarf es wiederum objektiver Lernleistungskontrollen.

Die angeführten Beispiele zeigen, daß es bislang und auch wohl in absehbarer Zukunft kein Bildungssystem gibt, das auf die Funktion der Leistungsbeurteilung in dieser oder jener Form verzichten könnte. Die unterschiedlichen Methoden schulischer Leistungsbeurteilung und ihre Einsatzmöglichkeiten im Dienste der hier getrennt aufgewiesenen, in Wirklichkeit jedoch verzahnten, Aufgabenfunktionen stehen deshalb im Mittelpunkt der folgenden Ausführungen. Dabei sollen theoretische und praktische Probleme gleichermaßen Beachtung finden.

3. Formen und Methoden der Leistungsbeurteilung im Bildungswesen

3.1. Testtheoretische Grundlagen

3.1.1. Die klassische Testtheorie und ihre Kritik

Die klassische Testtheorie, auch Meßfehlertheorie genannt, bildet nach wie vor die Hauptgrundlage standardisierter (formeller und informeller) Schulleistungstests. Jeder, der sich solcher Meßverfahren im Rahmen der Schülerbeurteilung bedient, sollte deshalb – schon um unkritischer Anwendung und Fehleinschätzungen vorzubeugen – die theoretischen Voraussetzungen testdiagnostischen Vorgehens in etwa kennen. Da die klassische Testtheorie unabhängig vom Inhalt der einzelnen Verfahren (z. B. Intelligenztest, Leistungstest usw.) gilt, maß man diesem Konzept in der pädagogisch-psychologischen Diagnostik bis in die jüngste Vergangenheit praktisch uneingeschränkte Bedeutung bei. Obwohl heute die Relativierung dieses Anspruchs unbestritten ist, finden sich zunehmend Tendenzen, die Theorie als solche überhaupt abzulehnen. Abgesehen davon, daß man damit das Kind mit dem Bad ausschüttet, entarten solche Versuche nicht selten in absurde ideologische Verstrickungen, wie ein kürzlich erschienener Artikel in der Deutschen Schule eindrucksvoll demonstriert (vgl. *Neander* 1973). Damit ist m. E. weder der theoretischen Neubesinnung noch den praktischen Bedürfnissen in irgendeiner Weise gedient.

Das Wort „Test" bedeutet ursprünglich „Prüfung", „Stichprobe", oder „Zeugnis", wobei sich der Begriff in der Testdiagnostik vor allem auf standardisierte Verhaltensstichproben bezieht, d. h. auf Prüfungen, die unter genau festgelegten Bedingungen durchgeführt werden. Solche Prüfverfahren oder *Tests* messen bestimmte Merkmalsausprägungen (z. B. Intelligenz, Schulleistung, Ängstlichkeit usw.), indem sie *inter*individuelle Unterschiede (Unterschiede zwischen den einzelnen Individuen) oder *intra*individuelle Unterschiede (Unterschiede in bezug auf dieselbe Person, z. B. im Verlauf der Ontogenese) erfassen. Bei den sog. standardisierten Schulleistungstests steht der erste Aspekt im Vordergrund, d. h. der Vergleich eines individuellen Meßwertes im Test mit der Gruppennorm, die als Testleistungsmaßstab in Form von Alters- und/oder Schul- bzw. Klassennormen operational definiert ist.

In dem testtheoretischen Beitrag von *Langfeldt* (1974) werden zunächst die Begriffe „Messen" (Testen) und „Meßskala" sowie die wichtigsten Anforderungs-

kriterien in bezug auf Messungen, auch „Gütekriterien" genannt, erläutert. *Messen* (Testen) wird hier als ein Vorgang des Vergleichens anhand eines Maßstabs (Testnormen) aufgefaßt, wobei die Meßwerte durch unterschiedliche Skalenniveaus repräsentiert sein können. Die wichtigsten Skalentypen sind: 1. die *Nominal-* oder *Klassifikationsskala* (unterste Ebene des Messens), die nur einem Kriterium, dem der Äquivalenz, genügt; 2. die *Ordinal-* oder *Rangskala* (nächsthöhere Ebene), die zwei Kriterien, dem der Äquivalenz und dem der rangmäßigen Beziehung, genügt; 3. die *Intervallskala,* die drei Kriterien, der Äquivalenz, der Rangbeziehung und der Intervallkonstanz, genügt; 4. die *Rational-* oder *Verhältnisskala* (oberste Ebene des Messens), die zusätzlich zu den aufgeführten Kriterien einen absoluten Nullpunkt aufweist, also vier Kriterien genügt. Je höher das repräsentierte Skalenniveau ist, desto besser und vielfältiger sind die Verarbeitungsmöglichkeiten entsprechender Kennwerte und damit ihr Informationsgehalt. Schulleistungstests messen in der Regel auf Rangskalenniveau, selten auf Intervallskalenniveau (z. B. Intelligenztests).

Unabhängig vom jeweiligen Skalenniveau wird der Interpretationsspielraum noch durch die sog. *Testgütekriterien* (Objektivität, Reliabilität, Validität) beeinflußt. *Objektivität* meint hier die Unabhängigkeit der Testergebnisse von der Person des Testleiters sowohl in bezug auf die Testanweisung (Instruktion) als auch in bezug auf die Testauswertung und Testinterpretation. Die Durchführungsbestimmungen eines Tests müssen so präzise festgelegt sein, daß Intersubjektivität gewährleistet ist. Diese Forderung ist in der Praxis der Testdurchführung nicht immer leicht zu erfüllen. Die *Reliabilität* oder Zuverlässigkeit bezieht sich auf die Meßgenauigkeit. Ein Schulleistungstest ist beispielsweise dann reliabel (zuverlässig), wenn die Ergebnisse unabhängig vom Zeitpunkt der Messung zustandekommen, d. h. wenn die wiederholte Anwendung des betr. Tests bei derselben Klasse oder einzelnen Individuen in etwa zum gleichen Resultat führt. Es gibt verschiedene Aspekte der Reliabilität und dementsprechend unterschiedliche Methoden ihrer Kontrolle, die alle von *Langfeldt* besprochen werden. Die *Validität* oder Gültigkeit bezieht sich auf die Genauigkeit, mit der ein Test das mißt, was er messen soll. So sagt die Validität eines Rechentests etwas darüber aus, wie genau tatsächlich Rechenkenntnisse (und nicht etwa Konzentrationsfähigkeit oder andere Merkmale) erfaßt werden. Während sich also die Reliabilität auf die Zuverlässigkeit des Meßinstrumentes (Tests) selbst und damit auf die formale Meßgenauigkeit bezieht, informiert uns die Validität darüber, „welche psychodiagnostischen Schlußfolgerungen die numerischen Resultate eines Tests zulassen und welchen Grad an Sicherheit solche Schlußfolgerungen aufweisen" (Michel 1964, S. 47). Damit ist die diagnostische Valenz oder Treffsicherheit eines Testverfahrens angesprochen. Was endlich den Zusammenhang zwischen Reliabilität und Validität betrifft, so gilt: Die Reliabilität ist die notwendige, aber keine hinreichende, Voraussetzung für die Validität eines Tests. Sehr oft erweist sich die Validierung eines Schultests als das schwierigste Unterfangen überhaupt, wobei man sich nicht selten mit der Bestimmung der curricularen Gültigkeit begnügt. Seltener kommen hier die Methoden der Kriterienvalidierung (z. B. zur Ermittlung der Übereinstimmungsvalidität von Testleistung und Schulnoten als „Außenkriterien") oder der Konstruktvalidierung (analog zur

Validierung „neuer" Intelligenztests; vgl. *Heller* 1973, S. 75) zum Einsatz. Bei der Konstruktion eines Schulleistungstests kommt man den Anforderungskriterien der Reliabilität und Validität dadurch entgegen, daß man schon frühzeitig – in der sog. Aufgaben- oder Itemanalyse – *Schwierigkeit* und *Trennschärfe* jeder einzelnen Testaufgabe ermittelt, um von vornherein die unbrauchbaren Aufgaben (Items) auszusondern.

Damit nun die Ergebnisse eines Schulleistungstests vernünftig interpretiert werden können, sind zwei weitere Voraussetzungen unerläßlich: die Normierung des Tests und die Ermittlung des sog. (Standard-)Meßfehlers. In den *Testnormen* (z. B. PR = Prozenträngе auf Ordinalskalenniveau versus Z, C, Abweichungs-IQ u. ä. Standardwert-Normen auf Intervallskalenniveau bzw. flächentransformierte T-Standard-Äquivalent-Normen) liegen hierfür relativierte Testwerte, d. h. auf die jeweilige Alters- oder Klassengruppe bezogene Vergleichsmaßstäbe vor. So besagt eine Schülerleistung auf dem 75. PR, daß der betreffende Schüler in dem untersuchten Schulfach bessere Leistungen erzielt als 75% seiner Bezugsgruppe (und schlechtere Leistungen als 25%), während eine Leistung auf dem 50. PR durchschnittliche Leistungsfähigkeit indiziert. Nun wäre es naiv anzunehmen, daß die ermittelte (beobachtete) Testleistung von PR = 75 exakt die tatsächliche (wahre) Leistungsfähigkeit des Schülers wiedergibt; vielmehr muß davon ausgegangen werden, daß in diesem Testergebnis – wie in jedem Meßwert – eine zunächst unbekannte Fehlerquote steckt, die zu Lasten der Irreliabilität bzw. mangelnden Objektivität einer Untersuchung geht. Diese Annahme trifft im Kern das *Meßfehler*konzept, das wichtigste Axiom der klassischen Testtheorie. Danach setzt sich jeder *beobachtete Wert* (Meßwert) aus dem „wahren Wert" (einem zeitlich konstanten Parameter) und einem „Fehlerwert" (Meßfehler) zusammen. Beide werden als unkorrelierte Größen postuliert, wobei der Meßfehler als Zufallsvariable auftreten soll. Mit dessen Hilfe werden dann – auf der Basis wahrscheinlichkeitstheoretischer Überlegungen – die sog. *Vertrauensintervalle* oder *Konfidenzintervalle* ermittelt, d. h. jene Bereiche, in denen mit einer bestimmten Wahrscheinlichkeit der wahre Wert erwartet wird. Auf diese Weise läßt sich die Interpretation der durch Tests erfaßten Schülerleistungen auf eine (meßtheoretisch) gesicherte Grundlage stellen – ein Vorteil gegenüber anderen Formen der Schülerbeurteilung (z. B. Zensurierung), der sowohl dem interindividuellen Vergleich als auch intraindividuellen Untersuchungszielen (z. B. der Erfassung von Lernleistungsfortschritten oder sog. Profilanalysen) zugute kommt (s. noch *Heller* 1973, S. 66 ff. u. 157 ff.).

Die *Kritik der klassischen Testtheorie* setzt an der Axiomatik an, d. h. an Widersprüchen zwischen einigen testtheoretischen Voraussetzungen und empirischen Befunden dazu. Die wichtigsten Voraussetzungen der klassischen Testtheorie lassen sich in folgenden drei Axiomen zusammenfassen. 1. *Existenzaxiom:* Zu jedem beobachteten (gemessenen) Wert existiert ein „wahrer" Wert im Sinne einer bestimmten individuellen Merkmalsausprägung (z. B. Höhe der Schulleistung). Diese wird als Konstante – wenigstens über einen gewissen Zeitraum hinweg – angenommen. 2. *Fehleraxiom:* Der Meßfehler einer Messung ist eine Zufallsvariable. Für diese gilt, daß die Summe bzw. das arithmetische Mittel der Fehlerwerte den Wert Null ergibt. 3. *Verknüpfungsaxiom:* Der beobachtete Wert (Meßwert) setzt sich additiv

aus wahrem Wert und Fehlerwert zusammen. Daraus kann eine Reihe von Sätzen abgeleitet werden, worauf wir nicht mehr eingehen wollen.

Die auf der Grundlage der klassischen Meßfehlertheorie konstruierten Schulleistungstests lassen vor allem folgende Phänomene ungeklärt: erstens die Beobachtung, daß – entgegen obiger Annahmen – die Fehlerwerte nicht unabhängig von den wahren Werten auftreten, d. h. „daß unterschiedliche wahre Werte auch unterschiedliche Fehlerwerte bedingen" können und diese nicht allein von der (mangelnden) Reliabilität des Tests, sondern auch von der jeweiligen Untersuchungspopulation beeinflußt werden; zweitens die (prinzipielle) Variabilität der wahren Werte, also die Tatsache, daß Merkmalsschwankungen (z. B. Intelligenz- oder Schulleistungszuwachs durch gezielte Fördermaßnahmen) auftreten können; drittens die damit verbundene Schwierigkeit, meßtheoretischen Anforderungen (vor allem der Reliabilität und Validität) *und* berechtigten schulpädagogischen (Lernleistungsförderung) gleichermaßen zu genügen. „Die klassische Testtheorie versucht, Aussagen über schon vorgegebene Meßwerte zu machen. Sie fragt nicht danach, *wie* diese Meßwerte zustande kamen" *(Langfeldt* 1974, S. 129). Inwieweit diese Probleme durch die „modernen" testtheoretischen Ansätze einer Lösung zugeführt werden können, soll im folgenden behandelt werden.

3.1.2. Neuere Modellansätze und ihre Problematik

Nach einigen terminologischen Vorklärungen befaßt sich *Büscher* (1974) mit der durch das *Rasch*-Modell eingeleiteten neuen Entwicklung einer *psychologischen* Meßtheorie, die vor allem am Problem der Populationsabhängigkeit klassischer Leistungsmessung ansetzt. „Für gewöhnlich werden die Eigenschaften eines psychologischen Tests durch individuelle Unterschiede innerhalb einer bestimmten Population definiert, und die Beurteilung jeder einzelnen Person ist mit der Referenzpopulation verknüpft. In vielen Fällen ist man jedoch daran interessiert, Individuen *per se* zu vergleichen, ohne sich dabei auf eine Population beziehen zu müssen. Die Frage etwa, ob sich ein Kind während eines oder mehrerer Jahre verbessert hat, kann unmöglich mit Hilfe verschiedener Tests beantwortet werden, die an Populationen unterschiedlich alter Kinder geeicht wurden. Im Rahmen eines neuen testtheoretischen Ansatzes versuchte *Rasch* (1960) derartige Probleme zu lösen, um das Studium einzelner Personen oder einzelner Testitems (Testaufgaben) unabhängig von Referenzpopulationen zu ermöglichen" *(Stene* 1968, S. 229).

Am Beispiel *kriterienbezogener* Leistungstests behandelt nun *Büscher* alle einschlägigen Probleme moderner testtheoretischer Provenienz. Ausführlich werden hierbei neue Verfahrensansätze zur Test- bzw. Itemanalyse erörtert, so die Itemanalyse von *Cox & Vargas, Cox & Graham, Popham, Fricke* u. a. oder die verschiedenen Methoden zur Reliabilitäts- und Validitätsbestimmung, z. B. nach *Carver, Livingston, Guttman, Cronbach, Jackson, Fricke* u. a. Deren praktische Relevanz im Hinblick auf die Schülerbeurteilung ist jedoch – vorerst noch – ungeklärt. Symptomatisch hierfür ist die abschließende Bewertung des *Rasch*-Modells durch *Büscher*. „Die praktische Anwendung des *Rasch*-Modells auf lehrzielorientierte Tests ist problematisch, da zum einen eine große Anzahl von Probanden erforderlich ist, zum andern der Konstruktions- und Rechenaufwand enorm ist und schließlich

das Modell nicht mehr angewandt werden kann, wenn alle Probanden das Lehrziel erreicht (bzw. nicht erreicht) haben. Das bedeutet, wenn überhaupt, dann ist das *Rasch*-Modell nur für normbezogene Messung sinnvoll" (S. 150 f.).

Damit, so könnte es scheinen, sind wir wieder an den Anfang unserer Problemdiskussion verwiesen. Dem ist nicht so. Vielmehr stehen wir mitten in einer Umorientierungsphase, in der neue Denkmodelle entwickelt (vgl. *Fischer* 1968, *Fricke* 1972) und die klassische Testtheorie relativiert (nicht verworfen) werden. Soviel steht jetzt schon fest, daß kriterienbezogene Messungen genauso wie normbezogene „ihre eigene bedeutsame pädagogische Funktion" haben. „Keineswegs wird das eine Meßverfahren das andere verdrängen" (a.a.O., S. 155).

3.1.3. Vorschläge zur Klassifikation von Schultests

So unterschiedlich wie die testtheoretischen Ansätze ist die in der Literatur verwendete Terminologie im Bereich der Schulleistungsmessung. Nach einer kritischen Analyse in sich oft widersprüchlicher Konzepte unterbreitete *Rosemann* (1974 b) eigene Vorschläge zu einer *pädagogisch* begründeten Klassifikation von Schultests. Während sich die konventionellen Bezeichnungen vorwiegend an meßtheoretischen Kriterien orientieren, begründet *Rosemann* sein Klassifikationskonzept auf den *Funktionen* schulischer Leistungsbeurteilung. Dabei ist die Unterscheidung von „Leistungsfeststellung" und „Leistungsbewertung" von Bedeutung. „Im ersteren Falle verschafft man sich lediglich Information darüber, was die Schüler im Verlaufe des Unterrichtsgeschehens gelernt bzw. nicht gelernt haben. Diese Informationen per se kann der Lehrer in vielfältiger Weise verwenden. Im zweiten Falle geht man einen Schritt weiter, man will die festgestellten Leistungen der Schüler bewerten, wobei verschiedene Bezugspunkte für die Bewertung in Betracht kommen können" (S. 163). Für Testverfahren, die im Dienste der erstgenannten Funktionseinheit stehen, schlägt der Autor die Sammelbezeichnung „Lernsteuerungstests" vor, für Verfahren der zweiten Kategorie die Bezeichnung „Lernkontrolltests". Ohne Zweifel werden mit dieser Einteilung zentrale Aufgabenfunktionen der Leistungsdiagnostik in der Schule getroffen, nämlich die „Lenkung und Steuerung des Lernprozesses" und die „Bewertung der Ergebnisse dieses Vorganges". Während die erste Zielsetzung dem „permanenten Informationsaustausch zwischen Lernenden und Lehrenden" dient und somit eine optimale Instruktion ermöglichen soll, liegt der entscheidende Vorteil der Lernkontrolltests in der „Objektivierung des Bewertungsvorganges", also der im Vergleich zur nicht-testgebundenen subjektiven Leistungsbeurteilung – Benotung nach schriftlicher (z. B. Klassenarbeiten) oder mündlicher Prüfung, Aufsatzbeurteilung usw. – größeren Zuverlässigkeit der Urteilsfindung.

Anhand des *Rosemann*schen Ordnungskonzeptes lassen sich nunmehr die verschiedenen Testformen folgendermaßen zusammenfassen: in *Lernsteuerungstets* (nach bisheriger Benennung die [informellen] kriteriumsbezogenen Schulleistungstests ohne Benotung und – gelegentlich auch – normbezogene Tests für kleinere Lerneinheiten) und *Lernkontrolltests* (die sog. standardisierten Schulleistungstests, informelle normbezogene Tests sowie [informelle] kriteriumsbezogene Tests mit Benotung). Der Autor selbst betont den Charakter seines Ordnungskonzeptes als

Diskussionsgrundlage. Die Vorschläge *Rosemanns* scheinen mir einer gründlichen Überlegung wert, vor allem im Hinblick auf die praktischen Bedürfnisse der Testanwendung in der Schule.

3.2. Objektive Verfahren schulischer Leistungsbeurteilung

3.2.1. Lernzieldefinition als Voraussetzung der Leistungsmessung

Anders als beim Programmierten Unterricht (PU), wo man sich von Anfang an vor die Notwendigkeit gestellt sah, *operationalisierte* Lehr-/Lernziele zu formulieren, sind die Probleme einer exakten Unterrichtsplanung im Sinne präziser Zieldefinitionen in der traditionellen Unterrichtslehre verhältnismäßig spät Gegenstand der Diskussion geworden. Virulent wurden solche Problemfragen eigentlich erst in dem Moment, wo das Bemühen um Überprüfung der Unterrichtsziele – analog zum PU – einsetzte, sei es im Rahmen curricularer Innovationen, bei der Erprobung neuer Schulmodelle (Gesamtschule) und damit zusammenhängenden Fragen der Unterrichtsdifferenzierung oder auch, um didaktische und schulpädagogische bzw. therapeutische Fördermaßnahmen auf leistungsdiagnostischer Grundlage zu sichern. Dazu bedarf es in jedem Falle operationalisierter Lernziele, d. h.: Lernleistungsprüfungen sind ohne (vorausgehende) Lernzielbestimmung weder pädagogisch sinnvoll noch meßtechnisch (via Kriteriumstests) durchführbar.

Operationalisierung von Lernzielen meint hier (nach *Horn* 1974 b) die genaue Festlegung der vom Schüler am Ende einer Unterrichtseinheit geforderten Verhaltensweisen (Operationen). „Diese Verhaltensweisen müssen direkt beobachtbar sein. Daher findet man bei Lernzielen häufig Formulierungen wie ‚Der Schüler soll ... nennen (aufschreiben, lösen usw.) können'." (a.a.O., S. 171).

Bei der Festlegung von Lernzielen orientiert man sich heute vielfach an der *Bloom*schen Taxonomie, die 6 verschiedene, hierarchisch geordnete Komplexitätsstufen enthält: Wissen, Verstehen, Anwendung, Analyse, Synthese, Evaluation. Die Kategorien 2 bis 6 betreffen dabei mehr „intellektuelle Fähigkeiten" und repräsentieren Formen zunehmender Komplexität des kognitiven Verhaltens, die auch als „Strategien des Problemlösens" *(Problemlösungsstrategien* = relativ inhaltsunabhängige „Grundmuster") aufgefaßt werden können. Für die praktische Anwendung des *Bloom*schen Klassifikationsmodells bei der Unterrichtsplanung gibt *Horn* (1972, 1974 b) wertvolle Hinweise und erläutert diese an einem Beispiel aus dem Naturkundeunterricht, wobei die Darstellung seines eigenen Lehrplan-Analyseschemas besondere Beachtung verdient. Ausführlich wird dann auf die Konstruktion von Prüfungsaufgaben eingegangen. Auch hierzu bietet der Autor eine Reihe praktischer Beispiele. Insgesamt wird somit deutlich, daß die Messung schulischer Lernleistungen vor allen Methodenfragen zunächst von exakten Lernzieldefinitionen abhängt, wobei der Operationalisierung der Lernziele vorrangige Bedeutung zukommt.

3.2.2. Informelle Tests (Lernkontrolltests)

Hauptanliegen des zweiten Beitrags von *Rosemann* (1974 a) ist die Information über Lernkontrolltests, aufgezeigt am Beispiel sog. informeller Leistungs-

messung. Dabei soll der Leser mit den Konstruktionstechniken und einigen Problemen informeller Tests soweit vertraut gemacht werden, daß er in der Lage ist, „den einen oder anderen Test selbst zu entwickeln bzw. bestehende Tests kritisch zu beurteilen" (S. 183).

Ein den Ausführungen vorangestelltes Ablaufdiagramm vermittelt einen schnellen Überblick über die einzelnen Arbeitsschritte. Nach der *Operationalisierung der Lernziele* (siehe oben) und der Erstellung einer *Spezifikationstabelle* (in der nicht nur die Anzahl der Testaufgaben oder Items festgelegt, sondern auch eine Entscheidung darüber getroffen wird, welche Lernziele überprüft und somit in den Test aufgenommen werden sollen) erfolgt die eigentliche *Aufgabenkonstruktion.* Hierzu werden vom Autor alle relevanten Item- und Antworttypen besprochen und an Hand einschlägiger Beispiele das Vorgehen veranschaulicht. Ausführlich erörtert *Rosemann* dann die einzelnen Schritte zur Entwicklung der sog. *Testvorform* (Aufgabengruppierung, Erstellung des Testaufgabenheftes einschließlich der Instruktion, Festlegung des Bewertungsschlüssels usw.) sowie die Vorbereitung und Durchführung der *Itemanalyse* (Schwierigkeits- und Trennschärfenanalyse jeder einzelnen Testaufgabe sowie Distraktorenanalyse, Aufgabenselektion bzw. -revision). Schließlich werden die Methoden zur Überprüfung der *Reliabilität* und *Validität* sowie Verfahren zur *Normierung* des Tests dargestellt, wobei eine Reihe von Illustrationsbeispielen wiederum die praktische Arbeit der Testkonstruktion und zugleich das Verständnis testtheoretischer Grundlagen – nunmehr am konkreten Objekt – wesentlich erleichtern dürfte. Der Weg von der Lernzieldefinition bis hin zur *Testendform* ist somit lückenlos beschrieben.

Die Lektüre dieses Beitrags dürfte jeden, der sich mit Fragen praktischer Schulleistungsmessung zu beschäftigen hat, verhältnismäßig rasch und umfassend über entsprechende Möglichkeiten (informeller Tests) informieren. Selbst derjenige Leser, der noch über keine oder nur wenige Kenntnisse und Erfahrungen auf diesem Gebiet verfügt, wird sich hierbei ohne größere Mühe zurechtfinden.

3.2.3. Standardisierte Schulleistungstests

Im letzten Beitrag zum Kapitel 4 (Objektive Verfahren) des hier referierten Sammelbandes gibt *Horn* (1974 a) einen Überblick über die wichtigsten zur Zeit verfügbaren sog. standardisierten Schulleistungstests und deren Anwendungsmöglichkeiten im Rahmen der Schülerbeurteilung. Die Unterschiede zwischen *informellen* und *standardisierten* (formellen) Schulleistungstests sind freilich nur gradueller Natur, wie die Ausführungen *Rosemanns* (1974 b, S. 158 ff.) deutlich gemacht haben. Demnach unterscheiden sich beide Verfahrensansätze vor allem bezüglich der vergleichsweise niedrigeren vs. höheren Reliabilität, des engeren vs. weiteren Anwendungsbereiches, der spezielleren vs. allgemeineren Testinhalte, der Normenfunktion, d. h. verfügbarer Testnormen für bestimmte Klassenstufen (derselben Schulart) vs. für verschiedene Klassen und Schularten sowie hinsichtlich einiger Konstruktionsaspekte, z. B. vorwiegend durch Lehrer vs. Testexperten erstellt. Die Bezeichnung „standardisierter" Schulleistungstest bezieht sich also keineswegs, wie oft angenommen wird, ausschließlich auf das Normenkriterium.

Unter *inhaltlichem* Gesichtspunkt lassen sich die standardisierten Schulleistungs-

tests folgendermaßen gruppieren: 1. *fächerübergreifende* Tests oder sog. Omnibusverfahren (z. B. Allgemeiner Schulleistungstest für 2. Klassen AST 2 bzw. für 3. Klassen AST 3 usw.); 2. *fachspezifische* Tests (z. B. Erdkundetest Deutschland ETD 5–7, Geschichtstest Neuzeit GTN 8–10, Naturlehretest NLT 9 usw. oder Diagnostischer Englisch-Leistungstest ELT 6–7, Französischer Wortschatztest FWS 9–12 usw.); 3. *lernbereichsorientierte* Tests, etwa Lesetests (z. B. Lesetest LT 2), Rechtschreibtests (z. B. Diagnostischer Rechtschreibtest DRT 3), Rechentests (z. B. Bruchrechentest BRT 6). Im Hinblick auf den *Anwendungsbereich* bzw. die Zielgruppe oder Untersuchungs*population* könnte man schließlich Tests für verschiedene Schulstufen und Schularten unterscheiden: 1. Tests für den *Primar-* vs. *Sekundarstufen*bereich bzw. die einzelnen Schul*klassen;* 2. Tests für die Schul*typen* des Gymnasiums, der Realschule und Hauptschule vs. Grundschule – seltener Gesamtschulen, die fast ausschließlich informelle Tests zur Lernleistungskontrolle verwenden (s. *Gaude & Teschner* 1970).

Neben zahlreichen Test- und Aufgabenbeispielen sowie praktischen Hinweisen für den schulischen Einsatz standardisierter Leistungstests bringt *Horn* am Ende seines Beitrags (1974 a, S. 228 f.) eine Gesamtübersicht der wichtigsten in der BRD lieferbaren Schultests (Stand 1973). Darüber hinaus kann sich jeder Leser an Hand der angeführten Verlagsanschriften durch Anforderung von Prospektmaterial bzw. Gesamtverzeichnissen jederzeit über die neuesten Testangebote informieren.

3.3. Subjektive Verfahren schulischer Leistungsbeurteilung

Zu den „subjektiven" Verfahren rechnen wir alle nicht-messenden Methoden der Schülerbeurteilung, d. h. Verfahrensweisen, die im Sinne der klassischen Meßfehlertheorie (vgl. Meß- bzw. Testgütekriterien) – erfahrungsgemäß – nicht objektiv sind und einen relativ geringen Grad an Zuverlässigkeit und Gültigkeit aufweisen. Im einzelnen fallen hierunter alle Formen konventioneller Notengebung, mündliche Prüfung, Aufsatzbeurteilung u. ä. Aber auch die in der Schulpraxis bislang viel zu wenig beachteten Methoden der (wissenschaftlichen) Verhaltensbeobachtung und einer Reihe von Beurteilungstechniken i. e. S. (vgl. *Heller* et al. 1974, S. 28 ff., bes. S. 37 ff.) sind der Kategorie der subjektiven Verfahren zuzuordnen. Damit befaßt sich nun der Beitrag von *Langhorst* (1974).

3.3.1. Verhaltensbeobachtung und Schülerbeurteilung

Über die Notwendigkeit der Verhaltensbeobachtung bzw. einzelner Techniken der Beurteilung i. e. S. im Unterricht sollte es eigentlich keine Diskussion mehr geben. Ohne den Einsatz dieser Methode(n) müßten wir oft auf wichtige Informationen verzichten, z. B. über das Schüler- und Lehrerverhalten im Interaktionsspiel des Unterrichtsablaufs, damit zugleich auf nicht via Tests erfaßbare Daten in bezug auf die Unterrichtsplanung, die Curriculumentwicklung, Lernzielbestimmung usw. *Stake* formulierte dies in einem anderen Zusammenhang sehr deutlich, wenn er sagt (zit. nach *Rosenshine* 1973, S. 201): „Ohne Informationen über die Lehrmethoden ist es weder möglich, das Wesen eines Curriculums zu verstehen, noch zu wissen, was als nächstes ausprobiert werden muß. In manchen Evaluationsuntersuchungen

werden die mit Hilfe der Unterrichtsbeobachtung gewonnenen Daten die wichtigsten und wertvollsten sein."

Nach einem kurzen Überblick über gängige Verfahren der Schülerbeobachtung und -beurteilung im Unterricht diskutiert *Langhorst* eine Reihe von Möglichkeiten, diese Verfahrensansätze zu systematisieren und somit auf eine objektivere Basis zu stellen. Dabei gilt es zunächst, die *Hauptphasen der Verhaltensbeobachtung*, nämlich die *Beobachtung i. e. S.* (auf den Beobachtungsgegenstand zentrierte, Unwesentliches selegierende Wahrnehmung), die *Beschreibung* oder *Deskription* (Protokollierung des beobachteten Verhaltens) und die eigentliche *Beurteilung* (Interpretation der Beobachtungsdaten), strikt auseinanderzuhalten. Nur so kann die Zahl der subjektiven (unkontrollierbaren) Einflüsse auf ein Minimum reduziert und die Aussagekraft der Ergebnisse erhöht werden.

Im zweiten Teil seines Beitrags geht der Autor ausführlicher auf die Möglichkeiten der Schülerbeurteilung mit Hilfe von *Schätzskalen* (rating scales) ein. Unterrichtsrelevante Formen dieses Vorgehens werden an Hand von praktischen Beispielen demonstriert und kritisch auf ihre Verwendungsmöglichkeiten hin untersucht. In diesem Zusammenhang nimmt die Erörterung der häufigsten *Beurteilungsfehler* (Hofeffekt, logischer Fehler, Mildefehler, Kontrastfehler, Ähnlichkeitsfehler, „Pygmalion"-Effekt u. ä.) einen wichtigen Platz ein, entscheidet doch ihre Kenntnis letzten Endes darüber, „ob der Lehrer die Vielzahl der subjektiven Urteilstendenzen ... beim Umgang mit seinen Schülern in etwa kontrollieren und steuern kann ... Entsprechende Anstrengungen sind schon deshalb lohnenswert, weil – erwiesenermaßen – im Fall einer gerechten Verhaltensbeurteilung sich der Schüler vom Lehrer besser verstanden fühlt" (S. 250).

3.3.2. Leistungsbeurteilung durch Notengebung

„Schulnoten sind Maßzahlen für erbrachte Leistungen der Schüler, die der Lehrer nach seinen Erfahrungen und Einschätzungen auf der Notenskala einstuft. Noten kommen also aufgrund eines Urteilsprozesses des Lehrers zustande und sind mit all den Mängeln behaftet, die man bei Urteilsprozessen nachgewiesen hat" *(Fingerhut & Langfeldt* 1974, S. 253). Dabei sind situative Einflüsse genauso beteiligt wie sozial- und persönlichkeitspsychologische Momente (vgl. Beurteilungsfehler). Daraus können wir folgern, daß Schulzensuren prinzipiell schlechte Indikatoren für Schulleistungen sind. Siehe noch *Ingenkamp* (1971).

Diese Vermutung läßt sich durch zahlreiche empirische Belege stützen. So kommen *Fingerhut* und *Langfeldt* in dem zitierten Beitrag, der die wichtigsten (neueren) Untersuchungen zu diesem Thema berücksichtigt, zu dem eindeutigen Ergebnis, daß die Lehrerzensuren herkömmlicher Art den meßtheoretischen Forderungen der Objektivität, Reliabilität und Validität nicht oder höchst unvollkommen genügen. „Dies kann den Lehrern jedoch nicht zum Vorwurf gemacht werden. In jeder Unterrichtsstunde wird von ihnen eine Vielzahl schneller und unabhängiger Entscheidungen und Beurteilungen verlangt, die sie nur bewältigen können, wenn sie bestimmte Urteils- und Verhaltensstrategien entwickeln. Diese notwendige Bildung von Stereotypen verhindert aber exakte Urteile" (a.a.O., S. 254). Daraus kann m. E. nur die Folgerung abgeleitet werden, dem Lehrer die nötigen Hilfen in Form

objektiver und zuverlässiger Methoden (z. B. formelle und informelle Tests) an die Hand zu geben, um auf diese Weise die subjektiven Urteile abzusichern. Nur unter dieser Voraussetzung wäre es – wenn überhaupt – sinnvoll, Schulnoten die fast alles im Bildungsgang entscheidende Funktion zuzuerkennen.

Die Absicherung subjektiver Lehrerurteile sollte also *mittelbar* durch Tests u. ä. erfolgen und nicht durch unkritische Angleichung an testtheoretische Modelle versucht werden, da ein solches Unterfangen mancherlei pädagogischen und erzieherischen Absichten widersprechen würde. So führt beispielsweise *Zielinski* (1973, S. 11) nicht weniger als 10 verschiedene Funktionen der Zensierung auf. Sowohl die Intention der Schule, Lernleistungen zu fördern und damit Schülerleistungen (positiv) zu verändern, als auch die Verquickung von pädagogischer Anreizfunktion und reiner Meßfunktion der Schulnoten lassen alle Bemühungen, konsistente (zuverlässige) Leistungsmessungen via Zensuren zu erzielen, praktisch im Ansatz scheitern. Entweder müßte man auf berechtigte schulpädagogische Anliegen hier verzichten, um – nicht weniger berechtigten – meßtheoretischen Anforderungen zu genügen, oder man sollte inkompatible Forderungen fallenlassen und schiedlich zwischen pädagogischen und Meßfunktionen bei der Notengebung trennen (was theoretisch möglich ist) bzw. ganz auf Noten verzichten (was vielen praktisch unmöglich erscheint). Zweifellos wäre die letzte Alternative die konsequenteste Haltung. Da wir aber ihre Realisierungschance momentan als gering erachten, bleibt wohl als *vorläufiger* Weg aus dem Dilemma nur die erste Alternative übrig, deren Gefahrenmomente (Vermischung der Funktionen, Beurteilungsfehler usw.) durch Testhilfen soweit als möglich kompensiert werden müssen. Hierzu werden die formellen und die informellen Schulleistungstests einschließlich sog. Standardarbeiten (vgl. *Wendeler* 1969) gleichermaßen beitragen können.

Prinzipiell wäre damit auch der Weg gewiesen, *einheitlichere Beurteilungsmaßstäbe* zu erzielen. Nach den Ergebnissen von *Fingerhut* und *Langfeldt* sowie zahlreicher anderer – in dem genannten Buch zitierter – Forscher werden Schulnoten fast durchweg fachspezifisch erteilt, wobei sich drei Gruppen herausbilden: „Fächer mit milder (musische Fächer und Religion), mit mittlerer (Nebenfächer) und mit strenger Beurteilung (Hauptfächer)" (a.a.O., S. 255 ff.). Die Möglichkeiten der Differenzierung sind außerdem bei milder Beurteilung stark reduziert. Hinzu kommen inter- versus intraindividuelle, d. h. lehrerabhängige, Einfluß- bzw. Fehlervariablen (siehe oben). Damit wäre erneut die Notwendigkeit *testunterstützter Notengebung* demonstriert.

Die Möglichkeiten zur objektiven Leistungsbeurteilung sind freilich trotz des wachsenden Testangebots begrenzt, und diese Grenzen können mitunter in der Spezifität des Gegenstandes selbst liegen. Als Paradebeispiel hierfür sei die Aufsatzbeurteilung genannt. Andererseits wird m. E zu Recht darauf hingewiesen, daß man – aus den verschiedensten Gründen – auf die Übung des Aufsatzes (und seine Bewertung) nicht verzichten sollte. Der Bedeutung des Gegenstandes angemessen beschäftigen sich die letzten zwei Buchbeiträge von *Nickel & Wieczerkowski* (1974) und *Wendeler* (1974) ausführlicher mit der Zensierung von Schüleraufsätzen.

3.3.3. Probleme der Aufsatzbeurteilung

In drei aufeinander aufbauenden Untersuchungsserien versuchten *Nickel* und *Wieczerkowski*, „in einer quasi-experimentellen Situation den Einfluß verschiedener vermutlich bedeutsamer Variablen auf die Bewertung von Aufsätzen zu erforschen" (a.a.O., S. 272). Die methodisch sehr sorgfältig angelegte Studie, in der eine Reihe unabhängiger Beurteilergruppen (Schüler, Studenten der Erziehungswissenschaften und Psychologie sowie Referendare und praxiserfahrene Lehrer) Schüleraufsätze von Viertklaßkindern unter weitgehend standardisierten Bedingungen beurteilte, erbrachte – bei systematischer, isolierender Variation einzelner Variablen – folgende Hauptergebnisse (siehe die vollständige Darstellung in Kap. 5.3 des zitierten Werkes).

Keinen Einfluß auf die Aufsatzzensur hat demnach das Geschlecht, und zwar weder des Schülers bzw. Aufsatzschreibers noch des Lehrers bzw. Aufsatzbeurteilers. Ebensowenig scheinen vorausgehende oder nachfolgende Arbeiten gleicher oder ähnlicher Art die Ausbildung individueller Bezugssysteme auf seiten des Beurteilers zu beeinflussen.

Demgegenüber konnten die Autoren folgende Einflußgrößen ermitteln: *Vor-* bzw. *Zusatzinformationen* über den Schüler, insbesondere über dessen sonstige Schulleistungen; *Praxiserfahrung* der Beurteiler, sowohl allgemein in bezug auf das Unterrichten als auch speziell im Hinblick auf Aufsatzbewertungen; verschiedene *Sprachkriterien* des Aufsatzschreibers (Aufsatzlänge, Sprachrichtigkeit vs. Fehlerhaftigkeit, Differenziertheit des Ausdrucksstils, Einfallsreichtum und Originalität, Handlungsganzheit u. ä.). „Eine Analyse des Einflusses verschiedener Sprachkriterien bei annähernder Konstanthaltung des Faktors ‚Länge' und ‚Sprachrichtigkeit' ergab, daß die drei Variablen ‚Originalität der Einfälle', ‚Differenziertheit des sprachlichen Ausdrucks' und ‚Flüssigkeit bzw. Abgeschlossenheit des Handlungsablaufs' allein für die Gesamtzensur bereits einen Schätzungseffekt von 52% besitzen. Wenn also die Richtigkeit und die Länge der Darstellung als Grundlage für eine Beurteilungsdifferenzierung entfallen, läßt sich die Aufsatzzensur in beträchtlichem Umfang bereits allein aufgrund dieser drei Sprachkriterien voraussagen" *(Nickel & Wieczerkowski* 1974, S. 305).

Diese Befunde gewähren einen guten Einblick in den Bedingungskomplex, unter dem subjektive Urteile über Schülerleistungen, aufgezeigt am Beispiel der Aufsatzbeurteilung, zustandekommen. Es ist denkbar, daß diesen Ergebnissen ein größerer Allgemeinheitsgrad zukommt, zumindest was die Beurteilung *sprachlicher* Leistungen betrifft. Darüber hinaus liefert die Arbeit wichtige Kriterien für eine einheitlichere Beurteilungspraxis – ein Thema, dem der nachfolgend referierte Beitrag gewidmet ist.

Das Ausmaß mangelnder Übereinstimmung in der Benotung von Schüleraufsätzen ist tatsächlich erschreckend, wenn man die von *Wendeler* (1974) zitierten zahlreichen Literaturquellen als repräsentativ für die heutige Praxissituation ansehen darf (genauer: muß). Nach Meinung des Autors ist das Problem der Aufsatzbeurteilung „keineswegs ein Spezialproblem des Sprachunterrichts, sondern das Problem dieser Prüfungsform als solcher". Dabei kann man sich den Gesamtvorgang zweiphasig vorstellen. „Der erste Schritt ist die *Leistungsermittlung*, bei der man die

Gesamtleistung oder bestimmte Teilleistungen nach einem Bewertungsschlüssel als ‚richtig', ‚falsch', ‚gut' oder ‚schlecht' beurteilt, meist mit Punktwerten versieht, und bei der man daraufhin einen Gesamtpunktwert bestimmt.

Der zweite Schritt ist die *Zensierung*, d. h. die Zuordnung einer Zensur zu der errechneten Punktzahl" (a.a.O., S. 311 ff.). Die Crux der Aufsatzbewertung liegt in der ersten Phase, nämlich in den *Kriterien* der Leistungserfassung und deren *objektiven* Bestimmung. Abweichende Urteile über ein und denselben Aufsatz sind demnach vor allem Folgen unterschiedlicher Kriterienmaßstäbe.

Im Bemühen, einheitliche (verbindliche) Kriterien bzw. Kriteriensysteme für die Aufsatzbeurteilung zu entwickeln, waren besonders einige angelsächsische Versuche erfolgreich. So erstellten *Diederich, French* und *Carlton* auf induktivem Wege ein Kategoriensystem für die Aufsatzbewertung, das folgende 5 (faktorenanalysierte) *Urteilsdimensionen* enthält: Ideen, innere Form (Gliederung), Lebendigkeit (Originalität), Sprachrichtigkeit und Wortwahl (Flüssigkeit). Darüber sowie über eine Reihe weiterer Ansätze berichtet *Wendeler* ausführlich in seinem Beitrag. Die Übereinstimmung zahlreicher dieser Befunde mit den im vorigen Abschnitt referierten ist unverkennbar und erhärtet ihre praktische Relevanz.

Im zweiten Teil seines Beitrags geht dann der Autor auf Fragen und Probleme der Objektivitätssteigerung im Kontext Aufsatzbeurteilung ein. Hierzu werden wiederum mehrere Ansätze zur Erfassung der *Urteilsobjektivität* referiert sowie zahlreiche praktische Maßnahmen zu ihrer Verbesserung vorgeschlagen. Die mit entsprechenden Beispielen versehenen Ausführungen sollten deutlich machen, „daß tatsächlich ein befriedigender Objektivitätsgrad erreichbar ist" (a.a.O., S. 321).

4. Ausblick

Die Schulleistungsdiagnostik ist Bestandteil einer umfassenderen *pädagogischen* Diagnostik und steht als solche im Dienste unterrichtlicher und erzieherischer Ziele. Sie kann und will nie Selbstzweck sein. Andererseits stellt sie in mannigfacher Weise die notwendigen Informationen für unumgängliche pädagogische Entscheidungen zur Verfügung, die ohne diese Hilfen nicht oder nur unzulänglich zu fällen wären. Diese Bewertung diagnostischer Funktionen und Möglichkeiten bedeutet keine Technokratisierung des Unterrichts, wie vielfach behauptet wird, sondern auf rationaler Basis begründete Lehr-/Lernprozesse, wofür wir uns allerdings aussprechen, nicht zuletzt unter dem Gesichtspunkt pädagogischer und bildungspolitischer Forderungen wie der der Verwirklichung von Chancengleichheit im Bildungsgang.

Demgegenüber gilt es, auf die Grenzen diagnostischer Möglichkeiten hinzuweisen. Diagnostische Hilfen in Form von Testergebnissen oder anderen Urteilen (z. B. Aussagen via Schätzskalen) können dem Lehrer die pädagogische Entscheidung selbst nie abnehmen. Diese Einschränkung gilt gleichermaßen im Hinblick auf die praktischen Funktionen und die theoretischen Voraussetzungen pädagogischer Leistungsbeurteilung, die sich ja keineswegs problemlos darstellt. Doch welcher Forschungs- oder Praxisbereich könnte dies heute schon für sich beanspruchen?

Daraus den Schluß zu ziehen, man könne auf die gebotenen – begrenzten – Möglichkeiten diagnostischer Hilfen in der Schule verzichten, wäre m. E. allerdings verhängnisvoll. Ich glaube nicht, daß die Praxis (hier: tagtäglich abgegebener Schüler-

beurteilungen) besser sein kann als der jeweilige Erkenntnisstand wissenschaftlicher Forschung hierzu. Beide sind vielmehr in einem gegenseitigen Bedingungsgefüge zu sehen; paradigmatisch verweise ich in diesem Zusammenhang auf die Diskussion um die sog. Selektions- versus Klassifikationsproblematik im Rahmen schulischer Differenzierungsmodelle (vgl. *Hopf* 1973 sowie bereits *Heller* 1970). Hier – wie so oft – sind Theorie und Praxis aufeinander angewiesen, will man sich nicht in praxisfernen Modelle oder kurzschlüssige Praktiken verlieren.

Ich möchte den kurzen Überblick nicht beschließen, ohne auf zwei Desiderata hingewiesen zu haben. Zum einen wünschte man sich endlich eine umfassende *Theorie der pädagogischen Diagnostik*, die es bislang nur ansatzweise gibt (z. B. *Guthke* 1972). Zum anderen dürften die Ausführungen die Notwendigkeit vor Augen geführt haben, *persönlichkeits-* und *sozialpsychologische* Fragestellungen in einer solchen Theorie – stärker als in den vorliegenden Ansätzen – zu berücksichtigen (vgl. *Ulich & Mertens* 1973). Das Forschungsfeld und die praktische Bewährung schulischer Leistungsbeurteilung harren weiterer Initiativen.

Literaturverzeichnis

Büscher, P.: Einige testtheoretische Aspekte kriterienbezogener Leistungsmessung. In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974.

Fingerhut, W., u. *Langfeldt, H.-P.:* Schülermerkmale, Lehrermerkmale und ihre Beziehungen zu Schulnoten. Unveröffentl. Diplomarbeit, Univ. Marburg, 1971.

— — Leistungsbeurteilung durch Notengebung. In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974.

Fischer, G. H. (Hrsg.): Psychologische Testtheorie. Huber, Bern 1968.

Fricke, R.: Über Meßmodelle in der Schulleistungsdiagnostik. Schwann, Düsseldorf 1972.

Furck, C. L.: Das pädagogische Problem der Leistung in der Schule. Beltz, Weinheim 1961, 1964 (2. Aufl.).

Gaedike, A.-K.: Determinanten der Schulleistung. In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974.

Gaude, P., u. *Teschner, W. P.:* Objektivierte Leistungsmessung in der Schule. Diesterweg, Frankfurt M. usw. 1970.

Guthke, J.: Zur Diagnostik der intellektuellen Lernfähigkeit. VEB Deutscher Verlag der Wissenschaften, Berlin 1972.

Heller, K.: Aktivierung der Bildungsreserven. Huber und Klett, Bern und Stuttgart 1970.

— Intelligenzmessung. Neckarverlag, Villingen 1973.

— (Hrsg.): Leistungsbeurteilung in der Schule. Quelle & Meyer, Heidelberg 1974.

—, *Rosemann, B.*, u. *Gaedike, A. K.:* Planung und Auswertung empirischer Untersuchungen. Klett, Stuttgart 1974.

Hentig, H. von: Systemzwang und Selbstbestimmung. Klett, Stuttgart 1968, 1970 (3. Aufl.).

Hofer, M., u. *Weinert, F. E.* (Hrsg.): Pädagogische Psychologie. Grundlagentexte 2 zum Funk-Kolleg. Päd. Psych., Fischer Taschenbuch, Frankfurt/M. 1973.

Hopf, D.: Möglichkeiten und Grenzen der Anwendung von Tests. In: *Hofer, M.*, u. *Weinert, F. E.* (Hrsg.): Pädagogische Psychologie. Frankfurt/M. 1973.

Horn, R.: Lernziele und Schülerleistung. Beltz, Weinheim 1972.

— Einsatz standardisierter Schulleistungstests. In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974 a.

— Leistungsmessung und Lernzieldefinition. In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974 b.

Ingenkamp, K. (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Beltz, Weinheim usw. 1971, 1972 (3. Aufl.).

— (Hrsg.): Tests in der Schulpraxis. Beltz, Weinheim usw. 1971, 1973 (3. Aufl.).
Langfeldt, H.-P.: Die klassische Testtheorie als Grundlage standardisierter Schulleistungstests. In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974.
Langfeldt, H.-P., u. *Fingerhut, W.:* Empirische Ansätze zur Aufklärung des Konstruktes „Schulleistung". In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974.
Langhorst, E.: Beobachtung und Beurteilung des Schülerverhaltens im Unterricht. In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974.
Michel, L.: Allgemeine Grundlagen psychometrischer Tests. In: Hb. d. Psychol., Bd. 6: Psychol. Diagnostik (Hrsg. *Heiß, R.).* Hogrefe, Göttingen 1964.
Neander, J.: Objektivierte Lernerfolgsmessung in der Gesamtschule — Fortschritt für wen? Die Deutsche Schule, *65* (1973) 35—47.
Nickel, H., u. *Wieczerkowski, W.:* Einflüsse auf die Beurteilung von Schüleraufsätzen — Ergebnisse einer quasi-experimentellen Versuchsreihe. In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974.
Rosemann, B.: Konstruktion und Einsatz von Informellen Tests zur Leistungsbeurteilung (Lernkontrolltests). In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule, Heidelberg 1974 a.
— Zur Problematik der Klassifikation von Schultests. In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974 b.
Rosenshine, B.: Die Beobachtung des Unterrichts in der Klasse. In: *Hofer, M.,* u. *Weinert, F. E.* (Hrsg.): Pädagogische Psychologie. Frankfurt/M. 1973.
Stene, J.: Einführung in *Raschs* Theorie der psychologischen Messung. In: *Fischer, G.* (Hrsg.): Psychologische Testtheorie. Bern 1968.
Symposium des Europarates in Berlin vom 11.—19. 11. 1971. Unveröffentl. Bericht der Arbeitsgruppe 2 (Tutor: Dr. *Weiß, R.,* Protokollant: Dipl.-Psych. *Büscher, P.).*
Ulich, D., u. *Mertens, W.:* Urteile über Schüler. Beltz, Weinheim usw. 1973.
Wendeler, J.: Standardarbeiten — Verfahren zur Objektivierung der Notengebung. Beltz, Weinheim usw. 1969, 1973 (5. Aufl.).
— Bemühungen um Vereinheitlichung der Aufsatzbeurteilung. In: *Heller, K.* (Hrsg.): Leistungsbeurteilung in der Schule. Heidelberg 1974.
Zielinski, W.: Die Beurteilung von Schülerleistungen. In: Funkkolleg „Pädagogische Psychologie" (Studienbegleitbrief 12). Beltz, Weinheim usw. 1973.

Prof. Dr. Kurt Heller
Päd. Hochschule Rheinland
Abt. Bonn, Psychol. Seminar
53 Bonn
Römerstraße 164