Praetorius, Anna-Katharina

Einschätzung von Unterrichtsqualität durch externe Beobachterinnen und Beobachter. Eine kritische Betrachtung der aktuellen Vorgehensweise in der Schulpraxis

Beiträge zur Lehrerbildung 31 (2013) 2, S. 174-185

BEITRÄGE ZUR LEHRERBILDUNG, 31 (2), 2013

Einschätzung von Unterrichtsqualität durch externe Beobachterinnen und Beobachter – Eine kritische Betrachtung der aktuellen Vorgehensweise in der Schulpraxis

Anna-Katharina Praetorius

Zusammenfassung Die Einschätzung von Unterricht durch externe Beobachterinnen und Beobachter ist seit jeher konstitutiver Bestandteil des Lehrberufs. In welchem Ausmass solche Beobachtereinschätzungen zuverlässige Aussagen über den Unterricht von Lehrkräften darstellen, wurde bislang jedoch kaum empirisch überprüft. Der vorliegende Beitrag gibt einen Überblick über Studien, die infrage stellen, inwiefern es sich bei Beobachtereinschätzungen um den «Königsweg» zur Erfassung von Unterrichtsqualität handelt. Der Beitrag schliesst mit Folgerungen für die Unterrichtspraxis. Dabei wird unter anderem die Notwendigkeit des Einsatzes mehrerer Beobachterinnen und Beobachter sowie einer Beobachtung von längeren Zeiträumen diskutiert.

Schlagwörter Unterrichtsqualität – Messung – Beobachterinnen und Beobachter – Ratings

Measuring Instructional Quality with Observer Ratings – A Critical Examination of Current Practice in Schools

Abstract Assessments of teaching by external observers have always been a constitutive part of the teaching profession. Still, the extent to which these ratings allow reliable statements about the observed teachers' instruction has rarely been empirically investigated. Against this background, the following article provides an overview of studies which challenge the assumption that observer ratings are the ideal method of measuring instructional quality. It concludes with some implications for instructional practice. Among other things, the discussion touches on the need to include several observers and to cover longer periods of observation.

Keywords instructional quality – measurement – observer – ratings

1 Einleitung

Unterrichtsbeurteilungen und -rückmeldungen sind seit jeher konstitutiver Bestandteil des Lehrberufs: Bereits in den Schulpraktika im Studium erhalten Lehramtsstudierende Rückmeldungen zu Stärken und Schwächen der von ihnen gehaltenen Unterrichtsstunden. Im Referendariat kommt zu diesen Rückmeldungen auch die (benotete) Bewertung von Unterrichtsstunden hinzu. Doch auch im Anschluss an die Ausbildung sind Lehrkräfte immer wieder mit Unterrichtseinschätzungen externer Beobachterinnen und Beobachter konfrontiert, u.a. von Schulleiterinnen und Schulleitern sowie Schulinspek-

torinnen und Schulinspektoren, aber auch bei gegenseitigen Hospitationen durch Kolleginnen und Kollegen. Diese Unterrichtsbesuche können zum einen zu Forschungszwecken erfolgen und zum anderen im Kontext der Lehrpersonenaus- und -fortbildung zu Zwecken der summativen oder formativen Evaluation eingesetzt werden. Bei summativen Einschätzungen steht die Bewertung (und Benotung) von Unterricht im Vordergrund. Dies spielt in der aktuellen Lehrpersonenausbildung eine grosse Rolle: Die Benotung von Unterrichtsstunden dient u.a. dazu, die am besten geeigneten Lehrkräfte für eine begrenzte Stellenanzahl auszuwählen. Bei formativen Einschätzungen hingegen steht die Rückmeldung zu Unterricht im Vordergrund. Solche Rückmeldungen zu den unterrichtlichen Stärken und Schwächen von Lehrkräften sollen als Basis für die Weiterentwicklung von Unterricht dienen – dies wird auch als evidenzbasierte Unterrichtsentwicklung bezeichnet (vgl. z.B. Altrichter, Messner & Posch, 2004; Helmke, 2009; Horster & Rolff, 2001). Warum eine solche evidenzbasierte Entwicklung von Unterricht wichtig ist, ist unmittelbar einsichtig: Nur wenn Lehrpersonen über den Ist-Stand informiert sind, können sie ihren Unterricht passgenau weiterentwickeln (vgl. dazu auch Helmke & Lenske, 2013 [in diesem Heft]).

Sowohl zu Forschungszwecken als auch in der Lehrpersonenaus- und -fortbildung werden zur Erfassung von Unterrichtsqualität oftmals hochinferente Ratingskalen verwendet. Bei diesen Skalen handelt es sich um eine Zusammenstellung von Items, die Schlussfolgerungen über das direkt Sichtbare hinaus erfordern. Ein Beispiel für ein solches hochinferentes Item lautet: «Mit Schülerinnen- und Schülerbeiträgen ist die Lehrkraft wertschätzend umgegangen.» Ein solcher wertschätzender Umgang kann nicht direkt aus den Verhaltensweisen der Lehrkraft abgelesen werden, sondern muss anhand diverser Indikatoren erschlossen werden (z.B. «Die Lehrkraft würdigt unvollständige/halb richtige Schülerinnen- und Schülerbeiträge, indem sie die richtigen Aspekte aufgreift»). Zur Beurteilung der Items werden dann vorgegebene Antwortmöglichkeiten (z.B. trifft nicht zu – trifft eher nicht zu – trifft eher zu – trifft zu) verwendet. Grund für den bevorzugten Einsatz solcher hochinferenter Items ist, dass diese Art von Items eine höhere Erklärungskraft für diverse Aussenkriterien (z.B. für die Leistungsentwicklung von Schülerinnen und Schülern) hat als niedriginferente Items, solche Items also, deren Beantwortung keine oder kaum Schlussfolgerungen über das Sichtbare hinaus erfordert (z.B. die Anzahl an Meldungen von Schülerinnen und Schülern) (Clausen, 2002; Sommer, 2011). Ein Nachteil von hochinferenten Items ist die Tatsache, dass bei deren Beantwortung Beurteilerfehler eine deutlich grössere Rolle spielen als bei niedriginferenten Items (vgl. auch Clausen, 2002).

In welchem Ausmass externe Beobachterinnen und Beobachter auf der Basis hochinferenter Ratinginstrumente zuverlässige Aussagen über den Unterricht von Lehrkräften treffen, wurde bislang kaum thematisiert (Hill et al., 2012; Pietsch & Tosana, 2008; Praetorius, in Druck). Dies ist erstaunlich, da die Vergabe von Stellen basierend auf Unterrichtsbeurteilungen nur dann sinnvoll ist, wenn diese Beurteilungen auch zuverlässig sind. Und auch evidenzbasierte Unterrichtsentwicklung ist lediglich in dem Ausmass

möglich, in dem die Informationen über den Ist-Stand den Ist-Stand tatsächlich widerspiegeln.

Im Folgenden wird zunächst auf die in der Literatur genannten Vorteile von externen Beobachtereinschätzungen eingegangen (Abschnitt 2). Anschliessend sollen Befunde der empirischen Unterrichtsforschung dargestellt werden, die zum Teil kritisch beleuchten, ob die Zuverlässigkeit der Unterrichtseinschätzungen externer Beobachterinnen und Beobachter tatsächlich gegeben ist (Abschnitt 3). Der Beitrag schliesst mit Schlussfolgerungen aus den berichteten Ergebnissen (Abschnitt 4).

2 Vorteile externer Beobachtereinschätzungen von Unterricht

Externe Beobachtereinschätzungen von Unterricht werden oft als «Königsweg» zur Erfassung von Unterrichtsqualität beschrieben (Helmke, 2009, S. 288; vgl. auch Clare, Valdés, Pascal & Steinberg, 2001; Pianta & Hamre, 2009; Petko, Waldis, Pauli & Reusser, 2003). Dafür gibt es diverse Gründe:

– Erstens durchlaufen Beobachterinnen und Beobachter – zumindest in der Unterrichtsforschung und im Kontext der Schulinspektion – in der Regel ein Training für die Einschätzung von Unterricht. Diese Schulung sollte es ihnen erlauben, die hohe Komplexität von Unterricht angemessen zu erfassen (Helmke, 2009; Petko et al., 2003).
– Zweitens wird argumentiert, dass Beobachterinnen und Beobachter – im Gegensatz zu Lehrkräften sowie Schülerinnen und Schülern – nicht Teil des Unterrichtsgeschehens sind und dadurch zum einen eher objektive Unterrichtseinschätzungen abgeben können und zum anderen ihre kognitiven Ressourcen vollständig für die Wahrnehmung des Unterrichts zur Verfügung haben (Rakoczy, 2008; Waldis et al., 2010).
– Drittens wird darauf verwiesen, dass Beobachterinnen und Beobachter in der Regel etliche Lehrkräfte einschätzen und dadurch in einem hohen Masse Vergleichsmöglichkeiten haben, während Lehrkräfte selbst vor allem ihren eigenen Unterricht kennen (Clausen, 2002; Rakoczy, 2008).

3 Nachteile externer Beobachtereinschätzungen von Unterricht

Aufgrund der in Abschnitt 2 genannten Vorteile werden externe Beobachtereinschätzungen oft als vergleichsweise «objektive» Möglichkeit zur Erfassung von Unterrichtsqualität beschrieben und als gegenüber anderen Methoden (vor allem Lehrkraftselbsteinschätzungen und Einschätzungen von Schülerinnen und Schülern) zu bevorzugen dargestellt (z.B. Clare et al., 2001; Helmke, 2009; Rakoczy, 2008). Studien zeigen jedoch, dass auch Beobachtungen von unbeteiligten, externen Personen nicht einfach «die Realität» und damit die tatsächliche Unterrichtsqualität abbilden (vgl. auch Clau-

sen, 2002; Hill, Charalambous & Kraft, 2012; Waldis et al., 2010). Dies liegt zum einen daran, dass auch externe Beobachterinnen und Beobachter diversen Beurteilerfehlern unterliegen (Abschnitt 3.1), zum anderen aber auch daran, dass externe Beobachtungen Restriktionen in Bezug auf die zur Verfügung stehenden Informationen unterliegen (Abschnitt 3.2).

3.1 Beurteilereffekte bei externen Beobachtereinschätzungen

Durchsucht man die Literatur in Bezug auf die Frage, wie zuverlässig bzw. genau die Einschätzungen externer Beobachterinnen und Beobachter sind, so finden sich erstaunlich wenige Studien. Die vorhandenen Untersuchungen (Clausen et al., 2003; Hill et al., 2012; Kobarg & Seidel, 2005; Matsumura, Garnier, Pascal & Valdés, 2002; Newton, 2010; Pietsch & Tosana, 2008; Praetorius, in Druck; Strong et al., 2011) zeigen, dass zwischen 0 und 41% der Variation in Ratings – und damit teilweise ein sehr hoher Anteil – auf die Beobachterinnen und Beobachter anstelle auf das zu messende Merkmal zurückzuführen sind. In der Untersuchung von Clausen et al. (2003) weist beispielsweise das Merkmal «Disziplinprobleme» einen sehr geringen Raterhaupteffekt (d.h. Unterschiede zwischen Raterinnen und Ratern in ihrer Strenge bzw. Milde) von 3% auf, das Merkmal «Klarheit» mit 41% hingegen einen sehr hohen entsprechenden Effekt.

Diesem zum Teil hohen Anteil an Ratereffekten können unterschiedlichste Ursachen zugrunde liegen. Wegen der begrenzten menschlichen Wahrnehmung unterscheiden sich Personen bereits darin, was sie beobachten (vgl. auch Lemons & Helsing, 2010). Aber auch in der Verarbeitung des Beobachteten treten deutliche Unterschiede auf (vgl. z.B. Bradburn, 2004): So werden Items aufgrund der Uneindeutigkeit von Sprache uneinheitlich interpretiert. Es werden unterschiedliche Informationen aus dem Gedächtnis abgerufen, diese Informationen dann verschieden gewichtet und kombiniert, und auch die Wahl einer geeigneten Antwortkategorie kann unterschiedlich ausfallen (vgl. auch Praetorius, in Druck). Besonders deutlich treten solche Unterschiede auf, wenn Beobachterinnen und Beobachter kein gemeinsames Training durchlaufen haben. Dies wird beispielsweise in einer Untersuchung von Praetorius, Lenske und Helmke (2010) deutlich, in der sich ungeschulte Lehramtsstudierende eine Unterrichtsstunde ansahen und anschliessend in einem Einzelinterview dazu befragt wurden. In Bezug auf ein ausgewähltes Item finden sich in Abbildung 1 die Aussagen derjenigen vier Studierenden, die bei diesem Item «stimme eher zu» gewählt haben. Wie aus den Begründungen deutlich wird, interpretieren die Studierenden den im Item enthaltenen Begriff «Diskussion» sehr unterschiedlich, beginnend bei einer Gleichsetzung mit dem Stillsein der Schülerinnen und Schüler, über deren Meldeverhalten bis hin zum Äussern der eigenen Meinung.

Um solche Beobachtereffekte so weit wie möglich zu minimieren, werden externe Beobachterinnen und Beobachter in der Unterrichtsforschung sowie im Kontext der Schulinspektion trainiert. Inwiefern diese Trainings tatsächlich dazu führen, dass die Einschätzungen in einem höheren Ausmass zuverlässig sind (= Reliabilität) und verstärkt

Item: «Die Schülerinnen und Schüler haben sich an der Diskussion im Unterricht beteiligt.»

Antwortkategorie bei allen Antworten: «stimme eher zu»

- «Die meisten schon oder waren zumindest leise und haben zugehört.»
- «Ja, nicht unbedingt jeder, aber doch sehr viele. Es gingen glaube ich die Finger nach oben, also ich stimme eher zu.»
- «Es haben sich eigentlich relativ viele Schüler dran beteiligt.»
- «Ja klar, die haben sich immer gemeldet. Diskussion, wurde es dann zum Schluss zur Diskussion – war es eine Diskussion oder war es einfach nur eine Wiedergabe? Also zum Schluss sollten sie ihre Meinung äussern, das war eine Diskussion, also stimme ich eher zu.»

Abbildung 1: Unterschiedliche Iteminterpretationen von ungeschulten Raterinnen und Ratern (Praetorius et al., 2010).

das messen, was sie messen sollen (= Validität), ist bislang kaum untersucht worden. Eine Untersuchung von Praetorius et al. (2012) weist darauf hin, dass die Validität der Einschätzungen durch entsprechende Trainings erhöht werden kann: Trainierte Raterinnen und Rater nehmen bei ihren Einschätzungen in einem stärkeren Ausmass Bezug auf Indikatoren, die laut Ratingmanual dem zu messenden Konstrukt entsprechen, und weisen zudem differenziertere Auseinandersetzungen mit den jeweiligen Items auf als untrainierte Raterinnen und Rater. In derselben Studie zeigt sich allerdings auch, dass die Reliabilität (in dieser Untersuchung gemessen über den relativen Generalisierbarkeitskoeffizienten im Rahmen der Generalisierbarkeitstheorie, vgl. Brennan, 2001) bei geschulten Raterinnen und Ratern entgegen den theoretischen Annahmen nicht höher ausfiel als bei ungeschulten Raterinnen und Ratern – und zum Teil sogar geringer ausgeprägt war als bei ungeschulten Lehramtsstudierenden (vgl. Abbildung 2). Ähnliche Ergebnisse zeigen sich auch in den wenigen weiteren Studien zu diesem Thema (z.B. Strong et al., 2011). Die Ergebnisse der Untersuchung von Praetorius et al. (2012) wei-

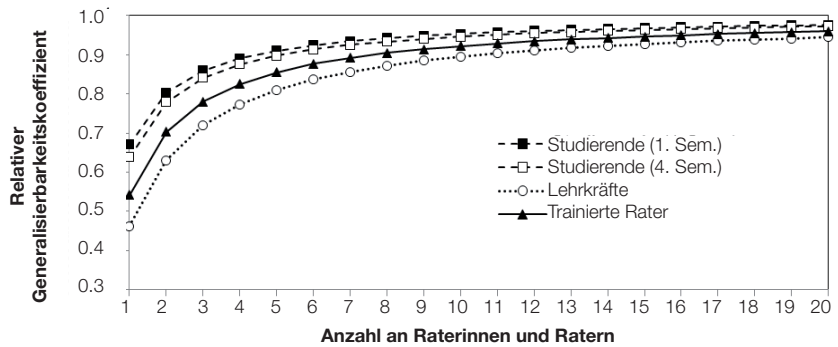

Abbildung 2: Entscheidungsstudien zur Unterrichtsdimension «Schülerorientierung» (Praetorius, in Druck, S. 125).

sen darauf hin, dass beispielsweise bei der Erfassung der Unterrichtsqualitätsdimension «Schülerorientierung» für eine Messgenauigkeit von .70 zwei Lehramtsstudierende des ersten Semesters, drei Lehramtsstudierende des vierten Semesters, drei Lehrkräfte oder zwei geschulte Raterinnen und Rater notwendig sind. Mit einer einzelnen Beobachterin oder einem einzelnen Beobachter – geschult oder ungeschult – lässt sich demnach keine hinreichende Messgenauigkeit erreichen. Möchte man individualdiagnostische Einschätzungen treffen, sind sogar Reliabilitätskoeffizienten von .90 oder höher erforderlich. Um einen solchen Koeffizienten zu erreichen, müssen deutlich mehr Raterinnen und Rater eingesetzt werden (z.B. elf ungeschulte Lehrkräfte oder acht trainierte Raterinnen und Rater).

3.2 Die eingeschränkte Informationsgrundlage externer Beobachterratings

Neben dem Auftreten von Beobachtereffekten werden Beobachterratings auch dahin gehend kritisiert, dass die den Beobachterinnen und Beobachtern zur Verfügung stehende Informationsgrundlage sehr beschränkt ist. So beobachten sie Unterricht in der Regel nur im Rahmen eines sehr kurzen Zeitraums (Abschnitt 3.2.1), Informationen über den Klassenkontext und die einzelnen Schülerinnen und Schüler stehen ihnen nur begrenzt oder gar nicht zur Verfügung (Abschnitt 3.2.2) und Lehrkräfte sowie Schülerinnen und Schüler verhalten sich unter Umständen unter Beobachtung anders als im alltäglichen Unterricht (Abschnitt 3.2.3).

3.2.1 Kurzer Beobachtungszeitraum

Lehrkräfte sowie Schülerinnen und Schüler können Aussagen über Unterricht bezogen auf einen langen Zeitraum treffen. Dies ist für Beobachterinnen und Beobachter in der Regel nicht möglich, da in der Unterrichtsforschung meist ein bis zwei Unterrichtsstunden pro Lehrkraft eingeschätzt werden (Übersicht in Praetorius, in Druck). Bei kollegialen Hospitationen, Besuchen durch Schulleitungsmitglieder oder bei Unterrichtsversuchen im Rahmen des Referendariats sind es üblicherweise ebenfalls maximal ein bis zwei Stunden pro Person, die beobachtet werden. Schulinspektorinnen und Schulinspektoren wie auch Schulleitungsmitglieder besuchen zum Teil sogar nur 20 bis 30 Minuten einer Unterrichtsstunde (vgl. z.B. Hill et al., 2012; Pietsch & Tosana, 2008). Einige Autorinnen und Autoren stellen daher infrage, ob auf der Basis eines solch kurzen Zeitraums Aussagen über die Qualität des Unterrichts einer Lehrkraft im Allgemeinen überhaupt möglich sind (Berliner, 2005; Brophy, 2006; Stigler et al., 1999). Tatsächlich zeigen empirische Untersuchungen (z.B. Hill et al., 2012; Newton, 2010; Praetorius, in Druck), dass bei manchen Unterrichtsqualitätsmerkmalen die Stabilität so gering ist, dass weitaus mehr Unterrichtszeit beobachtet werden müsste, um eine stabile Einschätzung dieser Merkmale zu erhalten. So zeigte sich beispielsweise in der Untersuchung von Hill et al. (2012), dass keine der drei untersuchten Dimensionen – Reichhaltigkeit der mathematischen Inhalte, Fehler und Unpräzision seitens der Lehrkraft sowie die Partizipation der Schülerinnen und Schüler an Begründungen – beim Einsatz von zwei Raterinnen und Ratern im Rahmen einer Unterrichtsstunde zuverlässig erfasst werden

konnte. Die Untersuchung von Praetorius (in Druck) weist darauf hin, dass für eine stabile Erfassung der Unterrichtsdimension «kognitive Aktivierung» sogar neun Unterrichtsstunden pro Lehrkraft notwendig sind. Die genannte Untersuchung zeigt jedoch auch, dass es Merkmale gibt, die so stabil sind, dass eine einzelne Unterrichtsstunde für deren Einschätzung ausreicht; die Klassenführung von Lehrkräften ist hierfür ein Beispiel.

3.2.2 Fehlende Kontextinformationen

In den letzten Jahren hat sich in der Unterrichtsforschung ein Modell etabliert, das unter dem Namen «Angebots-Nutzungs-Modell» bekannt wurde (Abbildung 3; vgl. Helmke, 2009). Dieses Modell besagt, dass Unterricht immer nur ein Angebot sein kann, das von den Schülerinnen und Schülern genutzt werden muss und nur bei adäquater Nutzung zu Lerngewinnen führt. Es wird dabei davon ausgegangen, dass Unterricht immer an die jeweilige Klasse (u.a. an ihren Leistungsstand) angepasst werden sollte. Unterricht ist damit nicht nur abhängig von der jeweiligen Lehrkraft, sondern auch von den Schülerinnen und Schülern (u.a. von deren Lernpotenzial) und von vielen weiteren Merkmalen (u.a. von diversen Kontextmerkmalen). Solche zusätzlichen Informationen über die Schülerinnen und Schüler und den Kontext liegen externen Beobachterinnen und Beobachtern in der Regel jedoch nicht oder nur begrenzt vor (Clausen, 2002; Waldis et al., 2010).

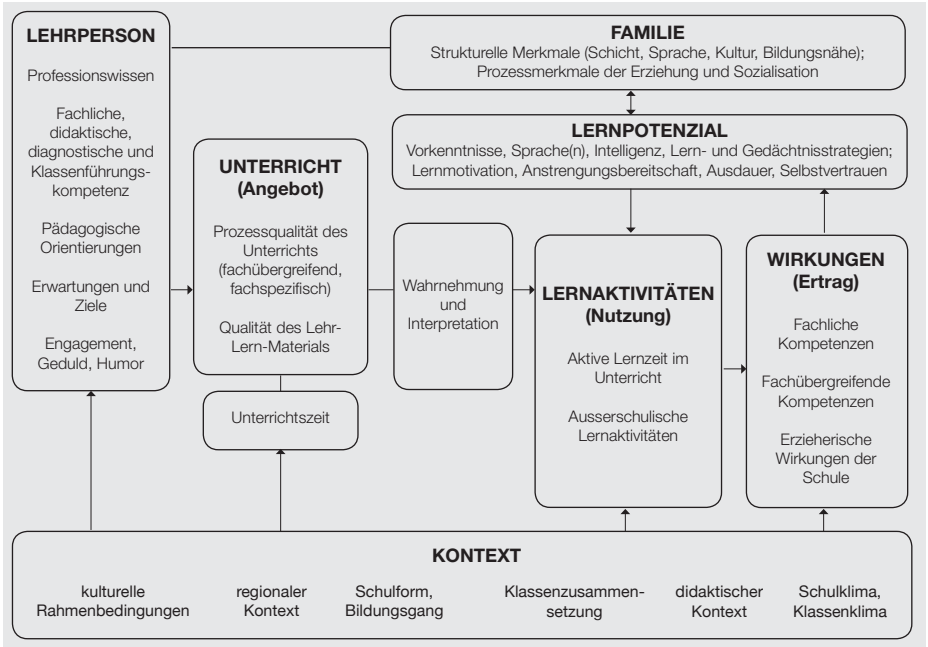

Abbildung 3: Angebots-Nutzungs-Modell des Unterrichts (Helmke, 2009, S. 73).

3.2.3 Reaktivitätseffekte

Wissen Lehrkräfte darum, dass eine bestimmte Unterrichtsstunde beobachtet werden soll, bereiten sie diese Stunde unter Umständen besonders gut vor oder setzen andere Methoden ein als in ihrem regulären Unterricht (Stigler, 1998; Stigler et al., 1999; Waldis et al., 2010). Diese Problematik wird als Reaktivitätseffekt bezeichnet: Personen verhalten sich unter Beobachtung anders, als sie es sonst tun würden. Stigler (1998, S. 141) beschreibt beobachtete Unterrichtsstunden daher als «[a] somewhat idealized version of what the teacher normally does in the classroom». Helmke (2009) hingegen verweist darauf, dass sich bestimmte Verhaltensweisen von Lehrkräften, insbesondere Routinen, kurzfristig kaum verändern lassen und daher auch beobachteter Unterricht einen guten Einblick in den regulären Unterricht bieten sollte. Beide Vermutungen, diejenige von Stigler (1998) sowie diejenige von Helmke (2009), scheinen plausibel zu sein. Welche tatsächlich besser auf Unterrichtsbeobachtungen zutrifft, ist ungeklärt: Bislang existieren keine empirischen Untersuchungen zu Reaktivitätseffekten in Bezug auf die Beobachtung von Unterricht.

4 Mangelnde Zuverlässigkeit von Beobachtereinschätzungen: Schlussfolgerungen

Externe Beobachterinnen und Beobachter werden oft als besonders geeignet zur Messung von Unterrichtsqualität beschrieben (Clare, Valdés, Pascal & Steinberg, 2001; Helmke, 2009; Pianta & Hamre, 2009). Bei näherer Betrachtung zeigen sich jedoch diverse Einschränkungen von Beobachterratings. So verfügen Beobachterinnen und Beobachter in der Regel lediglich über eine eingeschränkte Informationsgrundlage. Dies trägt dazu bei, dass ihre Einschätzungen die Unterrichtsqualität von Lehrkräften nur begrenzt abbilden können. Die in Abschnitt 3 dargestellten Studien zeigen zudem, dass externe Beobachtereinschätzungen zum Teil deutliche Ratereffekte aufweisen. Dies gilt in verstärktem Masse für ungeschulte Beobachterinnen und Beobachter wie beispielsweise Lehrkräfte.

Geht es Lehrkräften bei gegenseitigen Hospitationen lediglich darum, sich über die eigenen Vorstellungen von gutem Unterricht Gedanken zu machen und diese auszutauschen (vgl. Horster & Rolff, 2001), ist eine geringe Zuverlässigkeit der Einschätzungen nur von nachrangiger Bedeutung. Sollen die Einschätzungen hingegen als Unterrichtsdiagnose dienen und damit als Ist-Stand-Bestimmung die Grundlage für Unterrichtsentwicklung bilden, müssen hohe Massstäbe an die Messqualität gelegt werden. Verzerrungen in den Einschätzungen können sonst im ungünstigsten Fall zu einer Verschlechterung anstelle einer Verbesserung von Unterricht führen (vgl. Praetorius et al., 2012).

Welche Schlussfolgerungen lassen sich daraus nun für die Unterrichtspraxis ziehen? Müller und Pietsch (2011) folgern aus der Messfehlerbehaftetheit der Einschätzungen

von Schulinspektorinnen und Schulinspektoren, dass die in diesem Kontext üblichen Aussagen auf Schulebene gerechtfertigt seien, Individualdiagnostik hingegen vermieden werden sollte. Für die Unterrichtspraxis ist eine solche Schlussfolgerung insofern problematisch, als Unterrichtsentwicklung und auch Bewertungen von Lehrkräften Ist-Stand-Bestimmungen auf Individualebene erfordern. Da davon auszugehen ist, dass Einschätzungen von Unterrichtsqualität aus dem Bauch heraus noch ungenauer und zudem invalide sind (Helmke et al., 2011), erscheint ein Verzicht auf Unterrichtseinschätzungen mittels standardisierter Instrumente nicht sinnvoll. Über mögliche Alternativen, die in der Unterrichtspraxis umsetzbar sind, ist bislang jedoch kaum etwas bekannt. Zieht man die berichteten Befunde heran, erweisen sich für die Umsetzung im schulischen Kontext mehrere Alternativen als gangbar – wenngleich diese jeweils mit einigem Aufwand verbunden sind:

1) Unterrichtsbewertungen, aber auch Rückmeldungen im Rahmen kollegialer Hospitationen beruhen gegenwärtig oft auf dem Urteil einer einzelnen Person. Diese Urteile enthalten in der Folge in hohem Masse auch Informationen über die Urteilerin oder den Urteiler und sind daher invalide. Um diese Problematik abzumildern, sollten – zumindest wenn es nicht lediglich um den Austausch von Vorstellungen über guten Unterricht unter Kolleginnen und Kollegen geht – stets mehrere Beobachterinnen und Beobachter eine Unterrichtsstunde beurteilen und deren Einschätzungen im Anschluss daran gemittelt werden, um individuelle Fehleinschätzungen zu minimieren.
2) Da nicht alle Unterrichtsqualitätsmerkmale innerhalb einer Unterrichtsstunde beurteilbar sind (z.B. kognitive Aktivierung, vgl. Abschnitt 3.2.1), sollten Beobachtungen zudem nach Möglichkeit auf einen längeren Zeitraum angelegt sein. Die Einschätzungen dieser Einzelbeobachtungen können danach gemittelt werden, was analog zur Mittelung mehrerer Raterinnen und Rater zu einer Minimierung des Fehleranteils in den Ratings führen sollte.
3) Aus diversen Forschungsbereichen (Übersicht in Hoyt & Kerns, 1999) wissen wir, dass Einschätzungen ungeschulter Beobachterinnen und Beobachter oftmals eine nur unzureichende Qualität aufweisen. Dies ist darauf zurückzuführen, dass die oben erwähnten Ursachen von Beobachtereffekten hier noch verstärkter auftreten. Insbesondere bei Unterrichtsbewertungen, aber auch in Bezug auf Unterrichtsrückmeldungen kommt einer Schulung von Beobachterinnen und Beobachtern (z.B. hospitierende Lehrkräfte oder Schulleitungsmitglieder) zur Einschätzung von Unterricht daher hohe Relevanz zu. Eine solche Schulung hat den Befunden von Praetorius et al. (2012) zufolge zwar keine positiven Effekte auf die Reliabilität der Urteile, gleichwohl aber auf deren Validität.
4) Neben der Optimierung der Unterrichtseinschätzungen externer Beobachterinnen und Beobachter ist auch ein Rückgriff auf Schülereinschätzungen möglich (z.B. Clausen, 2002; Kunter & Baumert, 2006; Lenske, 2012; Lüdtke, Trautwein, Kunter & Baumert, 2006; Marsh & Roche, 1997). Neben etlichen Vorteilen (u.a. der Möglichkeit eines langen Beurteilungszeitraums) werden jedoch auch in Bezug auf Einschätzungen von Schülerinnen und Schülern einige Nachteile thematisiert

(u.a. deren fehlendes methodisch-didaktisches Wissen), weswegen es sich anbietet, Schülereinschätzungen ergänzend zu und nicht anstelle von externen Beobachtereinschätzungen einzusetzen. Ein ungelöstes Problem ist dabei jedoch, dass Unterrichtseinschätzungen unterschiedlicher Perspektiven in der Regel eher geringe Übereinstimmungen aufweisen (z.B. Clausen, 2002: $-.28 \leq r \leq .45$), sodass sich im konkreten Fall die Frage stellt, auf welche Aspekte welcher Perspektive sich Lehrkräfte verlassen sollten.

«Auf den Lehrer kommt es an» (Lipowsky, 2006) – dies wurde in den letzten Jahren und Jahrzehnten vielfach empirisch bestätigt. Die im vorliegenden Beitrag dargestellten Befunde und Einschränkungen in Bezug auf die Interpretierbarkeit von Unterrichtseinschätzungen externer Beobachterinnen und Beobachter zeigen darüber hinaus: Auch auf die Messung kommt es an! Denn nur dann, wenn wir Unterrichtsqualität hinreichend genau und valide messen, können auf dieser Messung basierende wissenschaftliche Schlussfolgerungen sinnvolle Erkenntnisse bieten und unterrichtspraktische Entwicklungen zu einer Verbesserung von Unterricht führen.

Literatur

Altrichter, H., Messner, E. & Posch, P. (2004). *Schulen evaluieren sich selbst: Ein Leitfaden*. Seelze Velber: Kallmeyer'sche Verlagsbuchhandlung.

Berliner, D.C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56* (3), 205–213.

Bradburn, N.M. (2004). Understanding the question-answer process. *Statistics Canada, 30* (1), 5–15.

Brennan, R.L. (2001). *Generalizability Theory*. New York: Springer.

Brophy, J. (2006). Observational research on generic aspects of classroom teaching. In P.A. Alexander & P.H. Winne (Hrsg.), *Handbook of educational psychology* (2. Auflage) (S. 755–780). Mahwah, NJ: Erlbaum.

Clare, L., Valdés, R., Pascal, J. & Steinberg, J. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools* (CSE Technical Report Nr. 545). Los Angeles: National Center for Research on Evaluation.

Clausen, M. (2002). *Qualität von Unterricht – Eine Frage der Perspektive?* Münster: Waxmann.

Clausen, M., Reusser, K. & Klieme, E. (2003). Unterrichtsqualität auf der Basis hoch-inferenter Unterrichtsbeurteilungen: Ein Vergleich zwischen Deutschland und der deutschsprachigen Schweiz. *Unterrichtswissenschaft, 31* (2), 122–141.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.

Helmke, A., Helmke, T., Lenske, G., Pham, G., Praetorius, A.-K., Schrader, F.-W. et al. (2011). Unterrichtsdiagnostik – Voraussetzungen für die Verbesserung der Unterrichtsqualität. In A. Bartz, M. Dammann, S. Huber, C. Kloft & M. Schreiner (Hrsg.), *PraxisWissen SchulLeitung, AL 28* (Kapitel 30.71). Köln: Wolters Kluwer.

Helmke, A. & Lenske, G. (2013). Unterrichtsdiagnostik als Voraussetzung für Unterrichtsentwicklung. *Beiträge zur Lehrerbildung, 31* (2), 214–233.

Hill, H.C., Charalambous, C.Y. & Kraft, M.A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41* (2), 56–64.

Horster, L. & Rolff, H.-G. (2001). *Unterrichtsentwicklung: Grundlagen einer reflektorischen Praxis* (2. Auflage). Weinheim: Beltz.

Hoyt, W.T. & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4* (4), 403–424.
Kobarg, M. & Seidel, T. (2005). Coding manual – Process-oriented teaching. In T. Seidel, M. Prenzel & M. Kobarg (Hrsg.), *How to run a video study. Technical report of the IPN Video Study* (S. 108–144). Münster: Waxmann.
Kunter, M. & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9* (3), 231–251.
Lemons, R.W. & Helsing, D. (2010). High quality teaching and learning: Do we know it when we see it (and when we don't)? *Education Canada, 48* (5), 14–18.
Lenske, G. (2012). *Schülerfeedback zur Unterrichtsqualität in der Grundschule: Studien zur Validität*. Unveröffentlichte Dissertation. Landau: Universität Koblenz-Landau.
Lipowsky, F. (2006). Auf den Lehrer kommt es an: Empirische Evidenzen für Zusammenhänge zwischen Lehrerkompetenzen, Lehrerhandeln und dem Lernen der Schüler. In C. Allemann-Ghionda (Hrsg.), *Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern* (S. 47–70). Weinheim: Beltz.
Lüdtke, O., Trautwein, U., Kunter, M. & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9* (3), 215–230.
Marsh, H.W. & Roche, L.A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52* (11), 1187–1197.
Matsumura, L.C., Garnier, H.E., Pascal, J. & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment, 8* (3), 207–229.
Müller, S. & Pietsch, M. (2011). Was wir messen, wenn wir Unterrichtsqualität messen: Inter-Beurteilerübereinstimmung und -Reliabilität bei Unterrichtsbeobachtungen im Rahmen von Schulinspektion. In S. Müller, M. Pietsch & W. Bos (Hrsg.), *Schulinspektionen in Deutschland – eine erste Zwischenbilanz* (S. 33–56). Münster: Waxmann.
Newton, X.A. (2010). Developing indicators of classroom practice to evaluate the impact of district mathematics reform initiative: A generalizability analysis. *Studies in Educational Evaluation, 36* (1–2), 1–13.
Petko, D., Waldis, M., Pauli, C. & Reusser, K. (2003). Methodologische Überlegungen zur videogestützten Forschung in der Mathematikdidaktik: Ansätze der TIMSS 1999 Video Studie und ihrer schweizerischen Erweiterung. *Zentralblatt für Didaktik der Mathematik, 35* (6), 265–280.
Pianta, R.C. & Hamre, B.K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38* (2), 109–119.
Pietsch, M. & Tosana, S. (2008). Beurteilereffekte bei der Messung von Unterrichtsqualität: Das Multifacetten-Rasch-Modell und die Generalisierbarkeitstheorie als Methoden der Qualitätssicherung in der externen Evaluation von Schulen. *Zeitschrift für Erziehungswissenschaft, 11* (3), 430–452.
Praetorius, A.-K. (in Druck). *Eignung von hoch-inferenten Beobachterratings zur Messung von Unterrichtsqualität*. Münster: Waxmann.
Praetorius, A.-K., Lenske, G. & Helmke, A. (2010). *Auf der Suche nach Ursachen für Beurteilungsdifferenzen – Ist die Methode des Lauten Denkens für die Analyse unterrichtsbezogener Urteilsprozesse ertragreich?* Vortrag gehalten an der 74. Tagung der Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF), Jena.
Praetorius, A.-K., Lenske, G. & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction, 22* (6), 387–400.
Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht: Unterricht aus der Perspektive von Lernenden und Beobachtern.* Münster: Waxmann.
Sommer, N. (2011). Unterrichtsqualität im Urteil der externen Schulevaluation. In S. Müller, M. Pietsch & W. Bos (Hrsg.), *Schulinspektionen in Deutschland – eine erste Zwischenbilanz* (S. 97–136). Münster: Waxmann.
Stigler, J. (1998). Video surveys: New data for the improvement of classroom instruction. In S.G. Paris & H.M. Wellman (Hrsg.), *Global prospects for education. Development, culture and schooling* (S. 129–168). Washington, DC: American Psychological Association.

Stigler, J., Gonzales, P., Kawanaka, T., Knoll, S. & Serrano, A. (1999). *The TIMSS-Videotape Classroom Study* (Technical Report). Los Angeles: University of California.
Strong, M., Gargani, J. & Hacifazlioglu, O. (2011). Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. *Journal of Teacher Education, 62* (4), 367–382.
Waldis, M., Grob, U., Pauli, C. & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 171–208). Münster: Waxmann.

Autorin

Anna-Katharina Praetorius, Dr., Universität Augsburg, Lehrstuhl für Psychologie,
anna.praetorius@phil.uni-augsburg.de