Jürgen Oelkers

Leistungen und Noten: Probleme der Schülerbeurteilung[*)]

Die „Fragwürdigkeit der Zensurengebung“ hiess ein in Deutschland und in der Schweiz sehr erfolg- und folgenreicher Sammelband, der 1971 in erster Auflage erschien und bereits fünf Jahre später sechs Auflagen erreicht hatte[1]. Die Befunde dieses Bandes beschäftigten seinerzeit die Medien und lösten auch heftige Reaktionen in der Lehrerschaft aus. Sie war überwiegend entsetzt über die Befunde der Forschung und fühlte sich zu Unrecht angegriffen. Oft waren die Reaktionen polemisch und überwiegend uneinsichtig. Der Herausgeber des Bandes, KARL-HEINZ INGENKAMP, berichtet darüber im Vorwort zur sechsten Auflage des Buches, also fünf Jahre nach Erscheinen des Bandes. Es heisst hier:

> „Ich war ... überrascht, dass fast kein Wort der Fäkaliensprache ausgelassen wurde, um meine ‚Nestbeschmutzung‘ zu kennzeichnen. Solche Stellungnahmen konnten aber einzelnen Aussenseitern zugerechnet werden. Deprimiert haben mich jedoch Briefe der Wortführer einer ‚schweigenden Mehrheit‘, die mir kollegial ins Gewissen redeten und in jeder Zeile offenbarten, dass sie noch nie etwas von den Fehlerquellen der Zensurengebung gehört hatten und auch nicht bereit waren, in entsprechenden Untersuchungen mehr als eine Zahlenspielerei unredlicher Statistiker zu sehen“ (INGENKAMP 1976, S. 5).

Inzwischen, heisst es weiter, habe sich diese Einstellung „bei sehr vielen Lehrern geändert“, nicht zuletzt „durch die Verbreitung dieses Buches“. Auch in der Ausbildung sei das Problem der Zensurengebung stärker gewichtet worden. Von der Forschung „unberührt“ geblieben sei „eigentlich nur die Praxis unserer Bildungsbürokratie“ (ebd., S. 5/6). So gesehen wäre der Band keine Leidens-, sondern eine Erfolgsgeschichte, uneinsichtige Bürokraten einmal beiseite gelassen.

26 Jahre später gibt es immer noch Noten, ist keineswegs, wie INGENKAMP seinerzeit forderte, ein „besseres diagnostisches Instrumentarium“ entwickelt worden und erscheint weiterhin „fragwürdig“, wie schulische Leistungen beurteilt werden und welche Folgen damit verbunden sind. Zwischen 1971 und 2002 sind diverse Varianten ausprobiert worden, es fehlt auch nicht an Vorschlägen für mehr oder weniger radikale Aenderungen der Beurteilungspraxis, aber das *hauptsächliche* Instrument der Beschreibung von Leistungen ist immer noch das Notenschema. Das gilt mit Zunahme des Alters der Schüler umso mehr: Je höher die Schulstufe, desto strikter wird das Notenschema verwendet.

Dieser Tatbestand kann nicht einfach auf die verstockte „Bildungsbürokratie“ zurückgeführt werden. Offenbar hat das Notenschema in der Praxis viele Vorteile, die in der psychologischen oder pädagogischen Kritik immer übersehen wurden. Die Kritik ist nicht

---

[*)] Vortrag anlässlich der Fortbildungstagung des Gymnasiums Hofwil im coop-Zentrum Muttenz am 11. Februar 2002.

[1] Neun Auflagen bis 1995.

neu, und sie ist auch kein Phänomen der Aufbruchstimmung zu Beginn der siebziger Jahre, als eine neue Schule erfunden werden sollte. Voten „gegen Prüfungen und Noten“ finden sich schon am Ende des 19. Jahrhunderts (SCHREIBER 1899), seinerzeit vorgebracht von Medizinern, die die Fachöffentlichkeit mit den schädlichen Folgen von Prüfungsstress konfrontierten. Seitdem sind die *Nachteile* von Prüfungen und Noten extensiv untersucht worden, ohne die einfache Frage zu beantworten, warum es angesichts der Nachteile immer noch Prüfungen und Noten gibt. Die Praxis kann nicht einfach als einzige Fehlhandlung hingestellt werden, das würde hundert Jahre Unsinn unterstellen.

Die *Vorteile* von Ziffernnoten sind etwa die folgenden:

1. Eine Skala von fünf oder sechs Noten erlaubt die Beschreibung einer Normalverteilung in der Klasse.
2. Die Beschreibung ist kurz und eindeutig.
3. Das Notenschema ist ein öffentlicher Standard und wird nicht nur in der Schule verwendet.
4. Das Schema lässt sich auf ökonomische Weise einsetzen und kommunizieren.
5. Probleme der Ausdeutung sind gering.

Demgegenüber haben die oft genannten Alternativen zur Notenskala erhebliche Nachteile.

- *Textliche Leistungsbeschreibungen* oder „Wortgutachten“ verlangen einen erheblich grösseren Aufwand, sind in stärkerer Weise interpretationsabhängig und haben Probleme vor allem bei der Formulierung negativer Beurteilungen[2]. Sie verstecken die Urteile oft hinter differenzierter Freundlichkeit.
- *Standardisierte Beurteilungsbögen* verwenden zumeist weiche und vage Kriterien, zielen auf „ganzheitliche“ Beurteilungen und müssen wiederum aufwendungsreich interpretiert werden.
- *Diskursive Verfahren*, etwa Gespräche mit Eltern und Schülern, konfrontieren die Notengeber, also die Lehrkräfte, mit Akzeptanzproblemen, die nicht selten Machtproben darstellen. Was begründet werden muss, verlangt erheblichen Aufwand und führt nicht immer zu einem glücklichen Ausgang.

Diese alternativen Verfahren mischen Notengebung und Austausch, ohne dass über die Noten wirklich verhandelt werden könnte. Letztlich taucht in allen Varianten das Notenschema wieder auf, weil in jedem Falle eine vergleichende Leistungsbeurteilung abgegeben werden muss. Warum also sollte man das bewährte Schema dann nicht gleich und in originaler Form verwenden?

Ich werde die offenbar heikle Frage der Leistungsbeurteilung nicht lediglich als *Problem*, sondern zugleich als *Chance* der Schulentwicklung verstehen. Die Idee einer nicht-selektiven Schule, die auf vergleichende Leistungsbeurteilung gänzlich verzichtet, wird sich mindestens im Gymnasialbereich nicht durchsetzen. Sie ist auch insofern fraglich, als Schülerinnen und Schüler in aller Regel eine solche Leistungsbeurteilung wünschen und wenn, dann Form und Verfahren kritisieren. Insbesondere monieren sie die mangelnde Transparenz, die dazu führt, die Notengebung wie ein undurchschaubares Schicksal zu erleben.

---

[2] ULBRICHT (1993).

Die Praxis der Leistungsbeurteilung aber kann entwickelt werden, wobei auffällig ist, dass diese Aufgabe in den heutigen Modellen der Schulentwicklung bislang kaum eine Rolle spielt. Das zentrale Problem, wie die Leistungen bewertet und beschrieben werden sollen, erhält überraschend wenig Aufmerksamkeit. Das gilt nicht für die Lehrkräfte, die hier einen hohen Belastungsfaktor sehen, wohl aber für die Reformdiskussion, in der die Zukunft der Schule auf seltsame Weise *notenfrei* oder *notenfern* gesehen wird. Aber wenn Schulen Leistungen abverlangen, müssen sie sie auch bewerten, und wenn Leistungen unterschiedlich ausfallen, muss dies dargestellt werden.

Meine Ausführungen haben drei Teile: Zunächst werde ich auf den Stand der Forschung eingehen und berichten, was seit INGENKAMPS „Fragwürdigkeit der Zensurengebung" geschehen und auch nicht geschehen ist (1). In einem zweiten Schritt gehe ich auf die Belastungsfaktoren heutiger Schularbeit näher ein, unter denen offenbar Leistungsbeurteilung und Notengebung einen hohen Stellenwert einnehmen (2). Abschliessend frage ich nach realistischen Möglichkeiten der Entwicklung, die in einem selektiven Bildungssystem nicht einfach darin bestehen kann, jeden Schüler nach seinem Talent für sich und ohne Vergleich mit Anderen zu fördern (3).

1. Ergebnisse der Forschung

Zeugnisse und Noten, dazu Prüfungen und alternative Leistungsdiagnosen, sind Forschungsthemen, die sich im deutschen Sprachraum deutlich auf die zwanziger und sechziger Jahre des 20. Jahrhunderts konzentrieren. Besonders zwischen 1950 und 1970 lässt sich ein starkes Forschungsaufkommen nachweisen, das danach, im Unterschied zu den angelsächsischen Verhältnissen, abnimmt. INGENKAMP fasste die vorliegenden Forschungen zusammen und beschrieb die Resultate im Blick auf vier Dimensionen:

- subjektive Fehlerquellen
- das klasseninterne Bezugssystem
- die Qualitätsunterschiede in den einzelnen Schulfächern
- den prognostischen Wert.

Die bis 1976 vorliegenden Arbeiten verweisen, so INGENKAMP, auf eine sehr weitgehende *Subjektivität* in der Zensurengebung. Zensuren und Noten sind persönlich gefärbte Urteile von Lehrkräften, nicht objektive Bewertungen, die auf einen gemeinsamen Massstab verweisen würden. Das gelte selbst oder gerade für den Mathematikunterricht. Eine deutsche Studie zeigte, dass einige Mathematiklehrkräfte von den Fehlern, andere von den richtigen Lösungen ausgingen, um zur Zensur zu gelangen. Die Ansichten über die Bewertung von Faktoren wie Sauberkeit, Schrift, formal-mechanisches Rechnen im Vergleich zum Finden des Lösungsansatzes oder zum mathematischen Denken variierten erheblich (HAECKER 1971). Für die Bewertung sei auch die „implizite Persönlichkeitstheorie" entscheidend (HOFER 1975), mit der die Lehrkräfte ihre Verhaltens- und Charaktererwartungen an die Schüler bestimmen. Diese Erwartungen werden nicht explizit formuliert und sind gleichwohl wirksam. Weitere Fehlerquellen im Urteil sind Stereotypen in der Geschlechtswahrnehmung[3] oder auch einfach individuelle Vorlieben, wie schon früh in der Bewertung von Aufsätzen gezeigt wurde (BOBERTAG 1933, ULSHÖFER 1949).

[3] Mädchen erhalten bessere Noten, als ihren in Tests erfassten Leistungen entspricht (INGENKAMP 1976, S. 71f.).

Die drei anderen Dimensionen der Fragwürdigkeit beziehen sich auf Bedingungen und Folgen der Subjektivität. Noten werden aufgrund des *klasseninternen Bezugssystems*, also der beobachteten und bewerteten Verteilung in einer bestimmten Klasse, gegeben, aber der Wert der Noten bezieht sich immer auf das gesamte Schulsystem und alle Berechtigungen. Unterstellt wird die Vergleichbarkeit der Notengebung an allen Orten, während die tatsächlichen Bewertungen von Schule zu Schule und von Fach zu Fach variieren, zum Teil erheblich und nicht selten ärgerlich. Ziffernnoten sind nie mit den Entstehungskontext versehen, also eine „5 _ in Französisch" erscheint *nicht* mit dem Signum „Gymnasium Hofwil/Maturitätsjahrgang 2002", zusätzlich Angaben über Qualitätsmassstäbe, Lehrmittel und Leistungsniveaus der Schüler im Vergleich.

In verschiedenen Fächern wird mit unterschiedlicher Strenge zensiert, die Strenge nimmt mit der Bedeutung des Faches in der Stundentafel zu, „stark selektive Hauptfächer" haben die strengste Notengebung (DE GROOT 1971). Gleichzeitig zeigen Untersuchungen der Häufigkeitsverteilung, dass Zensuren „keine Intervallskala mit gleichen Notenabständen" darstellen (INGENKAMP 1976, S. 232). Die Benotungsstrenge nimmt mit der Klassenstufe zu, bleibt in der Verteilung der Fächer aber gleich. Die Hauptfächer des Gymnasiums steigern die Anforderungen, die Zensurenverteilung in den Nebenfächern wird eher noch milder (HOPP/LIENERT 1976). In vielen Fällen, nicht zuletzt in bestimmten Universitätsfächern, werden nur die beiden obersten Ziffern der Skala wirklich genutzt.

Der prognostische Wert dieser Praxis sei gering, schrieb INGENKAMP 1976. Zwar korrelieren innerhalb eines Schultyps die Zensuren in aufeinander folgenden Klassenstufen relativ hoch, aber das sei für die Prognose von Erfolg oder Misserfolg in nachfolgenden Schulen oder Stufen kaum von Bedeutung. Die Qualität ist zu unterschiedlich: Primarschulzeugnisse gewichten Elementarisierungseffekte wie die Rechtschreibleistungen oder auch Faktoren wie Fleiss und Wohlverhalten erheblich stärker als intellektuelle Fähigkeiten. Die Masstäbe von Aufnahmeprüfungen variieren von Jahr zu Jahr und von Ort zu Ort wiederum erheblich. Zudem hängt der Schulerfolg von der Stärke des Jahrgangs und den Klassenfrequenzen ab (TENT 1969; SAUER/GAMSJÄGER 1996). Der stärkste Faktor für Erfolg oder Misserfolg aber ist die Schichtzugehörigkeit, das soziale Milieu, aus dem die Schüler stammen (INGENKAMP 1976, S. 274ff; für den Kanton Zürich auch MOSER/RHYN 2000).

Der Topos der „Fragwürdigkeit" zieht sich seitdem durch die deutschsprachige Forschungsliteratur. Die vier Dimensionen von INGENKAMP werden in einem Forschungsbericht von 1999 nochmals herausgestellt. Demnach ist die Benotung in verschiedenen Fächern unterschiedlich streng (ZIEGENSPECK 1999, S. 137ff.), wobei nicht nur der Rang des Faches eine Rolle spielt. In der Tendenz wird umso strenger zensiert,

- je mehr die Leistungen in schriftlichen Arbeiten überprüft werden,
- je mehr die Leistungen quantifizierbar sind
- und/oder je stärker die verbalen Anforderungen hervortreten[4].

Mädchen erhalten dabei bessere Zeugnisnoten als Jungen, sie repetieren erheblich weniger und haben inzwischen mindestens in Deutschland den grösseren Schulerfolg (RICHTER 1996). „Erfolg" wird in den diesbezüglichen Studien wie eine objektive Grösse betrachtet und mit Noten in Verbindung gebracht, zugleich wird der Topos der Fragwürdigkeit der Noten weiter verwendet. Auch 1999 ist der prognostische Wert von

[4] WEISS (1962); INGENKAMP (1976).

Zensuren oder Zeugnissen gering (ZIEGENSPECK 1999, S. 156ff.) und greift das „klasseninterne Bezugssystem", also die Relativierung der Notenaussagen im Blick auf Ort und Kontext, in dem sie entstehen. Noten sollen ortsübergreifend vergleichbare Leistungen beschreiben, aber was sie erfassen, ist lediglich die Verteilung in einer bestimmten Leistungsgruppe.

Grossen Raum nimmt wiederum die Beschreibung der „subjekiven Fehlerquellen" ein. Als Effekte werden unter anderem genannt:

- *Halo-Effekt*: Ein globaler Allgemeineindruck bestimmt die Wahrnehmung einzelner Merkmale
- *Beharrlichkeitstendenz*: Lehrkräfte rücken von einem bereits gefällten Urteil bei späteren Beurteilungen nicht ab
- *Reihungseffekt*: Unter dem Eindruck, „es können doch nicht alle gleich schlecht sein" werden bessere Noten gegeben
- *Kontrasteffekt*: Nach einer Serie von sehr guten Leistungen wird eine mittelmässige Leistung tendenziell als schlecht bewertet
- *Beurteilungstendenzen*[5]: Milde oder Strenge, „zentrale Tendenz" (Vermeidung von Extremwerten) und „motivierende" versus „selektive" Notengebung
- *Wissen-um-die Folgen-Fehler*: Mildere Beurteilung bei absehbar negativen Folgen für die Schüler, nicht umgekehrt.

Damit verbunden sind die seit Beginn des 20. Jahrhunderts immer wieder beschriebenen Beurteilungsunterschiede bei der Zensierung zwischen verschiedenen Lehrkräften oder zwischen derselben Lehrkraft zu verschiedenen Zeitpunkten (GREY 1913; ZIEGENSPECK 1999, S. 188ff.). Die Streuung der Aufsatznoten ist ein bekanntes Phänomen, weniger bekannt sind Befunde, die darauf hindeuten, dass die gleichen Beurteiler die gleiche Arbeit zu verschiedenen Zeitpunkten höchst unterschiedlich bewerten. Zudem gibt es eine „Verlaufskurve der Bewertung" (WEIDIG 1961), die sich bei langwierigen Korrekturarbeiten einstellt und im gleichen Fach individuell sehr verschieden ist. Die erste Arbeit einer Korrekturserie wird tendenziell anders bewertet als die letzte, wobei der Effekt mit der Länge der Gesamtkorrektur zunimmt. Schliesslich hat die Reihenfolge in der Qualitätswahrnehmung einen Einfluss: „Kontrasteffekt" meint in diesem Zusammenhang, dass dieselbe Arbeit, vom Lehrer nach Durchsicht vieler *guter* Proben korrigiert, ganz anders ausfällt als dies nach Durchsicht vieler *schlechter* Problem geschehen würde.

Implizite Persönlichkeits- und Charakterurteile spielen eine nicht zu unterschätzende Rolle, ebenso die Einschätzung der allgemeinen Qualität und Leistungsbereitschaft. Lehrerinnen und Lehrer beziehen sich in ihrem Unterricht mehr auf die guten Schüler, sie lassen gute wie schlechte Schüler direkt oder häufiger indirekt wissen, was sie von ihrer Leistungsfähigkeit und oft damit verknüpft von ihrer Person halten, und das Selbstbild der Schüler passt sich tendenziell dem Bild an, das die Lehrer von ihnen haben und mitteilen[6]. Die Qualitätshierarchie in einer Klasse steht relativ früh fest und ist nur schwer revidierbar.

Hinzukommen die Attribuierungen aufgrund von Erfolg oder Misserfolg. Beide können dem eigenen Vermögen oder dem eigenen Unvermögen zugeschrieben werden, wobei das schulische Selbstbild offenbar in beiden Hinsichten verstärkend wirken kann. Es scheint dann ein „vicious circle" aufzutreten, wer als Schüler ständig keinen oder nur mässigen Erfolg hat, traut sich künftig auch keinen Erfolg zu. Es gibt auch Effekte, die zeigen, wie abhängig

[5] Nach ULBRICHT (1993).
[6] „Pygmalion-Effekt" nach ROSENTHAL/JACOBSON (1971).

Lehrkräfte von der sozialen Situation und dem Entgegenkommen der Klasse sind, die sie unterrichten. Lehrkräfte, denen die Schüler *positive* Erwartungen entgegenbringen, stufen sich selbst im Blick auf ihre Tüchtigkeit höher ein als Lehrkräfte, bei denen die Erwartungen der Schüler *negativ* sind. Wenn ein Schüler konstante Leistungen zeigt, ist das auch eine Reaktion auf das Verhältnis zum Lehrer.

Ueber Lehrerkommentare zu Leistungen ist folgendes bekannt (KRAMPEN 1987):

1. *Sozial orientierte* Lehrerkommentare wirken bei leistungsschwächeren Schülern deutlich negativ, bei leistungsstärkeren neutral oder leicht positiv.
2. An einem *sachlichen Standard* orientierte Lehrerkommentare wirken in der Tendenz bei allen Schülern positiv, ohne dass eine bestimmte Leistungsgruppe deutlich von ihnen profitiert.
3. *Individuell orientierte* Lehrerkommentare wirken ebenfalls bei allen Schülern tendenziell positiv, am meisten profitieren davon die leistungsschwächeren.

Generell scheint die feste allgemeine Einschätzung eines Schülers oder einer Schülerin die Notengebung zu beeinflussen. Wer gute Leistungen zeigt, erhält und *hat* so auch Charaktervorteile. Die Noten einzelner Schüler in bestimmten Fächern sind über das Schuljahr verteilt erstaunlich konstant[7]. Wenn die generelle Einschätzung einmal gefasst ist, scheint sie kaum korigierbar zu sein. Tendenziell wird dem guten Schüler eine schlechte Leistung nicht nachteilig, dem schlechten Schüler eine gute Leistung nicht positiv vermerkt. Offenbar korreliert auch Sympathie auf Seiten des Lehrers mit besseren und Antipathie mit schlechteren Noten (ZIEGENSPECK 1999, S. 206). Zudem ist bekannt, dass

> dort, wo Schüler in zwei Fächern von *ein und derselben* Lehrkraft unterrichtet werden, sich ein bedeutsamer und hoher Zusammenhang zwischen den Zensuren in beiden Fächern ergibt, während dieser signifikante Fachnotenzusammenhang bei Schülern, die in den zwei betreffenden Fächern *verschiedene* Lehrer hatten, nicht nachgewiesen werden konnte (ebd., S. 207).

Angesichts dieser Resultate könnte man schliessen, es sei *unmöglich*, zu einer „gerechten Zensierung" der Leistungen von Schülern zu gelangen und dass dies der Lehrerschaft auch bewusst sei (DÖRING 1925, S. 177)[8]. Das schrieb der Jugendpsycholge OTTO DÖRING 1925, aber das wirft dann nochmals die Frage auf, warum immer noch weitgehend *Zifferennoten* die Praxis der Leistungsbeurteilung bestimmt. Es ist mindestens frappierend, wie vergleichbar gering die Entwicklungsarbeit gerade auf diesem Gebiet gewesen ist und wie stark die offenkundigen *Vorteile* der Notenskala das Geschäft bestimmt haben. Nochmals: Noten sind einfache Instrumente, deren Aufwand begrenzt ist und die leicht kommuniziert werden können.

Aber die Genauigkeit von Schülerbeurteilungen scheint sich auf das „*klasseninterne* Bezugssytem" zu beschränken. „Im Durchschnitt betrachtet", heisst es in einer gerade veröffentlichten Forschungsübersicht, können die Lehrkräfte „die Rangreihe der Leistungen innerhalb ihrer Klasse recht gut einschätzen", auch wenn mit „erheblichen Unterschieden" zwischen den Lehrkräften gerechnet werden muss (WEINERT 2001, S. 50). Die klasseninterne Rangreihe der Leistungen, also die Normalverteilung, entspricht *nicht* den tatsächlichen Schülerleistungen, wenn man diese *unabhängig* testet. Noten beschreiben den internen Rangunterschied. Aber auch wenn Lehrkräfte gut in der Lage sind, die Schüler ihrer eigenen

[7] Seit RANK (1962) und INGENKAMP (1963) ist das verschiedentlich nachgewiesen worden.
[8] Zum Problem der Gerechtigkeit WEISS (1962.

Klasse gemäss ihren Leistungen zu rangieren, heisst das nicht, „dass gleichen Noten in unterschiedlichen Klassen auch vergleichbare Leistungen zu Grunde liegen" (ebd.).

Daraus ergibt sich eine dreifache Forderung:

1. Die subjektive Beurteilung muss durch eine *objektive* ersetzt werden, die Standards oder einen übergreifenden „Bezugsmasstab" voraussetzt.
2. Das objektive Kriterium sind die vom Lehrplan geforderten *Lernziele* unabhängig von den in einer Klasse vorhandenen Leistungsunterschieden.
3. Dafür müssen am Lehrplan orientierte, *diagnostische Instrumente* entwickelt werden, die die Noten ersetzen.
   (ebd.).

Die beiden ersten Forderungen sind inzwischen in der bildungspolitischen Diskussion akzeptiert, wenngleich unklar ist, wie sie erfüllt werden sollen. Die dritte Forderung ist heikel, und dies nicht nur, weil sie die Profession noch mehr in Richtung Psychologie verändern würde. Solche neuen Instrumente haben zumeist nicht nur eine Dimension, nämlich die der Lernziele. Diese Dimension wird oft „kriteriale Norm" genannt, die in vielen alternativen Modellen der Leistungsbeurteilung gleichrangig ergänzt wird durch eine *Sozialnorm* und eine *Individualnorm*, also den Leistungsvergleich innerhalb der jeweiligen Lerngruppe und den individuellen Fortschritt, den die Leistung für den Schüler erbracht hat.

In trivialeren Konstrukten wird zwischen „Sozial-", „Selbst-"- und „Sachkompetenz" unterschieden, ohne dabei der Sachkompetenz das deutlich grössere Gewicht zu verleihen. *Selbstkompetez*, ohne dass klar wäre, was genau darunter verstanden werden soll, erhält den gleichen Rang wie *Sachkompetenz*, wobei nicht geklärt ist, wie Fachunterricht zu Kompetenz führen kann oder soll. Auffällig bei derartigen Vorschlägen ist aber insbesondere, dass sich der Aufwand durch Komplizierung und Vervielfätigung der Kriterien erhöht, ohne dass reale Belastungsfaktoren ins Spiel gebracht würden. Mit diesem Missverhältnis beschäftige ich mich in einem zweiten Schritt. Er gilt dem Phänomen, dass Reformen immer die Belastungen steigern, ohne sie je wirklich zu kalkulieren. Innovationen sind in diesem Sinne immer abstrakt und können sich doch als gutartig hinstellen.

2. Aufgaben und Belastungen der Lehrkräfte

Die historisch sehr stabile Notenskala erlaubt den Lehrkräften bei allen Belastungen immer noch eine halbwegs ökonomische Arbeitsweise, die offenbar im Blick auf das „klasseninterne Bezugsssystem" so schlecht nicht ist. Lehrkräfte beurteilen zutreffend, was sie beurteilen können, nämlich die Leistungen in der Klasse, die sie unterrichten. Dass die Urteile „subjektiv" sind, muss nicht heissen, dass sie sämtlich unvergleichlich sind. Auf der anderen Seite werden Noten gerne dann verwendet, wenn bestimmte Interessen objektiviert werden sollen: Die Beschreibungen des Schulerfolgs von Mädchen und jungen Frauen verwendet jenes Notensystem, das ansonsten verpönt ist. Aehnlich liegt das „fragwürdige" Notensystem allen Einschätzungen zugrunde, die - zu Recht - die Benachteiligung bestimmter Schülergruppen durch ihre Herkunft monieren. Misserfolg wird auch hier mit Noten beschrieben.

Die schwache Prognosefähigkeit muss im Blick auf den Schulerfolg relativiert werden, offenbar sind Noten nicht einfach nur subjektive Etikettierungen, die einzig in einer

bestimmten Klasse Geltung haben. Zudem liegen Langzeituntersuchungen kaum vor, die nicht einzelne Gruppen oder Kohorten befragen, sondern Erfolg und Misserfolg bestimmter Gruppen durch die Schulzeit und danach verfolgen. Eine der wenigen Studien, die so verfährt[9], zeigt etwa, dass Schüler, die bei Schuleintritt Lesen konnten und bestimmte Rechenoperationen beherrschten, mithin gleichsam automatisch zu den Besten ihrer Klasse zählten, zu einem grösseren Teil auch nach sechs Jahren noch diesen Rang einnahmen, trotz Klassen-, Stufen- und Lehrerwechsel sowie vielfacher Benotung. Entweder ist die Benotung dann so schlecht nicht, wie die Kritik unterstellt, oder andere Faktoren, etwa die Milieuzugehlrigkeit, sind stärker als alles, was Schulen unternehmen können.

Wie auch immer: Die empirische Notenforschung verfolgt Fragestellungen, die seit Beginn des 20. Jahrhunderts stabil sind und eigentlich von vornherein die Notengebung unter Verdacht stellen. Kritische Gegenevidenz ist unter dem Eindruck bestimmter Studien mindestens im deutschen Sprachraum nie erzeugt worden. Auffällig ist auch, dass die grundlegenden Datensätze kaum erneuert wurden, obwohl sich die Praxis der Notengebung seit den sechziger Jahren, aus denen die hauptsächlichen Untersuchungen stammen, erheblich verändert hat. Die Schärfe der Kritik hat auch zu tun mit dem Objektivitätsideal, das der Datenerhebung zugrunde liegt. Dieses Ideal ist nie wirklich hinterfragt worden, schon gar nicht mit unbefangener Empirie, während es genutzt wird, um a priori einen Verdacht aussprechen zu können.

Die „Mängel der traditionellen Notengebung" sind auch Thema eines Trendberichts der Schweizerischen Koordinierungsstelle für Bildungsforschung, der 1999 veröffentlicht wurde (VÖGELI-MANTOVANI 1999). Dieser Bericht trägt den programmatischen Titel

Mehr fördern,
weniger auslesen.
Zur Entwickung
der schulischen Beurteilung
in der Schweiz.

Die Kritik der Noten wiederholt im wesentlich den Topos der „Fragwürdigkeit", INGENKAMPS Thesen von 1971 spielen also auch 1999 noch eine zentrale Rolle in der Beurteilung des Problems. Die *Mängel* der Notengebung werden in sechs Aussagen zusammengefasst: Noten sind schlecht oder unbrauchbar, denn:

1. Verschiedene Lehrkräfte bewerten dieselbe Arbeit unterschiedlich.
2. Die Lehrkraft hat die Tendenz, dieselbe Arbeit zu verschiedenen Zeitpunkten unterschiedlich zu bewerten.
3. Es ist keineswegs klar, was mit einer Note zum Ausdruck gebracht wird.
4. Die gängige Benotungspraxis hat viele unerwünschte Nebeneffekte.
5. Noten sind zur Beurteilung bestimmter Sachverhalte ungeeignet.
6. Notenarithmetik ist mathematisch unzulässig.

Von möglichen *Vorteilen* ist dagegen keine Rede. Die ersten beiden Verdikte folgen Daten und Interpretationen aus dem Sammelband von INGENKAMP. Der dritten Aussage liegt eine Schweizer Studie aus dem Jahre 1971 zugrunde (FLAMMER 1971), die darauf verweist, dass Noten übermässig viele Funktionen erfüllen müssen[10], ohne klar definiert zu sein. Im

[9] Hinweise auf diese Studie verdanke ich MARGRIT STAMM.

[10] „Noten sollen u.a. den Schüler und andere über dessen Leistungsstand informieren, der Lehrkraft Rückschlüsse auf die Qualität ihres Unterrichts erlauben und ihr Planungshilfe sein, Voraussagen über die

Normalfall kennt der Schüler die lehrerspezifische Beurteilungsstrategie *nicht*, so dass die Note nur im Blick auf das Symbol, also die Ziffer, eindeutig ist. Das Zustandekommen ist zumeist intransparent, was aber natürlich den symbolischen Wert zunächst einmal nicht mindert.

Die vierte Aussage verwendet zwei Studien, die Ende der siebziger Jahre entstanden (HUBERMAN 1980; RHEINBERG 1980) und die zeigen, dass die Orientierung am Klassendurchschnitt sich ungünstig auf Leistungsmotivation, Anstrengungsbereitschaft, Ursachenzuschreibung von Schulleistungen sowie die Selbstwahrnehmung auswirkt. Aber damit ist eigentlich nur gesagt, dass Schüler auf Erfolge und Misserfolge unterschiedlich reagieren, während ein Leistungssystem davon bestimmt ist, dass nicht jeder gleich Erfolg haben kann. Die fünfte These hat keinen Bezug zu neueren Arbeiten und äussert den allgemeinen Verdacht, dass Standardisierung der Inhalte, also die Festlegung eindeutiger Kriterien, mit der Gefahr verbunden sei, die wünschbare Variationsbreite der Leistungen zu vermindern. Auf der Strecke blieben „Kreativität, Phantasie und Originalität" (VÖGELI-MANTOVANI 1999, S. 85). Die sechste These bezieht sich ein Problem, das ebenfalls immer wieder ins Feld geführt wurde (SACHER 1984), nämlich die „trügerische Sicherheit" der Durchschnittsberechnung.

> „Noten sind im besten Fall grobe Schätzwerte für den Leistungsrangplatz eines Schülers oder einer Schülerin innerhalb einer Klasse. Die Beurteilung lässt sich bei diesen ungenauen Ausgangsdaten auch durch exakte mathematische Prozeduren wie z.B. eine Durchschnittsberechnung nicht verbessern. Die Berechnung von Kommastellen und deren Interpretation vermitteln eine trügerische Sicherheit, die durch die *scheinbar* mathematische Genauigkeit suggeriert wird. Durchschnittsberechnungen mit Noten sind aber aus *mathematischen* Gründen unzulässig, weil Noten als Ordinalzahlen von unzureichender mathematischer Qualität sind, so dass keine rechnerischen Grundoperationen durchgeführt werden dürfen" (VÖGELI-MANTOVANI 1999, S. 86; Hervorhebung J.O.).

Mit der Logik dieses Arguments könnte weder eine Kantonsschul-Aufnahmeprüfung noch eine Maturitätsprüfung durchgeführt werden, die auf der anderen Seite aber weder beliebig verlaufen noch zu vollkommen willkürlichen Resultaten führen. Was in Rechnung gestellt werden muss, ist eine gewisse Unschärfe, die mehr oder weniger gross sein kann und die mit Belastungsfaktoren korreliert werden muss. Interessanterweise tun das die mir vorliegenden Studien zur „Fragwürdigkeit" der Notengebung sämtlich *nicht*. Sie kritisieren den psychologischen Wert von Noten, den pädagogischen Zweck oder den mit Noten verbundenen mathematischen Schein, während offenbar die Notengebung vor dem Hintergrund der zur Verfügung stehenden Zeit und so im Blick auf effizientes Arbeiten beurteilt werden muss (SCHRADER 1997, S. 664f.).

Inzwischen liegen einige seriöse Studien zu Belastungen der Lehrkräfte vor, darunter eine, die speziell für den Kanton Zürich in Auftrag gegeben wurde. Zumeist handelt es sich um Befragungen, aber auch um Interviews, Feldbeobachtungen und Dokumentenanalysen. Wenn etwa medizinische Gutachten ausgewertet werden, die Diagnosen enthalten, mit denen

---

zukünftige (Leistungs-) Entwicklung des Schülers ermöglichen und so als Grundlage von Schul- und Laufbahnentscheidungen dienen. Noten haben auch einen hohen Stellenwert bei Promotions- und Selektionsentscheiden. Zudem sollen sie zu guten Leistungen anspornen oder auch aufsässige Schüler disziplinieren „(VÖGELI-MANTOVANI 1999, S. 83).

die Frühpensionierung von Lehrkräften begründet werden, dann ergeben sich Daten[11], die auf Folgen von schweren Belastungen hinweisen, Depressionen, Gefühle des Ausgebranntseins, eine Zunahme psychosomatischer Erkrankungen und Aehnliches mehr. Qualitative Studien zeigen, dass amtierende Lehrkräfte[12] ihren hauptsächlichen Konflikt im Widerstreit zwischen dem eigenen pädagogischen Anspruch und dem, was sie alltäglich tun und erleben müssen, sehen (SCHÖNWÄLDER/PLUM 1998).

Belastend wirkt auch der berufseigene Idealismus, der auf anderen Seite unverzichtbar ist. Arbeitszeitstudien zeigen, dass die Lehrkräfte ihren Arbeitsaufwand als zu hoch einschätzen, während Entlastungsstunden kaum Wirkung zeigen. Die Gesamtarbeitszeit reduziert sich nicht proportional mit den reduzierten Stunden, was sich auch daran zeigt, dass Teilzeitangestellte Lehrkräfte in Relation mehr Arbeitszeit aufwenden als Vollzeitkräfte (Mummert+Partner 1999). Diese Resultate beziehen sich nicht auf bestimmte nationale Systeme, sondern gelten quer zu den Systemen, wenngleich mit typischen Variationen.

Eine grosse Untersuchung zur Arbeitszeit der Lehrpersonen in der deutschsporachigen Schweiz (LANDERT 1999) zeigte unter anderem folgende Befunde:

- Lehrkräfte unterschätzen ihre Arbeitszeit eher als dass sie sie überschätzen.
- Alle Wochentage sind belastet, die Wochenendarbeit variiert nach Schultyp und Schulstufe.
- Die durchschnittliche Arbeitszeit liegt ferienbereinigt höher, als im öffentlichen Dienst verlangt: Zwischen 44,6 und 47,3 Wochenstunden je nach Pensengrösse, zwischen 44,4 und 47,8 Stunden bezogen auf die Schulstufen.
- Die Jahresarbeitszeit konzenteriert sich auf das *Hauptgeschäft*, nämlich Unterrichten, Vor- und Nachbereitung sowie Planung und Auswertung.
- Für Betreuung und Beratung stehen 3% der durchschnittlichen Jahrsarbeitszeit zur Verfügung.

Wer in einem solchen Zeitrahmen tätig ist, wird jede unnötige Belastung vermeiden. Eine Steigerung der Belastung muss zum Aufgabenrahmen und zum Engagment passen. Das zeigt deutlich auch die Zürcher Studie (FORNECK/SCHRIEVER 2000), die Ende 2000 vorgelegen hat, also akute Daten präsentiert. Sie erhebt nicht nur zeitliche, sondern generell berufliche Belastungen. Lehrkräfte aller Kategorien geben dabei an, dass sie als besonders belastend empfinden

- Disziplinierungen angesichts von Sozialisationsdefiziten,
- Selektionsentscheide,
- Korrekturen/Prüfungen,
- Belastung durch Schulreform und Schulentwicklung,
- zu grosse Klassen.

Es gibt keine Zunahme der Arbeitszufriedenheit mit zunehmendem Alter, was darauf hindeutet, dass der vielzitierte Effekt der Routinisierung nicht vorhanden ist oder zumindest nicht die Zufriedenheit erhöht. Besonders belastend sind Beurteilungen an den Schnittstellen der schulischen Selektion, das Ungenügen der täglichen Unterrichtsvorbereitung oder - überproportional häufig in der Primarschule – „problembelastete Schüler". In der Mehrzahl

[11] Ich beziehe mich auf eine arbeitsmedizinische Studie der Universität Erlangen. Analysiert wurden 7100 Dokumente, in denen bayrische Lehrkräfte auf eine krankheitsbedingte Frühpensionierung untersucht wurden.
[12] An der Untersuchung waren Lehrkräfte von deutschen Grundschulen, Sonderschulen, Berufsschulen und eines Schulzentrums beteiligt.

der Kategorien wird die zur Verfügung stehende Zeit als *nicht ausreichend* angesehen. Und das hat auch etwas mit dem Beruf selbst zu tun: „Die Professionstätigkeit zeichnet sich dadurch aus, dass die in der Arbeit mit Menschen wahrgenommenen professionellen Herausforderungen *tendenziell unabschliessbar* sind" (ebd., S. 84; Hervorhebung J.O.).

Die an der Befragung beteiligten Mittelschul-Lehrkräfte gaben an, dass wohl auch *Unterrichtsvorbereitungen* ein Belastungsfaktor seien, vor allem aber, und dies im Unterschied zu sämtlichen anderen Schulformen, die *Korrektur von Prüfungen* sowie, damit zusammenhängend, wenngleich weniger gewichtet, die *Benotungen*. Das deckt sich mit der subjektiven Problemwahrnehmung, die gesondert erhoben wurde. Die Mittelschullehrerinnen und -lehrer empfinden die verfügbare Zeit für die Korrektur von Klassenarbeiten und so die Basis für die Notengebung als weitaus zu gering. Sie investieren im Vergleich mit den anderen Kategorien übermässig viel Zeit in schulische Sonderformen[13]. Die Lehrkräfte der Volksschule und des Kindergarten sind dagegen der Auffassung, zu wenig Zeit für die Beurteilung der Schülerinnen und Schüler zur Verfügung zu haben. Die Mittelschullehrkräfte geben an, dass dafür die Zeit bei ihnen ausreicht, vermutlich weil *Beurteilung* und *Leistungs*beurteilung sehr angenähert sind. Ihr Problem sind der besondere Stoffdruck und die besondere Zeitknappheit angesichts hoher Erwartungen, die sich kaum je sehr präzise fassen lassen.

Es gibt in der Zürcher Lehrerschaft, sozusagen entgegen anderslautenden Meldungen, insgesamt keine grundsätzlichen Widerstände gegen die laufenden Schulreformen, aber die dafür zur Verfügung stehende Zeit wird deutlich als ungenügend angesehen. Burned-Out-Symptome lassen sich nachweisen, allerdings mit schwach niederigeren Werten als in vergleichbaren deutschen Untersuchungen. Besonders hohe Erschöpfungswerte zeigen Lehrpersonen der Real- und der Oberschule. Beklagt wird durchgehend eine „höhere Arbeitsüberforderung" (ebd., S., 95), während die Unzufriedenheit mit dem Lehrerberuf insgesamt nicht besonders ausgeprägt ist. Das mag auch damit zusammenhängen, dass Fremdkontrolle kaum erlebt wird. Interessant ist schliesslich auch noch, dass die Selbstwirksamkeitseinschätzung mit der Schulstufe abnimmt. Die mit Abstand höchsten Werte für Selbstwirkamkeit zeigen sich bei den Kindergärtnerinnen, also die Gruppe, die weder Korrekturen vornehmen muss noch Noten vergibt.

Einige andere Befunde, die sich nicht mit dem Problem von Prüfungen und Noten in Verbindung bringen lassen, aber Aufschlüsse für das Gesamtbild ergeben, sind:

1. Mit einer Ausnahme arbeiten vollangestellte Lehrkräfte auch im Kanton Zürich mehr, als vom Personalgesetz gefordert.
2. Die schulformspezifischen Differenzen in der Jahresarbeitszeit sind eher gering, die individuellen Unterschiede dagegen hoch. *Innerhalb* einer Schulform arbeiten die Lehrkräfte "höchst unterschiedlich lang".
3. Teilzeitarbeitskräfte arbeiten deutlich mehr, als ihr Anstellungsgrad verlangt.
4. Eine Reduktion der Pflichtlektionen führt nicht zu einer Reduktion der Arbeitszeit. Vielmehr verwenden die Lehrkräfte dann grössere Zeiteinheiten in die einzelnen Tätigkeiten.
5. Lehrkräfte in Leitungsfunktionen arbeiten in allen Schulformen signifikant mehr, eine höhere Belastung ist auch bei Lehrkräften feszustellen, die an der Schulentwicklung beteiligt sind.

[13] Arbeitswochen oder Projektwochen mit Lehrtätigkeit, Mitarbeit bei Projekten, Exkursionen, Schulreisen, Prüfungstätigkeit ohne Prüfungen in eigenen Klassen, Mitwirkung an Sporttagen u.ä.

Insgesamt lässt sich festhalten: Im Mittelpunkt der Arbeit der Lehrkräfte stehen die *unterrichtsbezogenen* Tätigkeiten. Für Lehrerinnen und Lehrer an Mittelschulen bedeutet das, Fächer mit wissenschaftlichem Anspruch zu unterrichten und darauf bezogen Leistungen zu bewerten. Sie haben einen weit höheren Aufwand im Bereich der Korrekturen, Prüfungen und Notengebung als andere Kategorien. Dafür steht deutlich zu wenig Zeit zur Verfügung, während der tatsächliche Arbeitsaufwand eher unter- als überschätzt wird. Alles, was den Aufwand steigert, ohne den Ertrag zu verbessern, wird in dieser Praxis keine Verwendung finden.

Das gilt auch und sozusagen spiegelbildlich für die andere Seite, die der Schülerinnen und Schüler. In der dritte Ehemaligenbefragung (2001) des Kantons Zürich, die insgesamt eine hohe Zufriedenheit ergeben hat[14], wird die „zeitliche Belastung durch die Schule“ zentral gewichtet, die kaum Raum lasse für die Entwicklung persönlicher Interessen (Befragung 2001, S. 22). Die Mittelschulen werden vornehmlich als Institutionen für die Vermittluing *kognitiver* Fähigkeiten wahrgenommen, wobei der Ausbildungsstand in den obligatorischen weit besser beurteilt wird als der in den fakultativen Fächern, was mit drei Faktoren zu tun hat:

- wesentlich kürzerer Lektionenumfang,
- weniger starke Wissensförderung
- und keine selektive Benotung
  (ebd., S. 18).

Was nicht selektiv benotet wird, erhält von den Schülern weniger Beachtung; was sich nicht auf Wissen bezieht, erhält einen geringeren Rang; was zeitlich nicht ausreichend bedient wird, erhält den Status von Nebentätigkeiten. Die „Erwartungen der Schülerschaft“ an ein Schulfach spielen eine erhebliche Rolle beim Zustandekommen der Leistung, wobei die Erwartungen sich wesentlichen darauf beziehen, wann ein Fach ernst zu nehmen ist und wann nicht. Offenbar spielen dabei Prüfungen und Noten ein zentrale Rolle, was in den erwähnten Studien so gut wie nie erfasst oder auch nur berücksichtigt worden ist.

Meine abschliessenden Ueberlegungen, die Leistungsbeurteilungen als *Chance* der Schulentwicklung verstehen, haben diese Voraussetzung. Sie gehen vom System aus und urteilen nicht abstrakt mit Idealnormen, wie fast immer in der bisherigen Diskussiond es Problems.

3. Leistungsbeurteilung und Schulentwicklung

Das historische Notensystem[15] ist überwiegend ein Fünferschema. Auffällig ist dieser Befund, weil auch in einem Sechsersystem faktisch nur fünf Noten genutzt werden. Der „Gothaer Schulmethodus” von 1685 sah zum ersten Male ein einheitliches Schema für die Unterrichtsgegenstände vor. Gleichzeitig sollten Ingenium und Mores, also die Geistesgaben und das sittliche Verhalten, der Schüler beurteilt werden. Für das Ingenium genügten vier Stufen, die beiden anderen Bereiche sollten mit dem Fünfersystem bewertet werden.

Ingenium	*Unterrichtsgegenstände*	*Mores*

[14] 87% der Ehemaligen äussern sich positiv zu ihrer Mittelschule (die kantonstätliche Vorgabe liegt bei 85%) (Befragung 2001, S. 10).
[15] PRINZ VON HOHENZOLLERN/LIEDTKE (1991).

Sehr fein	fein	fromm
Gut	fertig	fleissig
Ziemlich	ziemlich	still
Schlecht	etwas/wenig	unfleissig
	schlecht	ungehorsam

Offenbar ist dieses Schema, zeitgemäss verändert, geeignet, Leistungs- und Verhaltensdifferenzen innerhalb einer bestimmten, grösseren Lerngruppe zu beschreiben. Die vergleichsweise intensive Forschung hat bislang keine Alternativen geliefert, die praktikabler wären.

Ein allein auf Lehrerurteile gestütztes selektives Schulsystem gilt in der Literatur als „dysfunktional" (LANGFELDT/TENT 1999, S. 77), aber Vorschläge etwa derart, Zeugnisse als „entwicklungsbezogene Berichte" aufzufassen (ebd., S. 80) oder in grösserem Masse Leistungstests einzusetzen, haben sich bislang nicht durchgesetzt. Standardisierte Tests wie in England, die verknüpft sind mit Inspektoraten, erhöhen den statistischen Wert der Prüfungen, aber zugleich auch den Stress der Lehrkräfte (TRAVERS/COOPER 1996), die auf permanente Prüfungen hin, die nicht sie selbst bestimmen, unterrichten müssen. Andererseits helfen auch allgemeine Hinweise auf die gesellschaftlichen Risiken schlechter Noten nicht recht weiter (GRÜNIG u.a. 1999, S. 78ff.). Eher ist die globale Notenkritik selbst ein Belastungsfaktor für die Lehrkräfte, weil sie gesagt bekommen, was sie falsch machen, ohne zugleich gesagt zu bekommen, wie es unter den gegebenen Umständen besser gemacht werden kann.

Die meisten Vorschläge, die die Lehrkräfte als „Diagnostiker" (JÄGER 2000) aufwerten und ihnen zusätzliche Aufgaben aufbürden, erhöhen nur den Aufwand, ohne die reale Zeitverteilung in Rechnung zu stellen. Nach den vorliegenden Schweizer Daten konzentriert sich die Jahresarbeitszeit der Lehrkräfte mit durchschnittlich zwischen 80 und 90 Prozent auf die *unterrichtsbezogenen* Tätigkeiten. Den verbleibenden Rest einer stark gestressten Zeit müssen sich Betreuung und Beratung, Weiterbildung oder Gemeinschaftsarbeit und alles Uebrige teilen (LANDERT 1999). Es ist dann ziemlich grotesk, Listen mit allerlei diagnostischen Tätigkeiten zu lesen, die ungewichtet sind und die zeitlichen Belastungen unberührt lassen (JÄGER 2000, S. 101).

Das Grundproblem von Aufwand und Effekt ist nicht gelöst, zumal nicht in einem Berufsfeld, das vom individuellen Engagment lebt und sich in zeitlicher Hinsicht nicht standardisieren lässt. Wer die erhöhte Arbeitszeit künstlich belastet, gefährdet den Schulerfolg. Auf der anderen Seite müssen, ich wiederhole diesen zentralen Punkt, Reformvorschläge daraufhin überprüft werden, ob sie einfach nur die Belastungen erhöhen, ohne zugleich den Effekt zu verbessern. Wer, wie im Trendbericht der SKBF beschrieben (VÖGELI-MANTOVANI 1999, S. 78), gleichgewichtige Aussagen über

- Sozialform
- kriteriale Norm und
- Individualnorm

machen soll, muss über viel zusätzliche Zeit, geeignete Instrumente und viel Geduld bei der Auswertung von aufwendig erhobenen Daten verfügen, ohne dass gewiss wäre, ein besseres Resultat zu erzielen.

Chancen für die Schulentwicklung ergeben sich erst dann, wenn nicht der ideale Notengeber vor Augen steht, den die Kritik zugleich voraussetzt und ablehnt, sondern wenn die Bedingungen des Praxisfeldes kalkuliert werden. Versteht man unter „Noten“ die Beschreibung von Leistungen im Blick auf Aufgaben[16] während einer bestimmten Periode der Beurteilung, dann stellen sich eigentlich nur drei wirkliche Probleme:

1. Wie *vergleichbar* sind die Aufgaben?
2. Wie *transparent* sind die Beurteilungen?
3. Wie berechnet sich die *Lernzeit* und so die Chance, die Aufgabe zu lösen?

An der Kritik der Zensurengebung ist der Hinweis gerechtfertigt, dass Noten sich auf ein „klasseninternes Bezugssystem“ beziehen und aber Allgemeingültigkeit beanspruchen. Die erwähnte „Fünfeinhalb“ in Französisch kann an verschiedenen Orten verschiedenes bedeuten, ohne auf eine annähernd einheitliche Qualität schliessen zu lassen, die aber bei der Verwendung der Noten vorausgesetzt ist. Das Problem lässt sich nie ganz auflösen, wohl aber reduzieren, indem fachliche *Standards* und damit verbunden *Leistungsniveaus* eingeführt werden, die die Notengebung bestimmen (OELKERS 2001).

Bis heute beziehen sich Noten nicht auf festgelegte Inhalte, sondern zumeist auf die Stoffverteilung im Programm einer bestimmten Lehrkraft. Das ist gesetzeskonform, denn mindestens im Promotionsreglement für die Gymnasien des Kantons Zürich ist davon keine Rede. Aber es wäre ein Stück Schulentwicklung, die Inhalte anzupassen und Aufgaben vergleichbar zu bestimmen, denn immerhin verhelfen die Noten zu gleichen Berechtigungen. Im Promotionsreglement wird bestimmt, dass die Lehrperson die Klasse „über die Art der Leistungsbeurteilung“ im Fach informiert (Promotionsreglement § 7, 2). Transparent sind Beurteilungen aber nur dann, wenn zugleich das Lernpensum kalkuliert werden kann, und dies vor Beginn der Beurteilungsperiode. Die Fächer müssten die Anforderungen an die Schüler abstimmen, die dann individuell ihren eigenen Lernaufwand, von dem der Lernerfolg massgeblich abhängig ist, abklären und einteilen könnten.

Nach unseren Daten geschieht das heute zumeist nicht. Die Schüler kennen nicht nur die lehrerspezifischen Beurteilungsstrategien nicht, ausgenommen, was sie davon im Unterricht erleben, sie wissen vor allem nicht, was im Beurteilungszeitraum genau auf sie zukommt. Hier wären transparente Lernprogramme mit möglichst realistischer Zeiterwartung zu entwickeln und zwischen den Schulen vergleichbaren Typus abzustimmen. Das ist *nicht* Praxis und verlangt zum Teil aufwändige Zusatztätigkeit, die abgekürzt werden kann, wenn externe Experten mit Reviews und Angleichung einzelner Programme beauftragt werden.

Die Notenpraxis wird vermutlich auf absehbare Zeit das Fünferschema nicht überwinden. Es kommt eher darauf an, möglichst präzise zu fassen, was Noten in bestimmten Fächern im Blick auf welche Standards genau beschreiben sollen. Auch das fehlt überwiegend. Die Praxis der Notengebung ist eine der einzelnen Lehrkräfte, die die Noten nach eigenem Urteil festlegen, ohne sich in der Regel mit Anderen auszutauschen. Gelegentlich wird vorgeschlagen, Notenagenturen ausserhalb der Schule damit zu beauftragen, aber vermutlich verbessern solche Agenturen das Resultat *nicht*, weil Noten sich auf *fortlaufende* Leistungen beziehen, die am Ende unter dem Eindruck der Beurteilungsperiode gesamthaft bewertet werden, wozu nur die verantwortliche Lehrkraft wirklich imstande ist.

[16] *Tasks and achievements*: PETERS (1967).

Ein Resultat der Forschung war, dass Lehrkräfte dieses Geschäft recht gut beherrschen, so dass hier auch die Entwicklung einsetzen muss. Daher ist nicht die „implizite Persönlichkeitstheorie“ das Problem, die nicht zu vermeiden ist und aber vermutlich mehr Vorteile hat, als die Forschung zugesteht; vielmehr wird es darum gehen, das Engagement des Notengebers und seine Kompetenz zu nutzen und zugleich die Gefahr zu grosser Streuungen zu begrenzen. Das gelingt am besten durch klar formulierte Standards oder Bezugsnormen, und zwar sowohl der Inhalte als auch der Noten selbst. Vor allem hier ist Entwicklung nötig, die abrückt vom einfachen Gegensatz: „Noten oder nicht?“

Die gegenwärtigen Tendenzen auf Volksschul- und Sekundarschulstufe zeigen im übrigen, dass das Notenprinzip längst nicht mehr in der Absolutheit gilt, die die Kritik unterstellt (VÖGELI-MANTOVANI 1999, S. 89ff.). Inzwischen gibt es nicht nur „Noten“,

- sondern Noten mit und ohne explizite Bezugsnormen,
- Lernberichte,
- fakultative wie nicht-fakultative Beurteilungsgespräche,
- Orientierungsarbeiten zur Standortbestimmung,
- Selbstbeurteilungen der Schülerinnen und Schüler,
- Zeugnisse mit lernzielbezogenen Wortetiketten,
- Zeugnisse mit lernzielbezogenen Wortetiketten für Beurteilung des Lernprozesses und der Leistung.

Im Blick auf die Kantone zeigen sich grosse Unterschiede: In Basel gibt es „prognostische Noten“ erst ab der sechsten Klasse, reguläre Noten erst nach dem Uebertritt in der achten Klasse. In Baselland gibt es reguläre Noten ab der sechsten Klasse, zuvor wahlweise Noten oder Lernberichte. Im Aargau gibt es reguläre Noten ab dem zweiten Beurteilungszeitraum der ersten Klasse, in Bern werden bis zur sechsten Klasse Beurteilungsgespräche geführt, Lernberichte erstellt und in der dritten sowie sechsten Klasse lernzielorientierte Noten erteilt. In Solothurn sind die ersten drei Jahrgänge notenfrei, in Freiburg gibt es Verbalzeugnisse, Lernberichte und Beurteilungsgespräche von der ersten Klasse an - Diese Entwicklung zeigt, dass in Zukunft eher die Wahrung des Notenschemas das Problem ist, weil Teile der Schulreformdiskussion die Meinung von KARLHEINZ INGENKAMP übernommen hat, wonach Noten und Zeugnisse *an sich* höchst fragwürdige Phänomene seien und daher abgeschafft oder ersetzt werden müssten.

Der Ersatz ist immer eine Reduktion, meistens von Leistungsanforderungen, die mit dem Verzicht auf Noten ihre Klarheit verlieren und mit endlosen Differenzierungen immer neue Aequivalente erzeugen. Letztlich geht es bei dieser Entwicklung um die Frage, ob eine selektive Leistungsschule gewollt ist oder nicht. Ist sie gewollt, verlangt sie möglichst eindeutige, zugleich faire Systeme der Beschreibung und Beurteilung von Leistungen, also weder des Charakters noch der Person, sondern im Blick auf das, was Lehrkräfte offenbar gut können, nämlich die Verteilung der individuellen Leistungsfähigkeit im Blick auf ihre Klasse, vorausgesetzt Bezugsnormen, die den Rahmen einer bestimmten Klassen übersteigen, also inhaltliche Standards des Faches darstellen. Dass Urteile fehlbar sind, ist trivial; aber offenbar ist das nicht das Problem. Vielmehr wäre zu fragen, wie die Praxis der Notengebung organisiert ist und wie sie verbessert werden kann. In diesem Sinne ist Notengebung eine Chance der Schulentwicklung. Das gilt insbesondere dann, wenn die Urteile objektive Unterstützung erfahren und sich auf anerkannte Fachnormen beziehen können.

Literatur

Befragung ehemaliger Mittelschülerinnen und Mittelschüler. Erstellt vom Statistischen Amt des Kantos Zürich. Zürch 2001.
BOBERTAG, O.: Leistungsschätzung und Leistungsmessung in der Volksschule. In: Zeitschrift für Pädagogische Psychologie 34 (1933), S. 377-393.
DE GROOT, A.D.: Fünfen und Sechsen. Weinheim 1971.
DÖRING, O.: Untersuchungen zur Psychologie des Lehrers. Leipzig 1925.
FLAMMER, A.: Zur Definition der Notenskala. In: Schweizerische Zeitschrift für Psychologie 30 (1971), S. 204-218.
FORNECK, H.J./SCHRIEVER, F.: Die individualisierte Profession. Untersuchung der Lehrer/ -innenarbeitszeit und -belastung im Kanton Zürich. Ms. Zürich 2000.
GREY, C.F.: Variations in the Grades of High School Pupils. Baltimore 1913.
GRÜNIG B. u.a.: Leistung und Kontrolle. Die Entwicklung von Zensurengebung und Leistungsmessung in der Schule. Weinheim/München 1999. (= Erziehung im Wandel. Hrsg.v. H. RAUSCHENBERGER, Bd. 4)
HAECKER, H.: Subjektive Faktoren im Leistungsurteil der Lehrer. In: Schule und Psychologie 18 (1971), S. 74-84.
HOFER, M.: Die Validität der impliziten Persönlichkeitstheorie von Lehrern. In: Unterrichtswissenschaft Heft 2 (1975), S. 5-18.
HOPP, A.-D./LIENBERT, G.A.: Eine Verteilungsanalyse von Gymnasialzensuren. In: KH. INGENKAMP (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte. 6., überarb. u. erw. Aufl. Weinheim/Basel 1976, S. 250-263.
HUBERMAN, M.: Das Selbstkonzept. Eine Untersuchung über die Wirkung von Noten, Ranglisten und Preisen auf Kinder der Genfer Primarschule. Genève: FAPSE 1980.
INGENKAMP, KH.: Zur Problematik der Auslese und ihrer Bewährungskontrolle. In: KH. INGENKAMP (Hrsg.): Pädagogisch-psychologische Untersuchungen zum Uebergang auf weiterführende Schulen. Weinheim 1963, S. 7-54.
INGENKAMP, KH. (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte. 6., überarb. u. erw. Aufl. Weinheim/Basel 1976. (erste Aufl. 1971)
JÄGER, R.S.: Von der Beobachtung zur Notengebung. Ein Lehrbuch. Diagnostik und Benotung in der Aus-, Fort- und Weiterbildung. M. e. Beitrag v. U. LISSMANN. Landau 2000.
KRAMPEN, G.: Effekte von Lehrerkommentaren zu Noten bei Schülern. In: R. OLECHOWSKI/E. PERSY (Hrsg.): Fördernde Leistungsbeurteilung. Wien/München 1987, S. 297-227.
LANGFELDT, H.P./TENT, L.: Pädagogisch-psychologische Diagnostik. Band 2: Anwendungsbereiche und Praxisfelder.Göttingen/Bern/Toronto/Seattle 1999.
LANDERT, CH.: Die Arbeitszeit von Lehrpersonen in der Deutschschweiz. Zürich 1999.
Mummert+Partner: Untersuchung zur Ermittlung, Bewertung und Bemessung der Arbeitszeit der Lehrerinnen und Lehrer im Land Nordrhein-Westfalen. Bd. I/II. Düsseldorf 1999.
MOSER, U./RHYN, H.: Lernerfolg in der Primarschule. Eine Evaluation der Leistungen am Ende der Primarschule. Aarau 2000.
OELKERS, J.: Was und wie sollen Jugendliche im Jahr 2006 auf der Sekundarstufe II lernen? Expertise zuhanden des Mittelschul- und Berufsbildungsamtes der Bildungsdirektion des Kantons Zürich. Zürich 2001.
PETERS, R.S.: What is an Educational Process? In: R.S. PETERS (Ed.): The Concept of Education. London: Routledge&Kegan Paul 1967, S. 1-23.

PRINZ VON HOHENZOLLERN, J.G./LIEDTKE, M. (Hrsg.): Schülerbeurteilungen und Schulzeugnisse. Historische und systematische Aspekte. Bad Heilbrunn/Obb. 1991. (= Schriftenreihe zum Bayerischen Schulmuseum Ichenhausen, hrsg. v. Bayerischen Nationalmuseum, Bd. 10)
Promotionsreglement für die Gymnasien des Kantons Zürich vom 10. März 1998.
RANK, TH.: Schulleistung und Persönlichkeit. München 1962.
RHEINBERG, F.: Leistungsmessung und Lernmotivation. Göttingen 1980.
RICHTER, S.: Unterschiede in den Schulleistungen von Mädchen und Jungen. Geschlechtspezifische Aspekte des Schriftsprachenerwerbs und ihre Berücksichtigung im Unterricht. Regensburg 1996.
ROSENTHAL, R./JACOBSON, L.: Pygmalion im Unterricht. Lehrererwartungen und Intelligenzentwicklung der Schüler. Weinheim 1971.
SACHER, W.: Praxis der Notengebung. Hilfen für den Schulalltag. Bad Heilbrunn/Obb. 1984.
SAUER, J./GAMSJÄGER, E.: Ist Schulerfolg vorhersehbar? Die Determinanten der Grundschulleistung und ihr prognostischer Wert für den Sekundarschulerfolg. Göttingen/Bern/Toronto/Seattle 1996.
SCHÖNWÄLDER, H.-G./PLUM, W.: Pädagogische Arbeit der Lehrer und Lehrerinnen – terra incognita der Bildiungspolitik! Bericht über eine Expert(innen)-befragung in Nordrhein-Westfalen, Bremen, Münster. Ms. GEW Nordrhein-Westfalen. Düsseldorf 1998.
SCHRADER, F.-W.: Lern- und Leistungsdiagnostik im Unterricht. In: F. E. WEINERT (Hrsg.): Psychologie des Unterrichts und der Schule. Göttingen/Bern/Toronto/Seattle 1997, S. 659-699.
SCHREIBER, H.: Gegen Prüfugen und Noten. In: Zeitschrift für Philosophie und Pädagogik 6 (1899), S. 31-38.
TENT, L.: Die Auslese von Schülern für weiterführende Schulen. Göttingen 1969.
TRAVERS, C.J./COOPER, C.L.: Teachers under Pressure. Stress in the Teaching Profession. London/New York: Routledge 1996.
ULSHÖFER, R.: Zur Beurteilung von Reifeprüfungsaufsätzen. In: Der Deutschunterricht 1 (1949), S. 84-102.
ULBRICHT, H.: Wortgutachten auf dem Prüfstand. Eine empirische Untersuchung zur verbalen Beurteilung in der 1. und 2. Klasse der Grundschule mittels Elternbefragung und Zeugnisanalyse. Münster/New York 1993.
VÖGELI-MANTOVANI, U.: Mehr fördern, weniger auslesen. Zur Entwicklung der schulischen Beurteilung in der Schweiz. Aarau 1999. (= SKBF Trendbericht, Nr. 3)
WEIDIG, E.R.: Die Bewertung von Schülerleistungen. Weinheim 1961.
WEINERT, F.E. (Hrsg.): Leistungsmessungen in Schulen. Weinheim/Basel 2001.
WEISE, G.: Leistungsmessung. In: J. PETERSEN/G.-B. REINERT (Hrsg.): Pädagogische Positionen. Ein Leitfaden für Lehrer aller Schulen. 2. Aufl. Donauwörth 1990, S. 216-230.
WEISS, R.: Zensur und Zeugnis. Beiträge zu einer Kritik der Zuverlässigkeit und Zweckmässigkeit der Ziffernbenotung. Linz 1965. (= Wissenschaftliche Veröffentlichungen des Bundes in Oberösterreich, Bd. 3)
WOLF, K.: Die Gerechtigkeit des Erziehers. München 1962.
ZIEGENSPECK, J.: Handbuch Zensur und Zeugnis in der Schule. Historischer Rückblick, allgemeine Problematik, empirische Befunde und bildungspolitische Implikationen. Ein Studien- und Arbeitsbuch. Bad Heilbrunn/Obb. 1999.