Göllner, Richard; Wagner, Wolfgang; Klieme, Eckhard; Lüdtke, Oliver; Nagengast, Benjamin; Trautwein, Ulrich

Erfassung der Unterrichtsqualität mithilfe von Schülerurteilen: Chancen, Grenzen und Forschungsperspektiven

Bundesministerium für Bildung und Forschung [Hrsg.]: Forschungsvorhaben in Ankopplung an Large-Scale-Assessments. Berlin : Bundesministerium für Bildung und Forschung 2016, S. 63-82. - (Bildungsforschung; 44)

Richard Göllner, Wolfgang Wagner, Eckhard Klieme, Oliver Lüdtke, Benjamin Nagengast, Ulrich Trautwein

Erfassung der Unterrichtsqualität mithilfe von Schülerurteilen: Chancen, Grenzen und Forschungsperspektiven

1 Einleitung

Schülerurteile stellen eine wichtige Datenquelle zur Messung lern- und leistungsförderlicher Qualitätsaspekte des Unterrichts dar (Clausen, 2002; Klieme, Schümer & Knoll, 2001; Fraser & Walberg, 1991; Seidel & Shavelson, 2007). Schülerinnen und Schüler gelten als Experten des Unterrichts. Sie sind in der Lage, Vergleiche mit anderen Lehrkräften anzustellen, sie können ihr Urteil auf einen ausgedehnten Beobachtungszeitraum stützen, und sie haben die Möglichkeit, selbst über seltene Ereignisse im Unterricht Auskunft zu geben. Zudem können mit Schülerbefragungen eine Vielzahl von Beurteilern im Hinblick auf einen Beurteilungsgegenstand befragt und somit eine hohe Informationsdichte (und gegebenenfalls -vielfalt) erzielt werden. Gerade in Hinblick auf große Schulleistungsstudien wie PISA, IGLU oder TRAIN sind Schülerurteile von potenziell hoher Bedeutung, da ihre Erfassung relativ kostengünstig ist und wenig Zeitressourcen in Anspruch nimmt (Clausen, 2002; Lüdtke, Trautwein, Kunter & Baumert, 2006).

Aber wie gut sind diese Schülerurteile? Nach wie vor ist zu wenig über die psychometrische Güte (wie Reliabilität und Validität) von Schülerurteilen über den Unterricht bekannt. Können Schülerinnen und Schüler theoretisch distinkte Facetten des Unterrichtsgeschehens zuverlässig und zutreffend einschätzen? Inwiefern sind die Einschätzungen von Schülerinnen und Schülern über unterschiedliche Kontexte (z. B. Unterrichtsfächer oder Schulklassen) hinweg vergleichbar? Inwiefern beeinflusst die sprachliche Gestaltung der Items den Beurteilungsprozess? Diese bislang wenig untersuchten Fragen standen im Mittelpunkt des durch das BMBF geförderten Projekts „Erfassung der Unterrichtsqualität in Large-Scale-Studien: Optimierung der Modellierung und Itemauswahl“ (01LSA008; Trautwein, Lüdtke, Klieme, Nagengast & Wagner, 2011). In dem vorliegenden Beitrag werden zentrale Ergebnisse des Projekts vorgestellt. Zunächst wird ein allgemeiner Überblick über die Erfassung der Unterrichtsqualität gegeben, um dann genauer auf Chancen und Grenzen von Schülerbeurteilungen des Unterrichts anhand von empirischen Ergebnissen des Projekts einzugehen.

2 Unterrichtsqualität und ihre Erfassung

Aktuelle Konzepte der Unterrichtsqualitätsforschung betrachten das Unterrichtsgeschehen als ein Zusammenspiel von Merkmalen, welche die Lerngeschichte von Schülern nachhaltig beeinflussen können (Hattie, 2009; Helmke, 2010; Klieme et al., 2008; Scheerens & Bosker, 1997). Inzwischen liegen sowohl national als auch international empirisch begründete Beschreibungssysteme relevanter Merkmale vor. Diese umfassen sogenannte Basismerkmale der Unterrichtsqualität, welche dann auf untergeordneten Hierarchien das konkrete Lehrerverhalten beschreiben. Zu den in der Literatur genannten Basismerkmalen bzw. Dimensionen gehören die Klassenführung, die Schülerorientierung und die kognitive Aktivierung. Diesen Modellen zufolge ist ein Unterricht lern- und leistungsförderlich, der klar strukturiert ist, die zur Verfügung stehende Lernzeit nutzt und dabei kognitiv anregend und individuell unterstützend gestaltet ist (Hamre & Pianta, 2010; Klieme, Schümer & Knoll, 2001).

Die Erfassung der Unterrichtsqualität erfolgt in der Unterrichtsforschung meist anhand von Schülerbeurteilungen, Selbstberichten der Lehrkräfte oder Beobachtungen durch externe Beobachter, die entweder unmittelbar in der Klasse oder auf der Basis von Videoaufnahmen eine Beurteilung vornehmen bzw. Unterrichtskonzepte oder Unterrichtsmaterialien bewerten (Clausen, 2002; Fraser & Walberg, 1991). Alle genannten Zugänge weisen ihre spezifischen Vor- und Nachteile auf, insbesondere dann, wenn es um ihren Einsatz in Large-Scale-Studien geht. So gelten Beobachtungen durch Experten zwar allgemein als der „Goldstandard" in der Lehr-Lern-Forschung, sie sind aber mit sehr hohem zeitlichen und finanziellen Aufwand verbunden, und die Experteneinschätzungen entsprechen nicht immer den üblichen psychometrischen Standards. Insbesondere die Erfassung sogenannter hoch inferenter Beobachtungen erfordert ein hohes Maß an Schulung der Beurteiler. Hoch inferente Beobachtungen verlangen von den Beobachtern, über das konkret beobachtete Verhalten hinaus auf abstrakte Sachverhalte oder allgemeine Verhaltenstendenzen zu schließen. Demgegenüber beschränken sich die sogenannten niedrig inferenten Beurteilungen auf spezifische, beobachtbare Verhaltensweisen, welche einfach und objektiv zu codieren sind (Rosenshine, 1970). Die Codierung und Auswertung einer Unterrichtsstunde beansprucht häufig ein Vielfaches der eigentlichen Beobachtungsdauer (Fauth, Decristan, Rieser, Klieme & Büttner, 2014). Trotz dieses enormen Aufwands wird eine zufriedenstellende Messgüte in vielen Fällen nur über eine Anpassung der untersuchten Unterrichtsmerkmale erreicht (Derry et al., 2010).

Deutlich weniger aufwendig ist die Verwendung von Lehrerselbstberichten (Clausen, 2002; Desimone, Smith & Frisvold, 2010). Selbstberichte stellen eine gut umsetzbare Möglichkeit dar, lern- und leistungsrelevante Qualitätsmerkmale des Unterrichts zu erfassen. Selbstberichte von Lehrkräften haben in den vergangenen Jahren wieder an Bedeutung gewonnen und sind inzwischen aus dem Forschungsfeld der Unterrichtsqualität, der Lehrerexpertise und des Professionswissens (Fachwissen, pädagogisches Wissen und fachdidaktisches Wissen) nicht mehr wegzudenken (Baumert & Kunter, 2006; Blömeke, 2004). Lehrkräfte sollten aufgrund ihrer beruflichen Qualifikation auch komplexe Sachverhalte ihres beruflichen Handelns beschreiben können. Allerdings weisen empirische Befunde der vergangenen Jahre auch auf eine Vielzahl von „Pro-

blemen“ hin (Clausen, 2002; Wubbels, Brekelmans & Hooymayers, 1992). Es findet sich eine nur geringe Übereinstimmung von Lehrerselbstberichten mit anderen Datenquellen (z. B. Beurteilungsdaten) und nur wenige Hinweise auf eine prädiktive Validität für relevante Zielkriterien des Unterrichts. Erklärt wird dies durch die Involviertheit der Lehrkraft in den täglichen Unterricht, die zu selbstwertschützenden Verzerrungen oder auch sozial erwünschten Beurteilungstendenzen führen kann (Clausen, 2002). Zudem ist aus dem Bereich der Persönlichkeitspsychologie bekannt, dass nicht alle Aspekte des eigenen Verhaltens einer Selbstbeschreibung in gleicher Weise zugänglich sind und entsprechend durch außenstehende Personen akkurater beschrieben werden können (z. B. Funder, 2001; Vazire & Solomon, 2015).

Eine dieser Außenperspektiven und die wohl am häufigsten eingesetzte Methode der Datengewinnung zur Erfassung der Unterrichtsqualität in Large-Scale-Assessments sind Schülerbeurteilungen des Unterrichts. Üblicherweise werden hierbei Fragebogenitems für theoretisch distinkte Dimensionen der Unterrichtsqualität verwendet, die keinen konkreten Stundenbezug aufweisen, sondern rückblickend auf einen längeren Unterrichtszeitraum (z. B. das vergangene Schuljahr) nach Merkmalen des Unterrichts fragen. Aus forschungspraktischer Perspektive stellen Schülerbeurteilungen des Unterrichts eine höchst effektive Form der Datengewinnung dar (Clausen, 2002; Fraser & Walberg, 1991). Neben dem geringeren Erhebungsaufwand und der ökonomischen Durchführung ist insbesondere die hohe Reliabilität von Schülerdaten aufgrund der Zusammenfassung mehrere Einzelurteile zu nennen (Lüdtke, Trautwein, Kunter & Baumert, 2006). Doch auch die Verwendung von Schülerurteilen wird mitunter kritisch gesehen. Schülerinnen und Schüler sind nicht nur Beurteiler des Unterrichts, sondern auch stark involvierte Akteure. Zudem verfügen sie trotz ihrer Erfahrungen nicht über eine didaktisch-pädagogische Expertise im engeren Sinne. Empirische Studien der frühen 1990er-Jahre zeigten beispielsweise, dass das Ausmaß einer positiven Beurteilung der Lehrkraft bzw. des Unterrichts im Kontext universitärer Lehre sowohl vom Geschlecht, dem Leistungsstand, dem Interesse oder auch individuellen Urteilstendenzen (z. B. Milde-Strenge-Effekte) der einzelnen Schülerinnen und Schüler abhängig ist (z. B. Gigliotti & Buchtel, 1990; Greenwald & Gillmore, 1997; Marsh & Roche, 1997). Einige Arbeiten weisen für die Schülersicht sogenannte Halo-Effekte nach, die als Überstrahlungseffekte zu einer weniger detaillierten Beschreibung verschiedener Qualitätsdimensionen führen können (Clausen, 2002; Fisicaro & Lance, 1990).

Darüber hinaus ist die Beantwortung der Frage, inwieweit Schülerinnen und Schüler die Unterrichtsqualität einer Lehrkraft zuverlässig und zutreffend beurteilen können, erheblich durch die Vielzahl an vorliegenden Instrumenten erschwert. Bereits eine oberflächliche Betrachtung der Items aus den bekannten Large-Scale-Studien (z. B. PISA, KESS, COACTIV, BIJU, DESI usw.) zeigt, dass sich die eingesetzten Fragebogeninstrumente nicht nur im Hinblick auf die Auswahl und Operationalisierung der Qualitätsdimensionen unterscheiden, sondern auch eine erstaunlich hohe Variabilität der konkreten Itemformulierungen aufweisen. Welche (unbeabsichtigten) Auswirkungen spezifische Itemformulierungen auf die psychometrische Qualität von Schülerbeurteilungen des Unterrichts haben, ist eine nach wie vor offene Frage, die im Rahmen des vom BMBF geförderten Projekts „Erfassung der Unter-

richtsqualität in Large-Scale-Studien: Optimierung der Modellierung und Itemauswahl“ untersucht wurde. In den folgenden Abschnitten erfolgt eine Zusammenfassung zentraler Befunde.

3 Die Validität von Schülerbeurteilungen des Unterrichts

Schülerbeurteilungen der Unterrichtsqualität versprechen eine hohe Effektivität für die Erfassung lern- und leistungsförderlicher Qualitätsmerkmale des Unterrichts, sie sind im Rahmen von Large-Scale-Assessments flexibel einsetzbar und zeigen in einer Vielzahl von Studien substanzielle Zusammenhänge mit Kriterien des Schulerfolges auf (Clausen, 2002; Fraser & Walberg, 1991; Klieme et al., 2008). Andererseits scheint die Verwendung von Schülerbeurteilungen nicht grundsätzlich eine bessere Erfassung relevanter Unterrichtsaspekte zu versprechen. Die Kritikpunkte umfassen die mangelnde Fähigkeit von Schülerinnen und Schülern zur Differenzierung verschiedener Facetten der Unterrichtsqualität sowie die Vergleichbarkeit (bzw. Messäquivalenz) von Schülerbeurteilungen über kontextuelle Rahmenbedingungen, wie etwa Schulfächer oder unterschiedliche Schulklassen. Beide Aspekte wurden im Rahmen einer Studie von Wagner, Göllner, Helmke, Trautwein und Lüdtke (2013) näher untersucht, deren Ergebnisse im Folgenden dargestellt werden.

3.1 Die Fähigkeit von Schülern zur dimensionalen Beschreibung der Unterrichtsqualität

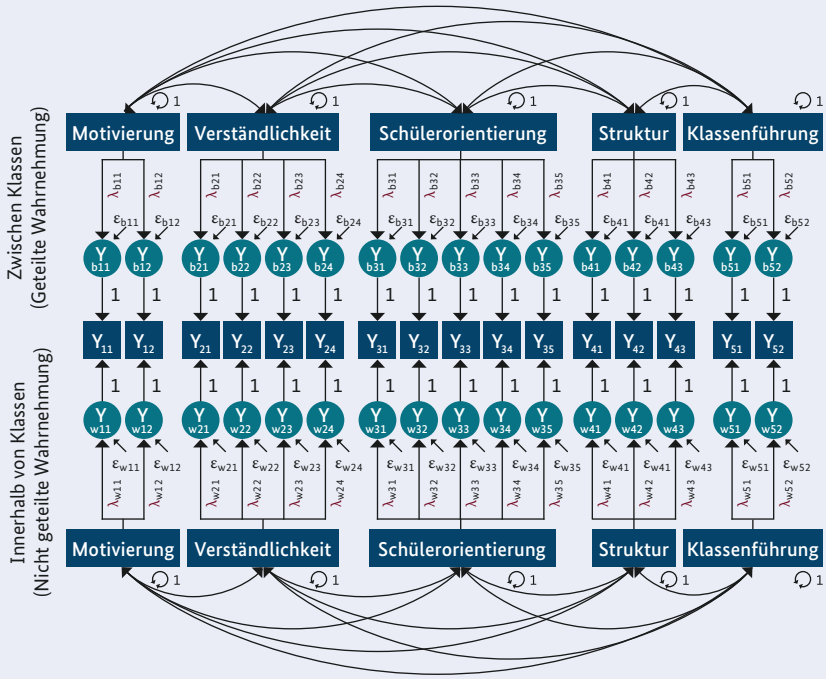
Im Allgemeinen ist die Frage nach der dimensionalen Struktur des Unterrichts eng mit der Anwendung faktorenanalytischer Verfahren verbunden (z. B. Nunnally, 1978). Faktorenanalytische Modelle geben Auskunft darüber, ob und wie sich ein Untersuchungsgegenstand anhand von Variablen (z. B. Fragebogenitems) mehrdimensional beschreiben lässt. Die entsprechenden Analysen erfolgen entweder erkundend im Sinne einer Exploration oder konfirmatorisch vor dem Hintergrund eines theoretischen Modells. Die für die empirische Bildungsforschung charakteristische Datenstruktur (Schülerinnen und Schüler innerhalb von Klassen und Schulen) erfordert die Anwendung sogenannter Mehrebenen-Strukturgleichungsmodelle (ML-SEM; Lüdtke, Marsh, Robitzsch & Trautwein, 2011; Marsh et al., 2013; Mehta & Neale, 2005). Diese Modelle ermöglichen es, die Faktorstruktur von Instrumenten zur Erfassung von Unterrichtsqualität simultan sowohl auf der Ebene der individuellen Schülerinnen und Schüler als auch auf Ebene der Klasse zu modellieren (Hox, Maas & Brinkhuis, 2010; Mehta & Neale, 2005), wobei im Falle von Schülerbeurteilungen der Unterrichtsqualität der Fokus der Analyse auf der Klassenebene liegt (Lüdtke, Robitzsch, Trautwein & Kunter, 2009).

Ein solches Mehrebenen-Strukturgleichungsmodell wurde in einer Studie von Wagner et al. (2013) auf Schülerbeurteilungen des Unterrichts in den Fächern Deutsch und Englisch angewandt. Grundlage der Untersuchung waren Daten der Studie Deutsch-Englisch-Schulleistungen International (DESI), die in den Jahren

2001 bis 2006 von einem interdisziplinären Konsortium unter der Federführung des Deutschen Instituts für Internationale Pädagogische Forschung (DIPF) durchgeführt wurde. DESI stellte eine bundesweite Untersuchung dar, an der sich allgemeinbildende Schulen aus sämtlichen Ländern der Bundesrepublik beteiligten. Ein Schwerpunkt des DESI-Projektes war die Untersuchung von Unterrichtsmerkmalen im Hinblick auf die Leistungsentwicklung von Schülerinnen und Schülern. Insgesamt nahmen 427 Klassen an der Untersuchung teil. Um im Rahmen der Analysen die Schülerbeurteilungen des Unterrichts für die Fächer Englisch und Deutsch parallel analysieren zu können, beschränkte sich die Auswertungsstichprobe auf 280 Schulklassen (N = 6.909 Schülerinnen und Schüler). Eingeschlossen wurden zudem nur Klassen, die in beiden Fächern durch unterschiedliche Lehrkräfte unterrichtet wurden. Das durchschnittliche Alter der Schülerinnen und Schüler betrug M = 15,8 Jahre. Das eingesetzte Instrument zur Beurteilung der Unterrichtsqualität aus Schülersicht umfasste eine große Zahl von Unterrichtsmerkmalen (siehe Klieme, Jude, Rauch, Ehlers, Helmke, Eichler et al., 2008). Aufgrund der Komplexität der Modellierung beschränkten sich die Analysen auf die fünf Unterrichtsdimensionen Schülermotivierung, Verständlichkeit, Schülerorientierung, Strukturiertheit und Klassenführung.

Das zugrunde liegende Faktormodell für beide Analyseebenen ist in Abbildung 1 dargestellt. Die Modellierung orientierte sich an dem von verschiedenen Autoren (z. B. Muthén, 1994) vorgeschlagenen Vorgehen.

Abb. 1: Mehrebenen-Strukturgleichungsmodelle (ML-SEM) zur Analyse von Schülerbeurteilungen des Unterrichts anhand von fünf Unterrichtsmerkmalen



Die Analyseergebnisse bestätigten eine dimensionale Struktur der zugrunde liegenden Schülerbeurteilungen des Unterrichts. Nicht nur auf der Ebene „innerhalb von Klassen", sondern auch „zwischen Klassen" waren die fünf Merkmalsdimensionen empirisch trennbar. Wenngleich die statistischen Zusammenhänge der einzelnen Unterrichtsmerkmale auf der Klassenebene vergleichsweise hoch ausfielen (Deutsch: $r = .70$ bis $r = .95$; Englisch: $r = .54$ bis $r = .94$), können Schülerurteile genutzt werden, um die Unterrichtsqualität einer Lehrkraft sehr spezifisch und detailliert beschreiben zu können. Dieser Befund wurde zusätzlich erhärtet durch die simultanen Analysen beider Unterrichtsfächer. Es zeigten sich nur mäßige Zusammenhänge der Unterrichtsmerkmale über beide Fächer (und unterschiedliche Lehrkräfte) hinweg. Die Zusammenhangsmaße auf der Klassenebene variierten für die fünf betrachten Merkmale zwischen $r = -.14$ und $r = .35$ (insgesamt 25 Korrelationen), wobei sich nur für drei Merkmale statistisch signifikante Zusammenhänge finden ließen. Demnach ist das geteilte Urteil von Schülerinnen und Schülern in einem Fach mehr oder weniger unabhängig von dem geteilten Urteil in dem anderen Fach (bezogen jeweils auf unterschiedliche Lehrkräfte). In anderen Worten spiegelt das Urteil von Schülerinnen und Schülern nicht primär stabile oder transkonsistente Anteile der Beurteiler wider, sondern beschreibt die spezifische Qualität einer Lehrkraft entlang verschiedener Qualitätsdimensionen.

3.2 Generalisierbarkeit von Schülerbeurteilungen des Unterrichts

Ein zweiter wichtiger Aspekt der Studie von Wagner et al. (2013) umfasste die Prüfung der Vergleichbarkeit von Schülerbeurteilungen des Unterrichts über unterschiedliche Kontexte (z. B. Fächer oder unterschiedliche Schulklassen) hinweg. Ohne Zweifel ist die Prüfung der faktoriellen Struktur von Konstrukten vor dem Hintergrund einer theoretischen Struktur ein unverzichtbarer erster Schritt zur Beantwortung der Frage, inwiefern Schülerinnen und Schüler die Unterrichtsqualität von Lehrkräften zutreffend bzw. valide beurteilen können. Ein sich daran anschließender und notwendiger zweiter Schritt umfasst die Frage nach der Vergleichbarkeit der Messung über verschiedene kontextuelle Bedingungen wie etwa Unterrichtsfächer oder Schulklassen. Im Kern ist zu klären, inwieweit Messergebnisse für diese unterschiedlichen Kontextbedingungen auf einem identischen Messmodell beruhen und folglich einen (unmittelbaren) Vergleich dieser Messwerte überhaupt erst rechtfertigen. So wäre bei Verletzung der Äquivalenzannahme beispielsweise ein Vergleich der Variabilität der Klassenführung von Lehrkräften in den Fächern Deutsch und Englisch nicht zulässig, da die Messungsunterschiede nicht sinnvoll auf Unterschiede bezüglich einer einzigen Dimension reduzierbar sind (sondern je nach Item mehr oder weniger stark variieren). Das Beispiel macht deutlich, wie stark auch die Unterrichtsqualität von Lehrkräften eine adäquate Messung im Sinne eines äquivalenten Messmodells voraussetzt.

Die Äquivalenzannahmen bei Schülerbeurteilungen der Unterrichtsqualität wurden im Beitrag von Wagner et al. (2013) sowohl für die an der Studie teilnehmenden Unterrichtsklassen als auch die beiden Unterrichtsfächer überprüft. Die Analysen

erfolgten in Anlehnung an das Verfahren multipler Gruppenanalysen, in der über eine sukzessive Fixierung verschiedener Parameter des Messmodells geprüft wird (Meredith, 1993; Widaman & Reise, 1997). Jedoch basiert die Grundidee der Äquivalenztestung über Klassen auf einem Vergleich der ebenenspezifischen Ladungsmuster (siehe Jak, Oort & Dolan, 2013; Wagner, 2008; Wagner et al., 2013). Inwieweit die Annahme eines äquivalenten Messmodells über Klassen hinweg haltbar ist, kann geprüft werden, indem Faktorladungen über Ebenen hinweg (deskriptiv) verglichen und mögliche Unterschiede inferenzstatistisch getestet werden. Die Anwendung eines solchen Verfahrens auf die Daten der DESI-Studie zeigte für die Merkmale Motivierung, Verständlichkeit und Schülerorientierung die größten Unterschiede. Für diese drei Merkmale ist ein Vergleich der Messwerte über verschiedene Schulklassen hinweg nur bedingt zu rechtfertigen. Vielmehr scheinen die entsprechenden Messungen durch jeweils klassenspezifische Eigenschaften der Schüler mitbestimmt. Hingegen bestätigte sich die Annahme äquivalenter Messmodelle für die Merkmale Strukturierung und Klassenführung. Für beide Merkmale können demnach Messwerte in sinnvoller und inhaltlich valider Weise über Klassen hinweg miteinander verglichen werden. Auf diesen ebenenspezifischen Ladungsmustern beruhend, wurde abschließend zusätzlich die fächerbezogene Messäquivalenz überprüft. Für die beiden Merkmale Strukturierung und Klassenführung konnte eine vollständige Ladungsäquivalenz über beide Analyseebenen (innerhalb und zwischen Klassen) und Fächer (Deutsch, Englisch) hinweg als gerechtfertigte Annahme bestätigt werden.

Zusammenfassend zeigen die Ergebnisse der Studie von Wagner et al. (2013), dass Schülerinnen und Schüler in differenzierter Weise das Unterrichtsgeschehen bzw. die Qualität des Unterrichts auf unterschiedlichen Dimensionen beurteilen können, wenngleich Einschränkungen hinsichtlich der Vergleichbarkeit über Klassen hinweg zu beachten sind. Insbesondere Merkmale, die durch einen hohen Schülerbezug (d. h. Schüler als Bestandteil eines gelingenden Unterrichts) gekennzeichnet sind (d. h. Motivierung, Schülerorientierung und Verständlichkeit), sind demnach nur bedingt für die Vergleiche verschiedener Lehrkräfte geeignet. Bei der Beurteilung solcher Unterrichtsmerkmale spielen offenbar Schülermerkmale eine nicht zu unterschätzende Rolle, die sich in verschiedenen Klassen unterscheiden.

4 Die Übereinstimmung unterschiedlicher Datenquellen

Neben der Überprüfung der faktoriellen Struktur und Messäquivalenz von Schülerbeurteilungen der Unterrichtsqualität sind weitere Kriterien zur Prüfung der Validität von Schülerbeurteilungen des Unterrichts formuliert worden. Hier ist in erster Linie die Übereinstimmung von Schülerbeurteilungen mit weiteren Datenquellen (z. B. Selbstberichte von Lehrkräften und Beobachtungen) zu nennen. Empirische Arbeiten zur Übereinstimmung unterschiedlicher Datenquellen sind rar und legen die Annahme lediglich moderater Zusammenhänge zwischen den Beurteilungsperspektiven nahe. So berichtete Clausen (2002) kleine bis mittlere relative Übereinstimmungskoeffizienten von $r = -.28$ bis $r = .42$ zwischen Schülerbeurteilungen und Lehrerselbstberichten, von $r = -.22$ bis $r = .45$ zwischen Schüler- und Videobeurteilern

und von $r = -.04$ bis $r = .43$ zwischen Lehrerberichten und Videobeurteilern. Untersuchungsergebnisse der Studie COACTIV zeigten nur unwesentlich höhere Übereinstimmungskoeffizienten zwischen Schülerinnen und Schülern und Lehrkräften ($r = .21$ bis $r = .31$; Kunter et al., 2008). Auch eine neuere Studie zur Unterrichtsbeurteilung von Grundschülern zeigt vergleichbare Ergebnisse für unterschiedliche Dimensionen der Unterrichtsqualität (Fauth et al., 2014). Worin liegen die Ursachen dieser vergleichsweise geringen Übereinstimmung unterschiedlicher Perspektiven? In Anlehnung an Clausen (2002) ist zunächst festzuhalten, dass keine der Datenquellen als prinzipiell „besser" oder „schlechter" zu betrachten ist. Vielmehr weisen alle Perspektiven spezifische Anteile auf, die je nach Beurteilungsgegenstand bzw. der betrachteten Unterrichtsmerkmale näher an der Unterrichtswirklichkeit liegen (Clausen, 2002; Kunter & Baumert, 2006).

Doch es gibt vermutlich weitere Erklärungsansätze, die insbesondere kognitive Anforderungen bei der Beurteilung als Ursache mangelnder Datenqualität identifizieren. Aus diesem Grund wurden im Rahmen des BMBF-Projektes verschiedene Anforderungsmerkmale von Fragebogenitems definiert und im Hinblick auf die Übereinstimmung verschiedener Urteilsperspektiven untersucht.

4.1 Anforderungsmerkmale von Items

Fraglos ist die Beantwortung eines Items für die Schülerinnen und Schüler nicht trivial. Es besteht Übereinkunft darin, dass die Beantwortung von Fragebogenitems entlang verschiedener Prozessschritte erfolgt, die je nach konkretem Iteminhalt teilweise anspruchsvolle kognitive und motivationale Anforderungen an den Beurteiler stellen (Lenske, 2011; Tourangeau, Rips & Rasinski, 2000). Diese Schritte umfassen unter anderem das Verständnis und die Interpretation des jeweiligen Iteminhalts, die Sammlung relevanter Informationen zu dessen Beantwortung, die Verdichtung der ermittelten Informationen zu einem Urteil und schließlich die Auswahl einer passenden Antwortalternative (z. B. Tourangeau et al., 2000). Bereits ein oberflächlicher Blick auf Fragebogeninstrumente zur Erfassung der Unterrichtsqualität zeigt, dass allein das sprachliche Verständnis und die Interpretation des Iteminhalts in hohem Maße anfordernd bzw. herausfordernd sein können. Aus theoretischer Perspektive und nach einer Systematisierung von Fragebogeninstrumenten aus großen deutschen Schulleistungsstudien können neben der sprachlichen (d. h. orthografischen, grammatikalischen und linguistischen) Komplexität drei weitere Anforderungsmerkmale unterschieden werden. Hierzu gehören a) der Adressatenbezug eines Items (bestimmt den Adressat des Lehrerverhaltens: individueller Schüler versus sämtliche Schüler der Klasse; den Brok, Brekelmans & Wubbels, 2006), b) die Wahrnehmungsperspektive (bestimmt die Sichtweise, aus der beurteilt wird: individuelle Sicht des Schülers oder Wir-Perspektive; den Brok, Brekelmans & Wubbels, 2006) und c) der Zeitbezug eines Items (bezieht sich das Urteil auf die letzte Stunde oder ist ein über mehrere Unterrichtsstunden aggregiertes Urteil verlangt; Tourangeau & Rasinski, 2000). Je nach Ausprägungsgrad stellt die Itembeantwortung eine mehr oder weniger hohe Anforderung an die Schülerinnen und Schüler dar. So müssen Schülerinnen

und Schüler in vielen Fällen nicht nur auf der Grundlage ihrer eigenen individuellen Erfahrungen mit der Lehrkraft urteilen (z. B. „Mein Lehrer achtet darauf, dass ich im Unterricht mitkomme"), sondern müssen in vielen Fällen die Erfahrungen der Mitschüler in ihr Urteil integrieren (z. B. „Der Lehrer unterstützt uns zusätzlich, wenn wir Hilfe brauchen"). Es ist zu vermuten, dass damit eine erhebliche Anforderung an die Schülerinnen und Schüler gestellt wird, welche die Güte der Beurteilung erheblich einschränken kann.

Eine im Rahmen des BMBF-Projektes durchgeführte Klassifikation von insgesamt 533 Fragebogenitems aus sieben deutschen Large-Scale-Studien zeigte, dass ein großer Teil von Items derartige Anforderungen stellt, wenngleich Bezüge und Perspektiven in vielen Fällen nicht eindeutig definiert sind und somit im Ermessen der einzelnen Schülerinnen und Schüler liegen (Göllner, Wagner, Klieme & Trautwein, 2014). Beispielsweise ist nur für einen kleinen Teil der gesichteten Schülerfragebogen der genaue Zeitbezug für die Beurteilung explizit gegeben. Die vorgenommenen Systematisierungen können als Basis für die Zusammenstellung von Items für zukünftige Schulleistungsstudien dienen, die beispielsweise eine größere Einheitlichkeit aufweisen als bisher genutzte Instrumente.

4.2 Die Bedeutung von Zeitbezügen für die Übereinstimmung von Schülern und Lehrkräften

Inwieweit der Zeitbezug der Beurteilungen der Unterrichtsqualität, der in den systematisch erfassten Items der deutschen Schulleistungsstudien erheblich variiert, die Übereinstimmung verschiedener Datenquellen beeinflusst, war Gegenstand einer vertiefenden Untersuchung von Wagner et al. (im Druck).

In vielen Fällen wird die Unterrichtsqualität von Lehrkräften als eine statische Größe behandelt, die Einfluss auf die Lern- und Leistungsentwicklung ausüben kann, aber selbst keiner Veränderung (z. B. im Laufe eines Schuljahres) unterliegt. Dennoch spielt der Faktor „Zeit" für den Beurteilungsprozess eine nicht unerhebliche Rolle. Erstens unterliegen mit großer Wahrscheinlichkeit auch Konstrukte wie das unterrichtliche Handeln einer Lehrkraft einer Veränderung. Lehrkräfte könnten beispielsweise im Laufe eines Schuljahres auf spezifische Bedürfnisse der Schülerinnen und Schüler reagieren und entsprechende Unterstützungsformen der Schülerorientierung oder die Strukturgebung stärker betonen. Belege für die Veränderbarkeit des Lehrerhandelns finden sich darüber hinaus in einer zunehmend größer werdenden Zahl von Interventionsstudien (z. B. Brown, Jones, LaRusso & Aber, 2010). Zweitens unterliegt vermutlich jegliche Form des Verhaltens natürlichen Schwankungen, die nicht im Sinne systematischer Veränderungen zu erklären, sondern der Spezifität einer bestimmten Situation oder eines Zeitpunktes geschuldet sind. So werden zwar Unterrichtsbeobachtungen oft als die beste Möglichkeit zur neutralen Erfassung des Unterrichts bezeichnet, sie können aber letztlich nur einen zeitlich begrenzten Ausschnitt eines zeitvariablen Verhaltens abbilden (Clausen, 2002). Schließlich ist der zeitliche Bezug in der eigentlichen Befragungssituation ein weiterer wichtiger Zeitaspekt. Der Umstand, dass der Beurteilungszeitraum bei den meisten Instrumenten zur

Erfassung der Unterrichtsqualität nicht explizit genannt wird, verschärft dieses Problem erheblich. Je nachdem, welchen zeitlichen Bezug die Schülerinnen und Schüler in der Befragungssituation wählen, könnten – unter der Annahme von sich über die Zeit veränderndem Lehrerhandeln – sehr unterschiedliche Beurteilungsergebnisse resultieren. Ein Teil der Schülerinnen und Schüler könnte seine Beurteilungen auf das bereits vergangene Schuljahr beziehen, während andere Schülerinnen und Schüler ihr Urteil auf den letzten Eindruck stützen. Beide Fälle würden im Ergebnis nicht nur dazu führen, dass Schülerinnen und Schüler gegebenenfalls zu sehr unterschiedlichen Beurteilungen gelangen, sondern auch die Übereinstimmung der Beurteilungen mit anderen Datenquellen (z. B. Lehrerbericht) beeinträchtigen.

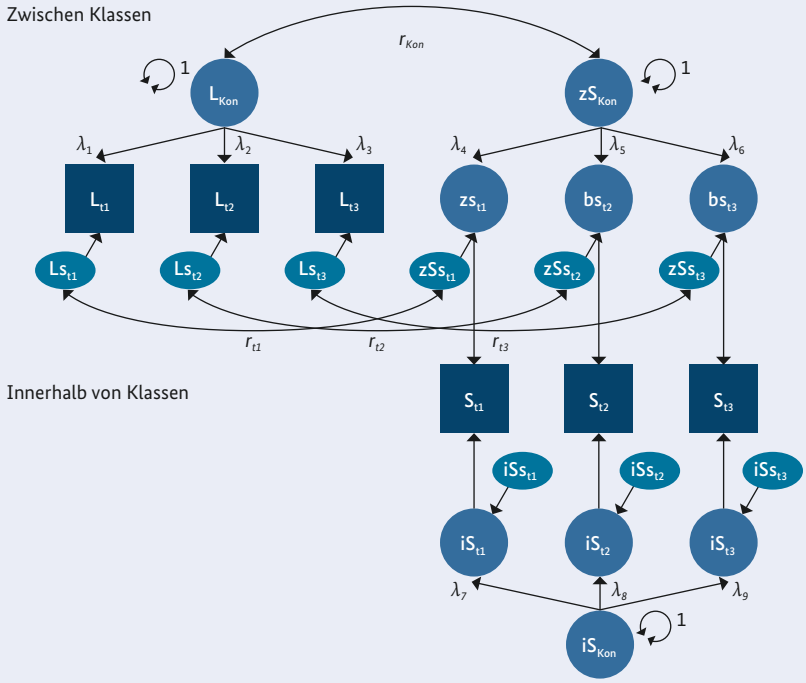
Die Bedeutung des Zeitbezuges bei der Beurteilung des Unterrichts wurde in dem BMBF-Projekt untersucht, indem Daten des ebenfalls vom BMBF geförderten Projekts „Lernen mit Plan" (vgl. Ogrin, Keller, Friedrich, Trautwein & Schmitz, im Druck) reanalysiert wurden. Um eine möglichst genaue Abschätzung möglicher Zeiteffekte zu ermöglichen, wurde ein Fragebogendesign verwendet, welches den Zeitbezug der Beurteilung für Schülerinnen und Schüler genau definierte. Es wurden $N = 686$ Schülerinnen und Schüler der fünften Jahrgangsstufe sowie ihre Mathematiklehrer aus insgesamt 74 Hauptschulklassen wiederholt zur Unterrichtsqualität im Fach Mathematik befragt. Das mittlere Alter der Schüler betrug 11,8 Jahre (48.3 Prozent Mädchen). Die Untersuchung erfolgte zu drei Messzeitpunkten über 13 Wochen hinweg (T1: Mitte des Schuljahres; T2: nach sechs Wochen; T3: nach weiteren sechs Wochen). Zum ersten Messzeitpunkt wurden sowohl die Schülerinnen und Schüler als auch die Lehrkräfte zur Qualität des Mathematikunterrichtes der Lehrkraft befragt. Zu den beiden folgenden Messzeitpunkten bezog sich die Einschätzung der Unterrichtsqualität auf die Zeit seit der letzten Befragung. Erfasst wurden neben der Klassenführung die Merkmale Autonomieunterstützung und Zielklarheit des Unterrichts. Die Instruktion und die gewählten Zeitpunkte der Untersuchung sind aus zweierlei Gründen interessant. Zum einen macht der vergleichsweise kurze Zeitraum der Untersuchung zur Mitte des Schuljahres eine „wirkliche" Veränderung der Unterrichtsqualität eher unwahrscheinlich. Die erste Messung fand nach Bekanntgabe der Halbjahresnoten statt, während die letzte Messung weit vor Ende des Schuljahres abgeschlossen wurde. Etwaige Einflussgrößen, wie die Unbekanntheit der Schülerinnen und Schüler mit der Lehrkraft oder auch eine mögliche Schwerpunktverschiebung des Unterrichts aufgrund bevorstehender Zeugnisnoten, können mit einiger Sicherheit ausgeschlossen werden. Zum Zweiten wurde der zu betrachtende Beurteilungszeitraum explizit definiert, um der Zeitspezifität der einzelnen Messzeitpunkte entsprechend gerecht werden zu können.

Die Ergebnisse eines State-Trait-Modells (siehe Abbildung 2) zeigten für die Schülerurteile (zSKon) mit Ausnahme der Klassenführung eine geringere Zeitspezifität als für Lehrerurteile (LKon). Nahezu die gesamte Variabilität in den Schülerbeurteilungen für die Merkmale Zielklarheit und Autonomieunterstützung auf Klassenebene konnte auf zeitkonsistente Urteilskomponenten zurückgeführt werden. Für das Merkmal Klassenführung lag der entsprechende Varianzanteil bei durchschnittlich 75 Prozent. Lehrerurteile variierten demgegenüber stärker über die Messzeitpunkte hinweg. Etwa drei Viertel der Unterschiede zwischen den Beurteilungen waren auf

zeitlich stabile Urteilskomponenten zurückzuführen, wobei die Unterschiede zwischen den Unterrichtsmerkmalen deutlich weniger variierten als bei den Schülerbeurteilungen. Mit Blick auf die Übereinstimmung der Beurteiler zeigten sich – mit Ausnahme der Autonomieunterstützung – höhere Beurteilerübereinstimmungen bezüglich der zeitkonsistenten Komponenten verglichen mit den „einfachen" messzeitpunktspezifischen Urteilen. Während sich für das Merkmal Klassenführung zu den einzelnen Messzeitpunkten mittlere bis hohe Korrelationen fanden ($r_{t1 =}$.72, r_{t2} = .58 und r_{t3} = .77), ergab sich für die zeitkonsistenten Urteilskomponenten eine nahezu perfekte Übereinstimmung von *r*Kon = 1.0. Ein ähnliches Bild zeigte sich für die Zielklarheit. Auch hier wiesen die zeitkonsistenten Urteilskomponenten der Lehrkräfte und Schülerinnen und Schüler höhere Übereinstimmungen (r_{Kon} = .50) als die einzelnen messzeitpunktspezifischen Urteile (r_{t1}= .45, r_{t2}= .35 und r_{t3}= .32) auf. Lediglich für das Merkmal Autonomieunterstützung konnten weder für die einzelnen Messzeitpunkte (r_{t1}= .05, r_{t2}= .11 und r_{t3}= .04) noch für die zeitkonsistenten Urteilskomponenten substanzielle Übereinstimmungen zwischen Schüler- und Lehrerperspektive nachgewiesen werden.

Zusammenfassend und mit Blick auf die Bedeutung des Zeitbezuges für die Unterrichtsbeurteilungen scheint die Erfassung zeitlich stabiler Unterrichtsmerkmale (im Sinne des „typischen" Unterrichts bei einer bestimmten Lehrkraft) durch Zusammenfassung mehrerer Messzeitpunkte vorteilhaft zu sein. Dies ist für die Unterrichtsforschung ausgesprochen aufschlussreich, da bei einer Vielzahl der bekannten Unterrichtsstudien die Erfassung des Unterrichts nur einmal erfolgte. Die Beurteilung des Unterrichts zum Zeitpunkt einer Messung beinhaltet demnach Messanteile, die für die Erfassung der postulierten Qualitätsmerkmale (z. B. generelle Strukturierung der einzelnen Unterrichtsstunden) weniger relevant sind. Dies zeigt sich sowohl für Lehrer- und Schülerbeurteilungen des Unterrichts, wenngleich für Schülerbeurteilungen das Ausmaß zeitspezifischer Anteile insgesamt geringer zu sein scheint. Die Annahme zeitspezifischer „Verunreinigungen" im Hinblick auf die Messung verschiedener Qualitätsaspekte des Unterrichts bestätigte sich im Rahmen weiterer Analysen zur Vorhersage relevanter Zielkriterien des Unterrichts (d. h. Entwicklung der Mathematikleistung und des Selbstkonzeptes über den Zeitraum der Untersuchung). Die Ergebnisse nachgeschalteter Analysen zeigten, dass beide Zielvariablen durch die drei untersuchten Unterrichtsmerkmale (insbesondere Schülerbeurteilungen) vorhergesagt werden konnten, wenngleich die Effekte in unsystematischer Weise über die Messzeitpunkte streuten. Es ließen sich keine „besseren" oder „schlechteren" Messzeitpunkte im Hinblick auf die Prädiktion der beiden Zielgrößen finden. Vielmehr wechselten sich die Messzeitpunkte sowohl in den Effektstärken als auch der statistischen Signifikanz in unsystematischer Weise ab. Demgegenüber ergab sich für die die zeitstabilen Beurteilungsanteile ein weitaus konsistenteres und theoriekonformeres Bild.

Abb 2: Latentes State-Trait-Modell zur Analyse von zeitkonsistenten und messzeitpunktspezifischen Urteilskomponenten



5 Die Bedeutung sprachlicher Komplexitätsmerkmale

Wird die Bedeutung des Iteminhalts für die psychometrische Qualität von Schülerbeurteilungen des Unterrichts in konsequenter Weise weitergedacht, bleiben mögliche Einflussgrößen nicht nur auf die Wahrnehmungsperspektive sowie den Adressaten- und Zeitbezug beschränkt, sondern umfassen alle schwierigkeitsgenerierenden Merkmale eines Fragebogenitems. So ist davon auszugehen, dass jede Form sprachlicher Komplexität mit der Güte von Schülerbeurteilungen assoziiert ist. Ein weiteres durch das BMBF gefördertes Projekt („Sprachliche Komplexitätsmerkmale von Fragebogenitems: Bedeutung für die psychometrische Qualität von Schülerbeurteilungen des Unterrichts aus Schülersicht und die Vorhersage des Lernerfolgs in Large-Scale-Assessments“, 01LSA1507) ermöglicht die Vertiefung und Erweiterung dieser Forschungsfrage.

5.1 Sprachliche Komplexitätsmerkmale

Modernen Konzepten der Leseforschung folgend, setzt auch die sprachliche Verarbeitung bzw. das sprachliche Verständnis eines Items die Fähigkeit voraus, die dargebotenen Informationen flüssig und sinnverstehend zu lesen. Zur Interpretation

muss der Itemtext in einzelne Elemente zerlegt, die einzelnen Wörter oder Chunks müssen decodiert und deren Bedeutung muss abgerufen werden. Anschließend sind die einzelnen Bestandteile entsprechend den grammatischen Relationen und semantischen Interpretationsprinzipien zu kombinieren, d. h., die Zusammenhänge zwischen den Satzteilen müssen erkannt werden. Diese Verarbeitungsschritte erfordern morphologische, syntaktische und semantische Kompetenzen sowie die Fähigkeit, das Item im Kontext adäquat pragmatisch zu interpretieren. Hierbei gilt es zu bedenken, dass Schriftsprache unabhängig von spezifischen Inhalten andere Chunks und Strukturen aufweist als gesprochene Sprache. Schriftsprachliches Material bzw. konzeptionell schriftliche Sprache zeichnet sich – im Gegensatz zur Alltagssprache bzw. zur konzeptionell mündlichen Sprache – unter anderem durch lange Nominalphrasen, die Verwendung von Passiv sowie lange und komplexe Sätze aus (für einen Überblick siehe Berendes, Dragon, Weinert, Heppt & Stanat, 2013).

Einige dieser Sprachmerkmale konnten in aktuellen Untersuchungen zur bildungssprachlichen Kompetenz von Schülerinnen und Schülern bereits als relevant für die Vergleichbarkeit von Testaufgaben identifiziert werden (z. B. Berendes, Wagner, Meurers & Trautwein, 2015). Es zeigte sich, dass für das Verständnis sprachlicher Inhalte, die in Situationen mit geringer sozialer und situativer Einbettung kommuniziert werden, bildungssprachliche Kompetenzen von direkter Relevanz sind. So fanden beispielsweise Shaftel, Belton-Kocher, Glasnapp und Poggio (2006) in einer Untersuchung mit Viert-, Siebt- und Zehntklässlern, dass Präpositionalphrasen und mehrdeutige Wörter in Testitems einen substanziellen Einfluss auf die Testleistung der Schülerinnen und Schüler ausüben (siehe auch Haag, Heppt, Stanat, Kuhl & Pant 2013).

5.2 Sprachmerkmale und psychometrische Güte von Schülerbeurteilungen des Unterrichts

Auch bei der Bearbeitung von Fragebogenitems zur Unterrichtswahrnehmung ist eher von einer geringen Kontextualisierung auszugehen. Zugleich enthalten Fragebogeninstrumente für Schülerbeurteilungen des Unterrichts durchaus komplexe morphologische Muster und Satzkonstruktionen sowie bildungssprachlich geprägte Begriffe. Auf der Grundlage der im Rahmen des Projektes durchgeführten Sichtung und Klassifikation von Fragebogenitems aus sieben deutschen Schulleistungsstudien wurde untersucht, inwieweit derartige Komplexitätsmerkmale mit der psychometrischen Güte von Schülerbeurteilungen des Unterrichts einhergingen (Göllner, Wagner, Meurers, Berendes & Trautwein, in Vorbereitung). Zu den für die Untersuchung relevanten Größen gehörten neben der relativen Übereinstimmung von Schülerinnen und Schülern weitere Maße der Itemstruktur, wie etwa die mittlere Iteminterkorrelation innerhalb der einzelnen Unterrichtsskalen. Die entsprechenden Indizes wurden anhand von Reanalysen der Originaldatensätze ermittelt. Zudem erfolgte eine Codierung potenziell relevanter Sprachmerkmale. Die Auswahl der Komplexitätsmerkmale orientierte sich an aktuellen Konzepten der Leseforschung sowie an Theorien des Textverstehens (Lenhard & Artelt, 2009). Da es sich bei Items um sehr

kurzes Sprachmaterial auf Satzebene handelt, war die Variationsbreite für eine Vielzahl bekannter Sprachmerkmale erheblich eingeschränkt. Für die Analysen wurden daher nur acht Sprachmerkmale für insgesamt 98 der 533 Items herangezogen. Dazu gehörten die Anzahl der Zeichen pro Wort, die Wortanzahl pro Satz, die mittlere Wortanzahl in Haupt- und Nebensätzen, die Anzahl der Nebensätze, die Anzahl von Nominalphrasen sowie die Anzahl von Hypernymen, Hyponymen und Synonymen pro Sinneinheit. Die Auswahl der Items erfolgte anhand zweier Kriterien. Erstens entstammten alle Items aus Studien mit dem Fokus Mathematik und kamen zudem in mindestens zwei Studien zur Anwendung. Die Ergebnisse zeigten, dass sich die Items im Hinblick auf die codierten Sprachmerkmale deutlich unterschieden. Das in der Analyse genutzte Gütekriterium von Schülerbeurteilungen war der Intraklassenkorrelationskoeffizient (ICC) als Maß der relativen Übereinstimmung von Schülerinnen und Schülern einer Klasse (Snijders & Bosker, 1999). Die mittlere ICC der 98 Items betrug .14 (*Min* = .02, *Max* = .31; *SD* = 0.07).

Im Anschluss daran wurden systematische Analysen zur Frage durchgeführt, ob sprachliche Komplexität das Antwortverhalten der Schülerinnen und Schüler beeinflusst. Hierbei wurde von der Annahme ausgegangen, dass sprachlich komplexe Items von den Schülerinnen und Schülern einer Klasse unterschiedlich gut verstanden und interpretiert werden. Sollten diese Unterschiede bestehen, würden bei entsprechend sprachlich komplexen Items die Schülerurteile innerhalb einer Klasse relativ unterschiedlich ausfallen, selbst wenn in Wirklichkeit die Schülerinnen und Schüler eine eher einheitliche Meinung zur abgefragten Qualitätsdimension haben. Eine hohe sprachliche Komplexität würde sich dementsprechend in relativ niedrigen ICCs ausdrücken. Tatsächlich konnte ein entsprechendes Muster für die Anzahl von Nebensätzen und die Anzahl von Nominalisierungen gefunden werden, nicht aber für die anderen Sprachmerkmale. Die beiden Komplexitätsmerkmale blieben auch dann prädiktiv, wenn für die Herkunftsstudie – und somit auch für Unterschiede bezüglich der Stichprobenzusammensetzungen – des Items statistisch kontrolliert wurde.

6 Zusammenfassung und Ausblick

Schülerurteile der Unterrichtsqualität bieten nicht nur eine hohe Effektivität bei der Erfassung bzw. Messung lern- und leistungsförderlicher Qualitätsmerkmale des Unterrichts, sondern versprechen auch eine umfangreiche und detaillierte Beschreibung des Unterrichtsgeschehens. Die Tatsache, dass Schülerbeurteilungen in vielen Fällen deutlich stärker mit verschiedenen Zielkriterien des Unterrichts assoziiert sind als Selbstberichte von Lehrkräften oder Beobachterdaten, unterstreicht zusätzlich die Bedeutung von Schülerbeurteilungen der Unterrichtsqualität in Large-Scale-Assessments. Dem wurde vielfach entgegengehalten, dass Schülerinnen und Schüler keine adäquate didaktisch-pädagogische Expertise besitzen, um komplexe Unterrichtsprozesse angemessen einordnen zu können, und die eigene Involviertheit von Schülerinnen und Schülern zu einer mangelnden Objektivität während der Beurteilung führen könnte (Aleamoni, 1999; Clausen, 2002; Kunter & Baumert, 2006).

Die Frage, wie angemessen Unterrichtsqualität mithilfe von Schülerurteilen erfasst werden kann, war entsprechend das zentrale Anliegen des vom BMBF geförderten Projekts „Erfassung der Unterrichtsqualität in Large-Scale-Studien: Optimierung der Modellierung und Itemauswahl". In dem Projekt entstanden eine Reihe von empirischen Studien, welche a) die Konstruktvalidität von Schülerbeurteilungen, b) die Bedeutung des Zeitbezugs für die Schüler-Lehrer-Übereinstimmung und c) die Bedeutung sprachlicher Komplexitätsmerkmale für die relative Übereinstimmung von Schülerbeurteilungen näher untersuchten.

In der Rückschau zeigen die Ergebnisse, dass Schülerinnen und Schüler – zumindest für eine Reihe von zentralen Qualitätsdimensionen – die Qualität des Unterrichts nicht nur zuverlässig, sondern auch differenziert in der dafür notwendigen Detailliertheit beschreiben können. Befragungen von Schülerinnen und Schülern ermöglichen somit eine Quantifizierung verschiedener Qualitätsaspekte des Unterrichts, die als zentrale Elemente eines lern- und leistungsförderlichen Unterrichts gelten. Dabei erlauben Schülerbeurteilungen der Unterrichtsqualität nicht nur, das Geschehen in seiner Breite zu erfassen, sondern können auch zur Erfassung von weniger gut beobachtbaren Unterrichtsmerkmalen dienen. So ist mit Blick auf die durchgeführte Untersuchung zur faktoriellen Struktur von Schülerbeurteilungen eine vergleichbar umfangreiche Erfassung relevanter Qualitätsdimensionen im Rahmen einer Videostudie kaum oder nur durch enormen Aufwand vorstellbar. Darüber hinaus bieten Schülerbeurteilungen der Unterrichtsqualität in Kombination mit neueren Verfahren der Datenanalyse weitreichende Möglichkeiten, die zentralen Gütemaßstäbe pädagogisch-psychologischer Forschung erweiternd zu prüfen. Beispielsweise besitzt das im Rahmen des Projektes entwickelte Analysevorgehen zur Äquivalenzprüfung von Schülerbeurteilungen über die Analyseeinheit der Klassen hohe Relevanz für die Überprüfung zukünftiger Fragebogeninstrumente in der Unterrichtsforschung.

Derartige Verfahren zeigen jedoch auch die Grenzen von Schülerbeurteilungen des Unterrichts und die Notwendigkeit weiterer Entwicklungsanstrengungen im Bereich der Unterrichtsforschung. Insbesondere sprachliche Anforderungs- und Komplexitätsmerkmale sind in ihrer Bedeutung für den Beantwortungsprozess zu berücksichtigen. Existierende Fragebogenverfahren weisen eine erstaunlich hohe Variation sprachlicher Anforderungsmerkmale auf, welche Einfluss auf die psychometrische Qualität von Schülerbeurteilungen des Unterrichts ausüben können. Beispielsweise lässt sich die Verletzung der Messäquivalenzannahme für einige der im Projekt untersuchten Unterrichtmerkmale nicht einseitig auf spezifische Charakteristika des Merkmals zurückführen, sondern könnte gleichfalls Ausdruck unterschiedlicher kognitiver Anforderungen während der Beurteilung sein. Die für die fünf Merkmale verwendeten Items zeichnen sich ausnahmslos durch einen hohen Inferenzgrad aus, welcher die Abhängigkeit des Messergebnisses unter Umständen deutlich erhöht.

Doch auch die Frage, ob die psychometrische Qualität von Schülerbeurteilungen darüber hinaus von weiteren Anforderungsmerkmalen abhängig ist, kann vorsichtig mit einem Ja beantwortet werden. Es wurde gezeigt, dass eine mangelnde Übereinstimmung zwischen Schülerinnen und Schülern und Lehrkräften in ihrer Beurteilung der Unterrichtsqualität zu einem substanziellen Anteil auf die Spe-

zifität des Messzeitpunkts zurückgeführt werden kann. Die geschieht sogar dann, wenn der Beurteilungszeitraum in der Fragebogeninstruktion explizit definiert ist. Die Situationsabhängigkeit einer Messung resultiert zumindest für einige der untersuchten Merkmale in einer Unterschätzung der Übereinstimmung und somit in einer weniger validen Messung. Die Tatsache, dass in der Mehrzahl angewendeter Fragebogeninstrumente kein Zeitbezug gegeben ist, verleiht notwendigen Anstrengungen zur Verbesserung von Fragebogeninstrumenten zusätzliches Gewicht. Diese Anstrengungen sollten sich jedoch nicht nur auf Anforderungsmerkmale wie etwa Perspektive, Adressaten- oder Zeitbezug in der Fragebogenkonstruktion beschränken, sondern auch weitere Komplexitätsmerkmale der Formulierung umfassen. Dies gilt insbesondere dann, wenn die Erfassung der Unterrichtsqualität nicht nur auf Schülerinnen und Schüler höherer Klassenstufen beschränkt bleiben soll. Die im Rahmen des bereits genannten Projektes „Sprachliche Komplexitätsmerkmale von Fragebogenitems" gefundenen Ergebnisse bezüglich der Urteilsübereinstimmung von Schülerinnen und Schülern lässt vermuten, dass die psychometrische Güte der Befragungsinstrumente bei jüngeren Schülerinnen und Schülern noch deutlich stärker beeinflusst ist. Entsprechende Weiterentwicklungen sind erforderlich, um Untersuchungsinstrumente für Schulleistungsuntersuchungen zu generieren, mit deren Hilfe noch differenzierter als bisher die Determinanten und Konsequenzen guten Unterrichts (vgl. Kunter & Trautwein, 2013) identifiziert werden können.

Literaturverzeichnis

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal Of Personnel Evaluation In Education, 13,* 153–166.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9,* 469–520.

Berendes, K., Dragon, N., Weinert, S., Heppt, B. & Stanat, P. (2013). Hürde Bildungssprache? Eine Annäherung an das Konzept „Bildungssprache" unter Einbezug aktueller empirischer Forschungsergebnisse. In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik. Perspektiven aus Psychologie, Sprachwissenschaft und empirischer Bildungsforschung* (S. 17–41). Münster: Waxmann.

Berendes, K., Wagner, W., Meurers, D. & Trautwein, U. (2015). Grammatikverständnis von Kindern unterschiedlicher sprachlicher und sozioökonomischer Herkunft. *Frühe Bildung, 4 (3),* 126–134.

Blömeke, S. (2004). Empirische Befunde zur Wirksamkeit der Lehrerbildung. In S. Blömeke, P. Reinhold, G. Tulodziecki & J. Wildt (Hrsg.). *Handbuch Lehrerbildung* (S. 59–91). Bad Heilbrunn: Klinkhardt.

Brown, J. L., Jones, S. M., LaRusso, M. D. & Aber, J. L. (2010). Improving classroom quality: Teacher influences and experimental impacts of the 4Rs program. *Journal of Educational Psychology, 102 (1),* 153–167.

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

den Brok, P., Brekelmans, M. & Wubbels, T. (2006). Multilevel issues in reasearch using students' perceptions of learning environments: The case of the questionnaire on teacher interaction. *Learning Environments Research, 9,* 199–213.

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L., Sherin, M. G. & Sherin, B. L. (2010). Conducting Video Research in the Learning Sciences: Guidance on Selection, Analysis, Technology, and Ethics. *Journal of The Learning Sciences, 19,* 3–53.

Desimone, L. M., Smith, T. M. & Frisvold, D. E. (2010). Survey Measures of Classroom Instruction: Comparing Student and Teacher Reports. *Educational Policy, 24,* 267–329.

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg. *Zeitschrift für Pädagogische Psychologie, 28,* 127–137.

Fisicaro, S. A. & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14,* 419–429.

Fraser, B. J. & Walberg, H. J. (1991). *Educational environments: Evaluation, antecedents and consequences.* Elmsford, NY: Pergamon Press.

Funder, D. C. (2001). Accuracy in personality judgment: Research and theory concerning an obvious question. In B. W. Roberts, R. Hogan (Hrsg.), *Personality psychology in the workplace* (S. 121–140). Washington, DC: American Psychological Association.

Gigliotti, R. J. & Buchtel, F. S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology, 82,* 341–351.

Göllner, R., Wagner, W., Klieme, E. & Trautwein, U. (2014). *Die Erfassung des Unterrichts aus Schülersicht: Ergebnisse einer Systematisierung nationaler Fragebogeninstrumente.* Vortrag auf der 2. Tagung der Gesellschaft für Empirische Bildungsforschung, Frankfurt.

Göllner, R., Wagner, W., Meurers, D. W., Berendes, K. & Trautwein, U. (2016). *When Questions Unintentially Shape the Answers: Psycholinguistic Item Features Predict Student Ratings of Instructional Quality.* Manuskript in Vorbereitung.

Greenwald, A. G. & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209–1217.

Haag, N., Heppt, B., Stanat, P., Kuhl, P. & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction, 28,* 24–34.

Hamre, B. K. & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. L. Meece & J. S. Eccles (Hrsg.), *Handbook of Research on Schools, Schooling and Human Development* (S. 25–41). New York, NY: Routledge.

Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses relating to achievement.* London: Routledge.

Helmke, A. (2010). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts.* Seelze: Klett-Kallmeyer.

Hox, J. J. & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8,* 157–174.

Jak, S., Oort, F. J. & Dolan, C. V. (2013). A Test for Cluster Bias: Detecting Violations of Measurement Invariance Across Clusters in Multilevel Data. S*tructural Equation Modeling, 20,* 265–282.

Klieme, E., Eichler, W., Helmke, A., Lehmann, R., Nold, G., Rolff, H. G., Schröder, K., Thomé, G. & Willenberg, H. (Hrsg.). (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI).* Weinheim: Beltz.

Klieme, E., Jude, N., Rauch, D., Ehlers, H., Helmke, A., Eichler, W., et al. (2008). Alltagspraxis, Qualität und Wirksamkeit des Deutschunterrichts. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 319–344). Weinheim: Beltz.

Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: Aufgabenkultur und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.), *TIMSS-Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 43–57). München: Medienhaus Biering.

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht.* 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,* 231–251.

Kunter, M. & Trautwein, U. (2013). *Psychologie des Unterrichts.* Stuttgart: UTB.

Lenhard, W. & Artelt, C. (2009). Komponenten des Leseverständnisses. In W. Lenhard & W. Schneider (Hrsg.), *Diagnose und Förderung von Leseverständnis und Lesekompetenz* (S. 1–18). Göttingen: Hogrefe.

Lenske, G. (2011). Pupils as raters of instructional quality: Does it work in primary school? In K. Ruhl (Hrsg.), *Das Poster in der Wissenschaft. Zum Stellenwert des Posters in der Nachwuchsförderung am Beispiel der Universität Koblenz-Landau.* Gießen: Johannes Herrmann.

Lüdtke, O., Marsh, H. W., Robitzsch, A. & Trautwein, U. (2011). A 2 x 2 Taxonomy of Multilevel Latent Contextual Models: Accuracy-Bias Trade-Offs in Full and Partial Error Correction Models. *Psychological Methods, 16,* 444–467.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings in multilevel modeling. *Contemporary Educational Psychology, 34,* 120–131.

Lüdtke, O., Trautwein, U., Kunter, M. & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment – A re-analysis of TIMSS data. *Learning Environments Research, 9,* 215–230.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47,* 106–124.

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,* 1187–1197.

Mehta, P. D. & Neale, M. C. (2005). Peoples are variables too: Multilevel structural equation modeling. *Psychological Methods, 3,* 259–284.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22 (3),* 376–398.

Nunnally, J. C. (1978). *Psychometric theory* (2. Auflage). New York, NY: McGraw-Hill.

Ogrin, S., Keller, S., Friedrich, A., Trautwein, U. & Schmitz, B. (im Druck). Entwicklung und empirische Prüfung einer Lehrkräftefortbildung zur Förderung von Selbstregulationskompetenz und mathematischer Kompetenz bei Schülerinnen und Schülern der Haupt- und Werkrealschule („Lernen mit Plan"). In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals. Interdisziplinäre Betrachtungen, Befunde und Perspektiven.* Berlin: Springer Online.

Rosenshine, B. (1970). Evaluation of classroom instruction. *Review of Educational Research, 40,* 279–300.

Scheerens, J. & Bosker, J. (1997). The foundations of educational effectiveness. Oxford: Elsevier.

Seidel, T. & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77,* 454–499.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11,* 105–126.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Tourangeau, R., Rips, L.J. & Rasinski, K. (2000). *The psychology of survey response.* New York, NY: Cambridge University Press.

Trautwein, U., Lüdtke, O., Klieme, E., Nagengast, B. & Wagner W. (2011). *Erfassung der Unterrichtsqualität in Large-Scale-Studien: Optimierung der Modellierung und Itemauswahl.* Unveröffentlichter BMBF-Antrag. Tübingen: Eberhard Karls Universitat Tübingen.

Vazire, S. & Solomon, B. C. (2015). Self- and other-knowledge of personality. In M. Mikulincer, P. R. Shaver, M. L. Cooper & R. Larsen (Hrsg.), *APA handbook of personality and social psychology, Vol. 4 Personality processes and individual differences* (S. 261–281). Washington, DC: American Psychological Association.

Wagner, W. (2008). *Methodenprobleme bei der Analyse der Unterrichtswahrnehmung und -wirksamkeit – am Beispiel der Studie DESI (Deutsch-Englisch-Schülerleistungen-International) der Kultusministerkonferenz.* Dissertation. Universität Koblenz-Landau, Campus Landau, Fachbereich Psychologie.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U. & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimen-

sionality and generalizability of domain-independent assessments. *Learning and Instruction, 28,* 1–11.

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B. & Trautwein, U. (im Druck). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*.

Widaman, K. F. & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Hrsg.), *The science of prevention: Methodological advances from alcohol and substance abuse research.* Washington, DC: American Psychological Association.

Wubbels, T., Brekelmans, M. & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching And Teacher Education, 8,* 47–58.