

Wenn die Schulklasse die Note mitbestimmt

Über die Bedeutung des Referenzgruppenfehlers bei Karlheinz Ingenkamp

Klassiker

Vor über **50 Jahren** veröffentlichte Karlheinz Ingenkamp eine Studie über die Vergleichbarkeit von Schulnoten (Textauszug in alter Rechtschreibung, s. Seite 27–29).

Die Ergebnisse waren durchaus spektakulär. Denn die Daten legten nahe, dass Noten ihre Gültigkeit verlieren, sobald sie das Klassenzimmer verlassen. Der schulische Alltag geht von der stillschweigenden Annahme aus, dass Noten eine universelle Gültigkeit besitzen. Die zumeist gleiche Notenskala suggeriert, dass Noten an einem bestimmten Schulort das Gleiche bedeuten wie im Nachbarort, dass sie in der einen Schulklasse das Gleiche bedeuten wie in der anderen. Ingenkamp kann aber mit bestürzender Deutlichkeit zeigen, dass das nicht der Fall ist.

Seine Daten sind ein Paukenschlag. Denn sie entziehen den Schulnoten ihre vermeintlich sichere Grundlage. Die Folgen, etwa bei Versetzungs- oder Übertrittsgesprächen, sind unabsehbar. Wenn Noten überall etwas anderes bedeuten können, wird damit die Basis schulischer Selektionsentscheidungen brüchig. Letztlich ist die individuelle Bildungsbiografie davon abhängig, in welchen Schulklassen man gesessen hat.

Die Beobachtungen von Ingenkamp sind eine Zäsur in der Geschichte der Leistungsbeurteilung. Um die möglichen Verunsicherungen nachzeichnen zu können, die das skizzierte Problem auslöst, wird zunächst die damalige Publikation in leicht gekürzter Fassung wiedergegeben. Danach soll die Frage erörtert werden, ob sich in den letzten 50 Jahren etwas geändert hat.

Karlheinz Ingenkamp

Sind Zensuren aus verschiedenen Klassen vergleichbar?

Die Vergleichbarkeit der in verschiedenen Klassen des gleichen Jahrganges erteilten Zensuren wird in unserem Jahrgangsklassensystem, das Schüler nach gleichen Lehrplänen und gleichen Stundentafeln unterrichtet, prinzipiell vorausgesetzt. Praktische Konsequenzen werden ständig daraus gezogen, wenn z. B. Schüler mit einem bestimmten Notendurchschnitt von Aufnahmeprüfungen für das Gymnasium dispensiert, Abiturienten bei Zulassungsbeschränkungen nur immatrikuliert werden, wenn sie in einigen oder allen Fächern bestimmte Abiturnoten aufweisen und eine Behörde oder Firma bei starkem Bewerbungsandrang Kandidaten mit bestimmten Zensuren nicht einstellt. Sehr häufig ist die Zensur bei derartigen Entscheidungen das einzige Leistungsmaß, das nicht durch andere Methoden überprüft wird. Die Vergleichbarkeit der Zensuren wird – abgesehen von geringeren Abweichungen – auch in der Praxis vorausgesetzt, denn sonst wäre unserem gesamten Berechtigungs- und Versetzungswesen die Basis entzogen. Es gibt in Deutschland zwar Autoren, die ernsthafte Zweifel an der Vergleichbarkeit der Zensuren äußern, aber nur einige Untersuchungen, die bei kleineren Stichproben erhebliche Notendiskrepanzen in verschiedenen Klassen bei gleicher Leistung feststellten.¹

Wir stießen auf dieses Problem, als wir bei allen 6. Klassen eines Berliner Bezirks Testergebnisse und Zensuren vergleichend auswerteten.² Von 37 Normalklassen lagen u. a. die Ergebnisse eines Rechtschreib- und eines Rechentests vor. Nach den Ergebnissen der Tests unterschieden sich die Leistungsverteilungen der Klassen erheblich. Dabei zeigte sich, daß in jeder der drei schlechtesten Klassen drei Viertel aller Schüler nur Testleistungen erreichten, die unter dem niedrigsten in den beiden besten Klassen überhaupt vorkommenden Wert lagen. Die besten Schüler aus den schlechtesten Klassen kamen zu Testpunkten, die in den besten Klassen mittleren Leistungen entsprachen. Die Mittelwerte in den drei schlechtesten Klassen waren etwa halb so groß wie in den drei besten Klassen.

Diese extremen Leistungsunterschiede spiegelten sich in den Rechenzensuren auch nicht annähernd wider. Alle Klassen hatten etwa die gleiche Zensurenverteilung. Die Zensur II wurde in den drei besten Klassen durchschnittlich etwas häufiger erteilt, kam aber z. B. in der drittbesten Klasse seltener vor als in der drittschlechtesten Klasse. Die Zensur VI wurde in keiner Klasse vergeben, aber paradoxerweise erhielten mehr Schüler in den besten Klassen die Zensur V als in den schlechtesten Klassen. In der schlechtesten Klasse erhielt kein Schüler die Zensur V. Derartige Diskrepanzen regen zur Frage nach der Validität des Tests an: Untersucht der Test das, was der Lehrer beurteilt, oder gehen beide von unterschiedlichen Beurteilungsaspekten aus? Wir sind diesen Fragen im Zusammenhang mit detaillierteren Fragen nach der Vergleichbarkeit der Zensuren weiter nachgegangen.³

Der Rechentest untersuchte das, was der Lehrplan forderte. Er hatte Lehrplangültigkeit. Die Aufgabenstellung entsprach der, wie sie in Rechenarbeiten vorkommt. Nur die Antwort-Auswahl-Methode und die größere Aufgabenzahl unterschieden den Test äußerlich von üblichen Klassenarbeiten. Zwischen den Rohpunkten des Tests und der Rechenzensur ließ sich für die Gesamtpopulation ($n = 1250$) eine Produkt-Moment-Korrelation von $r = 0,613$ errechnen, die für eine gute empirische Gültigkeit des Tests spricht. Ordnete man die Klassen aber nach ihren Mittelwerten im Rechentest in eine Rangreihe und nach ihren Mittelwerten in der Rechenzensur in eine andere und korrelierte beide Reihen, so ergab sich eine Rangkorrelation von 0,25, die nicht signifikant von 0 verschieden war.

Die Differenz in der Höhe beider Korrelationen deutete darauf hin, daß die Zensuren nach klasseninternen Maßstäben vergeben wurden. Innerhalb jeder Klasse erhalten die nach dem Rechentest besseren Schüler auch die besseren Zensuren und die im Rechentest schlechteren Schüler auch die schlechteren Zensuren. Extrem gute und extrem schlechte Schüler tragen zur relativ hohen Korrelation in der Gesamtpopulation bei, wo im Mittelbereich viele Überschneidungen zu erwarten sind. Da die Maßstäbe zwischen verschiedenen Klassen nicht übereinstimmen, ergibt die Rangkorrelation keine gesicherten Zusammenhänge. Diese – verkürzt dargestellte – Überlegung überprüften wir zunächst, indem wir die Rangkorrelationen für die drei Klassen berechneten, bei denen die größte Rangplatzdifferenz zwischen Testleistung und Zensur aufgetreten war (s. Grafik 1).

Klasse Kennziffer	Rangplatz Rechentest	Rangplatz Rechenzensur	Rangkorrelation
8	27	3	0,47
9	31	6	0,80
16	31	11	0,66

Grafik 1

Obwohl so erhebliche Rangplatzdifferenzen aufgetreten waren, bestehen also in jeder dieser Klassen gesicherte und teilweise hohe Zusammenhänge zwischen den Ergebnissen des Rechentests und der Rechenzensur.

Um unsere Hypothese zu bestätigen, haben wir für jede Klasse getrennt die Mittelwerte im Rechentest für Schüler mit der Zensur II, III oder IV berechnet. (Da die Zensuren I, V und VI zu selten vergeben wurden, war hier die Errechnung von Mittelwerten nicht angemessen.) Wenn unsere Hypothese zuträfe, müßte sich in jeder Klasse die Korrespondenz von Testergebnis und Zensur anschaulich erkennen lassen, während zwischen den Klassen sich gleichen Testwerten ganz unterschiedliche Zensuren zuordnen lassen müßten. Die Ergebnisse sind in **Abb. 1** dargestellt.

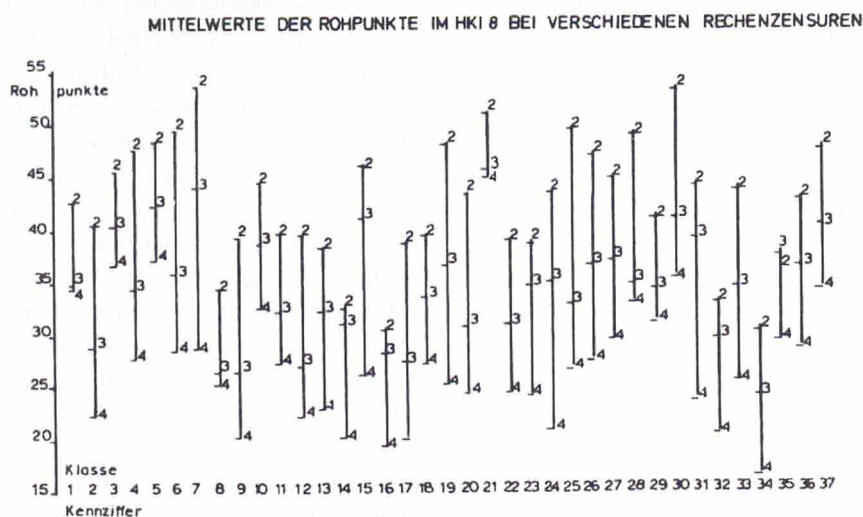


Abb. 1

Zur Erläuterung sei zunächst gesagt, daß in der Klasse mit der Kennziffer 2 die Schüler mit der Rechenzensur II im Rechentest durchschnittlich 40,5 Rohpunkte erzielten, die Schüler mit der Zensur III etwa 28,8 und die mit der Zensur IV etwa einen Mittelwert von 22,5. Überblickt man die **Abb. 1**, so lassen sich vor allem folgende Aussagen machen:

1. Mit Ausnahme der Klasse 35 haben in allen Klassen die Schüler mit der Rechenzensur II den höchsten und die Schüler mit der Zensur IV den niedrigsten Mittelwert im Rechentest. Innerhalb jeder einzelnen Klasse ist also ein deutlicher Zusammenhang zwischen Zensur und Testleistung zu erkennen.
2. In verschiedenen Klassen entsprechen der gleichen Rechenzensur ganz unterschiedliche Mittelwerte im Rechentest. In Klasse 7 haben die Schüler mit der Rechenzensur II einen Mittelwert von 53,5 Testpunkten, in Klasse 16 dagegen von 30,9 Testpunkten. Dieser Wert liegt ganz erheblich unter dem Mittelwert 44,8, den in Klasse 21 die Schüler mit der Rechenzensur IV erzielen. Der Mittelwert dieser Schüler mit einer IV liegt höher als die Mittelwerte, die in 23 anderen Klassen Schüler mit der Zensur II erreichten. In Klasse 34 kommen Schüler mit der Zensur IV nur auf einen Mittelwert von 16,4 Testpunkten.

Man darf bei der Darstellung in **Abb. 1** nicht vergessen, daß es sich um Klassen der gleichen 6. Klassenstufe der sechsjährigen Berliner Grundschule in einem Bezirk mittlerer Sozialstruktur handelt, die zum gleichen Termin untersucht wurden. Klassen wie die mit Nr. 34 und 35 sind Parallelklassen einer Schule mit gleichem Einzugsgebiet. Der naheliegende Schluß ist, daß es viel mehr von der zufälligen Klassenzugehörigkeit abhängt, ob ein Schüler eine II oder eine IV, vielleicht sogar eine V oder VI erhalte, als von der objektiv erfaßbaren Leistung.

Aus diesen Ergebnissen lassen sich folgende Schlußfolgerungen ableiten:

Die Lehrer orientieren sich bei ihrer Zensurenvergabe am Leistungsniveau der jeweils unterrichteten Klasse. Notendefinitionen oder die Erinnerung an vorher unterrichtete Klassen⁴ reichen nicht aus, um einen klassenübergreifenden Maßstab zu sichern.

Innerhalb einer jeden Klasse werden die Schüler durch Lehrer und Tests in eine relativ übereinstimmende Rangordnung gebracht. Damit kann einerseits die Gültigkeit der Tests (vom Kriterium Lehrerurteil her) nicht bestritten werden. Es sind

nicht unterschiedliche Beurteilungsgesichtspunkte, die zu den nachgewiesenen Differenzen führen. Andererseits zeigt sich, daß die Wirkung subjektiver Fehlerquellen bei der Zensurenggebung, so schwerwiegend sie für das Individuum sein mag, insgesamt nicht so stark ins Gewicht fällt wie erwartet.

In verschiedenen Klassen der gleichen Klassenstufe entsprechen der gleichen objektiv erfaßbaren Leistung aber ganz unterschiedliche Zensuren. Ob Schüler eines bestimmten Leistungsniveaus eine II oder eine IV oder – bei vorsichtiger Interpolation der Tafeln – eine I oder eine VI erhalten, hängt vor allem vom Zufall der Klassenzugehörigkeit ab. Die Zensuren haben über den Rahmen einer Klasse hinaus überhaupt keinen Vergleichswert.

Diese Erscheinungen lassen sich auch bei der Zusammenfassung mehrerer Zensuren und Tests nachweisen und wirken sich über die Zensuren auch auf Sitzenbleiben oder Versetzung, Empfehlung zur Hauptschule oder zum Gymnasium aus, wie wir an anderer Stelle gezeigt haben⁵. Wenn die Klassen als voneinander isolierte Gruppen durch die Schule gingen, könnte man die fehlende Vergleichbarkeit der Zensuren vielleicht tolerieren, obwohl es individuell unerträglich sein mag, mit einer Leistung in dieser Klasse nicht versetzt zu werden, während man in einer anderen Klasse mit guten Prädikaten weitergekommen wäre. Bei allen Gelegenheiten aber, wo die Zensur Funktionen für den Vergleich von Schülern aus vielen Klassen übernimmt, z. B. bei der Übergangsauslese, wirkt sich die vorausgesetzte, aber nicht vorhandene Vergleichbarkeit von Zeugniszensuren für das Individuum und für die Gesellschaft verhängnisvoll aus.

Diese Ergebnisse besagen nicht mehr und nicht weniger, als daß für unser gesamtes schulisches Berechtigungswesen keine sachliche Rechtfertigung besteht. Die Annahme, der Schüler, der seiner guten Zensuren wegen von der Aufnahmeprüfung dispensiert wird, weise bessere Schulleistungen auf als der Sitzenbleiber in einer anderen Klasse, ist bloße Spekulation. Wann sagen wir ehrlich, daß kein Lehrer aus den Zeugnissen eine vergleichbare Aussage über die Schulleistung von Schülern verschiedener Klassen entnehmen kann? Wann geben unsere Universitäten offen zu, daß es unsinnig ist, unter zwei Bewerbern aus verschiedenen Schulen, Städten oder gar Bundesländern einen nach den Abiturnoten auszuwählen? Man sollte diese Ergebnisse nicht damit abtun, daß die irgendwie nachweisbare Schulleistung gar nicht das Entscheidende sei und daß die Zensur in erster Linie die pädagogische Funktion des individuellen Ermunterns und Anspornens habe. Die ministeriellen Notendefinitionen, die praktischen und rechtlichen Konsequenzen, die Schulzeugnisse bei vielen Gelegenheiten haben, beweisen eindeutig, daß die Zensur immer wieder als Leistungsmaßstab angesehen wird.

Eine pädagogische Lösung dieser Schwierigkeiten ist nicht einfach, sofern auf Vergleichbarkeit der Zensuren und schulische Berechtigung nicht verzichtet werden kann, und bedürfte einer eigenen Darstellung. Hier kann nur einiges angedeutet werden. Keinesfalls empfehlen wir, Ergebnisse standardisierter Vergleichstests einfach als Kriterium für die Zensurenggebung zu übernehmen.⁶ Aber wir können vielleicht auch einmal das Verfahren der ausgleichenden Annäherung von Testwerten und Zensuren erproben, das in Schweden seit einem Vierteljahrhundert erfolgreich praktiziert wird.⁷

Die Tests sollen auch nicht einen einseitigen und vordergründigen Leistungsbegriff durchsetzen. Welche Leistungsebene im Test erfaßt und wie die Ergebnisse interpretiert werden sollen, muß Gegenstand durchdachter pädagogischer Entscheidung sein. Es gibt noch viele Möglichkeiten, die hier diskutiert werden könnten. Entscheidender Kernpunkt in diesem Zusammenhang ist jedoch, daß wir für das Lehrerurteil ein Korrektiv brauchen, das dem Lehrer zu bestimmten Zeiten einen Vergleichsmaßstab über viele Klassen hinweg bietet. Das beste gegenwärtig verfügbare Hilfsmittel für diesen Zweck stellen standardisierte Tests dar. Das deutsche Schulwesen wird sich in absehbarer Zeit entscheiden müssen, ob es sich ernsthaft mit diesem Hilfsmittel auseinandersetzen oder weiter mit der Fiktion der Vergleichbarkeit von Zensuren leben will.

Anmerkungen und Literatur

- ¹ Doye, P. und Ledig, H.-M.: Englischunterricht an der Berliner OPZ. Berliner Arbeitskreis Didaktik, Didaktische Informationen Nr. 6, Berlin o. J., S. 40 f. Schiefele, H.: Sind unsere Noten gerecht? In: Welt der Schule, Heft 6/1960, S. 251 – 257.
- Klink, J.-G.: Die Schülerleistung im Koordinatensystem der Zifferzensur. In: Lebendige Schule, Heft 9/1964, S. 375 – 383.
- Zielinski, W.: Erfahrungen mit einem Schulleistungstest für das 4. Schuljahr. In: Schule und Psychologie, Heft 1/1966, S. 14.
- ² Ingenkamp, K.: Untersuchungen zur Übergangsauslese. Weinheim, Berlin und Basel: Beltz 1968.
- ³ Ingenkamp, K.: Zur Problematik der Jahrgangsklasse. Weinheim, Berlin und Basel: Beltz 1969.
- ⁴ Ingenkamp, K.: Schulleistungen damals und heute. Weinheim und Berlin: Beltz 1967.
- ⁵ Ingenkamp, K.: Untersuchungen zur Übergangsauslese. Weinheim, Berlin und Basel: Beltz 1968.
- Ingenkamp, K.: Zur Problematik der Jahrgangsklasse. Weinheim, Berlin und Basel: Beltz 1969.
- ⁶ vgl. Ingenkamp, K.: Psychologische Tests für die Hand des Lehrers. 2./3. Auflage, Weinheim: Beltz 1964.
- ⁷ Henrysson, St.: Pädagogische Testanwendung in Schweden. In: Möglichkeiten und Grenzen der Testanwendung in der Schule, hrsg. von K. Ingenkamp u. Th. Marsolek. Weinheim, Berlin und Basel: Beltz 1968.

Zur Gültigkeit der Beobachtung von Ingenkamp in heutigen Schulklassen

Altlasten

Ingenkamp war nicht der Erste, der begründete Zweifel an der Aussagekraft von Schulnoten äußerte. Schon vor 130 Jahren stützte der irische Ökonom und Mathematiker Francis Ysidro Edgeworth sein tief sitzendes Misstrauen gegenüber schulischer Leistungsbewertung auf Ergebnisse aus seinen sorgfältigen Analysen. Eindringlich verwies er auf das „Element des Zufalls“ (1890, S. 461ff.) bei der Vergabe von Noten. Jahre später bemühte ein Schweizer Statistiker bei seinen Untersuchungen aus den 1950er-Jahren die besorgniserregenden Attribute zufällig, ungerecht und selbsttäuschend. Die beiden Beispiele sind einer langen Liste von Kritikern entnommen (vgl. ausführlich in Kronig 2007, S. 192 ff.). Die Rücksichtslosigkeit, mit der die jeweiligen Autoren die schulische Bewertungspraxis anprangern, ist weniger den teilweise scharfen Formulierungen zuzuschreiben. Vielmehr sind es die Daten selbst, welche rücksichtslos die Schieflagen von Schulnoten offenlegen.

Es wäre ein historisches Missverständnis, die Lehrkräfte anzuklagen. Vielmehr entlastet die Gesamtschau der Kritik sie in gewisser Weise. Denn diese zeigt, dass Noten sehr anfällig für eine Reihe von Verzerrungen sind und deshalb die Lehrkräfte an der eigentlichen Leistung vorbeurteilen. In einschlägigen Überblicken sind etwa Erwartungs-, Halo-, Reihungs- und Kontrasteffekte, generelle Beurteilungstendenzen, Beobachtungsmängel und Erinnerungsfehler sowie logische und mathematische Artefakte genannt (ebd. Ziegenspeck 1999). Es handelt sich nicht um Bagatellen, die man in der angespannten Hektik des schulischen Alltags achselzuckend in Kauf nehmen könnte. Jeder dieser Effekte hat für sich genommen das Potenzial, die Schulnoten dramatisch zu verfälschen. In ihrem Zusammenwirken bilden sie eine kaum durchschaubare Phalanx, die beinahe unüberwindbar einer leistungsgerechten Notengebung im Wege steht. Es ist ziemlich unwahrscheinlich, dass es in der Praxis selbst einer noch so sorgfältig kontrollierten und umsichtigen Beurteilung je gelingen könnte, alle möglichen Fehlerquellen gleichzeitig zu domestizieren. Den meisten Praktikerinnen und Praktiker dürften diese störenden Einflüsse, auch wenn vielleicht nicht namentlich, bekannt sein. Sie sind seit vielen Jahren fester Bestandteil der pädagogischen Ausbildung.

Kosmetik

Noten haben als „Währung“ des Bildungssystems eine lange Tradition. Fast ebenso alt sind auch die Einwände gegen sie. Ließen sich die Probleme leicht beheben, hätte man dies längst getan. An vielfältigen Versuchen hätte es nicht gefehlt. In mehreren Ländern gab es z. B. Bestrebungen, Ziffernnoten durch Buchstaben oder Wörter zu ersetzen. Noch verbreiteter ist die Praxis, Noten mit dem Ankreuzen vorgegebener Felder zu tauschen. In exotischeren Experimenten werden gegenwärtig auch Farben als Alternativen verwendet. Aber alle diese gut gemeinten Ansätze können die Probleme nicht überwinden, weil die möglichen Fehlerquellen nicht an die Form der Notengebung gebunden sind. Sie muten wie eine Sammlung von Illustrationen der Hilflosigkeit an, die Leistungszer-

tifikation gerechter zu gestalten. Vermutlich sind selbst ausführliche Lernberichte den Verzerrungseffekten schutzlos ausgeliefert.

Es kann gute Gründe für eine veränderte Form der Noten geben. Aber alle bisherigen Versuche finden gleichsam auf der Oberfläche statt. Die tieferliegenden Strukturen bleiben davon unberührt. Wenn der Zweck der Note ausschließlich in einer Leistungsrückmeldung an Kinder und Jugendliche liegen würde, wäre die gesamte Debatte um die Schulnoten viel entspannter zu führen. Aber Noten werden in hochselektiven Bildungssystemen vergeben. Erst in diesem Umfeld wird ihre soziale Brisanz dermaßen aufgeladen. Die Schule hat nicht nur die Aufgabe, Schülerinnen und Schüler in den als relevant angesehenen Fächern zu qualifizieren. Sie muss zwischen Schülerinnen und Schülern unterscheiden können. Und sie muss in der Lage sein, diese Unterschiede nach innen und nach außen zu dokumentieren. Diese zweite Aufgabe wandelt das unbeschwerte gemeinsame Lernen in einen rigorosen Wettbewerb um. Die Auswirkungen dieses widersprüchlichen Moments, der durch den doppelten gesellschaftlichen Auftrag ausgelöst wird, lässt sich auch empirisch beobachten.

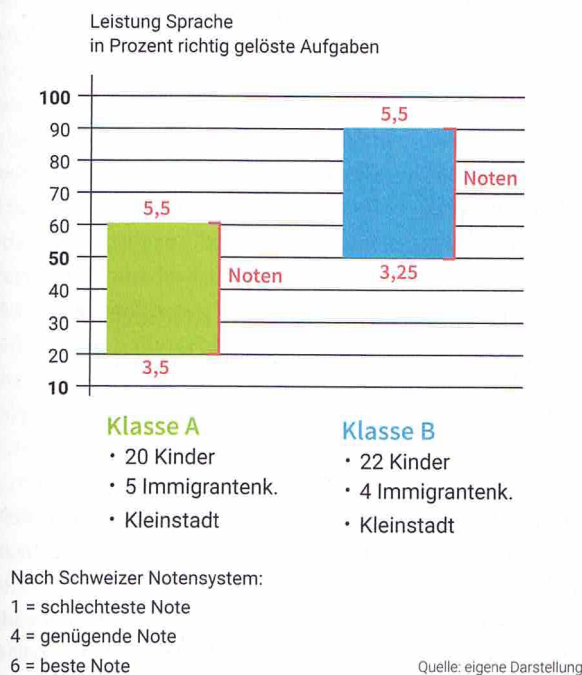
Während Kinder beim Lernen davon profitieren können, in einer leistungsstarken Schulklasse zu sitzen, mutiert in Prüfungssituationen der Vorteil der Leistungsstärke von Mitschülerinnen und Mitschülern zu einem Nachteil (Kronig 2007, S. 171 ff. und S. 199 ff.). In leistungsstarken Klassen ist es leichter, seine eigene Leistung zu steigern. Man bekommt dafür aber die schlechteren Noten. Umgekehrt belasten leistungsschwache Klassen tendenziell die Lernentwicklung. Dafür kommt man dort einfacher zu guten Noten.

Dass die Schule zwar den allgemeinen gesellschaftlichen Selektionsauftrag notorisch überinterpretiert, ist unter anderem an den regional und sogar kleinräumig unterschiedlichen Selektionsquoten abzulesen (Kronig 2007, S. 16). Aber man wird ihr schlecht vorhalten können, dass sie zwischen Gewinnern und Verlierern unterscheidet und ihnen gewissermaßen Rangplätze zuweist. Das muss sie tun. Ob dabei Ziffern, Buchstaben, Kreuzchen oder Wörter verwendet werden, ist unerheblich.

Referenzgruppenfehler

Bei einer Prüfung gibt die Schule den heterogenen Schülerleistungen eine hierarchische Ordnung. Bei dieser Übertragung von Leistung in Noten kommt es oft genug zu Übersetzungsfehlern. Und diese passieren mit einer gewissen Zwangsläufigkeit, der sich Lehrkräfte kaum entziehen können. Eindrücklich lässt sich das an jenem Effekt zeigen, der von Karlheinz Ingenkamp beschrieben und belegt worden ist. Er ist unter dem Namen „Referenzgruppenfehler“ oder „Schulklasseneffekt“ in die Literatur eingegangen. Im Prinzip geht es um das Problem, dass sich Lehrkräfte auch bei Klassen derselben Schulstufe vor ganz unterschiedliche Leistungsspektren gestellt sehen. Es kann sogar vorkommen, dass der stärkste Schüler einer Klasse eine schwächere Leistung aufweist als der schwächste Schüler einer anderen Klasse (vgl. Kronig 2007, S. 199). Dennoch sind Lehrkräfte weitgehend dazu gezwungen, dieselbe Bandbreite der Notenskala zu nutzen. Sie sind kaum in der Lage, diese Differenzen zwischen Schulklassen auf den gängigen Bewertungsskalen abzubilden. Sie müssten sonst in diesem Beispiel ihrem besten Schüler oder ihrer besten Schülerin eine sehr

1 Extrembeispiel der Auswirkung leistungsbezogener Schulklassenunterschiede auf die vergebenen Schulnoten im Unterrichtsfach Deutsch



schlechte Bewertung schreiben. Obschon sie nicht die Verursacher des Problems sind, hätten sie die heftigen Widerstände von Eltern und Behörden ganz allein zu tragen.

Die Verzerrungen, die allein aufgrund des Referenzgruppenfehlers zustande kommen, sind erheblich. In den Daten von rund 2.000 Schulkindern finden sich mehrfache Beispiele von Schülerinnen und Schülern, die eine ungenügende Leistungsbewertung in ihrem Zeugnis haben. Säßen sie in einer anderen Schulklasse, würden die gleichen Leistungen mit einer überdurchschnittlich hohen Leistungsbewertung ausgezeichnet werden (ausführlich ebd.).

Das Beispiel (s. Abb. 1) illustriert die Auswirkungen des Referenzgruppenfehlers. Der schwächste Schüler aus der Klasse B muss um seine Versetzung bangen, obschon seine Leistungen in der anderen Klasse mit Bestnoten bewertet würden. Seine Note unterschreitet mit einer 3,25 deutlich das geforderte Mindestniveau. Jedenfalls wird er künftig einem anspruchsärmeren Schultyp zugewiesen werden. Hätte er in der anderen Schulklasse gesessen, hätte seine weitere Bildungsbiografie mit großer Wahrscheinlichkeit einen völlig anderen Verlauf genommen.

Derartige Extrembeispiele geben aber nur einen Teil der Wirklichkeit wieder. Denn der Effekt erfasst beinahe alle Schulkinder, auch wenn das Ausmaß nicht bei allen gleich stark ausgeprägt ist (vgl. Kronig 2007, S. 200 ff.). Die Befunde gleichen den damaligen Ergebnissen von Ingenkamp. Darüber hinaus können Referenzgruppenfehler in sämtlichen Schulfächern auftreten. In der erwähnten Studie waren die Effekte in der Mathematik sogar etwas auffälliger als im Unterrichtsfach Deutsch. Das mag illustrieren, dass Mathe-

matikprüfungen nur vermeintlich zuverlässiger zu korrigieren sind als Aufsätze. Die verzerrte Leistungserfassung liegt hier nicht an Fehlurteilen der Lehrkräfte. Sie sind letztlich dem Effekt ebenso sehr ausgesetzt wie die Kinder ihrer Schulklassen.

Unfehlbarkeit

Der Aufwand für diese bizarre Leistungsdokumentation ist enorm. So werden selbst in einem kleinen Land wie der Schweiz, zurückhaltend gerechnet, zwischen Herbst und Weihnachten mehrere Millionen Noten an den Grundschulen vergeben. Woher kommt die unbeirrbar Resistenz der Schule gegen die mehr als hundertjährige Kritik an den Schulnoten? Die Beanstandungen wurden wiederholt vorgetragen und mit verschiedenen Argumenten versehen. Sie sind erfahrungswissenschaftlich durch eine erdrückende Zahl von Studien mit unterschiedlichen Methoden abgesichert. Und sie sind in ihrer Gesamtheit bislang unwidersprochen. Auch die Praxis begegnet den Noten zumindest teilweise mit Unbehagen. Davon zeugen z. B. die sporadischen Beiträge in der Presse, in denen Einzelvertreter oder ganze Verbände öffentlich deren Abschaffung verlangen.

Indes wird fortgesetzt minutiös und mit beträchtlichem Arbeitsaufwand für die Lehrkräfte zwischen den Leistungen von Schülerinnen und Schülern unterschieden. Dies verleiht den Noten eine fast schon magische Bedeutung. Allen Problemen zum Trotz hält sie weiterhin der Anschein einer gerechten, irrtumsfreien, zuverlässigen und punktgenauen Erfassung der Leistung von Schülern aufrecht. Die Verteilung von Schulnoten ist eine Zeremonie.

Wenn die Schule an ihr festhält, geschieht dies nicht aus angeordneter Willkür. Die gegenwärtige Praxis der Leistungsbeurteilung ist für das System funktional und organisatorisch sinnvoll. Sie erfüllt das Bedürfnis der Verwaltung und Lenkung von immer wieder neuen Schülerströmen in regional unterschiedlich ausgestaltete Strukturen. Dies zwingt die Schule dazu, die leistungsgerechte Notenvergabe eher zu inszenieren, als sich nach ihr zu richten. Die Schule kann gar nicht anders, als den von ihr gefällten Notenentscheidungen eine rituelle Glaubwürdigkeit zu verleihen, wenn sie nicht selbst unglaubwürdig werden will.

Wollte man tatsächlich den Schulnoten ihren Schrecken nehmen, müsste man die selektive Struktur des Systems antasten. Alle anderen Möglichkeiten scheinen inzwischen ausgeschöpft. Das würde die Verliererinnen und Verlierer des Bildungswettbewerbs entlasten. Ob die Gewinnerinnen und Gewinner damit einverstanden wären, ist aber sehr ungewiss.

Literatur

- Edgeworth, Francis Ysidro (1890). *The Element of Chance in Competitive Examinations*. In: *Journal of the Royal Statistical Society* 53 (3). S. 460 – 475.
- Ingenkamp, Karlheinz (1971). *Sind Zensuren aus verschiedenen Klassen vergleichbar?* In: Ingenkamp, Karlheinz: *Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte*. Weinheim und Basel: Beltz. S. 194 – 201.
- Kronig, Winfried (2007). *Die systematische Zufälligkeit des Bildungserfolgs. Theoretische Erklärungen und empirische Untersuchungen zur Lernentwicklung und Leistungsbewertung in unterschiedlichen Schulklassen*. Bern: Haupt.
- Ziegenspeck, Jörg (1999). *Handbuch Zensur und Zeugnis in der Schule. Historischer Rückblick, allgemeine Problematik, empirische Befunde und bildungspolitische Implikationen*. Bad Heilbrunn: Klinkhardt.

Mit Eltern über Leistung sprechen

Fallstricke im Elterngespräch vermeiden

Sprechen Lehrkräfte mit Eltern, dann geht es meist um die Leistungen ihrer Kinder. Dazu gehören die mündliche und schriftliche Mitarbeit, das Verhalten in der Schule, aber auch das (Nicht-)Erledigen von Hausaufgaben oder Üben von Vokabeln. In diesem Beitrag werden, basierend auf authentischen Gesprächen, die Situation und Perspektive von Eltern beschrieben.

Neben Elternabenden sind Elternsprechtage oft der einzige weitere Anlass für Eltern, in die Schule zu kommen. Anders als an den Elternabenden kommt es bei Elternsprechtagen zu einer dialogischen Kommunikationssituation, manchmal sind auch die Schülerinnen und Schüler anwesend. Für Gespräche an Elternsprechtagen werden, mit Blick auf unsere Gesprächsdaten (Wegner 2016), meist Zeitfenster von rund fünf bis zehn Minuten angesetzt. Die Schule gibt Gesprächstermine vor, die Eltern müssen ihre Teilnahme organisieren. Im Gegensatz zu den Lehrkräften, die die Verantwortung für das Gesprächssetting haben, ist die Situation für Eltern offener. Sie erscheinen als Gäste, mit oder ohne Agenda, innerlich gewappnet, gleichmütig oder mit Neugier und Vorfreude auf das Gespräch.

Elterngespräche analysieren

Im Gros der Gespräche geht es um den Leistungsstand und das (Sozial-)Verhalten der Schülerinnen und Schüler; mit Blick auf die Redeanteile dominieren die Lehrkräfte. Zwei Varianten der Rückmeldung sind zu unterscheiden: „Weiter so!“ oder „Nicht weiter so!“ Im ersten Fall ist die Lage einfach: Eltern und Lehrkraft können sich über den schulischen Erfolg des Kindes gemeinsam freuen. Beide haben offenbar alles richtig gemacht! Bisweilen kommen Eltern dann auch nur, um die Lehrkräfte

einmal kennenzulernen. Viele Gespräche finden jedoch statt, weil Situationen eingetreten sind, die von mindestens einer Partei als nicht zufriedenstellend eingeschätzt werden. Meist handelt es sich um defizitäre Leistungen und/oder inadäquates Sozialverhalten. Eltern haben dann ein besonders großes Informations- und Beratungsbedürfnis. Welche kommunikativen Schwierigkeiten in den Gesprächen zu beobachten sind, zeigen wir mithilfe einiger Gesprächsauszüge¹ auf.

Fall 1: Einfach kein Spaß und die im Stich gelassene Mutter

Mutter: *ich wollte fragen wie sie mit den sachen von tim zufrieden sind ob sich etwas gebessert hat oder obs schlechter geworden is*
Lehrer: *also so von seiner mitarbeit her tut er sich immer noch schwer es macht ihm einfach keinen spaß dieses zuhause sitzen und wirklich mal vokabeln lernen is nich sein ding*

Auf das Anliegen der Mutter hin liefert der Lehrer eine Zustandsbeschreibung (in der Mitarbeit tut er sich schwer) und eine Erklärung (keinen Spaß) mit vagem Informationsgehalt: Zum einen macht Schule bzw. die Mitarbeit im Unterricht sicherlich vielen Schülerinnen und Schülern keinen Spaß. Es handelt sich also um ein verbreitetes Phänomen, dessen Ursache auch im Unterricht selbst zu suchen wäre. Als Erklärung für die unzureichende Mitarbeit erweist es sich da-

mit als (zu) pauschal. Zum anderen verweist der Lehrer auf fehlendes Vokabelwissen. Damit gerät die Vorbereitung für die Stunde in den Blick, die zu Hause zu erfolgen hat. Die Grundlage für eine erfolgreiche Unterrichtsteilnahme wird im Elternhaus gelegt: zu Hause sitzen und wirklich lernen! Allerdings – und das macht die Situation verfahren – ist das „*nich sein ding*“. Die Lage für den Jungen scheint damit aussichtslos zu sein. Aus welchen Gründen er konkrete Handlungen nicht ausübt, verweigert oder schlicht nicht beherrscht, bleibt vage. Für diese Diagnose braucht es wahrlich keinen Experten! Im Unterricht bzw. von Seiten der Schule kann nichts für den Schüler getan werden. Damit liegt die Verantwortung für die Lösung bei ihm und seinen Eltern.

Fall 2: Das Sündenregister und die kollaborierende Mutter

Lehrer: *so du hast aber auch im deutschen machst du manchmal so die dinge nicht äh zum beispiel die ganzen die ganze zeit und personalformen da kommst du auch nich großartig weiter da darf ich ja im moment auch nicht fragen du arbeitest viel nach aber es von dir aus kommt da herzlich wenig also du bist nicht jemand der der mal aufzeigt und sagt äh so als erster vorprescht [...] in deutsch das letzte mal [...] haste auch nich geschafft dann sagste auch manchmal ich kann dat nich [...]*

Mutter: *der überblick fehlt mir jetzt n bisschen weil ich ja jetzt früher auch mit ihr haus-*

aufgaben gemacht hab und dat seh ich ja jetzt so nich mehr so ne

Lehrer: *ja, ja ne also da is auch im moment nich nich so viel los bei dir was deutsch an geht*

Mutter: *und du meinst du bist in deutsch so gut ne, hmh, ja*

Lehrer: *ja aber da bisse hasse im moment auch deine massiven schwierigkeiten oder nich*
Schülerin: *ein bisschen*

Lehrer: *ja da musst du du musst du einfach mal kucken dass du da dich festbeißt und mal mal n bisschen nachübst*

Der Lehrer beschreibt den fehlenden Lernfortschritt als Problem (*da kommst du auch nicht großartig weiter*). Anschließend listet er, im Transkript aus Platzgründen ausgespart, noch weitere Verfehlungen der Schülerin auf. Die Mutter antwortet und entschuldigt sich dafür, dass ihr der Überblick fehle, womit sie sich einerseits der Verantwortung entzieht, sich aber gleichermaßen einer gewissen „Vernachlässigung“ schuldig macht. Im Anschluss hält sie der Tochter eine Leistungsfehleinschätzung vor und verbündet sich mit dem Lehrer. Im Gesprächsverlauf wird die Verantwortung für die Problemlösung der Schülerin übertragen, *„da musst du einfach mal gucken“*. Wann und was konkret unternommen werden soll, bleibt wieder ungenau (gucken – festbeißen – nachüben).

Fall 3: Die moralische Verantwortung und der oppositionelle Vater

Vater: *und wie kann man dann mündliche noten machen*

Lehrerin: *bitte*

Vater: *wie könnt ich dann wenn ich jetzt schüler hier wäre und sie erklärn n thema neu da käm ich gar nicht auf den gedanken mündlich mitzuarbeiten, dann würd ich das (.) versuchen zu kapiern (.) was sie sagen*

Nachdem die Lehrerin die fehlende mündliche Mitarbeit moniert hat und Schülerin und Mutter im Gegenzug die Lehrerin und ihre Unterrichtsgestaltung für das Problem verantwortlich gemacht haben, ergreift der Vater das Wort. Er stellt Plausibilität in puncto Unterrichtsführung und Beurteilungspraxis infrage. Im weiteren Verlauf kommt es zu einer Fundamentalkritik. Die Lehrerin stellt Schülerinnen und Schüler vor unlösbare

Aufgaben und sanktioniere diese dann mit schlechten Noten für deren Nichterfüllung. Damit ist die Koalitionsbildung der Familie vollständig: Die Lehrerin trägt faktisch die Verantwortung für die schlechten Resultate der Schülerin, denn „gute“ Lehrkräfte konzipieren ihren Unterricht so, dass der Meldepflicht nachgekommen werden kann.

Resümee

Die Beispiele demonstrieren kommunikative Konstellationen, deren Ausgangspunkte in defizitären Leistungen liegen, womit zugleich eine Veränderungsnotwendigkeit angezeigt ist. Zwei Verantwortlichkeiten stehen dann zur Debatte: Wer ist verantwortlich für die Leistung? Und wer ist zuständig für die Lösung des Problems? In keinem Gespräch gelingt das Ideal einer partnerschaftlichen Zusammenarbeit zwischen Elternhaus und Schule. Wir können also resümieren, dass sich das Mitteilen von Leistungsproblemen in Gesprächen an Elternsprechtagen als prekär erweist, wenn ...

1. diese nicht im Rahmen diagnostischer Einschätzungen, pädagogischer oder fachdidaktischer Konkretisierungen thematisiert werden, die Informationen pauschal bleiben und nur laienhafte Lösungsangebote gemacht werden. All dies stellt die Existenz von Expertenwissen infrage. Hierin kann eine Ursache für den schlechten Ruf liegen, den Elterngespräche bei vielen Eltern genießen.
2. die Verantwortung für defizitäre Leistungen und deren Lösung einseitig auf die Schülerinnen und Schüler übertragen wird. Wenn Leistungsprobleme durch einseitige Anpassungsleistungen der Schülerinnen und Schüler gelöst werden sollen, besteht sicherlich keine gute Prognose für Veränderung. Einseitig Anpassung zu propagieren wird vermutlich eher Widerstand befördern und generationale Konflikte herausfordern.

Die Beispiele zeigen, dass ein Kernproblem der Gespräche in der Aufgabe gründet, eine für alle Beteiligten „gesichtsbedrohende“ (Brown/Levinson 2007) Bewertung vorzunehmen. Für die Leistungen von Schülerin-

nen und Schülern stehen immer auch die Eltern und die Lehrkräfte qua ihrer Mandate in Mitverantwortung. Eltern können sich mit der Lehrkraft verbünden, alternativ kann sich eine ganze Familie auch solidarisieren und sich gegen diese in Stellung bringen. Die Konsequenz bleibt identisch: Statt gemeinsam nach Lösungen für die Probleme zu suchen, etablieren die Beteiligten einen vorwurfsvollen, moralischen Diskurs um Schuld und Verantwortung, in dem es nicht mehr um Leistungsprobleme von Schülerinnen und Schülern geht, sondern um gutes Lehrer-/Lehrerinnen-/Eltern-/Schüler- und Schülerinnensein. Das Ringen um die Frage, wer die Verantwortung für die defizitäre schulische Leistungen trägt, führt gerade nicht zu einer konstruktiven Bearbeitung des Problems, sondern zu einer konfliktreichen Gesprächssituation, die am Ende keine moralischen Gewinner entlässt.

Diese in der geforderten Schule-Elternhaus-Kooperation strukturell angelegte Kernproblematik kann dabei weder auf fehlende Beratungskompetenzen von Lehrkräften noch auf die Existenz „schwieriger“ Eltern zurückgeführt werden. Die kommunikativen Aufgaben in diesen Gesprächen, die idealtypisch auf die (gegenseitige) Beratung und Information der Erziehungspartnerinnen und -partner zielen, sind daher nicht allein mit Techniken der Gesprächsführung zu bearbeiten (vgl. Bennewitz 2016). Dennoch dürfte insbesondere eine Sensibilität für diesen Umstand und die Einhaltung grundlegender Beratungskomponenten, wie z. B. eine Klärung des Anliegens, einem konstruktiven Gesprächsverlauf förderlich sein.

Anmerkung

¹ Die für die folgenden Analysen ausgewählten Gesprächsauszüge sind Bestandteil eines größeren Korpus authentischer Elternsprechtagsgespräche. Das Korpus wurde im Rahmen des Dissertationsprojekts von Wegner (2016) erstellt. Es umfasst Audioaufnahmen und Transkriptionen von 142 Elternsprechtagsgesprächen, die an verschiedenen Schulen (und Schulformen) in Nordrhein-Westfalen im Jahr 2010 stattfanden.

Literatur

- Bennewitz, H. (2016). *Beratung als Aufgabe des Lehrberufs*. In: Rothland, M. (Hrsg.). *Beruf Lehrer/Lehrerin. Ein Studienbuch*. Münster: Waxmann. S. 205 – 226.
- Brown, P./Levinson, S. (2007). *Gesichtsbedrohende Akte*. In: Herrmann, S. K./Krämer, S./Kuch, H. (Hrsg.). *Verletzende Worte*. Berlin: transcript Verlag. S. 59 – 88.
- Wegner, L. (2016). *Lehrkraft-Eltern-Interaktionen am Elternsprechtag. Eine gesprächs- und gattungsanalytische Untersuchung*. Berlin: de Gruyter.