Titel:

Zur Variabilität von Unterrichtsqualität zwischen und innerhalb von Unterrichtsstunden: Generalisierbarkeit generischer und fachspezifischer Merkmale im Mathematikunterricht

Variability of instructional quality between and within lessons: generalizability of generic and subject-specific characteristics in mathematics instruction

Autorinnen und Autoren:

Armin Jentsch, Fakultät Erziehungswissenschaft, Universität Hamburg, Von-Melle-Park 8, 20146 Hamburg, Festnetz: +49 40 42838 6422, E-Mail: armin.jentsch@uni-hamburg.de, ORCID-ID: 0000-0002-2423-3955

Dr. Gino Casale, Department Heilpädagogik und Rehabilitation, Humanwissenschaftliche Fakultät, Universität zu Köln, Klosterstr. 79c, 50931 Köln

Lena Schlesinger, Fakultät Erziehungswissenschaft, Universität Hamburg, Von-Melle-Park 8, 20146 Hamburg

Prof. Dr. Gabriele Kaiser, Fakultät Erziehungswissenschaft, Universität Hamburg, Von-Melle-Park 8, 20146 Hamburg

Prof. Dr. Johannes König, Empirische Schulforschung, Quantitative Methoden, Department Erziehungs- und Sozialwissenschaften, Humanwissenschaftliche Fakultät, Universität zu Köln, Gronewaldstr. 2a, 50931 Köln

Prof. Dr. Sigrid Blömeke, Center for Educational Measurement Oslo (CEMO), Gaustadalléen 30d, 0373 Oslo, Norwegen

Zusammenfassung:

Der Messung von Unterrichtsqualität durch Beobachterratings ist in den letzten Jahren große Aufmerksamkeit zuteilgeworden, wobei der Fokus eher auf generischen Merkmalen und der Variabilität der Unterrichtsqualität zwischen Unterrichtsstunden lag. Inwieweit jedoch auch innerhalb von Unterrichtsstunden Unterschiede in der Unterrichtsqualität bestehen und ob derartige Erhebungen zu verschiedenen Messzeitpunkten überhaupt vergleichbar sind, wurde bisher kaum untersucht. Des Weiteren liegen bisher nur wenige Befunde zur Variabilität fachspezifischer Merkmale der Unterrichtsqualität vor. Die im Artikel dargestellte Studie knüpft an diese Desiderate an und untersucht die Generalisierbarkeit von fachspezifischen und generischen Merkmalen der Unterrichtsqualität im Mathematikunterricht und zwar zwischen und innerhalb von Unterrichtsstunden. Es wurden jeweils zwei Doppelstunden von 37 Lehrpersonen beobachtet. Die Einschätzung der Unterrichtsqualität erfolgte viermal je Doppelstunde durch geschulte Beobachterinnen und Beobachter mittels hoch-inferenter Ratingskalen. Diese beschreiben insgesamt fünf Merkmale, von denen drei die bekannten Basisdimensionen der Unterrichtsqualität (Klassenführung, konstruktive Unterstützung und kognitive Aktivierung) und zwei zusätzliche fachspezifische Merkmale darstellen. Die Ergebnisse legen Messinvarianz zwischen und innerhalb Unterrichtsstunden sowie eine weitgehend akzeptable Generalisierbarkeit nahe.

Schlüsselwörter:

Unterrichtsqualität, Unterrichtsbeobachtungen, Messinvarianz, Generalisierbarkeitstheorie, Reliabilität

Abstract:

In order to measure instructional quality observer ratings have been used frequently within the last decade, mainly with a focus on generic dimensions and on between-lesson variability of instructional quality. However, there is a lack of research on within-lesson variability and on measurement invariance when instructional quality is evaluated on multiple time points. Furthermore, hardly any studies on subject-specific characteristics of instructional quality exist. The study presented in this paper therefore departs from these desiderata and focuses on the generalizability of both generic and subject-specific characteristics of instructional quality in mathematics instruction. Trained raters observed 37 teachers during two mathematics lessons. Instructional quality was evaluated four times per lesson by using high-inferent rating scales. We used three scales to measure the basic dimensions of instructional quality (classroom management, student support, cognitive activation) and two scales for subject-specific characteristics. Results indicate measurement invariance between and within lessons with mainly satisfying generalizability.

Keywords:

Instructional quality, observer ratings, measurement invariance, generalizability theory, reliability

1. Einleitung

Der Forschung zur Unterrichtsqualität ist in den letzten Jahren große Aufmerksamkeit zugekommen. Neben inhaltlichen Fragestellungen zur Beschaffenheit des Konstrukts (Ditton 2006; Klieme und Rakoczy 2008; Reusser 2009) und zur Wirksamkeit von Unterricht (Baumert et al. 2010; Brophy 2000; Seidel und Shavelson 2007) hat sich die empirische Unterrichtsforschung vermehrt mit methodischen Schwierigkeiten der Erfassung von Unterrichtsqualität durch Beobachterratings auseinandergesetzt (zusammenfassend Autorengruppe anonymisiert 2016). Zur Reliabilität und Validität solcher Beobachterratings liegen zahlreiche Befunde vor (u.a. Praetorius et al. 2012; Rakoczy 2008). Dabei ist mittlerweile gut belegt, dass Studienergebnisse auch von der Beobachterperspektive abhängen: So konvergieren die Einschätzungen von Lehrpersonen, ihren Schülerinnen und Schülern sowie externen Beobachterinnen und Beobachtern häufig nicht, wenn es um die Beurteilung derselben Unterrichtsstunde geht (Clausen 2002; Fauth et al. 2014).

Weniger bekannt ist dagegen, inwiefern Merkmale von Unterrichtsqualität über die Zeit variieren (Kuger et al. 2016; Praetorius et al. 2014). Das ist vor allem deshalb problematisch, weil geklärt werden müsste, ob beobachtete Unterschiede in der Unterrichtsqualität bei einer Messwiederholung auf die Instabilität des Merkmals oder auf Messfehler zurückzuführen sind (Kane 2011; Meyer et al. 2011). Die empirische Überprüfung dieser Frage erfolgt nach unserem Empfinden in der Literatur i. d. R. nicht.

Außerdem liegen unseres Wissens nach aktuell keine Untersuchungen vor, die die zeitliche Variation fächerübergreifender und fachspezifischer Merkmale von Unterrichtsqualität vergleichend analysieren. Das Ziel dieses Beitrags ist, die Variabilität von Unterrichtsqualität zwischen und innerhalb von Unterrichtsstunden zu untersuchen und dabei sowohl generische als auch fachspezifische Merkmale in den Blick zu nehmen. Dazu wird zunächst der Forschungsstand zur Konzeptualisierung und Stabilität von Unterrichtsqualität unter

Bezugnahme auf die drei bekannten Basisdimensionen (Klieme und Rakoczy 2008) sowie fachspezifische Ansätze für den Mathematikunterricht skizziert. Sodann werden zwei Fragestellungen zur Beschreibung und Erklärung der Variabilität von Unterrichtsqualität im Mathematikunterricht entwickelt, die in dieser Studie bearbeitet wurden.

2. Theorierahmen und Forschungsstand

2.1 Generische und fachspezifische Merkmale von Unterrichtsqualität

Unterricht kann sowohl hinsichtlich seiner Prozessqualität als auch hinsichtlich des Lernerfolgs von Schülerinnen und Schülern beurteilt werden (Ditton 2006). In der Unterrichtsforschung werden im Anschluss an die TIMSS-Videostudien (Third International Mathematics and Science Study, Hiebert et al. 2003) drei Merkmalsbereiche erfolgreicher Unterrichtsprozesse unterschieden. Diese werden auch als Basisdimensionen der Unterrichtsqualität bezeichnet (Klieme und Rakoczy 2008) und zielen auf die Beschreibung der Tiefenstruktur von Unterrichtsprozessen ab (Reusser 2009). Im Einzelnen sind das eine effiziente Klassenführung, kognitive Aktivierung und konstruktive Unterstützung (auch: Schülerorientierung).

Mit effizienter Klassenführung ist in Anlehnung an Kounin (1970) ein störungspräventives Verhalten der Lehrperson gemeint, das zu einer optimalen Nutzung der Lernzeit führen soll und Schülerinnen und Schülern überhaupt erst den organisatorischen Rahmen für erfolgreiche Lernprozesse bietet (Helmke 2012; Seidel und Shavelson 2007). Kognitive Aktivierung wird problemorientierten Lernsituationen attestiert, die auf das konzeptuelle Verständnis des Lerngegenstands abzielen (Baumert et al. 2010). Damit ist eine gemäßigt konstruktivistische Lerntheorie verbunden, gemäß der Wissen stets das Ergebnis individueller Konstruktionsleistungen ist (Reinmann-Rothmeier und Mandl 2006). Schließlich thematisiert die Basisdimension konstruktive Unterstützung in Anlehnung an die Selbstbestimmungstheorie (Deci und Ryan 1985; Rakoczy 2008), inwieweit Unterricht auf

motivationale Grundbedürfnisse von Schülerinnen und Schülern eingeht. Diese drei Basisdimensionen stellen eine fächerübergreifende Konzeptualisierung von Unterrichtsqualität dar.

In den letzten Jahren ist vermehrt die zusätzliche Erfassung fachspezifischer Aspekte von Unterrichtsqualität in den Blick geraten (Brunner 2017; Charalambous und Praetorius 2018; Autorengruppe anonymisiert 2016). So führt Blum (2012) an, dass die von Seiten der Bildungspolitik für den Mathematikunterricht geforderte Kompetenzorientierung sowie der damit verbundene fachliche Gehalt des Unterrichts keine Entsprechung in dem Modell der drei Basisdimensionen finde. Brunner (2017) bemängelt, dass insbesondere die fachliche Korrektheit der im Unterricht präsentierten Inhalte ein aus fachdidaktisch-normativer Sicht entscheidendes Qualitätsmerkmal sei, das im Modell der drei Basisdimensionen nicht berücksichtigt werde. Empirische Nachweise zum Zusammenhang zwischen fachspezifischen Aspekten von Unterrichtsqualität und Schülerleistungen liegen für den Mathematikunterricht aus dem amerikanischen Raum (z.B. Hill et al. 2005; Learning Mathematics for Teaching Project 2012) sowie aus der deutsch-schweizerischen Videostudie vor (Pythagoras-Studie, Drollinger-Vetter, 2011).

2.2 Generalisierbarkeitstheorie

Die Generalisierbarkeitstheorie (G-Theorie, Cronbach et al. 1972) hat sich insbesondere für Beobachtungsstudien als eine nützliche Erweiterung der klassischen Testtheorie (KTT) erwiesen, mit der die Variabilität von Unterrichtsqualität untersucht und die Reliabilität einer entsprechenden Messung geschätzt werden kann (z. B. Praetorius et al. 2014). Im Gegensatz zur KTT, bei der die Zerlegung einer Beobachtung in einen wahren Wert und einen zufälligen Fehlerterm vorgenommen wird, erlaubt die G-Theorie die simultane Untersuchung multipler Varianzquellen einer Messung im Rahmen eines varianzanalytischen Designs mit Zufallseffekten (Brennan 2001). Im Kontext der Forschung zur Unterrichtsqualität sind die

Untersuchungsobjekte (Shavelson und Webb 1991) in der Regel die beobachteten Lehrpersonen bzw. ihr Unterricht. Als weitere Varianzquellen kommen je nach Studiendesign und Forschungsinteresse z. B. Rater, Testaufgaben oder Messwiederholungen in Betracht. Inwieweit diese im Rahmen einer bestimmten Untersuchung als Fehlerquellen in die Analyse eingehen, ist abhängig von Entscheidungen des Forschenden (zusammenfassend Praetorius 2014). Die Schätzung der entsprechenden Varianzkomponenten erlaubt dann die Bestimmung eines für das Studiendesign spezifischen Generalisierbarkeitskoeffizienten (Cronbach et al. 1972), der analog zur Reliabilität in der KTT interpretiert wird.

Da im Rahmen einer G-Studie nicht nur zufällige, sondern auch systematische Fehlervarianz identifiziert werden kann (z. B. Rater-Bias, Praetorius et al. 2012), kann eine G-Studie auch als Validitätsprüfung verstanden werden (Hill et al. 2012; Mashburn et al. 2014). Dies ist dann möglich, wenn Forschende die Vergleichbarkeit zwischen verschiedenen Messgelegenheiten sicherstellen wollen, wenn also geklärt werden soll, ob diese als feste oder zufällige Effekte zu behandeln sind.

In der Regel wird sowohl beim Untersuchungsobjekt als auch bei den Fehlerquellen von zufälligen Effekten ausgegangen, d. h. es gibt beliebig viele mögliche Ausprägungen der entsprechenden Variablen und die durch die Stichprobe gegebene Realisation stellt eine Zufallsziehung mit verschwindendem Erwartungswert aus dem entsprechenden *Universum* dar (Shavelson und Webb 1991). Insbesondere weist eine solche Annahme darauf hin, dass von den Ausprägungen, die in einer gegebenen Untersuchung vorliegen, auf alle möglichen Ausprägungen geschlossen werden soll (also z. B. von zwei *beobachteten* auf *beliebige* Unterrichtsstunden). Dies ist bei der Spezifikation von festen Effekten in der G-Theorie nicht der Fall (Praetorius 2014); hier geht es darum, eine möglichst präzise Aussage für die in der Untersuchung vorliegenden Realisierungen einer Variable zu treffen (Unterrichtsstunde 1 vs. Unterrichtsstunde 2).

In der Unterrichtsforschung wird die Entscheidung eines Forschenden, Messwiederholungen als zufällige bzw. feste Effekte zu behandeln, kontrovers diskutiert. So bemängelt Kane (2011), dass in vielen Fällen unzulässig über Messgelegenheiten generalisiert würde, ohne die dazu notwendige Stabilität des Merkmals bzw. Messinvarianz geprüft zu haben. Messinvarianz zwischen Beobachtungszeitpunkten (z. B. Unterrichtsstunden) ist maßgeblich für weitere Analysen zur Variabilität der Unterrichtsmerkmale über die Zeit und daher insbesondere für die Frage, ob der Effekt einer Messwiederholung als zufällig oder fest betrachtet werden sollte.

2.3 Zeitliche Variabilität von Unterrichtsqualität

Bisher liegt keine zusammenfassende Untersuchung des Forschungsstandes zur Stabilität bzw. Variabilität von Unterrichtsqualität im Rahmen eines systematischen Literatur-Surveys oder einer Metaanalyse vor (Kuger et al. 2016). Mit wenigen Ausnahmen (Calkins et al. 1997; Hill et al. 2012; Newton 2010) fokussieren derartige Untersuchungen auf generische Merkmale von Unterrichtsqualität. Befunde liegen dabei insbesondere in Form von Generalisierbarkeitsanalysen (Cronbach et al. 1972) für das CLASS-Instrument (Classroom Assessment Scoring System, Pianta und Hamre, 2009) sowie das Beobachtungsinstrument zur deutsch-schweizerischen Videostudie (Pythagoras-Studie, Rakoczy und Pauli, 2006) vor, das eine Operationalisierung der drei Basisdimensionen darstellt.

Praetorius et al. (2014) untersuchten deren Variabilität bei 38 Lehrpersonen, von denen jeweils zwei Unterrichtseinheiten mit mehreren Unterrichtsstunden im Abstand von einigen Monaten in den Analysen berücksichtigt wurden. Dabei zeigte sich für Klassenführung und konstruktive Unterstützung eine geringe Variabilität zwischen Unterrichtsstunden (13% bzw. 5% der Gesamtvarianz), während kognitive Aktivierung stark zwischen Unterrichtsstunden variierte (fast 50% der Gesamtvarianz). Mashburn et al. (2014) fanden in einer Experimentalstudie mit 47 Lehrpersonen, bei denen jeweils drei Unterrichtsstunden während

eines Schuljahres mit dem CLASS-Instrument beurteilt wurden, dass die Variabilität zwischen und innerhalb von Unterrichtsstunden etwa 5 bis 24% der Gesamtvarianz betrug. Der Varianzanteil war dabei abhängig vom beobachteten Unterrichtsmerkmal und von der Experimentalbedingung, d. h. der Anzahl der Messzeitpunkte pro Unterrichtsstunde.

Für das fachspezifische Instrument MQI (Mathematics Quality of Instruction, Hill et al. 2012) weisen die beobachteten Merkmale zumeist weniger als 10% instabile Varianzanteile auf. Allerdings ist der Anteil der Residualvarianz hier hoch ausgeprägt (ca. ein Drittel der Gesamtvarianz) und kann instabile Varianzanteile in Form eines Interaktionseffekts höchster Ordnung enthalten (Brennan 2001). In der Studie von Newton (2010) wurde die Qualität von Mathematikunterricht in Primar- und Sekundarschulen mit neun Ratingskalen erfasst. Die Ergebnisse zeigen deutliche Unterschiede in der zeitlichen Variabilität zwischen diesen Dimensionen (0-40% der Gesamtvarianz), der Anteil der Residualvarianz fällt mit etwa 25% aber etwas niedriger aus als in der Untersuchung von Hill et al. (2012).

Allen bisher genannten Studien ist gemein, dass zeitliche Variabilität varianzanalytisch als Zufallseffekt behandelt wurde, die „wahre“ Unterrichtsqualität wurde also als über die Zeit unverändert angenommen. Die folgenden Studien nutzten einen anderen methodischen Ansatz, bei dem Zeitpunkte als feste Effekte behandelt werden. Meyer et al. (2011) argumentierten in ihrer Studie auf diese Weise gegen die Annahme einer unveränderten Ausprägung der Unterrichtsqualität und damit gegen eine Generalisierbarkeit auf das Universum aller Unterrichtsstunden, wenn zwischen den Messgelegenheiten große Zeiträume liegen. Sie untersuchten dazu die Unterrichtsqualität in 118 Klassen zu mehreren Zeitpunkten über ein Schuljahr. Casabianca et al. (2015) sowie Malmberg et al. (2010) untersuchten Unterrichtsqualität sogar über zwei Jahre.

Alle drei Studien verwendeten das CLASS-Instrument. In der Studie von Malmberg et al. (2010) wurden Novizen im Lehramt untersucht, bei denen sich über die verschiedenen

Messzeitpunkte hinweg erwartungsgemäß Unterschiede zeigten, die als Professionalisierung interpetiert wurden. Casabianca et al. (2015) diskutierten die zeitliche Variabilität von Unterrichtsqualität vor dem Hintergrund von Raterfehlern und schätzten Trendkurven für die Entwicklung der Unterrichtsqualität. Die instabilen Varianzanteile machten in dieser Untersuchung etwa 38% der Gesamtvarianz aus, zusätzlich entfielen 18% auf die Residualkomponente, was von der Autorengruppe insgesamt als mittlere Variabilität interpretiert wird (Casabianca et al. 2015). Patrick und Mantzicopoulos (2016) führten eine Analyse individueller Trends bei acht Lehrpersonen durch. Dabei wurden etwa neun Unterrichtsstunden pro Versuchsperson in einer Zeitspanne von ca. vier Wochen beurteilt. Die Autorinnen fanden zwar einen Effekt des Messzeitpunktes, d. h. die Versuchspersonen zeigten signifikante intraindividuelle Unterschiede, deren praktische Bedeutsamkeit stellte sich aber als eher gering heraus.

Einen etwas anderen Ansatz bietet die Studie von Curby et al. (2011), in der die Variabilität von Unterrichtsqualität innerhalb eines Tages zu acht Messzeitpunkten untersucht wurde. Die Ergebnisse einer autoregressiven Pfadanalyse interpretiert die Autorengruppe als mittlere bis hohe Stabilität für die betrachteten Merkmale ($.35 \leq \beta \leq .68$). Zudem ergab sich, dass die in den ca. 1500 Grundschulklassen beobachtete Variabilität der Unterrichtsqualität u. a. mit Merkmalen der Unterrichtsgestaltung (z. B. Sozialform) zusammenhing.

3. Forschungsfragen

Die Beschreibung des Forschungsstandes macht deutlich, dass bisherige Untersuchungen vor allem die Variabilität von Unterrichtsqualität *zwischen* Unterrichtsstunden in den Blick nahmen (im Überblick Praetorius et al. 2014). Inwieweit Unterschiede *innerhalb* von Unterrichtsstunden bestehen, blieb dabei meist unberücksichtigt. Diese Frage ist aber bedeutsam, weil dadurch Zusammenhangsanalysen zwischen Unterrichtsqualität und Oberflächenmerkmalen des Unterrichts möglich werden, die innerhalb von

Unterrichtsstunden variieren. Bis dato werden solche Untersuchungen häufig durch das Aggregieren solcher „Inszenierungsmuster“ auf Stundenebene durchgeführt (z.B. Hugener et al. 2007). Des Weiteren liegen bisher nur wenige Studien vor, die neben den bekannten Basisdimensionen die Variabilität fachspezifischer Merkmale der Unterrichtsqualität berücksichtigen. Wie zuvor geschildert, ist eine notwendige Voraussetzung zur Bearbeitung solcher Fragen das Vorliegen von Messinvarianz über die Zeit. Die genannten Desiderate stellen somit den Ausgangspunkt für die vorliegende Studie dar, die folgende Fragestellungen bearbeitet.

(1) Wie stark variiert die Unterrichtsqualität zwischen Lehrpersonen bzw. zwischen und innerhalb von Unterrichtsstunden und wie ist die Generalisierbarkeit der Erfassung fachspezifischer und generischer Merkmale von Unterrichtsqualität zu beurteilen?

Ziel der Entwicklung von Beobachtungsskalen ist die Erfassung von Unterschieden zwischen Lehrpersonen. Wir erwarten daher, dass alle fünf Merkmale in dieser Hinsicht signifikant variieren. Was Unterschiede zwischen Unterrichtsstunden und innerhalb dieser angeht, erwarten wir in Anlehnung an die Ergebnisse von Praetorius et al. (2014), dass die zeitliche Variabilität für die Basisdimension Klassenführung am geringsten ausfällt, während kognitive Aktivierung die höchste zeitliche Variabilität aufweisen sollte. Zur fachspezifischen Unterrichtsqualität und zur Variabilität innerhalb von Unterrichtsstunden liegen kaum Ergebnisse vor, darum sollen hierzu keine Hypothesen formuliert werden. Sollte der Varianzanteil *innerhalb* von Unterrichtsstunden jedoch substantiell ausfallen, ergibt sich in Anlehnung an Curby et al. (2011) die Frage nach Variablen, die diesen Varianzanteil zu erklären vermögen. In der vorliegenden Studie wurde daher ferner die folgende Fragestellung exploriert:

(2) Lässt sich die Variabilität von Unterrichtsqualität innerhalb von Unterrichtsstunden durch Zusammenhänge mit der Sozialform (Klassenunterricht vs. Schülerarbeitsphase) erklären?

In der Unterrichtsforschung besteht zwar weitgehend Einigkeit darüber, dass erfolgreicher Unterricht sich nicht durch eine bestimmte Methode auszeichnet (im Überblick Hugener et al. 2007; vgl. auch Hage et al. 1985). Aus der TIMS-Videostudie 1999 ist allerdings auch bekannt, dass gute Leistungen im Mathematikunterricht bisweilen mit bestimmten Unterrichtssettings einhergehen (Hiebert et al. 2003), obwohl die Tiefenstrukturen des Unterrichts als ursächlich für den Lernerfolg von Schülerinnen und Schülern angenommen werden (Ditton 2006; Reusser 2009). Insofern liegen zwar einige Befunde zum Zusammenhang zwischen Unterrichtsgestaltung und Unterrichtsqualität vor, aus theoretischer Sicht fehlt aber eine Klärung, wie sich Oberflächen- und Tiefenstrukturmerkmale des Unterrichts zueinander verhalten (Reusser 2009).

4. Methode

4.1 Stichprobe

Die Daten basieren auf einer Stichprobe von 37 Lehrpersonen der Klassenstufen 7 bis 9 und wurden im Schuljahr 2015-2016 erhoben. Bei jeder Lehrperson wurden im Abstand von ca. zwei Wochen Unterrichtsbeobachtungen in zwei Doppelstunden durchgeführt. Die Teilnahme der Lehrpersonen an den Unterrichtsbeobachtungen erfolgte auf freiwilliger Basis. Der Anteil der weiblichen Lehrpersonen betrug etwa 50%. Im Mittel (Median) waren die beobachteten Lehrpersonen 40 Jahre alt (min = 27, max = 71) und verfügten über rund zehn Jahre Unterrichtserfahrung (min = 1, max = 33). Das erste Staatsexamen wurde mit der Note 1,6 (Median, min = 1,0, max = 3,0) abgeschlossen, das zweite Staatsexamen mit der Note 1,9 (min = 1.0, max = 4,0). Etwa 80% der Lehrpersonen besaßen eine Lehrbefähigung für das Gymnasium, die übrigen für ein Lehramt der Sekundarstufe I (KMK-Lehramtstyp 2 bzw. 3).

4.2 Messinstrument und Datenerhebung

Zur Erfassung der Unterrichtsqualität wurde ein Beobachtungsinstrument mit 22 hoch-inferenten Items eingesetzt (vgl. Tab. 1), die eine Operationalisierung der drei Basisdimensionen von Unterrichtsqualität sowie zwei weiterer, fachdidaktischer Qualitätsdimensionen darstellen. Die Antwortskala entspricht einem vierstufigen Likert-Format (1: sehr niedrige Qualität bis 4: sehr hohe Qualität). Zur Erfassung der drei Basisdimensionen wurden die bereits in der Pythagoras-Studie verwendeten Skalen adaptiert (Autorengruppe anonymisiert 2018; Autorengruppe anonymisiert). Für diese sind Nachweise prognostischer und kriterialer Validität dokumentiert (im Überblick Praetorius et al. 2018).

Die Skalen zur Erfassung der fachdidaktischen Unterrichtsqualität wurden in Anlehnung an internationale (z.B. Learning Mathematics for Teaching Project 2012) und deutschsprachige mathematikdidaktische Studien (im Überblick Autorengruppe anonymisiert 2014) neu entwickelt. Im Einzelnen handelt es sich dabei um eine Skala zur stoffbezogenen Qualität des Unterrichts, mit der erfasst wird, inwiefern mathematische Unterrichtsinhalte in angemessener Tiefe und Präzision thematisiert werden. Außerdem wurde eine Skala zur unterrichtsbezogenen Qualität operationalisiert, die mathematikdidaktische Entscheidungen wie Repräsentationsformen und Übungsphasen in den Blick nimmt.

Für alle eingesetzten Skalen wurde eine mindestens befriedigende interne Konsistenz erzielt (vgl. Tab. 1). Die manifesten Interkorrelationen der eingesetzten Skalen betragen zwischen r = .19 für Klassenführung und unterrichtsbezogene mathematikdidaktische Qualität und r = .69 für kognitive Aktivierung und unterrichtsbezogene Qualität. Für die beiden Skalen zur fachspezifischen Unterrichtsqualität ergibt sich eine Interkorrelation von r = .66. Kriteriale Validität konnte durch Korrelate mit fachdidaktischen Kompetenzen von Lehrpersonen aufgezeigt werden (Autorengruppe anonymisiert im Review).

(Tabelle 1: Bitte hier einfügen.)

Die Unterrichtsbeobachtungen erfolgten durch jeweils zwei Rater und wurden während des Mathematikunterrichts (*in vivo*) durchgeführt. Von jeder teilnehmenden Lehrperson wurden zwei Doppelstunden mit jeweils 90-minütiger Länge beobachtet, in denen die Unterrichtsqualität jeweils viermal in gleichen zeitlichen Abständen eingeschätzt wurde, also nach ca. 22.5 Minuten.

Zur Datenerhebung standen sechs geschulte Beobachter bzw. Beobachterinnen zur Verfügung. Voraussetzung zur Rekrutierung als Rater ein abgeschlossenes Bachelorstudium in einem Lehramtsstudiengang mit Unterrichtsfach Mathematik, um auch fachdidaktische Merkmale von Unterrichtsqualität adäquat einschätzen zu können. Das Training der Rater umfasste ca. 30 Stunden und setzte sich zu ungefähr gleichen Teilen aus Videoanalysen, Diskussionen und Unterrichtsbeobachtungen zusammen. Ziel des Trainings war ein geteiltes theoretisches Verständnis von Unterrichtsqualität. Die Interrater-Reliabilität kann nach Ausschluss eines Items für gut befunden werden ($ICC > .80$, Wirtz und Caspar 2002).

4.3 Datenauswertung

Für die statistischen Analysen wurde ein Datensatz mit 296 Datenpunkten (37 Lehrpersonen× 2 Doppelstunden × 4 Messzeitpunkte) verwendet, der in drei Schritten sukzessiv ausgewertet wurde: 1) Prüfung der Messinvarianz über die beiden Doppelstunden hinweg, 2) Generalisierbarkeitsanalyse, 3) Test auf einen Effekt der Sozialform auf die Unterrichtsqualität. Das methodische Framework für die Bearbeitung der Forschungsfragen stellt die G-Theorie dar, wie sie in Abschnitt 2.2 beschrieben wurde. Einer Generalisierbarkeitsanalyse wird hier aber die Prüfung der Messinvarianz vorangestellt, um die Variabilität der Unterrichtsqualität über die Zeit beurteilen zu können. Letztere ist die Voraussetzung für die Spezifikation von Messwiederholungen als Zufallseffekt (Kane 2011), weil andernfalls nicht über die in der vorliegenden Untersuchung beobachteten Messgelegenheiten hinaus verallgemeinert werden kann (Meyer et al. 2011).

Zur Prüfung der Messinvarianz wurde für jedes der fünf erfassten Merkmale von Unterrichtsqualität auf der Ebene von 148 Messzeitpunkten (*within*) sowie auf der Ebene der 37 Lehrpersonen (*between*) dasselbe Modell spezifiziert. Für jedes Merkmal wurden zwei latente Faktoren (Unterrichtsstunde 1, Unterrichtsstunde 2) mit jeweils zwei Indikatoren (Testhälfte 1, Testhälfte 2) modelliert und bei gleichen Indikatoren zwischen Unterrichtsstunden eine Residualkorrelation als Methodenfaktor zugelassen (Geiser und Lockhart 2012). Die Konstruktion der Testhälften erfolgte durch Summenbildung über die Item-Rohwerte. Dieses Verfahren dient einerseits der Reduktion der Anzahl der zu schätzenden Parameter, die stets, aber insbesondere auf Grund der geringen Stichprobengröße in der vorliegenden Untersuchung so niedrig wie möglich ausfallen sollte (z. B. Little et al. 2002). Andererseits ist das Messniveau von Ratingskalen nicht eindeutig geklärt (Wirtz und Caspar 2002). Die Zusammenfassung von Rohwerten zu einer Testhälfte dient also auch dazu, annähernd normalverteilte metrische Indikatoren zu gewinnen (Bandalos 2002).

Die Modelle wurden sodann auf skalare Messinvarianz geprüft, d. h. Faktorladungen und Intercepts der Indikatoren wurden zwischen Unterrichtsstunden und auf beiden Ebenen gleichgesetzt. Die Analysen wurden mit dem Softwarepaket MPlus 7.4 (Muthén und Muthén 1998-2015) unter Verwendung des robusten MLR-Schätzers durchgeführt. Der Umgang mit fehlenden Werten erfolgte durch die in der Software implementierte FIML-Methode, wobei der Anteil fehlender Werte auf Messzeitpunktebene gering war (< 5%).

Zur Bestimmung der Variabilität innerhalb von Unterrichtsstunden und zur Einschätzung der Reliabilität für Forschungsfrage (1) wurde eine Generalisierbarkeitsanalyse mit anschließender Entscheidungsstudie durchgeführt. In der vorliegenden Studie stellen Lehrpersonen (l) das Untersuchungsobjekt dar (Shavelson und Webb 1991). Auf Grund von Messwiederholungen ergeben sich zwei weitere Varianzkomponenten, nämlich Unterrichtsstunden innerhalb von Lehrpersonen ($s : l$) und Messzeitpunkte innerhalb von

Unterrichtsstunden (*m* : *s* : *l*). Das Design entspricht dem Nullmodell einer Mehrebenenanalyse (Goldstein 2003), so dass die Varianzkomponenten *l*, *s : l* und *m : s : l* in MPlus 7.4 über die Funktion „type = threelevel“ geschätzt werden konnten.[1] Der G-Koeffizient ergibt sich dann über die Formel

$$G = \frac{\sigma_l^2}{\sigma_l^2 + \sigma_{Fehler}^2} = \frac{\sigma_l^2}{\sigma_l^2 + \frac{\sigma_{s:l}^2}{2} + \frac{\sigma_{m:s:l}^2}{8}}$$

(Shavelson und Webb 1991), weil pro Lehrperson zwei Doppelstunden und pro Doppelstunde vier Beobachtungen und damit acht Messzeitpunkte vorliegen.

Zur Bearbeitung von Fragestellung (2) wurde als Kontrollvariable der dichotome Prädiktor „Sozialform“ auf Ebene der Messzeitpunkte in das Modell eingeführt (in Anlehnung an Hugener 2006). Hiermit wurde für jeden Messzeitpunkt erfasst, ob es sich bei der einzuschätzenden Unterrichtsphase um Klassenunterricht (Code 0) oder eine Schülerarbeitsphase (Code 1) gehandelt hatte. Etwa 30% der Unterrichtsphasen wurden doppelt kodiert, die Beobachterübereinstimmung war zufriedenstellend (Cohens $\kappa = .75$, Wirtz und Caspar 2002). Bei 60% der Messzeitpunkte wurden Unterrichtsgespräche im Klassenverband beobachtet, so dass der Code 0 vergeben wurde. Bei 40% der Messzeitpunkte wurden Einzel-, Partner- oder Gruppenarbeitsphasen dokumentiert, so dass der Code 1 vergeben wurde.

5. Ergebnisse

5.1 Überprüfung der Messinvarianz

[1] Die in MPlus 7.4 verwendete FIML-Methode ist zur Schätzung von Varianzkomponenten bei kleinen Stichprobengrößen eigentlich nicht geeignet (Stegmueller 2013) und wurde lediglich aus pragmatischen Gründen so vorgenommen. Die Ergebnisse konnten aber mit marginalen Unterschieden sowohl durch einen Bayes-Schätzer (non-informative prior) als auch durch die REML-Methode in SPSS 23 repliziert werden.

Die Prüfung auf skalare Messinvarianz resultierte für die Basisdimension Klassenführung in einem gut passenden Modell (vgl. Tab. 2). Die Annahme gleicher Intercepts der Indikatoren für beide Unterrichtsstunden und gleicher Faktorladungen dieser Indikatoren zwischen und innerhalb von Unterrichtsstunden musste also nicht verworfen werden. Für die Dimension Konstruktive Unterstützung ergab die Prüfung der Messinvarianz ein ähnliches Ergebnis. Die Modellanpassung für die Dimension Kognitive Aktivierung fiel etwas schlechter aus, kann allerdings noch als akzeptabel beurteilt werden. In allen Modellen wurde nach einer ersten Schätzung die Residualvarianz für einen bzw. zwei Indikatoren auf null fixiert, weil ein geringfügig negativer Wert auftrat.

(Tabelle 2: Bitte hier einfügen.)

Für die unterrichtsbezogene mathematikdidaktische Qualität zeigte sich eine gute Passung des Modells mit skalarer Messinvarianz zwischen Unterrichtsstunden, während die Passung des Modells zur stoffbezogene Qualität akzeptabel war. Es wurden dazu drei Residualvarianzen auf null fixiert. Insgesamt kann damit von einer äquivalenten Messung zwischen Unterrichtsstunden für alle eingesetzten Skalen ausgegangen werden.

Tabelle 2 deutet bereits an, dass sich die Ratings zwischen den beobachteten Unterrichtsstunden in ihrer Ausprägung nicht wesentlich unterscheiden; in der Tat ist für keine Skala die Differenz der Stundenmittelwerte signifikant ($p > .17$). Die reliabilitätskorrigierten Korrelationen der jeweiligen Ratings zwischen beiden Unterrichtsstunden liegen bei $.59 \leq \Phi < 1.00$, wobei der schwächste Zusammenhang für Konstruktive Unterstützung und der stärkste für Kognitive Aktivierung vorliegt. Zusammenfassend kann der Effekt der Messgelegenheit also als zufällig gedeutet werden.

5.2 Generalisierbarkeits- und Entscheidungsstudie

In Tabelle 3 sind die Ergebnisse einer Varianzzerlegung gemäß Abschnitt 4.3 dargestellt. Stunden- und Messzeitpunkteffekte gehen als zufällige Komponenten in das Modell ein. Die

Varianz *zwischen Lehrpersonen* ist wie erwartet (und erwünscht, um die Unterschiede zwischen ihnen beschreiben zu können), für alle Skalen signifikant positiv ausgeprägt ($p < .01$) und nimmt für kognitive Aktivierung und die beiden fachspezifischen Skalen den jeweils größten Varianzanteil ein (etwa 40% der Gesamtvarianz). Für die Skala zur Klassenführung nimmt die Varianz zwischen Lehrpersonen einen ähnlichen relativen Anteil ein, lediglich für konstruktive Unterstützung ist dieser mit 26% deutlich niedriger ausgeprägt.

Was die Variabilität *zwischen Unterrichtsstunden* angeht, zeigt sich, dass diese für generische und fachspezifische Merkmale von Unterrichtsqualität unterschiedlich ausgeprägt ist: Während für die Skalen zur unterrichts- und stoffbezogenen mathematikdidaktischen Qualität ein beträchtlicher Anteil (etwa um 30%) hierauf entfällt, beträgt der Anteil für die generischen Merkmale höchstens 15%. Zudem ist diese Variabilität für konstruktive Unterstützung und kognitive Aktivierung nicht signifikant von Null verschieden ($p = .13$ bzw. $p = .11$).

Der mit der Residualvarianz konfundierte Anteil der Variabilität *innerhalb von Unterrichtsstunden* fällt ebenfalls durchweg positiv aus ($p < .01$) und nimmt für Klassenführung und konstruktive Unterstützung den größten Prozentanteil ein. Eine Zusatzanalyse zeigt, dass die Residualvarianz tatsächlich die Variabilität der Merkmale von Unterrichtsqualität innerhalb von Unterrichtsstunden widerspiegelt: Führt man als weitere Komponente den Rater-Effekt in das Modell ein, so lässt sich die Varianzquelle *m : s : l* von der Residualvarianz trennen und es ergeben sich abzüglich des Rater-Fehlers (für alle Skalen ca. 9%) ähnliche Prozentwerte für die Variabilität innerhalb von Unterrichtsstunden (Klassenführung: 39%, konstruktive Unterstützung: 58%, kognitive Aktivierung: 32%, unterrichtsbezogene mathematikdidaktische Qualität: 20%, stoffbezogene Qualität: 15%).

(Tabelle 3: Bitte hier einfügen.)

Eine Entscheidungsstudie ergibt, dass die Generalisierbarkeit damit für vier von fünf Qualitätsdimensionen als akzeptabel beurteilt werden kann, wenn man ein in den

Sozialwissenschaften übliches Reliabilitätskriterium von $G = .70$ ansetzt. Lediglich der Generalisierbarkeitskoeffizient für die stoffdidaktische Qualität fällt niedriger aus, was vermutlich mit der hohen Variabilität zwischen Unterrichtsstunden zusammenhängt.

5.3 Variabilität von Unterrichtsqualität innerhalb von Unterrichtsstunden

Die Ergebnisse der Generalisierbarkeitsanalyse zeigten eine signifikante Varianz der fünf Qualitätsmerkmale innerhalb von Unterrichtsstunden, und zwar insbesondere für generische Merkmale der Unterrichtsqualität. Die mittleren Ausprägungen für die Skalen zur Klassenführung und zur kognitiven Aktivierung zeigen, dass diese jeweils in beiden Doppelstunden zum ersten Messzeitpunkt am stärksten sind und dann monoton im weiteren Verlauf fallen (bei Klassenführung sehr deutlich, weniger stark bei kognitiver Aktivierung), wie Tabelle 3 zu entnehmen ist. Die Mittelwerte für konstruktive Unterstützung sind für die Messzeitpunkte 2 und 3 höher ausgeprägt als zu Beginn und zum Ende der Unterrichtsstunden, die Merkmalsausprägung ist also am stärksten im mittleren Teil der Doppelstunde. Für die eingesetzten Skalen zur fachspezifischen Unterrichtsqualität lassen sich über die vier Messzeitpunkte hinweg in keiner der beiden Doppelstunden systematischen Unterschiede feststellen. Die in Tabelle 3 dargestellten Autokorrelationen zwischen aufeinander folgenden Messzeitpunkten deuten auf eine hohe Variabilität der Skala konstruktive Unterstützung innerhalb von Unterrichtsstunden hin, während die übrigen Merkmale von Unterrichtsqualität als weitgehend zeitlich stabil im Sinne der Rangreihung betrachtet werden können.

(Tabelle 4: Bitte hier einfügen.)

Abschließend wurde untersucht, inwieweit die Variabilität innerhalb von Unterrichtsstunden mit unterschiedlichen Sozialformen im Unterricht zusammenhängt. Für die Skala zur Klassenführung zeigte sich kein Effekt der Sozialform ($\beta = -.06$, $p = .36$), ebenso wenig für die Skalen zur kognitiven Aktivierung und zur stoffbezogenen Qualität ($\beta = .12$, $p = .14$ bzw.

β = .02, p = .76). Für die Skala zur konstruktiven Unterstützung findet sich ein positiver Zusammenhang mittlerer Stärke (β = .49, p < .01), d. h. in Schülerarbeitsphasen findet signifikant mehr konstruktive Unterstützung statt. Der Effekt der Sozialform ist auch für die unterrichtsbezogene, mathematikdidaktische Qualität in der Tendenz positiv (β = .15, p = .05), also ist diese in Schülerarbeitsphasen tendenziell ebenfalls stärker ausgeprägt.

6. Zusammenfassung und Diskussion

Im vorliegenden Beitrag wurde die Variabilität von fachspezifischer und generischer Unterrichtsqualität auf der Basis von Daten von 37 Mathematiklehrpersonen analysiert, und zwar zwischen zwei Doppelstunden und innerhalb dieser. Die Ergebnisse zeigten zunächst, dass die Messung der Unterrichtsqualität mit einem zeitlichen Abstand von etwa zwei Wochen weitgehend äquivalent erfolgte. Dies bedeutet, dass sowohl die Struktur von Unterrichtsqualität (konfigurale Messinvarianz) als auch die Bedeutung der einzelnen Indikatoren (metrische Messinvarianz) und das Ausgangsniveau (skalare Messinvarianz) in diesem Zeitraum vergleichbar bleiben. Zudem ergab sich kein signifikanter Unterschied zwischen den mittleren Ratings der Unterrichtsqualitätsdimensionen zu den beiden Messgelegenheiten. Dieser Befund stützt die Überlegung von Meyer et al. (2011), die erst bei längeren Zeiträumen zwischen Messgelegenheiten von wahren Unterschieden in der Unterrichtsqualität ausgehen.

Die geschätzten Varianzanteile für Unterschiede zwischen Lehrpersonen und G-Koeffizienten fielen in der vorliegenden Studie akzeptabel und wie erwartet aus und entsprechen weitgehend denen in ähnlichen Beobachtungsstudien (für einen Überblick vgl. Praetorius et al. 2012). Die fünf Skalen zur Unterrichtsqualität sind damit zum einen geeignet, Unterschiede zwischen Lehrpersonen zu beschreiben. Zum anderen erwies sich die Variabilität zwischen den beobachteten zwei Doppelstunden pro Lehrperson überwiegend als gering. Eine Ausnahme stellten hier die Skalen zur mathematikdidaktischen Qualität dar, die

deutlich stärker variierten als die drei Basisdimensionen. Eine Erklärung dafür könnte sein, dass fachspezifische Merkmale einen stärkeren Bezug zum Unterrichtsinhalt aufweisen (Blum 2006), was Items wie „Erklären“ oder „Fachliche Korrektheit“ andeuten. Beides bezieht sich explizit auf den Unterrichtsinhalt und hängt auch mit dem fachdidaktischen Wissen der Lehrperson zusammen (Autorengruppe anonymisiert im Review; Learning Mathematics for Teaching Project 2011). Zukünftige Studien könnten sich daher mit der Frage befassen, inwieweit dem Unterrichtsinhalt eine vermittelnde Rolle zwischen professioneller Kompetenz von Lehrpersonen und fachspezifischer Unterrichtsqualität zukommt.

Die Variabilität innerhalb von Unterrichtsstunden erwies sich in der vorliegenden Studie als beträchtlich und fiel für generische Merkmale besonders stark aus. Klassenführung nahm im Verlauf der Doppelstunde substanziell ab. Dieses Ergebnis deutet darauf hin, dass Maßnahmen zur Störungsprävention und zur Strukturierung des Unterrichts möglicherweise vor allem zu Beginn einer Unterrichtsstunde erfolgen (oder zum Ende einer Unterrichtsstunde ausbleiben). Kognitive Aktivierung nahm im Unterrichtsverlauf ebenfalls signifikant, aber in geringem Ausmaß ab. Die monotone Abnahme in den Skalenwerten zur Klassenführung und zur kognitiven Aktivierung sollte in weiteren Studien näher untersucht werden.

Gezeigt hat sich auch, dass die Ratings zur konstruktiven Unterstützung in mittlerer Höhe mit der Sozialform kovariieren (vgl. auch Curby et al. 2011). Dieser Zusammenhang kann für die anderen Merkmale der Unterrichtsqualität nicht bzw. nur tendenziell gefunden werden, was nachvollziehbar erscheint, weil mit konstruktiver Unterstützung auch binnendifferenzierende Unterrichtsmethoden erfasst werden (Tab. 1; vgl. auch Rakoczy und Pauli 2006). Diese sind in Schülerarbeitsphasen anscheinend leichter umsetzbar, z. B. durch den Einsatz differenzierender Aufgaben oder durch individuell unterstützendes Verhalten der Lehrperson, wie es durch den Scaffolding-Begriff beschrieben wird (im Überblick van de Pol et al. 2010).

Die vorliegende Untersuchung trägt insgesamt in zweifacher Hinsicht zum wissenschaftlichen Diskurs bei: Erstens stellt die Erfassung von Unterrichtsqualität durch multiple Messzeitpunkte eine genauere Beschreibung von Lernangeboten im Unterricht dar (Helmke 2012), und zwar sowohl hinsichtlich des Informationsgehalts als auch hinsichtlich der Reliabilität. Bisher wurden mehrfache Ratings innerhalb von Unterrichtsstunden meist dafür verwendet, die Effizienz einer Untersuchung zu optimieren, mit dem Ziel, ein möglichst kurzes Segment einer Unterrichtsstunde einzuschätzen (Mashburn et al. 2014). Dies geschieht etwa vor dem Hintergrund ressourcenintensiver Hospitationen von Schulinspektionen (Pietsch und Tosana 2008). Alternativ könnte man mehrere Ratings innerhalb einer Unterrichtsstunde zu einem Summen- oder Mittelwert zusammenfassen, um Rater-Effekte zu minimieren. Beide Fälle resultieren in einem Wert pro Unterrichtsstunde, so dass die Beschreibung der Variabilität innerhalb einer Unterrichtsstunde wiederum nicht möglich ist und auch nicht für Zusammenhänge zur Unterrichtsgestaltung kontrolliert werden kann (Curby et al. 2011; Patrick und Mantzipopoulos 2016).

Zweitens unterbreitet die vorliegende Studie einen Vorschlag, wie Kanes (2011) methodologische Kritik an der Anwendung der G-Theorie mit Hilfe einer Überprüfung der Messinvarianz Rechnung getragen werden kann. Anstatt die Annahme der Äquivalenz der wahren Werte zwischen Messgelegenheiten fallen zu lassen (Meyer et al. 2011), kann dies vor einer G-Studie zunächst formal überprüft werden. Die Ergebnisse einer Analyse der Messinvarianz erlauben dann die Spezifikation von Messzeitpunkten als zufällige oder feste Effekte und eine entsprechende Generalisierung der Forschungsergebnisse (Kane 2011).

Einschränkend muss bemerkt werden, dass es sich bei der vorliegenden Stichprobe um eine Gelegenheitsstichprobe handelt, die vermutlich durch eine positive Verzerrung gekennzeichnet ist (vgl. berichtete Noten der Staatsexamina und Hill et al. 2012), von Repräsentativität kann daher nicht ausgegangen werden. Zudem untersuchten wir

Doppelstunden. Auch wenn diese häufig sind, wird Mathematik insbesondere in niedrigeren Klassenstufen oft in Einzelstunden unterrichtet. Weitere Studien erscheinen daher nötig, die die Ergebnisse daher möglichst an einer randomisiert gezogenen und ausreichend großen Stichprobe replizieren, welche eher die methodischen Anforderungen zur Schätzung multidimensionaler Mehrebenen-Modelle mit latenten Variablen erfüllen (Malmberg et al. 2010). Die Problematik der kleinen Stichprobengröße in der vorliegenden Studie zeigte sich bereits in Abschnitt 5.1, da bei der Schätzung der Modellparameter negative Residualvarianzen auftraten (u. a. Little et al. 2002). Die Ergebnisse laden also zu weiteren Forschungsvorhaben ein, die sich insbesondere mit dem Verhältnis der Variabilität innerhalb und zwischen Unterrichtsstunden beschäftigen könnten. Bisher liegen derartige Befunde auch fast ausschließlich für generische Merkmale von Unterrichtsqualität und das Fach Mathematik vor, so dass fachspezifische Ansätze für andere Unterrichtsfächer ein Desiderat bleiben.

Literaturverzeichnis

Autorengruppe anonymisiert (2014).

Autorengruppe anonymisiert (2016).

Autorengruppe anonymisiert (2018).

Autorengruppe anonymisiert (im Review).

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*, 78-102.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180.

Blum, W. (2006). Einführung. In W. Blum, C. Drüke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: Konkret. Sekundarstufe 1: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen* (pp. 14-32). Berlin: Cornelsen Scriptor.

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

Brophy, J. (2000). *Teaching.* Brüssel: International Academy of Education.

Brunner, E. (2017). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *Journal für Mathematikdidaktik.*

Calkins, D., Borich, G. D., Pascone, M., Kluge, S. & Marston, P. T. (1997). Generalizability of teacher behaviors across classroom observation systems. *Journal of Classroom Interaction, 13*, 9-22.

Casabianca, J. M., Lockwood, J. R. & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337.

Charalambous, C., & Praetorius, A.-K. (2018). Studying instructional quality in mathematics through different lenses: In search of common ground. *ZDM Mathematics Education, 50*(3).

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

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L. & Downer, J. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade. *The Elementary School Journal, 112*(1), 16-37.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior. Perspectives in social psychology.* New York: Plenum.

Ditton, H. (2006). Unterrichtsqualität. In K.-H. Arnold, U. Sandfuchs & J. Wiechmann (Hrsg.), *Handbuch Unterricht* (S. 235-243). Bad Heilbrunn: Klinkhardt.

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*(3), 127-137.

Geiser, C. & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent-state-trait-analyses. *Psychological Methods, 17*(2), 255-283.

Goldstein, H. (2003). *Multilevel Statistical Models.* London: Hodder Arnold.

Hage, K., Bischoff, H., Dichanz, H., Eubel, K., Oehlschläger, H. & Schwittmann, D. (1985). *Das Methodenrepertoire von Lehrern. Eine Untersuchung zum Unterrichtsalltag in der Sekundarstufe I.* Opladen: Leske + Budrich.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts.* Seelze: Klett-Kallmeyer.

Hiebert, J., Gallimore, R., Garnier, H., … Stigler, J. (2003). *Teaching Mathematics in Seven Countries. Results from the TIMSS 1999 Video Study.* Washington: National Center for Education Statistics.

Hill, H., Rowan, B. & Ball, D. L. (2005). Effects of teachers' Mathematical Knowledge for Teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Hill, H. C., Charalambous, C. & 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.

Hugener, I. (2006). Sozialformen und Lektionsdauer. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 3: Hugener, Isabelle; Pauli, Christine & Reusser, Kurt: Videoanalysen)* (S. 55-61). Frankfurt am Main: GFPF.

Hugener, I., Pauli, C. & Reusser, K. (2007). Inszenierungsmuster, kognitive Aktivierung und Leistung im Mathematikunterricht. Analysen aus der schweizerisch-deutschen Videostudie. In D. Lemmermühle, M. Rothangel, S. Bügeholz, M. Hasselhorn & R. Watermann (Hrsg.), *Professionell Lehren – Erfolgreich Lernen* (S. 109-121). Münster: Waxmann.

Kane, M. (2011). The errors of our ways. *Journal of Educational Measurement, 48*, 12-30.

Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik, 54*, 222–237.

Kounin, J. S. (1970). *Disciplin and group management in classrooms.* New York: Holt, Rinehart & Winston.

Kuger, S., Kluczniok, K., Kaplan, D. & Rossbach, H.-G. (2016). Stability and patterns of classroom quality in German early childhood education and care. *School Effectiveness and School Improvement, 27*(3), 418-440.

Learning Mathematics for Teaching Project (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education, 14,* 25–47.

Little, T. D., Cunningham, W. A., Shahar, G. & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighting the merits. *Structural Equation Modeling, 9*(2), 151-173.

Malmberg, L.-E., Hagger, H., Burn, Katharine, Mutton, T. & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology, 102*(4), 916-932.

Mashburn, A. J., Meyer, J. P., Allen, J. P. & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement, 74*(3), 400-422.

Meyer, J. P., Cash, A. H., Mashburn, A. (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment, 16,* 227-243.

Muthén, B. O. & Muthén, L. K. (2010). *Mplus user's guide (6th edition).* Los Angeles, CA: Muthén & Muthén.

Patrick, H. & Mantzicopoulos, P. (2016). Is effective teaching stable? *The Journal of Experimental Education, 84*(1), 23-47.

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., Lenske, G. & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfil what they promise? *Learning and Instruction, 6,* 387-400.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K. & Klieme E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018, im Druck). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM Mathematics Education, 50*(3).

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht – Unterricht aus der Sicht von Lernenden und Beobachtern.* Münster: Waxmann.

Rakoczy, K., & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 3: Hugener, Isabelle; Pauli, Christine & Reusser, Kurt: Videoanalysen)* (S. 189-205). Frankfurt am Main: GFPF.

Rakoczy, K., Klieme, E., Drollinger-Vetter, B. & Reusser, K. (2007). Structure as a quality feature in mathematics instruction: cognitive and motivational effects of a structured organisation of the learning environment vs. a structured presentation of learning content. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Program* (pp. 102-121). Münster: Waxmann.

Reinmann-Rothmeier, G., & Mandl, H. (2006). Unterrichten und Lernumgebungen gestalten. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie. Ein Lehrbuch* (S. 613-658). Weinheim: Beltz.

Reusser, K. (2009). Von der Bildungs- und Unterrichtsforschung zur Unterrichtsentwicklung. Probleme, Strategien, Werkzeuge und Bedingungen. *Beiträge zur Lehrerinnen und Lehrerbildung, 27*(3), 295-312.

Seidel, T. & Shavelson, R. (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.

Shavelson, R. & Webb, N. (1991). *Generalizability Theory: A Primer.* Thousand Oaks: Sage.

Stegmueller, D. (2013). How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches. *American Journal of Political Science, 57*(3), 748-761.

Van de Pol, J., Volman, M. & Beishuizen, J. (2010). Scaffolding in Teacher-Student Interaction: A Decade of Research. *Educational Psychology Review, 22*(3), 271-296.

Wirtz, M. A. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe.

Tabellenverzeichnis

Tabelle 1: Skalen zur Erfassung generischer und fachspezifischer Unterrichtsqualität

	# Items	Beispielitem	Cronbachs α
Klassenführung	5	Effektive Lernzeitnutzung	.86
Konstruktive Unterstützung	6	Umgang mit Heterogenität	.73
Kognitive Aktivierung	4	Herausfordernde Fragen und Aufgaben	.83
Unterrichtsbezogene Qualität	4	Repräsentationsformen	.69
Stoffbezogene Qualität	5	Fachliche Tiefe (u. a. Vernetzungen, Verallgemeinerungen, Strukturierung)	.77

Tabelle 2: Modellanpassung der konfirmatorischen Faktorenanalysen

	χ^2	*df*	*p*	*CFI*	*RMSEA*
Klassenführung	3.12	6	.79	1.00	.00
Konstruktive Unterstützung	6.73	7	.46	1.00	.00
Kognitive Aktivierung	12.80	9	.17	.95	.05
Unterrichtsbezogene Qualität	3.29	4	.51	1.00	.00
Stoffbezogene Qualität	11.60	10	.31	.97	.03

Anmerkungen: *CFI* = Comparative Fit Index, *RMSEA* = Root Mean Square Error of Approximation.

Tabelle 3: Ergebnisse der Generalisierbarkeitsanalyse mit Fehlervarianz und G-Koeffizient

	KF	KU	KA	UBQ	SBQ
Lehrperson, *l*	.088 (37,7%)	.049 (26,1%)	.079 (45,8%)	.090 (40,5%)	.047 (38,5%)
Unterrichtstunde, *s* : *l*	.029 (14,5%)	.012 (7,9%)	.008 (10,7%)	.061 (29,2%)	.046 (37,5%)
m : *s* : *l*, Residuum	.132 (47,8%)	.106 (66,0%)	.097 (43,5%)	.071 (30,3%)	.036 (24,0%)
Fehlervarianz	.031	.019	.016	.039	.028
G	.739	.720	.831	.700	.627

Anmerkungen: KF = Klassenführung, KU = Konstruktive Unterstützung, KA = Kognitive Aktivierung, UBQ = Unterrichtsbezogene Qualität, SBQ = Stoffbezogene Qualität, *G* = Generalisierbarkeitskoeffizient. Bis auf die Stundeneffekte (*s* : *l*) für die Skalen KU und KA sind alle Varianzkomponenten signifikant positiv ($p < .01$).

Tabelle 4: Deskriptive Statistiken für Messzeitpunkte (MZP) in Unterrichtsstunden.

	Unterrichtsstunde 1		Unterrichtsstunde 2		Produkt-Moment-Korrelation
	M	SD	M	SD	
Klassenführung	3.27	.48	3.29	.37	.61
MZP 1	3.43	.48	3.44	.36	-
MZP 2	3.38	.53	3.46	.41	.74
MZP 3	3.29	.58	3.32	.47	.75
MZP 4	3.00	.58	2.93	.52	.67
Konstruktive Unterstützung	1.97	.34	1.90	.30	.42
MZP 1	1.76	.42	1.73	.40	-
MZP 2	2.00	.41	1.98	.46	.38
MZP 3	2.05	.54	2.06	.44	.63
MZP 4	1.88	.42	1.83	.37	.47
Kognitive Aktivierung	2.60	.45	2.63	.32	.71
MZP 1	2.70	.46	2.74	.39	-
MZP 2	2.65	.54	2.76	.35	.73
MZP 3	2.59	.51	2.58	.43	.71
MZP 4	2.45	.55	2.42	.42	.65
Unterrichtsbezogene Qualität	2.41	.46	2.48	.43	.56
MZP 1	2.26	.52	2.41	.52	-
MZP 2	2.37	.54	2.55	.45	.73
MZP 3	2.50	.46	2.51	.49	.82
MZP 4	2.51	.48	2.43	.48	.78
Stoffbezogene Qualität	2.60	.37	2.69	.34	.55
MZP 1	2.56	.39	2.65	.36	-

MZP 2	2.61	.43	2.75	.37	.77
MZP 3	2.63	.389	2.72	.36	.83
MZP 4	2.59	.386	2.65	.39	.84

Anmerkung: Der eingerückte Eintrag in den Zeilen bezieht sich jeweils auf die gesamte Unterrichtsstunde. Die in der letzten Spalte von Tabelle 3 dargestellten Korrelationen werden zwar statistisch signifikant ($p < .01$), wurden aber ohne Berücksichtigung der Cluster-Struktur in den Daten berechnet. Korreliert wurde jeweils der Datenvektor für den Messzeitpunkt in der entsprechenden Tabellenzeile mit dem vorausgehenden (Autokorrelation erster Ordnung).