

Sylvia Rahn/Sabine Gruehn/Miriam Keune/Christoph Fuhrmann

Aus Schüleraussagen lernen?! – Auf dem Weg zu einer professionellen Feedbackkultur an Schulen

Zusammenfassung

Lehrkräfte sind aufgefordert, die Qualität ihres Unterrichts mithilfe von Schülerrückmeldungen zu überprüfen. In diesem Beitrag wird diskutiert, was unter einer „professionellen Feedbackkultur“ in Schulen im Hinblick auf die Verwendung des Schülerfeedbacks zur Messung von Unterrichtsqualität zu verstehen ist und auf welche Verwendungsweisen angesichts des empirischen Forschungsstands zur Validität von Schülerratings zur Unterrichtsqualität möglicherweise auch besser zu verzichten wäre.

Schlüsselwörter: Unterrichtsevaluation, Unterrichtsqualität, Schülerfeedback, Schülerratings, Validität, Biasfaktoren, Verzerrungsfaktoren

Learning from Students' Statements?! – Towards a Professional Feedback Culture at Schools

Summary

Teachers are encouraged to check the quality of their teaching by using student feedback. This paper discusses what a “professional feedback culture” in schools in the context of measuring teaching quality means. The second question is, whether some uses of student feedback are not useful in light of the empirical research literature on validity of student ratings of teaching quality.

Key words: course evaluation, course quality, student feedback, student ratings, validity, bias factors, distortion factors

1. Einleitung

Lehrkräfte sehen sich bereits seit längerem – z.B. in Empfehlungen zur internen schulischen Evaluation – dazu aufgefordert, den Wahrnehmungen und Einschätzungen ihrer Schülerinnen und Schüler zur Qualität des Unterrichts systematisch Aufmerksamkeit zu schenken, um hieraus produktive Konsequenzen für die Unterrichts- und eigene Professionalitätentwicklung abzuleiten. Erneutes Gewicht

erhält diese seit den 1990er-Jahren immer wieder formulierte Forderung nach dem Auf- und Ausbau einer „professionellen Feedbackkultur“ an Schulen etwa durch die Standards der Kultusministerkonferenz (KMK) für die bildungswissenschaftliche Lehrerbildung, in denen die Nutzung von „Rückmeldungen anderer“ als Ausbildungsziel genannt ist (KMK 2014, S. 13), oder durch die breite Rezeption der Hattie-Studie (Hattie 2013). Zwar wird Schülerfeedback an einigen Schulen bereits regelmäßig und systematisch genutzt; von einer flächendeckenden und selbstverständlichen Verwendung von Schülerrückmeldungen sind wir jedoch auch nach jahrelanger Diskussion um Evaluation, Qualitätsmanagement und Qualitätsentwicklung an Schulen noch weit entfernt. Hinzu kommt, dass sich die Art und Weise, wie das Schülerfeedback erhoben, ausgewertet und interpretiert wird, im inter- wie nationalen Vergleich der Verfahren erheblich unterscheidet.

Deshalb ist es das Ziel dieses Beitrages zu konkretisieren, was mit einer „professionellen Feedbackkultur“ in Schulen im Hinblick auf die Verwendung des Schülerfeedbacks zur Messung von Unterrichtsqualität im Einzelnen gemeint sein könnte und auf welche Verwendungsweisen angesichts des empirischen Forschungsstands zur Validität von Schülerratings zur Unterrichtsqualität möglicherweise auch besser zu verzichten wäre.

Zu diesem Zweck wird im Folgenden zunächst ein knapper Überblick darüber gegeben, wie Schüleraussagen derzeit in deutschen Sekundarschulen zu Evaluationszwecken erhoben und ausgewertet werden. Auf der Grundlage des internationalen Forschungsstandes sowie auf der Basis der Ergebnisse einer abgeschlossenen empirischen Untersuchung zu den „Determinanten des Schülerfeedbacks“ (Rahn/Gruehn/Böttcher 2015), in der Schülerratings zur Unterrichtsqualität in verschiedenen Fächern von insgesamt 2.008 Schülerinnen und Schülern der 11. Jahrgangsstufe an 38 beruflichen Gymnasien und 10 Gesamtschulen erhoben und ausgewertet wurden, wird sodann diskutiert, wie die Validität des Schülerfeedbacks einzuschätzen ist. Der Beitrag mündet schließlich in Empfehlungen, worauf bei der praktischen Verwendung von Schülerrückmeldungen zu achten wäre.

2. Der praktische Status quo – Schülerratings zur Unterrichtsqualität als Datenquelle in der Unterrichtsforschung und in der Schul- und Unterrichtsevaluation

Schüleraussagen werden auf vielfältige Weise zur Beschreibung, Diagnose, Evaluation und Entwicklung von Schule und Unterricht verwendet. Im Kontext der Unterrichtsforschung werden Schülerratings zur Unterrichtsqualität in der Regel fachbezogen erhoben und auch so zurückgemeldet. Als ein Beispiel hierfür lässt sich die Untersuchung „Qualitätssicherung in Schule und Unterricht“ (kurz: QuaSSU) anfüh-

ren, in der bayerische Hauptschul-, Realschul- und Gymnasialklassen zur Qualität ihres Mathematikunterrichts standardisiert befragt wurden (vgl. Ditton 2002). Wenn die Schülerinnen und Schüler der befragten Klassen dem zustimmten, konnten die Lehrkräfte der beteiligten Schulen Rückmeldungen zu den Ergebnissen der Schülerbefragungen erhalten. In diesen Feedbacks wurden die Mittelwerte für die jeweilige Schulklasse im Vergleich mit dem Mittelwert der Gesamtstichprobe der jeweils gleichen Schulart dargestellt.

Etwas andere Rückmeldungen erhielten die Lehrkräfte in der Studie „Deutsch Englisch Schülerleistungen International“ (kurz: DESI) für die Fächer Deutsch und Englisch (vgl. Klieme 2008). In diesem Zusammenhang erhielten Lehrkräfte, falls gewünscht, von den befragten Neuntklässlerinnen und Neuntklässlern Rückmeldungen zur Klassenführung sowie zur Motivation der Jugendlichen für fachbezogene Inhalte. Die Ergebnisse wurden lediglich den Lehrkräften der jeweils untersuchten Klasse mitgeteilt, ohne einen Referenzwert einer Vergleichsgruppe anzugeben.

In der Praxis der Schul- und Unterrichtsevaluation wird die wahrgenommene Unterrichtsqualität zwar auch in der Regel fachbezogen erhoben, aber – anders als in der Unterrichtsforschung – nicht immer fachbezogen ausgewertet. In Bezug auf die Rückmeldungen des Schülerfeedbacks an die Lehrkräfte finden sich verschiedene Ansätze von der individuellen Rückmeldung der Daten der eigenen Klasse(n) an einzelne Lehrkräfte über mehrperspektivische Vergleiche innerhalb einer Schule (beides fachbezogene Auswertungen) bis hin zu Vergleichen zwischen Schulen (fachübergreifende Auswertung).

Ein Beispiel für individuelle Rückmeldungen an Lehrkräfte ist das in NRW, Sachsen und Thüringen implementierte Konzept „Schüler als Experten für Unterricht“ (kurz: SEfU). Mit Hilfe von 35 Kernaussagen – basierend auf einem einschlägigen Qualitätsmodell der Schulforschung – wird die Schülersicht auf Unterricht online als „Expertenurteil“ für die Unterrichtsqualität erfasst und mit der Unterrichtswahrnehmung der jeweiligen Lehrperson über ein ebenfalls online verfügbares Auswertungstool verglichen (vgl. Kämpfe 2009; ten Venne/Nachtigall/Kämpfe-Hargrave 2010). Durch die wiederholte Nutzung können Lehrende das Schülerfeedback mehrerer Lerngruppen, die sie unterrichten, im Querschnitt oder Längsschnitt vergleichen.

Ein anderes, von der Kultusministerkonferenz gefördertes Angebot ist das Programm „Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung“ (kurz: EMU) (vgl. Helmke/Helmke 2014; Helmke et al. 2015). Das Programm bietet über eine Internetplattform frei zugängliche Instrumente an, mit denen Lehrkräfte die Qualität des eigenen Unterrichts mithilfe der Sichtweise eines oder mehrerer Kollegen und/oder der Schülersicht analysieren und mehrperspektivisch vergleichen können. Für die Auswertung steht ein Auswertungsprogramm zur Verfügung, das nach der

Dateneingabe die Verteilung der verschiedenen Perspektiven in Form von Profillinien anzeigt.

Während SefU und EMU stark auf den Vergleich verschiedener Perspektiven auf denselben Unterricht setzen, werden im „Netzwerk Schülerbefragung“, das als drittes Beispiel genannt werden soll, der fächerübergreifende Vergleich und der Vergleich zwischen verschiedenen Schulen betont. Bei diesem Zusammenschluss aus Schulen in Berlin und Brandenburg geben Lernende ein Feedback in allen unterrichteten Fächern und in sämtlichen Bildungsgängen ab (vgl. Wagner 2010, S. 292). Im „Netzwerk Schülerbefragung“ wird das Feedback der Klasse den Lehrenden im Vergleich mit dem „aggregierten Bewertungsprofil“ sowohl ihrer eigenen Schule als auch aller Netzwerkschulen mitgeteilt, und das unabhängig vom jeweils unterrichteten Fach (vgl. ebd.).¹

Neben den vorgestellten Programmen, die ausschließlich unterrichtsbezogene Schülerfeedbacks generieren, werden Schülerwahrnehmungen auch in nahezu allen auf die Schule als Organisation bezogenen Evaluationsverfahren und Qualitätsmanagementkonzepten genutzt. Dies gilt u.a. für das Konzept von „Selbstevaluation in Schulen“ (kurz: SEIS) der Bertelsmann Stiftung, das Modell der „European Foundation for Quality Management“ (kurz: EFQM) und die Konzepte zur Schul- und Unterrichtsentwicklung, die vom Institut für Schulentwicklungsforschung in Dortmund entwickelt wurden. Üblich ist im Zuge solcher Konzepte, dass die Schülerinnen und Schüler schriftlich zur Unterrichtsqualität an ihrer Schule befragt werden, ohne dass zwischen verschiedenen Fächern und Lehrpersonen unterschieden würde. Auf dieser Basis werden dann Schulmittelwerte berechnet, die zu den Durchschnittswerten der Unterrichtsqualität an allen Schulen derselben Schulform ins Verhältnis gesetzt werden. Die einzelne Lehrkraft bekommt keine individuelle Rückmeldung. Diese Vorgehensweise unterscheidet sich damit erheblich von den auf den Unterricht und das Feedback identifizierbarer Lerngruppen bezogenen Projekten der Unterrichtsforschung und von den Evaluationsverfahren, die das Schülerfeedback primär in den Dienst der individuellen Unterrichtsentwicklung der einzelnen Lehrkraft stellen (vgl. Thiel/Ulber 2007).

Per Saldo bietet der knappe Überblick über den Status quo zur Nutzung von Schüleraussagen in der Schul-, Unterrichts- und Lehrevaluation ein „buntes“ Bild, das die Frage nach den Kriterien und Ansprüchen aufwirft, denen die Evaluationsverfahren selbst genügen sollten.

1 Zu den Aktivitäten des „Netzwerks Schülerbefragung“ siehe neben den Veröffentlichungen von Wagner (2009, 2010) auch exemplarisch den Bericht eines Oberstufenzentrums. URL: http://www.peter-lenne-schule.de/images/pdfs/schuelerbefragung_2014/2014-03-26_ergebnisse_schuelerbefragung_2014.pdf; Zugriffsdatum: 28.01.2016.

3. Validität von Schülerratings zur Unterrichtsqualität im Spiegel empirischer Forschung

Was wissen wir nun aber darüber, wie valide Schülerratings im Zuge standardisierter Fragebogenerhebungen die Unterrichtsqualität messen? Zur Beantwortung dieser Frage werden im Folgenden empirische Forschungsbefunde zur – perspektivenspezifischen – Validität sowie zu den Bias- und Verzerrungsfaktoren der Schülerratings berichtet. Auf eine Darstellung der Ergebnisse zur prognostischen Validität, d.h. zur Vorhersagekraft der Schülerfeedbacks für zukünftige Schülerleistungen, wird hier – angesichts des begrenzten Raums – verzichtet, da diese Fragestellung zwar die Bildungsforschung durchaus beschäftigt hat (vgl. u.a. Clausen 2002; Gruehn 2000, S. 199; Kunter 2005; Lüdtke et al. 2006), in der schulischen Evaluationspraxis aber kaum eine Rolle spielt.

3.1 Validität des Schülerfeedbacks

Um einschätzen zu können, wie zuverlässig und valide die Prozessqualität des Unterrichts mithilfe von Schülerbefragungen erfasst werden kann, wird untersucht, wie gut das Schülerfeedback mit der Unterrichtswahrnehmung und -beurteilung von Lehrkräften einerseits und der von externen Beobachtern und Beobachterinnen andererseits übereinstimmt bzw. worin sich die drei Perspektiven jeweils unterscheiden. Der Perspektivenvergleich zeigt, dass die Ratings zwischen Schülerinnen und Schülern sowie externen Beobachterinnen und Beobachtern erheblich korrelieren, dass sie aber kein perspektivenübergreifendes Messmodell zur Unterrichtsqualität darstellen (vgl. Clausen 2002; Greimel 2002). Zwischen den Schüler- und Lehrkräfteeinschätzungen bestehen hingegen eher geringe Übereinstimmungen (vgl. Clausen 2002; Baumert/Kunter 2006), was u.a. daran liegt, dass die Angaben von Lehrkräften vielfach Selbstbeurteilungen darstellen und deshalb selbstwertdienlichen Verzerrungen unterliegen. Parallelen bestehen daher nur bei Unterrichtsmerkmalen, die eine längere gemeinsame Erfahrung unterrichtlicher Interaktion und Kommunikation voraussetzen wie etwa bei der Klassenführung (vgl. Clausen 2002; Baumert/Kunter 2006). Dies wurde unlängst auch in unserer einleitend erwähnten fächervergleichenden Untersuchung in der gymnasialen Oberstufe bestätigt: Aspekte der Klassenführung werden sowohl perspektiven- als auch fachübergreifend einheitlich wahrgenommen.

Für die meisten Unterrichtsmerkmale gilt, dass ihre Erfassung und Beurteilung nur perspektivenspezifisch valide möglich ist, wobei keine der drei Perspektiven grundsätzlich näher an der „wahren“ Unterrichtsqualität ist als die anderen. Für die perspektivenspezifische Validität des Schülerfeedbacks lässt sich der Forschungsstand dahingehend zusammenfassen, dass Unterrichtsbeschreibungen aus Schülersicht un-differenzierter ausfallen als dies wünschenswert wäre (vgl. Gruehn 2000, S. 142ff.;

Wagner 2010, S. 292). Überdies unterliegt das Schülerfeedback einem so genannten „affektiven Halo-Effekt“, d.h., der Gesamteindruck, den die Befragten vom Unterricht haben, wirkt sich auch auf die Einschätzung seiner einzelnen Prozessmerkmale aus (vgl. Clausen 2002; Wagner 2008; Wagner 2010). Insgesamt führt all dies jedoch nicht dazu, dass das Schülerfeedback es nicht erlauben würde, bei der Beschreibung und Bewertung der Unterrichtsqualität deutlich zwischen Unterrichtsmerkmalen zu unterscheiden. Im Gegenteil: Stärken- und Schwächenprofile sind im Mittelwertvergleich und in den sie visualisierenden Profillinien deutlich zu erkennen.

3.2 Der Einfluss von Bias- und Verzerrungsfaktoren

Zur Analyse der Validität des Schülerfeedbacks wurde zudem intensiv der Frage nachgegangen, inwieweit die Schülerratings durch Merkmale beeinflusst und verzerrt sind, die selbst nicht Ausdruck der Unterrichtsqualität sind (echte Biasfaktoren) oder aber sich der direkten Beeinflussung durch die Lehrperson entziehen (Fairnessvariablen) (vgl. Rindermann 2001). Bei der Beantwortung dieser Frage ist die Unterrichtsforschung in weiten Teilen der Forschung zur Evaluation der Hochschullehre gefolgt. Bislang wurde empirisch vor allem geprüft, inwieweit das Geschlecht der Lernenden, das Geschlecht der Lehrperson oder die Interaktion zwischen beiden, die Leistungsstände der Schülerinnen und Schüler und die Benotungspraxis der Lehrkräfte, die Sympathie mit der Lehrkraft sowie das Fachinteresse der Lernenden mit dem Schülerfeedback korrelieren.

Hinsichtlich der Bedeutung des Geschlechts der Lehrenden bzw. der Lernenden und der Interaktion zwischen beiden bleibt der Forschungsstand sowohl für den schulischen als auch für den hochschulischen Lehr-Lern-Kontext uneindeutig (vgl. Greimel-Fuhrmann/Geyer 2005, S. 115; Gruehn 2000, S. 78). Der Gedanke liegt nahe, dass die Geschlechtereffekte durch das mit dem Geschlecht der Lernenden kovariierende Interesse am Unterrichtsfach zustande kommen und nicht als Ausdruck allgemeingültiger geschlechtsspezifischer Beurteilungstendenzen zu verstehen sind. Genau dies war in der Untersuchung zu den Determinanten des Schülerfeedbacks in vier der fünf einbezogenen Unterrichtsfächer der Fall. In Deutsch, BWL/Rechnungswesen, Erziehungswissenschaft und Elektrotechnik ließ sich bei statistischer Kontrolle weiterer Schülermerkmale kein Geschlechtereffekt bei der Unterrichtsbeurteilung feststellen. Nur im Fach Mathematik zeigte sich trotz Berücksichtigung der Lernmotivation und des Fachinteresses der Schülerinnen und Schüler noch ein schwacher signifikanter Geschlechtereffekt.

Auch die Befürchtung, dass die Schüler- und Studierendenwahrnehmungen durch die erwarteten und faktisch erhaltenen Noten verzerrt werden, kann weder klar entkräftet noch zweifelsfrei empirisch gestützt werden (vgl. Aleamoni 1999; Böttcher/Grewe 2010; Rindermann 2001). Insbesondere die Geltung der so genannten „Grading-

Leniency-Hypothese“, wonach sich eine relativ zu den faktischen Leistungen milde Beurteilung der Schülerleistungen in einem ebenfalls günstigen Schülerfeedback niederschlägt (vgl. Marsh/Roche 1997; Wagner 2008, S. 23 und S. 124), bedarf noch eindeutiger empirischer Belege.

Klarer ist der Forschungsstand zum Zusammenhang zwischen Sympathie und Schülerfeedback. Er wurde sowohl für die Beurteilung der Lehrqualität im Fach Rechnungswesen an österreichischen Handelsakademien (vgl. Greimel-Fuhrmann/Geyer 2005) als auch am Beispiel der DESI-Daten für die Fächer Deutsch und Englisch belegt. In unserer Studie zum Schülerfeedback in der gymnasialen Oberstufe war der Zusammenhang zwischen der Sympathie mit der Lehrkraft und der wahrgenommenen Unterrichtsqualität in allen fünf Fächern signifikant und mit ähnlicher Effektstärke nachweisbar.

Besonders wichtig ist für das Schülerfeedback zur Unterrichtsqualität, welches Interesse die Schülerinnen und Schüler dem Unterrichtsfach entgegenbringen. In der Forschung zur Evaluation der Hochschullehre ist der Zusammenhang zwischen dem Interesse am Lehrinhalt und der wahrgenommenen Lehrqualität gut belegt (vgl. Spiel/Gössler 2000; Rindermann 2001, S. 179ff.). Dabei variiert der Stellenwert, den das Interesse für die studentischen Einschätzungen hat, nach Veranstaltungsart und Veranstaltungskontext (vgl. El-Hage 1996).

Damit ließe sich auch für schulische Kontexte erwarten, dass sich das Fachinteresse der Lernenden nicht nur auf das Schülerfeedback auswirkt, sondern dass es dies in Abhängigkeit vom Unterrichtsfach mit unterschiedlicher Stärke tut. Zwar konnte Ditton (vgl. 2002, S. 280) für den Mathematikunterricht an bayerischen Schulen der Klasse 9 mit Hilfe von Mehrebenenanalysen zeigen, dass sich die Interessantheit und Wichtigkeit des Faches nicht mehr verzerrend auf die wahrgenommene Unterrichtsqualität auswirkt, wenn statt der individuellen Werte die Klassenmittelwerte verwendet werden – sich durch die Aggregation der Daten also der Verzerrungseinfluss ausmittelt. Bei fächerübergreifenden Stichproben kann dieser Effekt allerdings nicht ohne weiteres vorausgesetzt werden. Denn nur wenn die Verzerrungsvariablen innerhalb der Klassen normalverteilt sind, sind Individualdaten verzerrt, Klassenmittelwerte aber nicht.

Dies ist im Falle von fächer- und bildungsgangübergreifenden Stichproben, wie sie in der Unterrichtsforschung selten, aber in Qualitätsmanagementverfahren häufig vorkommen, für den Einfluss des Fachinteresses aber nicht zu erwarten. Mit andauerndem Schulbesuch entwickeln Jugendliche ein sich zunehmend ausdifferenzierendes Interessenprofil. Erwartungsgemäß konnten deshalb in dem Projekt zu den Determinanten des Schülerfeedbacks in der Oberstufe im berufsbildenden Zweig ein signifikanter Effekt des Unterrichtsfachs auf das Fachinteresse und ein Interaktionseffekt mit dem besuchten Gymnasialzweig nachgewiesen werden. Im

Einzelnen bedeutet das, dass das Fachinteresse der Lernenden in den profilbildenden Fächern der beruflichen Gymnasien höher ist als in Mathematik oder Deutsch und dass sich Lehrende, die Mathematik im pädagogischen Gymnasialzweig oder Deutsch im technischen Gymnasium unterrichten, mit einem geringen durchschnittlichen Fachinteresse konfrontiert sehen. In der Folge konnte in der Oberstufenstudie neben dem Einfluss des individuellen Schülerinteresses auch ein signifikanter Effekt des Unterrichtsfachs als Klassenmerkmal auf die Gesamtbeurteilung des Unterrichts durch die Lernenden belegt werden.

Überdies wirkten sich nicht nur einzelne potenzielle Verzerrungsmerkmale, sondern auch die Unterrichtsmerkmale im Fächervergleich verschieden auf die wahrgenommene allgemeine Unterrichtsqualität aus, so dass die Frage, ob die Determinanten des Schülerfeedbacks über die Fächer hinweg vergleichbar sind (vgl. Ditton 2002, S. 282; Greimel-Fuhrmann 2003, S. 261), für die gymnasiale Oberstufe zu verneinen ist. Auch in DESI erwies sich das Schülerfeedback nicht als fachübergreifend valide (vgl. Wagner 2008, S. 131).

Somit spricht der Forschungsstand zwar insgesamt dafür, dass Schülerinnen und Schüler „durchaus“ als „Auskunftsgeber über unterrichtliche Bedingungen und Lernarrangements fungieren“ können (Gruehn 2000, S. 211; vgl. hierzu u.a. Baumert et al. 2004; Greimel-Fuhrmann 2003; Wittwer 2008), dass bei der Rezeption und Verwendung der Daten aber auch die skizzierten Einschränkungen bedacht sein sollten.

4. Kriterien und Empfehlungen zur Verwendung von Schülerratings an Schulen

Eine schulische Feedbackkultur, deren Ziel es ist, dass Lehrpersonen aus den Wahrnehmungen und Urteilen ihrer Schülerinnen und Schüler etwas für die eigene Professionalitätsentwicklung lernen können, setzt zweierlei voraus: Erstens sollten die Daten das gesamte Spektrum der Wahrnehmungen und Einschätzungen der Lerngruppe zeigen und zweitens ein sachlich hinreichend differenziertes Meinungsbild der Lernenden darstellen. Damit man ein Meinungsbild erhält, das die gesamte Verteilung der Schülerwahrnehmungen und Akzeptanzurteile in einer Klasse abbildet, sollten die Schülerinnen und Schüler ihre Wahrnehmungen niedrigschwellig äußern können, was sich durch leichte Zugänglichkeit und einfache Handhabbarkeit des Erhebungsinstruments sowie Gewährleistung der Anonymität der Lernenden erreichen lässt. Die eingangs beschriebenen Feedback-Instrumente zur individuellen Rückmeldung an Lehrkräfte erfüllen diese vier Anforderungen (vollständig hinsichtlich der Verteilung innerhalb der Lerngruppe, hinreichend differenziert, niedrig-

schwellig und anonym); allerdings kommt eine solche Feedbackkultur nicht ohne periodisch wiederkehrende standardisierte Fragebogenerhebungen aus.

Während Schülerinnen und Schüler die Möglichkeit haben müssen, ihre Unterrichtseinschätzungen „ohne Ansehen ihrer Person“ abgeben zu können, kann dies für Lehrkräfte nicht gelten. Die Feedbackkultur an Schulen sollte sich zwar auch, aber nicht nur an das gesamte Kollegium als so genannte „Professionellengemeinschaft“ richten. Auch die einzelne Lehrkraft als verantwortliche Akteurin/verantwortlicher Akteur ist anzusprechen. Wenn Schülerurteile zu Evaluationszwecken erhoben und ausgewertet werden, sollte mit anderen Worten eindeutig identifizierbar sein, welches Unterrichtsangebot und welche einzelne Lehrkraft konkret Gegenstand der Wahrnehmung und Beurteilung ist. Denn dies bildet die Grundlage dafür, dass Lehrende ein individuelles Stärken- und Schwächenprofil ihrer eigenen Instruktionsqualität erhalten.

Dies schließt die spätere Aggregation der Daten auf Abteilungs- und/oder Schulebene zwar nicht aus, aber ein rein organisationsbezogenes Qualitätsmanagement, das ausschließlich mit solchen aggregierten Daten arbeitet, erfüllt das Kriterium nicht. Eine solche Praxis droht stattdessen, den individuellen Unterricht als entscheidende Ebene der Leistungserbringung in den Schulen zu verfehlen. Sie bringt – ähnlich wie Ansätze der Klimaforschung – die einzelne Lehrkraft als Adressaten des Schülerfeedbacks gewissermaßen zum Verschwinden und erschwert oder verhindert sogar, dass die einzelne Lehrperson Konsequenzen für die eigene Professionalitätsentwicklung ziehen kann. Qualitätsmanagementverfahren, die stets die gesamte Schule in den Blick nehmen, können so zwar etwaigen Widerständen entgegen, die aus der Angst vor dem personenbezogenen Vergleich resultieren, verschenken aber auch die Chancen eines solchen Vergleichs und drohen damit notorisch ihren eigenen Anspruch auf Unterrichtsentwicklung zu verfehlen (vgl. Rahn 2008).

Der Entwicklung des Unterrichts dient es hingegen, wenn Lehrende die Chance haben, ein Schülerfeedback-basiertes Stärken-Schwächen-Profil und damit eine Rückmeldung mit individueller Bezugsnorm zu erhalten. Dass Leistungsrückmeldungen, in denen die Leistungen einer Person aufgrund der eigenen Leistungsentwicklung beurteilt werden, günstig für die Lern- und Leistungsmotivation von Lernenden sind, ist in der pädagogischen Psychologie gut belegt (vgl. Rheinberg 2001). Somit liegt die Hypothese nahe, dass auch die Lern- und Professionalitätsentwicklungsprozesse von Lehrpersonen bei der Rückmeldung von Schülerurteilen mit einer individuellen Bezugsnorm unterstützt werden können. Es wäre zu erwarten, dass sich die Selbstwirksamkeitserwartungen, d.h. die Überzeugung, aufgrund eigener Lehrkompetenz gute Lehrangebote erzeugen zu können, von Lehrkräften, die zunehmend besseres Schülerfeedback erhalten, positiv entwickeln.

Schülerfeedback birgt für die Lehrenden aber auch die Chance, das eigene Fähigkeitsselbstbild mit Hilfe der Schüleraussagen zu validieren und durch den sozialen Vergleich mit anderen Lehrkräften auf seinen Realitätsgehalt zu prüfen. Dies setzt allerdings voraus, dass in den Evaluationsverfahren ein Maßstab für solche sozialen Vergleiche, d.h., eine soziale Bezugsnorm zur Verfügung steht, was – wie oben gezeigt – nicht immer üblich ist. Wenn z.B. die Schülerinnen und Schüler einer Klasse das Lehrangebot einer Lehrperson hinsichtlich der Klarheit der Unterrichtssprache auf einer sechsstufigen, an der Notenfolge angelehnten Skala mit dem Mittelwert 2,8 bewerten – wie hat man diesen Wert dann zu interpretieren? Erst der soziale Vergleich zwischen Rückmeldungen verschiedener Lerngruppen und verschiedenen Lehrpersonen erlaubt es, einen konkreten Wert als relativ gut oder schlecht zu verorten. Insofern ist fraglich, ob eine professionelle Evaluationskultur an Schulen tatsächlich auf den sozialen Vergleich verzichten kann (siehe hierzu Thiel/Ulber 2007). Bekanntlich sind solche sozialen Vergleiche in der Leistungsbeurteilung einerseits wichtig, um eine realistische Rückmeldung über den eigenen Leistungsstand zu erhalten und einen angemessenen sachlichen Leistungsstandard zu justieren. Andererseits stoßen sie aber auch auf erheblichen Widerstand bei den Lehrenden und bergen Risiken für das Fähigkeitsselbstbild, d.h., auch für die Lern- und Leistungsmotivation des pädagogischen Personals.

Wenn aber verglichen wird – das gilt nicht nur für den Vergleich von Schülerleistungen –, dann muss der Vergleich fair sein. Und dies bedeutet, dass er – angesichts der empirischen Belege – nur innerhalb der einzelnen Unterrichtsfächer vorgenommen werden sollte. Andernfalls besteht die Gefahr, dass man die Leistungen von Lehrenden, die ihre Arbeit unter schwierigen Bedingungen erbringen müssen, systematisch unterschätzt und diejenigen zusätzlich belobigt, die bereits von guten Arbeitsbedingungen wie etwa einem hohen Fachinteresse der Lerngruppe profitieren.

5. Schlussbetrachtung und Ausblick

Schülerfeedback – so ist zu resümieren – sollte für die Lernenden anonym, hinreichend differenziert und mit Hilfe sorgfältig entwickelter Instrumente erhoben werden. Geeignete Verfahren stehen hierzu zur Verfügung. Individuelle Vergleiche sollten durch – faire – soziale Vergleiche, die aufgrund der oben skizzierten Befunde *innerhalb* von Unterrichtsfächern, aber nicht zwischen Unterrichtsfächern möglich sind, ergänzt werden können. Dies legt allerdings Modifikationen in den Rückmeldeverfahren einiger der bis dato implementierten Instrumente nahe.

Mehrheitlich – darauf ist abschließend ausdrücklich hinzuweisen – birgt das Schülerfeedback für die einzelne Lehrkraft, zumindest was die allgemeine Unterrichtseinschätzung betrifft, keine Überraschungen. Wen sollte das auch wundern:

Lehrpersonen kennen ihre Klassen, so dass der Vergleich zwischen dem tatsächlichen und dem von der Lehrkraft erwarteten Feedback etwa in der Studie zu den Determinanten des Schülerfeedbacks in der Oberstufe in 70 Prozent der Fälle nur geringe Abweichungen zeigte. Rund 30 Prozent der Lehrenden erlebten allerdings eine Überraschung, darunter ca. 12 Prozent eine positive, d.h., die Klasse beurteilte die Unterrichtsqualität besser als erwartet. Bedenkt man zudem, und dies sei an dieser Stelle ebenfalls betont, dass die Einschätzungen der Schülerinnen und Schüler zur Qualität der Lehrarbeit und des Unterrichts ohnehin überwiegend positiv, die Lernenden mit der Unterrichtsqualität also zufrieden sind, dann stellt die Erhebung von Schülerfeedback eine der Gelegenheiten dar, bei denen sich Lehrkräfte in ihrem Beruf als erfolgreich erleben können. In einem Beruf, der im Ruf steht, durch ein notorisches Defizit an sozialer Anerkennung gekennzeichnet zu sein, könnte das ein nicht zu unterschätzender Gewinn sein.

Wenn das Schülerfeedback aber, wie in der Oberstufenstudie bei rund 17 Prozent der Klassen, unerwartet negativ ausfällt, die Klassen der Lehrperson also einen Änderungsbedarf signalisieren, den sie selbst nicht wahrgenommen hat, kann dies im besten Fall den Impuls für einen Entwicklungsprozess geben – nicht mehr, aber auch nicht weniger.

Literatur und Internetquellen

- Aleamoni, L.M. (1999): Student Rating Myths versus Research Facts from 1924 to 1998. In: *Journal of Personnel Evaluation in Education* 13, H. 2, S. 153-166.
- Baumert, J./Kunter, M. (2006): Stichwort: Professionelle Kompetenz von Lehrkräften. In: *Zeitschrift für Erziehungswissenschaft* 9, H. 4, S. 469-520.
- Baumert, J./Kunter, M./Brunner, M./Krauss, S./Blum, W./Neubrand, M. (2004): Mathematikunterricht aus Sicht der PISA-Schülerinnen und -Schüler und ihrer Lehrkräfte. In: Prenzel, M./Baumert, J./Blum, W./Lehmann, R./Leutner, D./Neubrand, M./Pekrun, R./Rolf, H.-G./Rost, J./Schiefele, U. (Hrsg.): *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs*. Münster u.a.: Waxmann, S. 314-354.
- Böttcher, W./Grewe, C.M. (2010): Eine Untersuchung zur Wirksamkeit der studentischen Lehrveranstaltungskritik am Beispiel der Westfälischen Wilhelms-Universität Münster. In: Pohlenz, P./Oppermann, A. (Hrsg.): *Lehre und Studium professionell evaluieren: Wie viel Wissenschaft braucht die Evaluation?* Bielefeld: Universitäts Verlag Webler, S. 73-82.
- Clausen, M. (2002): Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität. Münster u.a.: Waxmann.
- Ditton, H. (2002): Lehrkräfte und Unterricht aus Schülersicht. Ergebnisse einer Untersuchung im Fach Mathematik. In: *Zeitschrift für Pädagogik* 48, H. 2, S. 262-286.
- El-Hage, N. (1996): *Lehrevaluation und studentische Veranstaltungskritik. Projekte, Instrumente und Grundlagen*. Bonn: Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie.

- Greimel, B. (2002): Lehrerevaluation durch Beurteilung der Lernenden – eine Analyse des Standes der Evaluationsforschung. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 98, H. 2, S. 197-224.
- Greimel-Fuhrmann, B. (2003): Evaluation von Lehrerinnen und Lehrern. Einflussgrößen auf das Gesamturteil von Lernenden. Innsbruck/Wien/München: Studienverlag.
- Greimel-Fuhrmann, B./Geyer, A. (2005): Die Wirkung von Interesse und Sympathie auf die Gesamtbeurteilung in der Lehrerevaluation. Direkte und indirekte Effekte unter Berücksichtigung des Lehrverhaltens. In: Empirische Pädagogik 19, H. 2, S. 103-120.
- Gruehn, S. (2000): Unterricht und schulisches Lernen. Schüler als Quellen der Unterrichtsbeschreibung. Münster u.a.: Waxmann.
- Hattie, J. (2013): Lernen sichtbar machen. Überarb. deutschsprachige Ausgabe von „Visible Learning“, besorgt von Wolfgang Beywl und Klaus Zierer. Baltmannsweiler: Schneider Verlag Hohengehren.
- Helmke, A./Helmke, T. (2014): Unterrichtsanalyse mit EMU (Evidenzbasierte Methoden der Unterrichtsentwicklung). In: Journal für Schulentwicklung 18, H. 1, S. 55-57.
- Helmke, A./Helmke, T./Lenke, G./Pham, G./Praetorius, A.-K./Schrader, F.-W./Ade-Thurow, M. (2015): Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung. Version 5.0 (15.01.2015). URL: <http://unterrichtsdiagnostik.info/>; Zugriffsdatum: 28.01.2016.
- Kämpfe, N. (2009): Schülerinnen und Schüler als Experten für Unterricht. In: Die Deutsche Schule 101, H. 2, S. 149-163.
- Klieme, E. (2008) (Hrsg.): Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie. Weinheim u.a.: Beltz.
- KMK (Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (2014): Standards für die Lehrerbildung: Bildungswissenschaften (Beschluss der Kultusministerkonferenz vom 16.12.2004 i.d.F. vom 12.06.2014). URL: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf; Zugriffsdatum: 28.01.2016.
- Kunter, M. (2005): Multiple Ziele im Mathematikunterricht. Münster u.a.: Waxmann.
- Lüdtke, O./Trautwein, U./Kunter, M./Baumert, J. (2006): Analyse von Lernumwelten. Ansätze zur Bestimmung der Reliabilität und Übereinstimmung von Schülerwahrnehmungen. In: Zeitschrift für pädagogische Psychologie 20, H. 1, S. 85-96.
- Marsh, H.W./Roche, L.A. (1997): Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility. In: American Psychologist 52, H. 11, S. 1187-1197.
- Rahn, S. (2008): Qualität, Qualitätssicherung und Qualitätsentwicklung im beruflichen Schulwesen – Charakteristika und Reichweite aktueller Konzepte. In: bwp Spezial Hochschultage Berufliche Bildung. URL: http://www.bwpat.de/ht2008/ft06/rahn_ft06-ht2008_spezial4.pdf; Zugriffsdatum: 28.01.2016.
- Rahn, S./Gruehn, S./Böttcher, W. (2015): Determinanten des Schülerfeedbacks: eine fächervergleichende Analyse von Schüleraussagen zur Unterrichtsqualität an allgemeinbildenden und beruflichen Gymnasien. Unveröffentlichter Abschlussbericht des gleichnamigen DFG-Projekts (RA 2380/1-1|GR 1951/2-1).
- Rheinberg, F. (2001): Bezugsnormen und schulische Leistungsbeurteilung. In: Weinert, F.E. (Hrsg.): Leistungsmessungen in Schulen. Weinheim: Beltz, S. 59-71.
- Rindermann, H. (2001): Lehrerevaluation. Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierter Unterrichts. Landau: Empirische Pädagogik.
- Spiel, C./Gössler, M. (2000): Zum Einfluss von Biasvariablen auf die Bewertung universitärer Lehre durch Studierende. In: Zeitschrift für Pädagogische Psychologie 14, H. 1, S. 38-47.

- ten Venne, M./Nachtigall, C./Kämpfe-Hargrave, N. (2010): SEfU – Schüler als Experten für Unterricht. In: Informationsschrift Recht und Bildung des Instituts für Bildungsforschung und Bildungsrecht 7, H. 1, S. 17-21.
- Thiel, F./Ulber, D. (2007): Unterrichtsentwicklung durch Evaluation. In: Bauer, K.-O. (Hrsg.): Evaluation an Schulen. Theoretischer Rahmen und Beispiele guter Evaluationspraxis. Weinheim u.a.: Juventa, S. 163-186.
- Wagner, C. (2009): Wissenschaftlich und dennoch praktikabel? Nutzbarkeit von Daten aus Fragebogenerhebungen im Rahmen der internen Evaluation an Schulen. In: Münk, D./Deißinger, T./Tenberg, R. (Hrsg.): Forschungserträge aus der Berufs- und Wirtschaftspädagogik. Probleme, Perspektiven, Handlungsfelder und Desiderata der beruflichen Bildung in der Bundesrepublik Deutschland, in Europa und im internationalen Raum. Opladen: Budrich, S. 40-49.
- Wagner, C. (2010): Unterrichtsentwicklung durch Evaluation am Beispiel des „Netzwerk Schülerbefragung“. In: Die berufsbildende Schule 62, H. 10, S. 291-297.
- Wagner, W. (2008): Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht. Am Beispiel der Studie DESI (Deutsch Englisch Schülerleistungen International) der Kultusministerkonferenz. Koblenz-Landau: Universität, FB Psychologie.
- Wittwer, J. (2008): What Influences the Agreement among Student Ratings of Science Instruction? In: Prenzel, M./Baumert, J. (Hrsg.): Vertiefende Analysen zu PISA 2006. Wiesbaden: VS, S. 205-220.

Christoph Fuhrmann, M.A., geb. 1963, wissenschaftlicher Mitarbeiter am Institut für Bildungsforschung in der School of Education der Bergischen Universität Wuppertal.
E-Mail: christoph.fuhrmann@uni-wuppertal.de

Miriam Sharon Keune, Dr., geb. 1980, wissenschaftliche Mitarbeiterin am Institut für Bildungsforschung in der School of Education der Bergischen Universität Wuppertal.
E-Mail: miriam.keune@uni-wuppertal.de

Sylvia Rahn, Prof. Dr., geb. 1967, Professorin für Berufsbildungsforschung am Institut für Bildungsforschung in der School of Education der Bergischen Universität Wuppertal.
E-Mail: sylvia.rahn@uni-wuppertal.de

Anschrift: Bergische Universität Wuppertal, Gaußstraße 20, 42119 Wuppertal

Sabine Gruehn, Prof. Dr., geb. 1966, Professorin für Erziehungswissenschaft mit dem Schwerpunkt Schultheorie/Schulforschung an der Westfälischen Wilhelms-Universität Münster.
E-Mail: sabine.gruehn@uni-muenster.de

Anschrift: Westfälische Wilhelms-Universität Münster, Bispinghof 5/6, 48143 Münster