Warum haben Noten als Feedback eine begrenzte Aussagekraft?

Henriette Klotzsche & Slavka Zanev

Die Leistungsbeurteilung in unserer Gesellschaft ist für uns aufs engste verknüpft mit dem mathematischen Ausdruck der Note. Die pädagogische Diagnostik beschäftigt sich unter anderem eingehend mit dem Problem der Zensurengebung. Zahlreiche Studien und Untersuchungen sind im Zusammenhang mit dieser Problematik entstanden. Das Handout fasst wichtige Bestandsaufnahmen zusammen und unternimmt den Versuch zu einer weiterführenden Diskussion anzuregen.

Ein Versuch einer Definition

Zensieren bedeutet, die Leistung der Schüler oder Studenten in eine quantifizierbare Form zu bringen. Diese Form erlaubt es, Leistungen zu vergleichen und einzustufen.

Im Folgenden wird versucht zu erläutern, welche Kriterien an die Benotung gestellt werden. Das Instrument der Benotung sollte jedoch in diesem Rahmen verstanden werden und kann auch nur in diesem wirken. Für die Entwicklung der Schüler und Studenten kommt man aber nicht an qualitativen Bewertungen vorbei, welche ergänzend zu den Noten hinzugezogen werden müssen.

Die Notenskala

Die Vereinbarung der Kultusministerkonferenz der Länder vom 3.10.1968 über die Bewertung von Schülerleistungen hat noch heute ihre unveränderte Gültigkeit:

Auf Beschluss der Kultusministerkonferenz vom 3. 10. 1968 sind den Noten die folgenden, bis heute gültigen Definitionen zugrunde gelegt worden (vgl. §25 der AschO):	
sehr gut (1)	Die Note "sehr gut" soll erteilt werden, wenn die Leistung den Anforderungen in besonderem Maße entspricht.
gut (2)	Die Note "gut" soll erteilt werden, wenn die Leistung den Anforderungen voll entspricht.
befriedigend (3)	Die Note "befriedigend" soll erteilt werden, wenn die Leistung im allgemeinen den Anforderungen entspricht.
ausreichend (4)	Die Note "ausreichend" soll erteilt werden, wenn die Leistung zwar Mängel

	aufweist, aber im ganzen den Anforderungen noch entspricht.
mangelhaft (5)	Die Note "mangelhaft" soll erteilt werden, wenn die Leistung den Anforderungen nicht entspricht, jedoch erkennen lässt, dass die notwendigen Grundkenntnisse vorhanden sind und die Mängel in absehbarer Zeit behoben werden können.
ungenügend (6)	Die Note "ungenügend" soll erteilt werden, wenn die Leistung den Anforderungen nicht entspricht und selbst die Grundkenntnisse so lückenhaft sind, dass die Mängel in absehbarer Zeit nicht behoben werden können.

„Bei der Bewertung von Schülerleistungen ist der Eigenart der Schulgattung und des Faches sowie dem Alter des Schülers Rechnung zu tragen. Der Begriff Anforderungen bezieht sich auf den Umfang sowie auf die selbständige und richtige Anwendung der Kenntnis und auf die Art der Darstellung."

Die Erläuterungen der Notenstufen sowie der Begriff der Anforderungen bleiben der subjektiven Interpretation überlassen.

Empirische Untersuchung zur Funktionswahrnehmung von Noten

In einer empirischen Untersuchung von 1978/79, welche von Krampen durchgeführt wurde, sollte festgestellt werden, welche Funktionen schulische Zensuren in den Augen von Lehrern, Schülern und Lehramtskandidaten haben und inwieweit sich die Funktionswahrnehmung dieser Gruppen unterscheiden. Den Probanten wurde die nachstehende Liste der Funktionswahrnehmungen (nach Zielinski 1974) vorgelegt und die Aufgabe erteilt, sie nach Wichtigkeit zu ordnen.

- **Rückmeldefunktion für den Lehrer über seinen Unterrichtserfolg**
- → *am höchsten von Lehrern und am geringsten von Studenten bewertet*
- **Rückmeldefunktion für den Schüler über seinen Leistungstand**
- → *von Schülern am geringsten, von Studenten am höchsten bewertet*
- **Berichtsfunktion für die Eltern**
- → *von Schülern am höchsten und von Studenten am geringsten bewertet*
- **Anreizfunktion für den Schüler** (Motivation zur Arbeit)
- **Disziplinarfunktion** (Belohnung oder Bestrafung)
- **Sozialisierungsfunktion** (der Schüler lernt durch Noten die Leistungsstandards der Gesellschaft kennen)
- → *von Studenten und Lehrern höher bewertet als von Schülern*
- **Klassifikationsfunktion** (zur inneren Differenzierung, z. B. zur Gruppierung von Schülern in Leistungskurse)
- → *von Lehrern höher bewertet als von Schülern und Studenten*
- **Selektionsfunktion** (zur äußeren Differenzierung, z. B. Auswahl für weiterführende Schulen)
- → *von Schülern am höchsten, von Lehrern am geringsten bewertet*

- **Zuteilungsfunktion** (Chancen für bestimmte Berufe)
- → *von Schülern und Studenten höher bewertet als von Lehrern*
- **Chancenausgleichsfunktion** (dann, wenn Noten dazu dienen die Chancengleichheit in der Schule zu optimieren)

Eine Studie zum Einfluss der Bewertungsform auf die schulische Entwicklung von Schülern

Ziel der Längsschnittstudie NOVARA („Noten oder Verbalbeurteilung? Akzeptanz, Realisierung, Auswirkungen"), welche durch die Humboldt–Universität Berlin durchgeführt wurde, war es, die Auswirkungen unterschiedlicher Bewertungsform auf die Leistungs- und Persönlichkeitsentwicklung von Grundschülern der 2. – 4. Klassenstufe zu untersuchen. Dazu wurden Schüler aus 41 Klassen von 23 Schulen in 15 Berliner Bezirken befragt.
Die Stichprobe beinhaltete
„20 Klassen aus dem Ostteil, von denen am Beginn der Untersuchung im 2. Schuljahr elf verbal beurteilt wurden, und 21 Klassen aus dem Westteil der Stadt, unter ihnen 15 mit verbaler Beurteilung in der 2. Klasse" (Wagner & Valtin. 2003)
Laut Wagner und Valtin fand im Verlauf der Untersuchung ein starker Rückgang der Klassen, welche verbal beurteilt wurden, statt. Dies entspräche jedoch dem allgemeinen Trend an Grundschulen. Im Folgenden werden für das Thema relevante Ergebnisse dargestellt, andere so z. B. regionale Unterschiede oder Unterschiede zwischen Mädchen und Jungen werden außer Acht gelassen.

Ergebnisse der Studie

1. Allgemeine und fächerspezifische Lernfreude:

- Zum ersten MZP (Messzeitpunkt) konnten keine Unterschiede zwischen den verbal beurteilten und den Notenkindern festgestellt werden.
- Von der 2. zur 4. Klasse verringerte sich die allgemeine Lernfreude bei beiden Gruppen.

2. Die Fähigkeitsselbstbilder in den verschiedenen Lernbereichen

- Von der Mitte der 2. Klasse an war das Selbstkonzept der Notenkinder höher als das der verbal beurteilten Kinder
- Bei den leistungsstärkeren Kindern hatten die verbal bewerteten durchweg niedrigere Selbstkonzepte als die Notenkinder
- Die Selbstkonzepte der leistungsschwächeren Notenkinder und der verbal beurteilten Kinder näherten sich mit zunehmendem Alter an.

3. Leistungsmotivation

- Zu Beginn der 2. klasse war die „Hoffnung auf Erfolg" stärker ausgeprägt als die „Furcht vor Misserfolg" – Es wurden keine Unterschiede zwischen Notenkindern und verbal beurteilten gefunden.

4. Schulleistungen

- In Schulleistungsteste wurden nur geringe Unterschiede zwischen Kindern mit Notengebung und Verbalbeurteilung gefunden.

Schlussfolgerungen

Durch die Studie konnte nur zu einem geringen Teil eine beeinträchtigende Wirkung der Notengebung nachweisen. Es kann geschlussfolgert werden, dass die Einführung der verbalen Beurteilung nicht automatisch die Mängel der Zensurengebung beseitigen kann. Pekrun hat in einer Stellungnahme zur Studie dazu geäußert:

„Ob Leistungsurteile anhand von Zahlen oder von Worten ausgedrückt werden […] ist offenkundig von ganz und gar oberflächlicher Art. Andere Merkmale von Leistungsurteilen […] sind bei weitem wichtiger: die individuelle und öffentliche Sichtbarmachung eines Urteils, der ihm zugrunde liegende Urteilsmaßstab, seine objektive und subjektive Valenz […] etc." (Pekrun. In: Wagner & Valentin 2003)

Der prognostische Wert von Noten (nach Zielinski 1974)

In Hinsicht auf das Problem einer prognostischen Gültigkeit von Zensuren trifft Zielinski folgende Aussagen:

- die Zusammenhänge zwischen Zensur und Schulstufe sind bei aufeinanderfolgenden Jahrgängen weitestgehend zufriedenstellend hoch
- mit zunehmender zeitlicher Distanz ist jedoch eine laufende Abnahme zu verzeichnen
- besonders gravierend wirken sich Wechsel in andere Schulsysteme aus

Infolge von Untersuchungen ließ sich feststellen, dass der Zusammenhang zwischen Grundschulzeugnis und dem Erfolg auf weiterführenden Schulen nach drei bis sechs Jahren als auf keinen Fall übereinstimmend zeigte. Durch Grundschulzeugnisse kann mit nur 10 % Sicherheit das zukünftige Leistungsverhalten des Schülers vorausgesagt werden. Bezogen auf die oben genannten Funktionen von Zensuren ist somit die Eignung dieses Zeugnisses als Auslesefunktion für den Übertritt in weiterführende Schulen hinfällig. So auch zeigen Abiturzeugnisse eine nur unzureichende prognostische Gültigkeit für Studien- oder Berufserfolg. In der Diskussion um den Numerus- Clausus für Studienfächer stellt das Abiturzeugnis also einen unzureichenden Wert dar und die somit verbundene Selektionsfunktion wird ebenfalls hinfällig.
Der prognostische Wert von Zensuren ist als unzureichend zu betrachten insofern sie über den Klassenrahmen oder das Schulsystem hinauszielen.

Fachspezifische Probleme der Zensierung

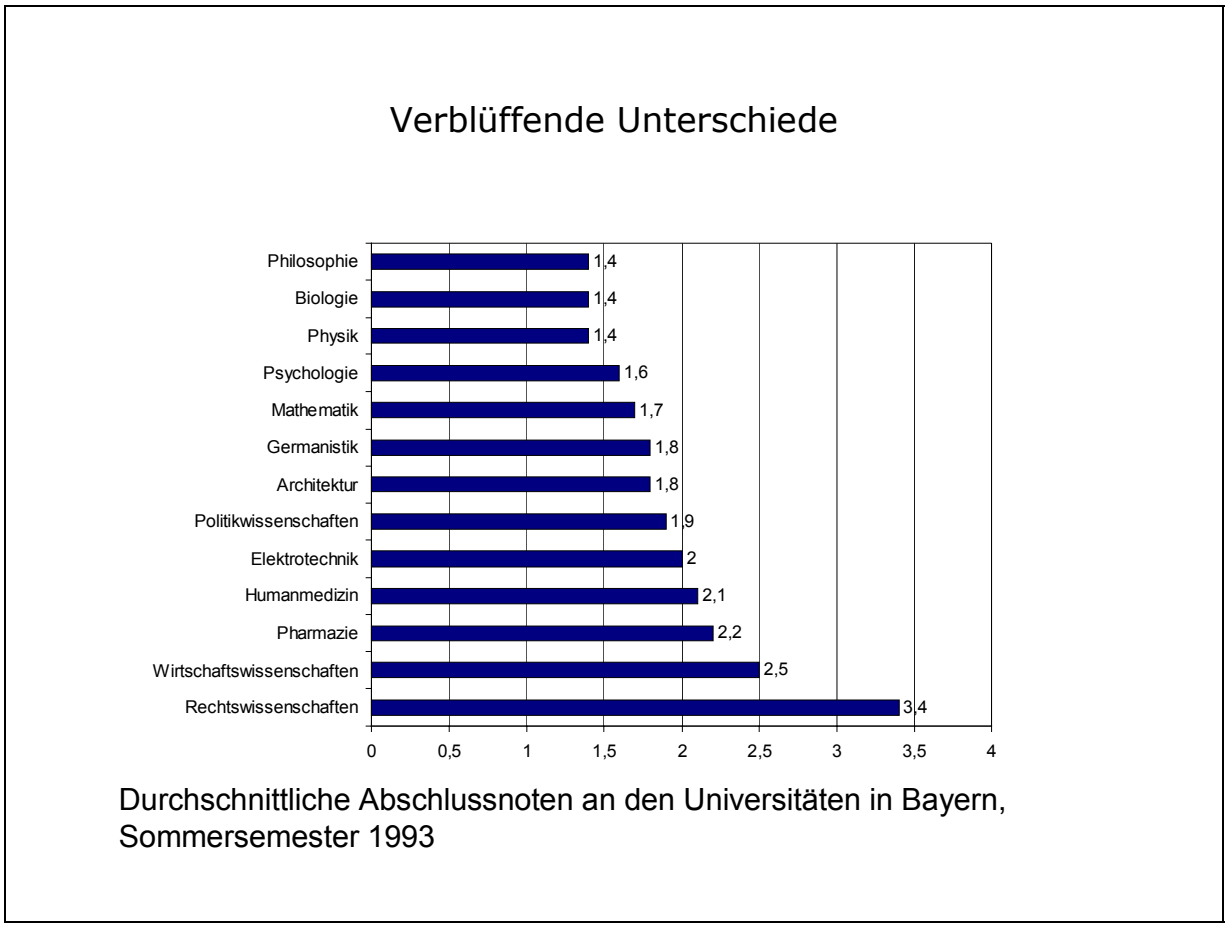
Nach der oben genannten Definition der Noten des Kultusminsteriums von 1968 ergibt sich ein Problem mit dem erläuterten Begriff der „Anforderung". Es fehlt hier eine genaue Aufschlüsselung der jeweiligen Anforderungen in einem Fach, denn man befand sich in der Annahme, dass sich die Notenverteilung in verschiedenen Fächern und Klassenstufen nicht unterscheiden dürften. Die Maßstäbe und Anforderungen jedoch, die dann tatsächlich gestellt wurden, differieren.

Nach zahlreich ausgewerteten Untersuchungen (Ingenkamp 1971) steht fest, dass in verschiedenen Fächern mit unterschiedlicher Strenge zensiert wird. Unterschieden wird dabei in folgende **Gruppen von Fächern**:

- **Haltungsfächer**
- **Leistungsfächer**
- **Beteiligungs- und Gestaltungsfächer**
- **musische Fächer**

- **Haupt- oder Nebenfächer.**

Für die **Haltungsfächer** (Betragen, Mitarbeit) sowie für Religion, Musik und Sport wurden die mildesten Durchschnittszensuren gefunden. Die **Leistungs-** und **Hauptfächer**, zumeist Deutsch, Mathematik und Fremdsprachen, wurden am strengsten zensiert. Daraus ergibt sich eine Klassifizierung der Fächer, die somit eine unterschiedlich stark ausgeprägte Auslesefunktion zeigen. Als möglichen Grund dafür führen die Autoren an, dass die Tendenz vorherrscht, Fächer strenger zu zensieren, wenn die Leistungen vorwiegend schriftlich geprüft werden. Die sehr unterschiedlichen Maßstäbe in den einzelnen Fächern lassen sich jedoch nicht befriedigend begründen.
Das nachstehende Diagramm zeigt diese Gegebenheit an universitären Fächern auf.



Anforderung an die Messqualität

Aufgrund der Berechtigungsfunktion von Zeugnissen ist vor allem die Zuverlässigkeit der Beurteilungen gefordert. Es ist erwiesen, dass Ziffernzensuren nur scheinbar zuverlässig sind. Objektivität und Zuverlässigkeit können aber nur ü-

ber Messwerte hergestellt werden. Die Zensur kann also kein exaktes Messergebnis sein, sondern sie ist eine Klassifikation.
Schulnoten sind ordinal skaliert. Das heißt, dass Skalenniveau erlaubt keine Rückschlüsse über die verhältnismäßige Wertigkeit zueinander. Das ordinale Skalenniveau bringt die Werte in eine Reihenfolge und es kann höchstens der Median aus diesen Werten gebildet werden. Die Berechnung des arithmetischen Mittels ist unzulässig (obwohl es in der Praxis häufig gemacht wird). Die Aussagekraft und somit das Feedbackpotential ist beschränkt. Die Wertigkeit der Noten obliegt der Fachkonferenzen der Schulen und ist auch von Sekundarstufe I zu Sekundarstufe II unterschiedlich. So ist der Verlauf von 15 bis 4 Punkte zum Beispiel in der Sek II in den Naturwissenschaften in Brandenburg linear (5% pro Punkt). Es handelt sich jedoch um keinen linearen Verlauf über das ganze Notenspektrum, welcher eine metrische Skalierung nach sich ziehen würde.
Die Aufspreizung der Notenskala 1-6 in eine 15 Punkte- Skala verweist auf einen guten Ansatz in Hinsicht auf die Vergleichbarkeit von Noten.

Diskussion: Für und wider die Ziffernzensur

Es wird auch vielfach die Meinung vertreten, dass Noten oft zu Unrecht kritisiert werden. Noten sind eine Mitteilungsform schulischer Leistungsergebnisse. Für Noten ist es kennzeichnend, dass sie (mit unterschiedlicher Stufenfolge und -definition) schon seit dem 16./17. Jahrhundert im Schulwesen gebräuchlich sind.

- Ihre jetzige Stufenfolge (von "sehr gut" bis "ungenügend") gilt seit 1968, ist jedermann bekannt und von der Öffentlichkeit weitgehend akzeptiert.
- Sie stellen ein relativ eindeutiges und unmissverständliches Urteil dar, weil ihre Interpretation durch die amtliche Notendefinition festgelegt ist.
- Sie eignen sich in besonderer Weise zur Zusammenfassung der Einzelbeobachtungen und -urteile des Lehrers, da sie knapp und prägnant die jeweilige Schülerleistung darstellen.

Aber:

- Zensuren implizieren keine Aussagekraft, da nachweislich die Vergleichbarkeit von Noten nicht gegeben ist.
- Auf die Probleme der Benotung wurde schon ansatzweise reagiert, zum Beispiel mit der Einführung der 15er Skala und mit einer Linearisierung der Notenverläufe um dem metrischen Ideal näher zu kommen.
- Trotz aller mathematischen Verbesserungen bleibt die Leistung der Benotung doch hauptsächlich in den Händen der Lehrer, welche nie objektiv handeln können. Noten haben symbolischen Charakter, der zur Entwicklung der Schüler und Studenten beitragen soll.

Literatur:

Wagner, C. & Valtin, R. (2003) Noten oder Verbalbeurteilungen? ZEPP, 35 (1), 27 – 36.

Ingenkamp, K. (1971) Die Fragwürdigkeit der Zensurengebung.

Ingenkamp, K. (1985) Lehrbuch der Pädagodischen Diagnostik.