VI. Leistungsbewertung in der Gesamtschule

1. Prinzipien der Leistungsbewertung

Es ist durch verschiedene Forschungen in Deutschland und im Ausland gut belegt, daß Lehrer zwar die *Rangfolge* der Schulleistungen *innerhalb* der Klasse (Gruppe) recht zuverlässig bestimmen können, ihre Zensuren jedoch als Vergleichsmaßstab zwischen *verschiedenen* Klassen (Gruppen) kaum brauchbar sind, Die Ziffernbewertung von Arbeiten unterliegt allgemein der Kritik, da die Beurteilungsmaßstäbe von Lehrern weder übereinstimmen noch über Zeiträume hinweg konstant sind.

In einer nach Leistung differenzierenden Gesamtschule mit Niveau- und Neigungsgruppen ist es unerläßlich, gruppenübergreifende und verläßliche Standards der Leistungsbeurteilung zu finden. Die Stimmigkeit der Differenzierung und die Durchlässigkeit zwischen den Gruppen hängen davon ab, daß bei den Zuweisungsentscheidungen nicht gruppenimmanente Bewertungsverfahren verwendet werden. Außerdem setzt die in der Gesamtschule notwendige, umfangreiche Schullaufbahnberatung (guidance) die Existenz objektiver, verläßlicher und gültiger diagnostischer Verfahren und die Kenntnis der Leistungsnormen der jeweiligen Bezugsgruppe voraus,

Im folgenden werden verschiedene Möglichkeiten gruppenübergreifend gültiger Beurteilungsmaßstäbe diskutiert. Als optimale Lösung wird die Entwicklung einer Itembank (s. S. 98) vorgeschlagen, mit deren Hilfe der Lehrer die Tests selbst zusammenstellen kann, Solche Tests gestatten den Vergleich des einzelnen Schülers mit der gesamten Bezugspopulation; sie können entweder die einzige Grundlage einer Entscheidung darstellen oder als „Reference Tests" der Korrektur anderer Leistungsurteile dienen. Auf die intensive Verwendung informeller Schulleistungstests kann nicht verzichtet werden.

Solange keine Itembank existiert, sollte die Leistungsbewertung auf den Ergebnissen informeller Tests basieren. Verfügbare standardisierte Tests können gelegentlich als Bezugstests (Reference Tests) herangezogen werden.

Die Beurteilung der nicht-kognitiven Dimensionen des Schülerverhaltens mit Hilfe von Schätzskalen (rating scales) sollte grundsätzlich getrennt von der Leistungsbewertung vorgenommen werden.

2. Die einzelnen Beurteilungsverfahren

Voraussetzung einer soliden Leistungsbeurteilung ist die Kongruenz der Lernziele (und -inhalte), die bei der Entwicklung des Bewertungsverfahrens Berücksichtigung gefunden haben, mit denen des Unterrichts, der der Beurteilung vorangegangen ist. Die Überlegenheit der meisten im folgenden beschriebenen Bewertungsverfahren beruht zu einem guten Teil darauf, daß sie auf einer sorgfältigen operationalen Definition von Lernzielen und -inhalten beruhen.

2.1 Standardisierte Schulleistungstests in der üblichen Form

Als Instrumente der Leistungsmessung sind standardisierte Tests das zuverlässigste Mittel. Sie setzen jedoch, um inhaltlich gültig und für alle Schüler gleich fair zu sein, ein homogenes Angebot an Unterrichtsgegenständen voraus. Bei inhomogenem Angebot könnte zwar bedingt gesichert werden, daß die Tests gleich unfair für alle sind, doch würde dadurch die inhaltliche Gültigkeit der Tests eingeschränkt. Da standardisierte Schulleistungstests in Deutschland kaum vorhanden sind und für die vielen Fächer, Altersstufen und Gruppierungen innerhalb der Fächer auf absehbare Zeit nicht in hin-

reichendem Maße vorliegen werden, entfällt diese Möglichkeit weitgehend. Sie würde zwar für das Schulsystem im ganzen außerordentlich nützlich sein; für die spezifischen Bedürfnisse der Differenzierung in Gesamtschulen, die zugleich noch ihr eigenes Lehrangebot (Curriculum) teils entwickeln, teils modifizieren müssen, sind Schulleistungstests, die in der Gesamtpopulation standardisiert wurden, allerdings nicht zureichend.

2.2 Standardisierte Schulleistungstests, ad hoc für Klassen (Gruppen) zusammengestellt: die Itembank

Mit Hilfe einer Itembank können für jede Gruppe Tests entwickelt werden. Eine Itembank enthält für alle wesentlichen Lernziele und -inhalte eine große Zahl von trennscharfen Aufgaben (items), deren inhaltliche Gültigkeit und Schwierigkeitsgrad für die relevanten Bezugspopulationen bekannt sind. Der Lehrer läßt sich für die von ihm gewünschten Stoffgebiete und Lernziele Tests von der Itembank ad hoc zusammenstellen; er ist dann in der Lage, die Leistungen seiner Schüler mit den Leistungen derjenigen Gruppe, an der die Normen gewonnen werden (Eichpopulation), zu vergleichen.

Die Resultate können dazu dienen, vorliegende Leistungsurteile, die beispielsweise mit informellen Tests gewonnen wurden, zu den Leistungen einer sehr viel größeren Bezugsgruppe in Beziehung zu setzen. Damit wäre die Voraussetzung für die Vergleichbarkeit von Leistungsurteilen auch über den Rahmen der einzelnen Schule hinaus geschaffen; außerdem können die Differenzierungs- und Übergangsentscheidungen innerhalb der Schule auf eine verläßliche Basis gestellt werden.

Bis eine arbeitsfähige Itembank bereit steht, ist eine langfristige und umfangreiche Entwicklungsarbeit erforderlich. Da jedoch die Konstruktion einer größeren Zahl der üblichen Tests kaum weniger Zeit in Anspruch nehmen dürfte, gebührt wegen ihrer Anpassungsfähigkeit an die spezifischen Bedürfnisse des Lehrers den Itembank-Tests der Vorzug. Das in Kapitel VIII („Die wissenschaftliche Kontrolle von Versuchen mit Gesamtschulen") genannte Institut für Schulforschung und Schulentwicklung wäre der geeignete Ort für die Entwicklung einer Itembank.

Im Hinblick auf Abschlußprüfungen stellt die Verwendung einer Itembank das Problem gemeinsamer bzw. äquivalenter Inhalte, über die Rahmenentscheidungen getroffen werden müssen. Insbesondere ergibt sich die Schwierigkeit adäquater Gewichtung intensiv behandelter Inhalte gegenüber einer größeren Zahl im Überblick dargebotener Stoffe. Zudem wird in der Gesamtschule eine Lehrplanrevision notwendig, die ebenfalls bei der Leistungsbewertung Berücksichtigung finden muß. Eine Itembank ist am ehesten geeignet, diese Probleme auf befriedigende Weise zu lösen und eine rigide Standardisierung des Lehrangebots überflüssig zu machen. Damit wird der Weg für die Erprobung gruppenspezifischer Lehrangebote und Lehrmethoden freigehalten und somit der Entwicklung der Gesamtschule ein größerer Spielraum gegeben.

2.3 Informelle Tests für die Klasse

Informelle Tests können das Lehrerurteil objektivieren und absichern. Notwendig ist, daß den Lehrern die leicht zu handhabenden Hilfen für die Herstellung solcher, ad hoc konstruierbarer und situationsangemessener Tests zur Verfügung stehen. Doch reichen informelle Tests nicht aus, um das Problem der gruppenübergreifenden Standards zur Sicherung der Mobilität zwischen Kursgruppen zu lösen, wenn sie nur an kleinen Gruppen geeicht werden. Wenn sie dagegen innerhalb einer Schule für eine ganze Altersstufe konstruiert worden sind, so können sie sehr wohl dazu beitragen, die Differenzierungsentscheidungen verläßlicher zu treffen als die herkömmlichen Zensuren. Allerdings können sie nicht die Präzision und Verläßlichkeit von Tests erreichen, die aus einer Itembank stam-

men, da den Lehrern nicht der bei der Konstruktion standardisierter Tests übliche Aufwand beim Abfassen der Aufgaben zugemutet werden kann.

Auch hier besteht das Problem der homogenen Unterrichtsinhalte zwischen Kursen, wenn die Tests gültig und fair sein sollen, besonders bei den Fächern, in denen bestimmte Lerninhalte Voraussetzung der weiteren Arbeit sind und nicht in wechselnder Reihenfolge angeboten werden können.

Informelle Tests müssen von den Lehrern selbst hergestellt werden. Dazu ist intensive Kooperation erforderlich (vgl. 3.). Außerdem ist eine Anleitung für die Herstellung solcher Tests (Operationalisierung der Lernziele, Aufgabenkonstruktion, Aufgabenanalyse) unerläßlich. Für die Anleitung kommen psychologisch vorgebildete Lehrer, Schulpsychologen etc. in Frage. Ausbildung und Koordinierung könnten Aufgabe eines Testforschungs- und Entwicklungsinstituts werden (vgl. Kapitel VIII).

2.4 Normarbeiten - in Verbindung mit gruppenspezifischen Notenskalen (Äquivalenzskalen)

Die Normarbeit setzt, um für mehr als eine Klasse (Gruppe) gültig zu sein, ebenfalls ein standardisiertes Lehrprogramm voraus, das die verschiedenen Gruppen absolvieren. In diesem Falle besitzt sie gewisse Eigenschaften von informellen Tests für die betroffenen Gruppen. Wegen der größeren Objektivität der Auswertung, die eine der Voraussetzungen für die Verläßlichkeit einer Bewertung darstellt, ist der informelle Test jedoch der Normarbeit überlegen. Außerdem besteht auch gegen die Normarbeit der Einwand, daß sie die Standardisierung des Lehrangebots voraussetzt.

Gegen äquivalente, jedoch gruppenspezifische Zensurenskalen kann der Einwand erhoben werden, daß die Zahlenwerte auf den Skalen zwar äquivalent und konstant bleiben, jedoch die Distanz zwischen diesen Werten in den verschiedenen Gruppen wahrscheinlich nicht gleichbleibt. Solche Skalen können Ungleichheiten in der Beurteilung eher verdecken. Ein psychologischer Nachteil ist ferner, daß die Schüler sehr bald wissen werden, daß ihre „guten" Noten „eigentlich" nur mäßige oder gar schlechte Zensuren sind und wegen der Beurteilung in den schwächeren Kursen kaum je „befriedigende" Zensuren erreicht werden können.

2.5 Nicht-kognitive Dimensionen des Schülerverhaltens

Um den Stellenwert eines Leistungsurteils adäquat interpretieren zu können, ist es notwendig, andere, Leistungsverhalten und Erfolg des Schülers beeinflussende Merkmale wie Arbeitshaltung, individueller Fortschritt, Interesse usw. zu berücksichtigen. Wenn ein Leistungsurteil solche Komponenten unausgesprochen enthält, verliert es viel von seinem Informationswert. Deshalb müssen die genannten Dimensionen gesondert aufgeführt werden.

Es genügt nicht, die Merkmale global zu bewerten, da eine Übereinstimmung über ihren Begriffsinhalt zwischen den Lehrern nicht vorausgesetzt werden kann. Vielmehr muß jedes einzelne Merkmal operational definiert werden, so daß die Einschätzung des Schülerverhaltens mit Hilfe einer vorgegebenen Beobachtungs- und Merkmalsliste in Form einer Schätz-Skala (rating scale) vorgenommen werden kann.

2.6 Schülerberichte

Es wird zahlreiche Fälle geben, in denen Testergebnisse und Skalenwerte mit der Selbsteinschätzung eines Schülers nicht übereinstimmen, bzw. wichtige Entwicklungsmomente unberücksichtigt bleiben. So können zum Beispiel schwache Leistungsergebnisse in den Tests mit hohen Werten für bestimmte nicht-kognitive Verhaltensbereiche („Interesse", „Arbeitshaltung"), usw. gemeinsam auf-

treten oder umgekehrt. Solche Differenzen verlangen nach einer Interpretation, denn sie können sehr verschiedene Gründe haben; unter Umständen könnten nämlich schwache Leistungen einen guten Erfolg im Rahmen einer individuellen Schulbiographie darstellen. Die Leistungsergebnisse müssen daher in den Kontext der individuellen Entwicklung gestellt werden, damit Schüler und Eltern die notwendigen Konsequenzen ziehen können, unnötige Entmutigungen und Konflikte unterbleiben und die erforderlichen pädagogischen Akzente gesetzt werden. Solche Interpretationen könnten in Form kurzer halbstandardisierter Berichte niedergelegt werden, die an Hand eines Katalogs relevanter Dimensionen vom Klassenlehrer nach Absprache mit den Fachlehrern und dem „Beratungsexperten" abgefaßt werden sollten.

2.7 Diagnose durch Schulleistungstests

Sowohl bei der Konstruktion der Tests wie bei ihrer Anwendung sollte der Gesichtspunkt der zuverlässigen Fehleranalyse zum Zweck der Diagnose Beachtung finden, damit Leistungsschwächen einzelner Schüler oder Schülergruppen festgestellt und entsprechend gezielte Förderungsmaßnahmen ergriffen werden. Gerade an dieser Stelle wird deutlich, daß objektive Testverfahren nicht auf die Funktion beschränkt sind, die Leistungsbewertung zu normieren, sondern für den Lehrer auch wichtige individualpädagogische Hilfsmittel sind.

3. Leistungsbeurteilung im Entwicklungsstadium der Gesmtschulen – Zeugnisse

Da für die Entwicklung einer Itembank in der Bundesrepublik eine Reihe von Jahren nötig ist, wird für das Experimentalprogramm die folgende Übergangslösung vorgeschlagen.

Für die Schüler bleibt das Bedürfnis bestehen, an ihrer Gruppe gemessen zu werden. Für den Lehrer bleibt die Forderung, daneben einen gruppenübergreifenden Maßstab zu gewinnen. Sobald bessere Instrumente, besonders die mit Hilfe einer Itembank zusammengestellten Tests vorliegen, sollen diese bei der Beurteilung herangezogen werden. Inzwischen wird folgendermaßen zu verfahren sein:

3.1 Zensuren

Die fachliche Leistung des Schülers sollte durch informelle Tests bestimmt werden, die an der gesamten Bezugspopulation, die dieselbe Schule besucht (also zum Beispiel an allen Schülern derselben Jahrgangsstufe und in demselben Fach), geeicht worden sind.

Aus der Verteilung der Testergebnisse in jeder einzelnen Klasse (Gruppe) wird die Position jedes Schülers nach dem Innenkriterium bestimmt.

Die Testergebnisse können innerhalb jeder Klasse in die gewohnten Zensuren umgesetzt werden, wobei nach einem mindestens für die einzelne Schule verbindlichen Schlüssel eine Zuordnung von Testpositionen (zum Beispiel Prozentrangplätzen) und Zensuren getroffen wird.

3.2 Übergänge

Übergangsentscheidungen zwischen Kursen werden nach Maßgabe der Testergebnisse getroffen.

3.3 Zeugnisse

Im Zeugnis werden die in der unter 3.1 beschriebenen Weise ermittelten klassenimmanenten Zensuren unter Hinzufügung der Kursbezeichnung mitgeteilt. Allerdings ist ein solches Zeugnis unverständlich, solange keine Kenntnis der Kurseinteilungen und -bezeichnungen in der Öffentlichkeit vorausgesetzt werden kann. Dafür ist die Übereinstimmung der Bezeichnung zwischen den Schulen auf Bundesebene, mindestens aber auf Landesebene, notwendig.

Bis diese Bedingungen erfüllt sind, muß entweder ein Aufriß des Differenzierungssystems der Schule im Zeugnis enthalten sein oder – vorzugsweise – zusätzlich zu den gruppenimmanenten Zensuren die relative Position des Schülers in seiner Bezugsgruppe angegeben werden (zum Beispiel Prozentrangplatz, bezogen auf die Leistung der Jahrgangsstufe der Schule in demselben Fach).

Zensurenäquivalenzen dürfen aus den oben unter 2.4 genannten Gründen allenfalls unter Beziehung auf einen gruppenübergreifenden Maßstab hergestellt werden, keinesfalls jedoch mechanisch (zum Beispiel C2 = B3 = A4) erfolgen. Die Mitteilung gruppenimmanenter Zensuren in der beschriebenen Verbindung mit der Testposition ist jedoch in jedem Falle vorzuziehen.

Nicht-kognitive Dimensionen werden durch Schätzskalen (rating scales), erfaßt und stets getrennt von den Leistungsurteilen mitgeteilt. Jedes Zeugnis mit Ausnahme des Abgangszeugnisses enthält außerdem einen Schülerbericht.

3.4 Übergangslösung für die Leistungsbewertung in der Oberstufe

Die Entwicklung objektivierender Beurteilungsverfahren ist für die komplexeren und vielseitigeren Lerninhalte der Oberstufe methodisch sehr viel schwieriger und arbeitsmäßig sehr viel aufwendiger als für die Beurteilungsverfahren der Mittelstufe. Daher wird man in der Oberstufe noch länger auf solche objektivierenden Beurteilungsverfahren verzichten müssen. Die Beurteilung der Leistungen von Oberstufenschülern wird noch lange Zeit mit den im bisherigen Schulsystem verwendeten Methoden erfolgen müssen. Diese Methoden können durch die Entwicklung eines Punktesystems für die Beurteilung (credit system) wenigstens etwas zuverlässiger und für Schüler und Lehrer durchschaubarer gemacht werden. Allerdings wird man schrittweise auch für die Oberstufe wenigstens informelle Tests entwickeln und insbesondere in den obligatorischen Fächern, in denen Lernziele stärker festgelegt sind, für die Leistungsbeurteilung sinnvoll verwenden können. Darüber hinaus könnte eine Itembank auch die meisten Beurteilungsprobleme in der Oberstufe lösen, ihre Entwicklung für diese Stufe bleibt aber aus den genannten Gründen ein Fernziel.

3.5 Kooperation der Lehrer

Die Entstehung eines die Gruppen übergreifenden Standards setzt voraus, daß die Lehrer kooperieren. Für die Zwecke der Leistungsbewertung ist die regelmäßige Hospitation in den Nachbarkursen und die Diskussion zwischen den Lehrern über Unterrichtsgestaltung und Unterrichtsinhalte, Leistungsniveau und Leistungsbewertung in den Kursen Voraussetzung für die Entwicklung gemeinsamer Standards und Beurteilungsnormen, die über die jeweils unterrichtete Gruppe hinausreichen. Die eigene Erfahrung mit Nachbargruppen wird dem Lehrer zu einer relativierten Einschätzung der Leistungsfähigkeit seiner Schüler verhelfen, die eine Grundvoraussetzung für die Mobilität von Schülern zwischen Gruppen ist.

Neben der regelmäßigen Hospitation und der Diskussion zwischen Lehrern muß der Austausch zwischen ihnen in pädagogisch-didaktischen Konferenzen (Fach- und Differenzierungskonferenzen) institutionalisiert werden, die das Problem der differenziellen Leistungsbewertung thematisieren.

Hospitation und Konferenz sind bisher bewährte organisatorische Mittel, um die Leistungsbewertung über verschiedene Gruppen hinweg sicherer und die Beurteilungskriterien transparent zu machen. Informelle Tests, Schätzskalen (rating scales) und der Bericht sind die entscheidenden Hilfsmittel für die Objektivierung der Beurteilung. Diese Lösung ist ein Notbehelf, solange gesicherte Verfahren nicht vorliegen. Im Rahmen des Experimentalprogramms, für die Dauer der Entwicklungsphase von Differenzierungsformen, müssen institutionell-organisatorische Vorkehrungen getroffen werden, um den Lehrern die gegenseitige Hospitation zu ermöglichen.

3.6 Punktsystem (credit system)

Eine stärkere differenzierende und flexiblere Schulorganisation muß auch neue Formen des Leistungsnachweises und der Erteilung von Berechtigungen entwickeln. Da die Schüler Kurse mit verschiedenen inhaltlichen Schwerpunkten und unterschiedlicher Intensität wählen können und viele Fächer leistungsdifferenziert unterrichtet werden, sollte eine Gesamtbeurteilung durch ein Punktsystem (credit system) erfolgen. Es kann besser als die herkömmlichen Zeugnisse die vom einzelnen erworbenen wie die für bestimmte Abschlüsse erwarteten Qualifikationen ausweisen.

Sobald gültige und verläßliche Tests vorhanden sind, wird die Einrichtung eines solchen Punktsystems sehr erleichtert.

3.7 Nachbemerkung: Pädagogische Bewertungen

Wie aus den vorangegangenen Ausführungen hervorgeht, können Ergebnisse informeller Tests und Schätzungen gleichwohl die pädagogische Beurteilung der Schüler durch die Lehrer nicht ersetzen.

Keinesfalls sollte die Objektivierung der Leistungsbewertung dazu führen, in Zukunft von der pädagogischen Beurteilung der vom einzelnen Schüler erreichten Leistungen abzusehen. Die hier vorgeschlagenen Verfahren bieten dem Lehrer indes Hilfen zur Präzisierung seines Urteils, die er im Rahmen von Gesamtschulen, aber nicht nur dort, dringend benötigt.