J Math Didakt (2021) 42:97–121
https://doi.org/10.1007/s13138-020-00168-x

ORIGINALARBEIT/ORIGINAL ARTICLE

Erfassung der fachspezifischen Qualität von Mathematikunterricht: Faktorenstruktur und Zusammenhänge zur professionellen Kompetenz von Mathematiklehrpersonen

Armin Jentsch · Lena Schlesinger · Hannah Heinrichs · Gabriele Kaiser · Johannes König · Sigrid Blömeke

Eingegangen: 3. März 2018 / Angenommen: 2. Mai 2020 / Online publiziert: 2. Juni 2020

Zusammenfassung Die deutschsprachige Unterrichtsforschung unterscheidet bereits seit längerem drei Basisdimensionen der Unterrichtsqualität: Effiziente Klassenführung, konstruktive Unterstützung und Potential zur kognitiven Aktivierung. Da die drei Basisdimensionen als fächerübergreifende Konzeptualisierung der Unterrichtsqualität gelten, stellt sich aus fachdidaktischer Sicht die Frage, inwieweit dieses Modell die Charakteristika von Mathematikunterricht hinreichend abbilden kann. Vor diesem Hintergrund wurde ein Beobachtungsinstrument zur Erfassung der Unterrichtsqualität im Mathematikunterricht der unteren Sekundarstufe entwickelt, das sowohl generische als auch fachspezifische Merkmale erfassen soll. Die vorliegende Studie untersucht die Faktorenstruktur dieses Beobachtungsinstruments sowie Zusammenhänge zu fachspezifischen Kompetenzfacetten der beobachteten Mathematiklehrpersonen zur Validitätsprüfung. Qualität von Mathematikunterricht wurde

A. Jentsch · L. Schlesinger · G. Kaiser
Fakultät Erziehungswissenschaft, Universität Hamburg, Von-Melle-Park 8, 20146 Hamburg, Deutschland
E-Mail: armin.jentsch@uni-hamburg.de

L. Schlesinger (✉)
Landesinstitut für Lehrerbildung und Schulentwicklung Hamburg, Hamburg, Deutschland
E-Mail: lena.schlesinger@uni-hamburg.de

H. Heinrichs
Stadtteilschule Bergedorf, Bergedorf, Deutschland

G. Kaiser
Australian Catholic University, Brisbane, Australien

J. König
Universität zu Köln, Köln, Deutschland

S. Blömeke
Centre of Educational Measurement, Universität Oslo, Oslo, Norwegen

durch hoch-inferente Ratingskalen operationalisiert und von geschulten Beobachtenden in 156 Unterrichtsstunden eingeschätzt. Die deskriptiven Kennwerte weisen auf eine akzeptable Messqualität der Ratingskalen hin. Eine explorative Faktorenanalyse zeigt, dass dem Beobachtungsinstrument eine vierdimensionale Struktur zugrunde liegt, die als die drei Basisdimensionen sowie eine vierte Dimension, fachdidaktische Strukturierung, gedeutet wird. Für drei der vier Qualitätsdimensionen finden sich mindestens schwach positive Korrelationen mit den fachspezifischen Kompetenzfacetten der Lehrpersonen, allerdings nicht für die Klassenführung. Die Ergebnisse werden in Bezug auf die Fachspezifität des Beobachtungsinstruments und die Konzeptualisierung von Unterrichtsqualität diskutiert.

Schlüsselwörter Unterrichtsqualität · Mathematikunterricht · Ratings · Beobachtungsinstrument

Measuring the subject-specific quality in mathematics instruction: factor structure and relations to mathematics teachers' professional competence

Abstract German educational research has already for a long time distinguished three basic dimensions of instructional quality: classroom management, constructive support and cognitive activation. Since the three basic dimensions are regarded as a generic conceptualization of instructional quality, recently the question has arisen as to how far this model can adequately reflect the characteristics of mathematics teaching. Against this background, an observation instrument was developed for assessing the quality of mathematics instruction at the lower secondary level, which claims to measure both generic and subject-specific characteristics. The present study examines the factor structure of the observational instrument as well as correlations with subject-specific competence facets of the observed mathematics teachers as validity criteria. Instructional quality was operationalized by high-inferent rating scales and assessed by trained raters in 156 lessons. Descriptive statistics indicate an acceptable psychometric quality of the rating scales. An explorative factor analysis yielded a four-dimensional structure for the observational instrument, which is interpreted as the three basic dimensions and a fourth dimension, called mathematics educational structuring. With the exception of classroom management, all dimensions correlated at least weakly positively with subject-specific teacher competence. We discuss our findings with regard to the subject-specifity of the observational instrument and the conceptualization of instructional quality.

Keywords Instructional quality · Quality of mathematics instruction · Ratings · Observational instrument

1 Einleitung

Eine hohe Unterrichtsqualität wird sowohl in der wissenschaftlichen als auch in der öffentlichen Diskussion als zentrale Voraussetzung für die Leistung von Lernenden und deren Leistungsfortschritte angesehen (z. B. Helmke 2012). Es stellt sich daher die Frage, wie Unterrichtsqualität beschrieben und erfasst werden kann.

Im Kontext der Studien TEDS-Unterricht und TEDS-Validierung als Folgeuntersuchungen der TEDS-M-Studie (Teacher Education and Development Study in Mathematics) wurde ein Beobachtungsinstrument entwickelt, das neben den drei Basisdimensionen auch die fachspezifische Qualität von Mathematikunterricht erfassen soll (Schlesinger et al. 2018; vgl. auch Charalambous und Praetorius 2018). Im vorliegenden Beitrag wird die Frage untersucht, inwieweit dieser Ansatz empirisch gelungen ist. Die Untersuchung erfolgt über eine Analyse der Konstruktvalidität (über eine Prüfung der Faktorenstruktur des Beobachtungsinstruments) sowie der konvergenten und diskriminanten Validität (über eine Prüfung von Zusammenhängen mit Kompetenzfacetten der beobachteten Mathematiklehrpersonen).

1.1 Drei Basisdimensionen der Unterrichtsqualität

Im Anschluss an Untersuchungen zur empirischen Wirksamkeit von Unterricht im Rahmen der TIMS-Videostudie (Hiebert et al. 2003) unterscheidet die deutschsprachige Bildungsforschung drei Basisdimensionen der Unterrichtsqualität in einem fächerübergreifenden (generischen) Modell: *Effiziente Klassenführung, konstruktive Unterstützung* und *Potential zur kognitiven Aktivierung* (Fauth et al. 2014; Klieme et al. 2006; Klieme und Rakoczy 2008; Kunter und Ewald 2016; Praetorius et al. 2018).

Effiziente Klassenführung zielt darauf ab, die zur Verfügung stehende Lernzeit durch eine entsprechende Steuerung des Unterrichts optimal zu nutzen. In Anlehnung an Kounin (1970) werden unter effizienter Klassenführung vor allem Strategien zur unterrichtlichen Störungsprävention und Lernorganisation subsumiert (König 2015; Kunter und Ewald 2016). *Konstruktive Unterstützung* bezieht sich darauf, inwieweit im Unterricht auf Grundbedürfnisse der Lernenden eingegangen wird (Klieme und Rakoczy 2008; vgl. auch Deci und Ryan 1985). Damit sind Maßnahmen zur Individualisierung und Differenzierung, aber auch adaptive Hilfestellungen und Rückmeldungen durch die Lehrperson angesprochen (u. a. van de Pol et al. 2010). Potential zur *kognitiven Aktivierung* bietet Unterricht, wenn Lernende zu vertieftem Nachdenken über Unterrichtsinhalte angeregt werden (Kunter und Voss 2011). Damit wird angestrebt, dass Schülerinnen und Schüler durch die Auseinandersetzung mit komplexen Problemstellungen zu einem konzeptuellen Verständnis des Lernstoffs gelangen (Mayer 2004).

In den letzten Jahren ist vor allem aus fachdidaktischer Sicht die Frage aufgeworfen worden, inwiefern das Modell der drei Basisdimensionen die Charakteristika von Mathematikunterricht abbilden kann (u. a. Blum 2006; Brunner 2018; Lipowsky et al. 2018; Leuders und Holzäpfel 2011). Ausgehend von der Erkenntnis, dass schulisches Lernen in einer Domäne stattfindet (Baumert et al. 2010; Weinert 1994), dürfte die Erfassung *fachspezifischer* Unterrichtsqualität über die Basisdimensio-

nen hinausgehend für den Lernerfolg von Schülerinnen und Schülern bedeutsam sein. Fachspezifische Unterrichtsqualität könnte etwa die Besonderheiten curricularer Sequenzierung von Fachinhalten im Schulunterricht beschreiben sowie Interaktionsmuster und Kommunikationsstrukturen, die für das Unterrichten eines Faches typisch sind (für den Mathematikunterricht vgl. Lindmeier und Heinze 2020).

Beispielsweise dürfte der Umgang mit Fehlern in der Fremdsprachendidaktik aufgrund der Einheit von Inhalt und Medium (Borg 2006) anderen Zielsetzungen folgen als in Übungssituationen des Mathematikunterrichts (Vollrath und Roth 2012). Darüber hinaus können fachspezifische Dimensionen fachlich bedingte Lernprozesse von Schülerinnen und Schülern abbilden, die Lehrkräfte für ihr Fach kennen sollten (z. B. Grundvorstellungen im Mathematikunterricht, kreatives Schreiben im Fremdsprachenunterricht oder Umgang mit divergierenden Wertvorstellungen im Religionsunterricht, Cramer 2012; Eichelmann et al. 2012).

Eine offene Frage ist allerdings, ob die Basisdimensionen bereits ein Gerüst darstellen, das nur fachspezifisch ausdifferenziert werden müsste, oder ob ergänzend zu den Basisdimensionen zusätzliche Merkmale erfasst werden sollten, die additiv fachspezifische Unterrichtsqualität konzeptualisieren (für den Mathematikunterricht vgl. Leuders 2001). Die Basisdimensionen haben den Anspruch und gleichzeitig den Vorteil, dass sie auf unterschiedliche Domänen bezogen werden können (zumindest messtechnisch: Praetorius et al. 2016), also *generisch* sind (Praetorius et al. 2018). Es stellt sich daher, möglicherweise für jede Domäne getrennt, die Frage nach einer ökologisch validen Interpretation von Fachbezogenheit, die für den Mathematikunterricht beispielsweise verschiedene mathematische Themengebiete berücksichtigt (Bruder et al. 2015).

In diesem Kontext diskutieren Charalambous und Praetorius (2018) neben einer generischen auch die *hybride* Erfassung von Unterrichtsqualität, bei der sowohl generische als auch fachspezifische Merkmale abgebildet werden. Während die Operationalisierung der drei Basisdimensionen durch hoch-inferente Ratingskalen etabliert ist (u. a. Praetorius et al. 2012, 2014; Rakoczy und Pauli 2006), liegen allerdings erst wenige fachbezogene oder hybride Beobachtungsinstrumente für den Mathematikunterricht vor, für die Nachweise der psychometrischen Güte zudem recht unterschiedlich ausfallen (Schlesinger und Jentsch 2016).

1.2 Fachbezogene Erfassung der Unterrichtsqualität

Die Diskussion um die zusätzliche Erfassung fachspezifischer Qualitätsmerkmale wird vor allem in Bezug auf den Mathematikunterricht geführt. Blum (2006) argumentiert beispielsweise dafür, dass eine *fachlich gehaltvolle Unterrichtsgestaltung* als Qualitätsmerkmal neben den drei Basisdimensionen zu berücksichtigen sei, um den Kompetenzerwerb im Mathematikunterricht gemäß der KMK-Bildungsstandards (2003) abzubilden. Darunter versteht Blum „vielfältige Gelegenheiten zu kompetenzbezogenen Tätigkeiten“ sowie die Herstellung von Vernetzungen innerhalb und außerhalb der Mathematik (2006, S. 29). Brunner (2018) erweitert dieses Spektrum an Qualitätsmerkmalen, indem sie anmerkt, dass die drei Basisdimensionen die aus fachdidaktisch-normativer Perspektive relevante *fachliche Korrektheit* der präsentierten Unterrichtsinhalte nicht enthalten, obwohl diese neben einer effi-

zienten Klassenführung als notwendige Bedingung für einen lernwirksamen Mathematikunterricht zu verstehen sei.

Als Fazit einer international angelegten Überblicksstudie schlussfolgern Charalambous und Praetorius (2018), dass Lernerfolg im Mathematikunterricht durch fachspezifische und generische Merkmale der Unterrichtsqualität gemeinsam besser erklärt werden könne als durch generische Merkmale allein. Beispielsweise wurden im Rahmen der MET Study (*Measures of Effective Teaching*, Kane und Staiger 2012) verschiedene Beobachtungsinstrumente zur Erfassung der Qualität von Mathematikunterricht eingesetzt, von denen einige ausschließlich für Mathematikunterricht entwickelt wurden und andere in verschiedenen Fächern einsetzbar sind. Die durch die unterschiedlichen Beobachtungsinstrumente erfassten Ratings korrelierten mit $0{,}67\ \rho < 0{,}88$ zwar hoch, dennoch lieferte jedes Beobachtungsinstrument auch spezifische Informationen über den beobachteten Mathematikunterricht (Charalambous und Praetorius 2018).

In Deutschland haben Lipowsky et al. (2018) den Zusammenhang zwischen den drei Basisdimensionen und einer empirisch ermittelten Dimension zur *fachdidaktischen Qualität unterrichtlicher Theoriephasen* bei der Einführung des Satzes von Pythagoras untersucht (Lipowsky et al. 2009). Eine positive Korrelation mittlerer Stärke zeigte sich zwischen der Klassenführung und struktureller Klarheit als Subdimension der fachdidaktischen Qualität. Weitere Zusammenhänge konnten an Hand der Gesamtstichprobe nicht nachgewiesen werden. Eine getrennte Analyse zeigte für Real- und Sekundarschulklassen ein ähnliches Muster wie für die Gesamtstichprobe, während sich bei Gymnasialklassen positive Korrelationen zwischen generischen und fachspezifischen Merkmalen ergaben. Die Autorengruppe schlussfolgerte daher wie Kane und Staiger (2012), dass sich die erfassten Merkmale der Unterrichtsqualität nicht gegenseitig kompensieren können (Lipowsky et al. 2018).

Eine offene Frage ist, inwieweit diese Ergebnisse wegen ihres Bezugs zum Satz des Pythagoras auf andere Themen des Mathematikunterrichts übertragbar sind. Das bezieht sich vor allem auf die Operationalisierung der fachdidaktischen Unterrichtsqualität (Subdimensionen „strukturelle Klarheit", „Repräsentationsformen" und „Verstehenselemente"[1], Drollinger-Vetter 2011). So ist es auch denkbar, dass die Zusammenhänge niedrig ausfielen, weil die fachdidaktische Unterrichtsqualität im Gegensatz zu den drei Basisdimensionen inhaltspezifisch konzeptualisiert und erfasst wurde.

2 Validierung der Erfassung fachspezifischer Unterrichtsqualität

2.1 Argumentatives Verständnis von Validität

Validität stellt das wichtigste, aber auch das am schwierigsten nachzuweisende Gütekriterium empirischer Sozialforschung dar (AERA et al. 1999). Validität beschreibt

[1] Lipowsky et al. (2018, S. 191, vgl. auch Drollinger-Vetter 2011) bezeichnen „Verstehenselemente" als „diejenigen Teilelemente und inhaltlichen Bausteine des Satzes von Pythagoras (...), welche man verstanden haben muss, um durch Verknüpfungsleistungen das Konzept als Ganzes verstehen zu können.".

„the appropriateness, meaningfulness, and usefulness of specific inferences made from test scores“ (AERA et al. 1999, S. 9). Dies bezieht sich in der Unterrichtsqualitätsforschung beispielsweise auf Schlussfolgerungen, die aus Ratings gezogen werden (Bell et al. 2012). Solche Schlussfolgerungen betreffen die inhaltlich und statistisch angemessene Bewertung des beobachteten Verhaltens, die Generalisierbarkeit oder Vorhersagekraft von Ratings sowie bildungspolitische Entscheidungen, die auf der Basis von Ratings getroffen werden (ausführlich Kane 2013; Hartig et al. 2008).

Zur Validierung dieser Schlussfolgerungen empfehlen Bell et al. (2012) sowie Taut und Rakoczy (2016), Nachweise der inhaltlichen und faktoriellen Validität zu erbringen und Generalisierbarkeitsstudien durchzuführen (Cronbach et al. 1972). Für das Beobachtungsinstrument, das in den Studien TEDS-Unterricht und TEDS-Validierung eingesetzt wurde, liegen erste Ergebnisse zur Inhaltsvalidität (Schlesinger et al. 2018) und zur Generalisierbarkeit vor (Jentsch et al. 2019). Es fehlen bislang belastbare empirische Hinweise auf die Faktorenstruktur des Beobachtungsinstruments.

Bell et al. (2012) diskutieren ferner Zusammenhangsanalysen mit geeigneten Kriterien (z. B. Kompetenzfacetten von Lehrpersonen) als Bestandteile einer Validierungsstrategie. Nach Bromme (1995) wird erwartet, dass Lehrpersonen für die unterrichtliche Implementation fachspezifischer Merkmale auf fachliches oder fachdidaktisches Wissen zurückgreifen (Shulman 1986; vgl. auch Praetorius und Charalambous 2018). Dieses Wissen ist dagegen nicht erforderlich, wenn generische Unterrichtsmerkmale in den Blick genommen werden. Diese Annahme erlaubt es, die Interpretation von Unterrichtsqualitätsdimensionen als generisch oder fachspezifisch durch Zusammenhangsanalysen mit Kompetenzfacetten der Lehrpersonen zu validieren (im Sinne konvergenter und diskriminanter Validität, Campbell und Fiske 1959; Hartig et al. 2008). Nachfolgend skizzieren wir die empirische Befundlage solcher Zusammenhangsanalysen für den Mathematikunterricht.

2.2 Zusammenhang zwischen Unterrichtsqualität und professioneller Kompetenz

Hill et al. (2012) führten eine Studie mit 34 Mathematiklehrpersonen durch und stellten hohe[2] manifeste Korrelationen zwischen der Qualität von Mathematikunterricht und *Mathematical Knowledge for Teaching* (MKT) fest, welches sowohl fachliche als auch fachdidaktische Aspekte beinhaltet (Hill et al. 2012). Die Unterrichtsqualität wurde mit einem Beobachtungsinstrument erhoben, das ausschließlich für den Mathematikunterricht entwickelt wurde. Mit diesem Beobachtungsinstrument wurden die Dimensionen „Richness of the mathematics“, „Errors and imprecision“, „Working with students and mathematics“ und „Common Core aligned student practices“ erfasst (Learning Mathematics for Teaching Project 2011).

[2] Wir folgen gängigen Konventionen (u. a. Cohen 1992, Bortz und Döring 2006) und bezeichnen die praktische Bedeutsamkeit von Effektstärken <0,30 als klein, zwischen 0,30 und 0,50 als moderat und >0,50 als groß.

In anderen Studien (Hill und Chin 2018; Kelcey et al. 2019) fielen diese Zusammenhänge deutlich niedriger aus, möglicherweise aufgrund einer komplexeren statistischen Modellierung, in der für zahlreiche Hintergrundeffekte auf Klassen-, Lehrpersonen-, Schul- und Distriktebene kontrolliert wurde. Kelcey et al. (2019) fanden in einer Studie mit 302 Mathematiklehrkräften keinen statistisch signifikanten Zusammenhang zwischen Klassenführung und MKT sowie schwache oder mittlere Korrelationen zu den fachspezifischen Merkmalen der Unterrichtsqualität („Ambitious mathematics" bzw. „Errors and imprecision"). Hill und Chin (2018) fanden lediglich eine schwache Korrelation zwischen der Urteilsgenauigkeit von 284 Mathematiklehrkräften und der Dimension „Working with students and mathematics".[3]

In einer Studie von Kersting et al. (2012) wurde MKT bei 38 Mathematiklehrpersonen erfasst. Zudem wurde ein Instrument zur Erfassung der fachspezifischen Unterrichtsqualität eingesetzt, das inhaltlich mit der fachdidaktischen Qualität unterrichtlicher Theoriephasen bei Lipowsky et al. (2018) korrespondiert. Es fanden sich erwartungswidrig keine Zusammenhänge zwischen dem fachspezifischen Wissen der Lehrpersonen und der Unterrichtsqualität. Mit einem videobasierten Instrument, das fachliche, fachdidaktische und pädagogische Aspekte der professionellen Kompetenz auf einer integrierten Skala erfasst, wurden jedoch hohe Zusammenhänge zur Unterrichtsqualität und moderate Korrelationen zum fachlichen und fachdidaktischen Wissen der Lehrpersonen mit MKT nachgewiesen (Kersting et al. 2012).

In der COACTIV-Studie (Baumert et al. 2010) wurden Zusammenhänge zwischen dem Professionswissen von Mathematiklehrpersonen und den Basisdimensionen der Unterrichtsqualität untersucht. In einer Mediationsanalyse fand die Autorengruppe einen kleinen Effekt des mathematikdidaktischen Wissens auf die konstruktive Unterstützung und einen moderaten Effekt auf das Potential zur kognitiven Aktivierung. Dabei ist zu berücksichtigen, dass das Potential zur kognitiven Aktivierung fachbezogen durch Aufgabenanalysen erhoben wurde. Eingeschätzt wurden die Subdimensionen „Typ mathematischen Arbeitens", „mathematisch Argumentieren" und „innermathematisches Modellieren". Konstruktive Unterstützung wurde dagegen stärker als generisches Merkmal verstanden.

Zusammenfassend zeigt die Befundlage, dass sich fachspezifische Kompetenzfacetten dazu eignen dürften, die Unterscheidung zwischen fachspezifischen bzw. generischen Unterrichtsqualitätsmerkmalen zu stützen. Deutlich geworden ist aber auch, dass es von den Messinstrumenten abhängig ist, inwieweit Merkmale der Unterrichtsqualität als fachspezifisch zu interpretieren sind (vgl. auch Praetorius und Charalambous 2018). Die Befunde der Studie von Hill et al. (2012) weisen ferner darauf hin, dass zwischen verschiedenen Arten von Kompetenzfacetten – stärker dispositionaler oder situationsbezogener Art (Blömeke et al. 2015; Kaiser et al. 2015) – unterschieden werden sollte.

[3] Die Lehrpersonen sollten für jede von insgesamt 37 Mathematikaufgaben einschätzen, wie viele ihrer Schülerinnen und Schüler die jeweilige Aufgabe korrekt lösen würden. Aus diesen Einschätzungen und den tatsächlichen Lösungshäufigkeiten wurde ein Differenzmaß gebildet, das in den Zusammenhangsanalysen eingesetzt wurde.

3 Zur vorliegenden Studie

3.1 Erfassung der Unterrichtsqualität in TEDS-Unterricht und TEDS-Validierung

Der Beitrag entstand im Rahmen der Studien TEDS-Unterricht und TEDS-Validierung, welche untersuchen, inwieweit intra- und interindividuelle Leistungsunterschiede von Lernenden der unteren Sekundarstufe im Fach Mathematik mit der professionellen Kompetenz der unterrichtenden Lehrpersonen zusammenhängen und inwieweit diese Zusammenhänge durch die Qualität des Mathematikunterrichts vermittelt werden. Die Unterrichtsqualität wurde durch ein Beobachtungsinstrument mit hoch-inferenten Ratingskalen erhoben, das sowohl die drei Basisdimensionen als auch stärker fachbezogene Merkmale abbildet (Schlesinger et al. 2018).

Effiziente Klassenführung wurde in Anlehnung an das Beobachtungsinstrument der Pythagoras-Studie (Rakoczy und Pauli 2006) durch Indikatoren des Zeitmanagements, der Störungsprävention und strukturierter Unterrichtsführung operationalisiert (vgl. Tab. 1 in Abschn. 4.3). Die Operationalisierung der konstruktiven Unterstützung erfolgte ebenfalls in Anlehnung an Rakoczy und Pauli (2006) und bezieht sich auf Angebote zur Individualisierung und Differenzierung, unterrichtliches Autonomieerleben und konstruktive Rückmeldungen der Lehrperson. Das Potential zur kognitiven Aktivierung wurde in Anlehnung an Maier et al. (2010), Lotz (2015) und Helmke (2012) fächerübergreifend, insbesondere ohne thematische Eingrenzung, operationalisiert und bezieht sich auf die unterrichtliche Problemorientierung und Wissenssicherung.

In Bezug auf die fachspezifische Unterrichtsqualität wurden nach einer systematischen Literaturrecherche (Schlesinger und Jentsch 2016; Schlesinger et al. 2018) Merkmale operationalisiert, die sich einerseits auf die fachlich kohärente Strukturierung der präsentierten Unterrichtsinhalte beziehen (*stoffbezogene, mathematikdidaktische Qualität*) und andererseits auf solche, die das fachspezifische Potential zur

Tab. 1 Indikatoren der Ratingskala „Herausfordernde Fragen und Probleme“ zur Erfassung der Basisdimension Potential zur kognitiven Aktivierung (vgl. Rakoczy und Pauli 2006)

Ratingskala	Herausfordernde Fragen und Probleme			
Dimension	Potential zur kognitiven Aktivierung			
Anmerkung	Ein fehlender Wert wird vergeben, wenn in einer Unterrichtsphase überwiegend keine Aufgaben zu bearbeiten sind (z. B. Gespräch über organisatorische Dinge)			
Indikatoren	Die Lehrperson stellt offene Fragen, die zum Nachdenken anregen			
	Das Frageverhalten der Lehrperson ist *nicht* kleinschrittig			
	Es wird an anspruchsvollen Aufgaben gearbeitet, die über die Verwendung von Lösungsalgorithmen hinausgehen			
	Die Lehrperson gibt geeignete Impulse zur vertieften Auseinandersetzung mit dem Unterrichtsinhalt (z. B. ein Problem, Dilemma, etc.)			
Ausprägung	1 – sehr niedrig	2 – eher niedrig	3 – eher hoch	4 – sehr hoch
	Kein Indikator trifft zu	Genau ein Indikator trifft zu	Mehrere, aber nicht alle Indikatoren treffen zu	Alle Indikatoren treffen zu

kognitiven Aktivierung in den Blick nehmen (*unterrichtsbezogene, mathematikdidaktische Qualität*, vgl. auch Lipowsky et al. 2018).

3.2 Forschungsfragen

In diesem Beitrag werden drei Fragestellungen bearbeitet. Die Beantwortung dieser Forschungsfragen versteht sich als Teil einer umfangreichen Validierungsstudie, die neben den hier dargestellten Aspekten auch die inhaltliche Angemessenheit und Generalisierbarkeit der Ratings untersucht. Die ersten beiden Fragestellungen beziehen sich auf die Struktur des Instruments:

1. *Welche Faktorenstruktur liegt dem Beobachtungsinstrument zu Grunde?*
2. *Wie stark interkorrelieren die empirisch ermittelten Dimensionen der Unterrichtsqualität?*

Wir gehen davon aus, dass sich wie in der Literatur drei Basisdimensionen identifizieren lassen. Offen ist aber, ob sich wie bei Lipowsky et al. (2018) eine eigenständige Dimension ermitteln lässt, die die fachspezifischen Unterrichtsmerkmale zusammenfasst, was einem additiven Ansatz entsprechen würde. Unabhängig von dieser Frage erwarten wir, dass die Qualitätsdimensionen *ein* gemeinsames Konstrukt abbilden, so dass sie im Einklang mit der Forschungslage mindestens schwach positiv korrelieren sollten.

Die vorliegende Studie zielt auf die Erfassung der Qualität alltäglichen Mathematikunterrichts ab, also ohne Einschränkung auf einen bestimmten Unterrichtsinhalt, und unterscheidet sich daher in einem wesentlichen Punkt vom methodischen Vorgehen bei Lipowsky et al. (2018). Es wäre somit auch denkbar, dass stärkere Zusammenhänge zwischen generischen und fachspezifischen Merkmalen der Unterrichtsqualität beobachtet werden, als dies bei Lipowsky et al. (2018) der Fall war.

3. *Wie stark korrelieren die empirisch ermittelten Dimensionen der Unterrichtsqualität mit fachspezifischen Kompetenzfacetten der Lehrpersonen?*

Mit der dritten Fragestellung wird thematisiert, in welchem Ausmaß die durch das Beobachtungsinstrument erfassten Dimensionen der Unterrichtsqualität als fachspezifisch gelten können. Fachspezifische Unterrichtsmerkmale zeichnen sich nach Bromme (1995) dadurch aus, dass Lehrpersonen für ihre Implementation auf fachspezifische Kompetenzfacetten zurückgreifen (vgl. auch Praetorius und Charalambous 2018). Zur konvergenten und diskriminanten Validierung der Qualitätsdimensionen werden daher Zusammenhänge zum mathematischen und mathematikdidaktischen Wissen sowie zur professionellen Wahrnehmung der Lehrpersonen von Mathematikunterricht untersucht (im Sinne einer stärker situationsspezifischen Kompetenzfacette, vgl. Blömeke et al. 2015; Kaiser et al. 2015).

Wir gehen davon aus, dass Merkmale der Klassenführung nicht mit den fachspezifischen Kompetenzfacetten der Lehrpersonen zusammenhängen (diskriminante Validität). In Anlehnung an den Forschungsstand (Baumert et al. 2010; Hill et al. 2012; Kersting et al. 2012) ist ferner anzunehmen, dass fachspezifische Kompetenzfacetten und fachspezifische Unterrichtsmerkmale moderat interkorrelieren und

umso höher ausfallen, je proximaler die Kompetenzfacetten in Bezug auf das Unterrichtshandeln der Lehrpersonen erfasst werden (konvergente Validität). Für konstruktive Unterstützung und das Potential zur kognitiven Aktivierung können wir auf Grund der unterschiedlichen Operationalisierungen in der Literatur dagegen keine Annahme treffen.

4 Methode

4.1 Stichprobe

Die im vorliegenden Beitrag dargestellte Untersuchung verwendet Daten aus den Studien TEDS-Unterricht und TEDS-Validierung, an denen 76 Lehrpersonen der unteren Sekundarstufe aus vier Bundesländern teilgenommen haben. Von den teilnehmenden Lehrpersonen waren 41 (54 %) weiblichen Geschlechts und 47 (62 %) unterrichteten an einem Gymnasium. Das mittlere Alter betrug 42 Jahre ($SD = 10{,}5$ Jahre) und die mittlere Berufserfahrung 14 Jahre ($SD = 10$ Jahre). Die Lehrpersonen hatten ihr erstes Staatsexamen mit der Note 1,8 ($SD = 0{,}6$) und das zweite Staatsexamen mit der Note 1,9 ($SD = 0{,}7$) bestanden. Die Teilnahme erfolgte auf freiwilliger Basis, es handelt sich daher um eine Gelegenheitsstichprobe.

4.2 Messung der fachspezifischen Kompetenzfacetten

Die fachspezifischen Kompetenzfacetten der Mathematiklehrpersonen wurden aus forschungsökonomischen Gründen online erfasst. Die Testung wurde zeitlich beschränkt, damit zur Aufgabenbearbeitung keine Nachschlagewerke eingesetzt werden konnten. Insgesamt betrug die Testdauer etwa 90 min.

Das mathematische Fachwissen (MCK) und das mathematikdidaktische Wissen (MPCK) wurde mit digitalisierten Papier-und-Bleistift-Tests erfasst. Es handelte sich dabei um verkürzte Versionen der TEDS-M-Tests (Blömeke et al. 2010), die in einer Nachfolgestudie von TEDS-M entwickelt und bereits mehrfach eingesetzt wurden. Die Testitems haben mit wenigen Ausnahmen ein Multiple-Choice-Antwortformat. Der MCK-Test besteht aus 26 Items aus den Bereichen Arithmetik, Algebra, Geometrie und Stochastik und beinhaltet die kognitiven Anforderungsniveaus Wissen, Anwenden und Begründen (Beispielitem in Anhang Abb. 1). Der MPCK-Test umfasst 29 Items und erfasst vor allem stoffdidaktisches, curriculares und planungsbezogenes Wissen (Beispielitem in Anhang Abb. 2; vgl. auch Buchholtz et al. 2014).

Ferner wurde die Fähigkeit der Lehrpersonen zur professionellen Wahrnehmung von Mathematikunterricht in den drei Subfacetten „Perception“, „Intepretation“ und „Decision-Making“ erhoben (M_PID, Blömeke et al. 2014; Kaiser et al. 2015). Dazu wurden ihnen drei Videovignetten vorgelegt, zu denen insgesamt 31 Items bearbeitet werden sollten. Diese bestanden etwa zur Hälfte aus offenen bzw. Multiple-Choice-Aufgaben. Die drei Videovignetten dauern zwischen 2,5 und 4 min und zeigen Mathematikunterricht in den Jahrgangsstufen 8–10. Die Szenen wurden gestellt, um eine möglichst hohe Dichte an Ereignissen zu erreichen und konnten nur einmal angesehen werden. Beides wurde den Probandinnen und Probanden zu Beginn der

Testung mitgeteilt. Die Lehrpersonen erhielten außerdem Informationen zu den mathematischen Inhalten sowie zur Zusammensetzung und zu den Vorkenntnissen der Klasse.

Wir beschreiben beispielhaft eine der Videovignetten. Es werden Szenen aus einer Mathematikstunde in einer leistungsheterogenen 9. Gymnasialklasse gezeigt. Die Schülerinnen und Schüler führen eine Partnerarbeit zu einer Geometrieaufgabe durch. Nach der Partnerarbeit findet ein Plenumsgespräch zum Ergebnisaustausch statt. Die Probandinnen und Probanden sollten nun Items zur Videovignette bearbeiten, die auf ihre Fähigkeit zur Wahrnehmung des Unterrichtsgeschehens (Perception) sowie auf ihr eigenes Handlungsrepertoire abzielen (Decision-Making, vgl. Anhang Abb. 3 für ein Beispielitem). Offene Antworten der Lehrpersonen wurden nach einem umfangreichen Manual kodiert, das Grenzfälle und Ankerbeispiele enthält. Für eine ausführliche Darstellung der Erfassung der Interpretationsfähigkeit (Interpretation) und zur Inhaltsvalidität verweisen wir auf Blömeke et al. (2014).

Die Tests zu den Kompetenzfacetten MCK, MPCK und M_PID ließen sich mit akzeptablen Reliabilitäten nach dem Raschmodell skalieren ($0{,}66 < WLE \leq 0{,}80$) und interkorrelieren messfehlerbereinigt stark positiv ($0{,}61 < r \leq 0{,}79$). Die Spannweite korrekt gelöster Items auf Personenebene betrug für alle drei Tests 25–82 %.

4.3 Zum Rating der Unterrichtsqualität

Für das in Abschn. 3.1 beschriebene Beobachtungsinstrument wurde ein Rating-Manual entwickelt, in dem verhaltensnahe Indikatoren für alle Unterrichtsqualitätsmerkmale formuliert sind (vgl. als Beispiel für Indikatoren Tab. 1, für eine Übersicht über alle Qualitätsmerkmale siehe Tab. 2 im Ergebnisteil). Die Einschätzung der Merkmale erfolgte auf vierstufigen, hoch-inferenten Ratingskalen (1 = „sehr niedrig“ bis 4 = „sehr hoch“). Die Erfassung der Unterrichtsqualität erfolgte durch Live-Ratings, welche eine leichtere Zugänglichkeit zum Forschungsfeld bieten und von denen angenommen wird, dass sie einen geringeren Einfluss auf das Unterrichtsgeschehen nehmen.

Insgesamt zehn Beobachterinnen und Beobachter wurden für die Ratings eingesetzt. Diese hatten mindestens einen Bachelorabschluss in einem Lehramtsstudium mit Unterrichtsfach Mathematik inne. Die Beobachtenden wurden vor der Datenerhebung umfangreich geschult. Die Schulung umfasste etwa 30 h und bestand zu ähnlichen Anteilen aus Videoanalysen, Live-Ratings und Gruppendiskussionen.

Jeweils zwei Beobachtende führten bei jeder teilnehmenden Lehrperson zwei Unterrichtsbeobachtungen im Mathematikunterricht durch. Der zeitliche Abstand zwischen den beiden Unterrichtsbeobachtungen betrug etwa zwei Wochen. Vornehmlich wurden Doppelstunden beobachtet (90 min), bei zwei Lehrpersonen gingen aus organisatorischen Gründen vier Einzelstunden in die Analysen ein. Da Ratings der Unterrichtsqualität als kognitiv anspruchsvoll und daher fehleranfällig gelten (Prae-

Tab. 2 Rotierte Ladungsmatrix einer explorativen Faktorenanalyse für das Beobachtungsinstrument ($n = 156$)

Ratingskalen	Faktor 1: *Klassenführung*	Faktor 2: *Konstruktive Unterstützung*	Faktor 3: *Kognitive Aktivierung*	Faktor 4: *Fachdidaktische Strukturierung*
Klassenführung				
Effektive Lernzeitnutzung	**0,51**		–0,25	
Störungsprävention	**0,92**			
Arbeitsatmosphäre	**0,92**			
Strukturierung				**0,60**
Konstruktive Unterstützung				
Individuelle Förderung		**0,55**		
Umgang mit Heterogenität		**0,77**		
Selbststeuerung		**0,75**		
Rückmeldungen				**0,56**
Kooperatives Arbeiten		0,45	0,39	
Kognitive Aktivierung				
Herausfordernde Fragen			**0,51**	0,32
Ko-Konstruktion				0,34
Qualität der Methoden	0,26		0,32	0,38
Wissenssicherung		–0,25		**0,63**
Unterrichtsbezogene, mathematikdidaktische Qualität				
Darstellungsformen			0,49	
Intelligentes Üben		0,26	**0,59**	
Qualität der Beispiele			0,47	
Relevanz (Sense-Making)			**0,65**	
Fachliche Tiefe			**0,51**	0,33
Stoffbezogene, mathematikdidaktische Qualität				
Umgang mit Fehlern				**0,58**
Fachliche Korrektheit				**0,73**
Erklärungen der Lehrperson				**0,77**
Kompetenzorientierung				

Anmerkungen: Faktorladungen <0,20 wurden zur besseren Lesbarkeit nicht abgedruckt. Faktorladungen ≥0,50 wurden fettgedruckt. Rotationsverfahren: geomin, Schätzverfahren: MLR

torius et al. 2012), wurden diese viermal innerhalb einer Doppelstunde bzw. zweimal innerhalb einer Einzelstunde durchgeführt.[4] Die Ratings wurden anschließend durch Mittelwertbildung auf Stundenebene aggregiert. Ziel der Verwendung von zwei Ratern und zwei Ratings pro Unterrichtsstunde war eine Reduktion der Fehlervarianz (Mashburn et al. 2014; Pietsch und Tosana 2008).

Nach jeder Unterrichtsbeobachtung fand eine Nachbesprechung statt, in der die Beobachterinnen und Beobachter mögliche Fehleinschätzungen reflektierten. Im Rahmen dieser Nachbesprechungen konnten Ratings verändert werden. Auch wenn dies bedeutet, dass die Einschätzungen nicht unabhängig voneinander vorgenommen wurden, erschien dieses Vorgehen als angemessen, um die Inhaltsvalidität der Ratings zu erhöhen (vgl. dazu König 2015).

Das Beobachtungsinstrument wurde im Frühjahr 2015 bei 13 Lehrpersonen aus drei Bundesländern pilotiert und einer Beurteilung durch Expertinnen und Experten unterzogen. Hierbei wurden mehrere Indikatoren umformuliert oder neu entwickelt. Fünf Ratingskalen wurden nach der Pilotierung wegen niedriger Interrater-Reliabilität („Regeln und Routinen“, „Klarheit“ und „Wertschätzung durch die Lehrperson“, $ICC < 0{,}65$) oder Trennschärfe ausgeschlossen („Feedback der Klasse an die Lehrperson“ und „Förderung von Metakognition“, $r_{it} < 0{,}15$). Die Interrater-Reliabilität war in der vorliegenden Studie insgesamt zufriedenstellend ($ICC > 0{,}80$, Wirtz und Caspar 2002).

4.4 Statistische Analysen

Die statistischen Analysen wurden mit $n = 156$ gemittelten Ratings durchgeführt. Zur Bearbeitung der ersten Fragestellung wurde zunächst die Anzahl der zu extrahierenden Faktoren durch eine Parallelanalyse (Horn 1965) mit dem R-Paket psych (Revelle 2018) bestimmt. Im Anschluss wurde eine explorative Faktorenanalyse mit schiefwinkligem Rotationsverfahren durchgeführt, da die extrahierten Faktoren als Facetten eines gemeinsamen Konstrukts gedeutet werden. Fehlende Werte wurden modellbasiert mit der in MPlus 7.4 (Muthén und Muthén 2010) implementierten FIML-Methode geschätzt (Full Information Maximum Likelihood). Die Cluster-Struktur der Daten, die sich durch Messwiederholungen ergibt, wurde durch eine Korrektur der Standardfehler berücksichtigt (MLR-Schätzer). Die Güte der Modellpassung wird durch das Verhältnis $\chi^2 / df \leq 2{,}50$ und die Kennwerte Root Mean Square Error of Approximation (RMSEA) $\leq 0{,}08$ und Standardized Root Mean Residual (SRMR) $\leq 0{,}05$ ausgewiesen (Hu und Bentler 1999).

Zur Bearbeitung der zweiten und dritten Fragestellung wurden Produkt-Moment-Korrelationen zwischen den Dimensionen der Unterrichtsqualität und den Kompetenzfacetten MCK, MPCK und M_PID geschätzt. Dazu wurden Faktor-Werte für die 156 Unterrichtsstunden bestimmt und diese anschließend durch Mittelwertbildung auf der Ebene der Klassen bzw. Lehrpersonen aggregiert. Auf Grund des

[4] Den Beobachtenden wurde angeraten, das Rating-Intervall von 22,5 min nach Absprache um einige Minuten zu strecken oder kürzen, wenn ihnen dies aus didaktischen Gründen sinnvoll erschien (z. B. auf Grund eines Wechsels von Sozialform oder Methode).

explorativen Charakters der vorliegenden Studie legen wir das Signifikanzniveau für statistische Tests auf $p = 0{,}10$ fest.

5 Ergebnisse

5.1 Faktorenstruktur des Beobachtungsinstruments

Die zuerst durchgeführte Parallelanalyse weist auf eine vierdimensionale Faktorenstruktur des Beobachtungsinstruments hin. Die Ergebnisse der im Anschluss durchgeführten explorativen Faktorenanalyse sind in Tab. 2 dargestellt. Die Gütemaße indizieren eine akzeptable Modellanpassung ($\chi^2 = 322{,}86$, $df = 149$, $p < 0{,}01$, $\chi^2/df = 2{,}17$, RMSEA = 0,08, SRMR = 0,05). In den meisten Fällen ergibt sich durch hohe Faktorladungen auf einem Faktor eine eindeutige Zuordnung der Unterrichtsqualitätsmerkmale. Nur für die Ratingskalen „Kooperatives Arbeiten“ und „Qualität der Methoden“ wurden Mehrfachladungen in vergleichbarer Größenordnung geschätzt, die keine unmittelbare Zuordnung erlauben. In diesen Fällen wurde nach inhaltlicher Diskussion von der Autorengruppe entschieden, „Kooperatives Arbeiten“ dem Faktor 2 und „Qualität der Methoden“ dem Faktor 3 zuzuordnen. Die Ratingskala „Kompetenzorientierung“ wurde ausgeschlossen, da keine Ladung auf einem Faktor feststellbar war.

Der erste Faktor besteht aus drei Ratingskalen, die sich entsprechend unserer Konzeptualisierung der Basisdimension Klassenführung zuordnen lassen. Der Faktor bildet Unterrichtsmerkmale zum Zeitmanagement und zur Disziplin ab, schließt aber unerwartet nicht die Ratingskala zur Strukturierung ein. Faktor 2 besteht aus vier Ratingskalen und lässt sich – wie konzeptualisiert –der Basisdimension konstruktive Unterstützung zuordnen. Er beschreibt Merkmale unterrichtlicher Differenzierung und Individualisierung, schließt aber unerwartet nicht die Ratingskala „Rückmeldungen“ ein. Faktor 3 fasst zwei Ratingskalen zur kognitiven Aktivierung und mehrere Merkmale mathematikdidaktischer Qualität zusammen, unter anderem „intelligentes Üben“ und „fachliche Tiefe“. Dieser Faktor verdichtet also die Unterrichtsmerkmale des generischen und fachbezogenen Potentials zur kognitiven Aktivierung.

Faktor 4 fasst schließlich die übrigen Ratingskalen zur mathematikdidaktischen Qualität und die unerwartet nicht einem der anderen drei Faktoren zugeordneten Unterrichtsqualitätsmerkmale zusammen. Er beschreibt stoffbezogene („fachliche Korrektheit“, „Erklärungen der Lehrperson“ und „Ko-Konstruktion“), strukturierende („Strukturierung“), evaluierende („Wissenssicherung“) und diagnostische Qualitätsmerkmale („Rückmeldungen“ und „Umgang mit Fehlern“). Wir bezeichnen diesen Faktor nachfolgend als *fachdidaktische Strukturierung*.

Tab. 3 präsentiert deskriptive Kennwerte der eingesetzten Ratingskalen und ihre finale Zuordnung zu den empirisch ermittelten Dimensionen der Unterrichtsqualität. Alle Ratingskalen weisen zufriedenstellende Trennschärfen auf. Klassenführung weist eine hohe Ausprägung und interne Konsistenz auf (drei Items, $M = 3{,}50$, $SD = 0{,}44$, $\alpha = 0{,}87$). Konstruktive Unterstützung zeigt eine niedrige Ausprägung und eine akzeptable interne Konsistenz (vier Items, $M = 1{,}59$, $SD = 0{,}41$, $\alpha = 0{,}73$). Kog-

Tab. 3 Deskriptive Item-Kennwerte für das Beobachtungsinstrument: Mittelwerte, Standardabweichungen und standardisierte Faktorladungen ($n = 156$)

Ratingskalen	*M*	*SD*	Trennschärfe	Empirisch ermittelte Dimension der Unterrichtsqualität (Cronbachs α)
Effektive Lernzeitnutzung	3,67	0,36	0,51	*Klassenführung* (α = 0,87)
Störungsprävention	3,44	0,54	0,92	
Arbeitsatmosphäre	3,71	0,56	0,92	
Individuelle Förderung	1,99	0,60	0,55	*Konstruktive Unterstützung* (α = 0,73)
Umgang mit Heterogenität	1,29	0,47	0,77	
Selbststeuerung	1,43	0,56	0,75	
Kooperatives Arbeiten	1,64	0,62	0,45	
Herausfordernde Fragen	2,58	0,51	0,51	*Potential zur kognitiven Aktivierung* (α = 0,80)
Qualität der Methoden	3,00	0,53	0,32	
Darstellungsformen	2,26	0,74	0,49	
Intelligentes Üben	2,29	0,61	0,59	
Qualität der Beispiele	3,13	0,48	0,47	
Relevanz (Sense-Making)	2,07	0,57	0,65	
Fachliche Tiefe	2,36	0,46	0,51	
Strukturierung	3,21	0,61	0,60	*Fachdidaktische Strukturierung* (α = 0,81)
Rückmeldungen	3,10	0,45	0,56	
Ko-Konstruktion	2,60	0,59	0,34	
Wissenssicherung	2,70	0,59	0,63	
Umgang mit Fehlern	2,90	0,57	0,58	
Fachliche Korrektheit	3,77	0,49	0,73	
Erklärungen der Lehrperson	3,14	0,57	0,77	

nitive Aktivierung weist eine Ausprägung nahe des theoretischen Mittelwerts und eine hohe interne Konsistenz auf (sieben Items, $M = 2{,}50$, $SD = 0{,}38$, $\alpha = 0{,}80$). Die fachdidaktische Strukturierung ist etwas höher ausgeprägt, Streuung und Reliabilität fallen ähnlich wie bei Faktor 3 aus (sieben Items, $M = 3{,}06$, $SD = 0{,}36$, $\alpha = 0{,}81$).

5.2 Interkorrelationen der Qualitätsdimensionen und Zusammenhänge zu Lehrerkompetenz

Zur Bearbeitung der zweiten Forschungsfrage wurde eine Korrelationsanalyse durchgeführt (vgl. Tab. 4), die aufzeigt, dass konstruktive Unterstützung mit den übri-

Tab. 4 Interkorrelationen der Unterrichtsqualitätsdimensionen und Produkt-Moment-Korrelationen zwischen Unterrichtsqualität und fachspezifischen Kompetenzfacetten ($n = 76$)

	(1)	(2)	(3)	(4)	MCK	MPCK	M_PID
(1) *Klassenführung*	1,00	–	–	–	0,00	–0,05	0,06
(2) *Konstruktive Unterstützung*	0,22*	1,00	–	–	0,24*	–0,03	0,22*
(3) *Kognitive Aktivierung*	0,43**	0,30**	1,00	–	–0,02	0,13	0,36**
(4) *Fachdidaktische Strukturierung*	0,41**	0,15$^{@}$	0,54**	1,00	0,15$^{@}$	0,19$^{@}$	0,16$^{@}$

Anmerkungen: $^{@}$ $p < 0{,}10$, * $p < 0{,}05$, ** $p < 0{,}01$, einseitig

gen Qualitätsdimensionen eher schwach zusammenhängt, mit der fachdidaktischen Strukturierung sogar nur tendenziell. Dagegen fallen die übrigen Zusammenhänge stärker aus, insbesondere die zur kognitiven Aktivierung. Die höchste Interkorrelation ergibt sich zwischen dem Potential zur kognitiven Aktivierung und der fachdidaktischen Strukturierung.

Die zur Beantwortung der dritten Forschungsfrage untersuchten Zusammenhänge zwischen den empirisch ermittelten Dimensionen der Unterrichtsqualität und den fachspezifischen Kompetenzfacetten MCK, MPCK und M_PID spiegeln nur teilweise unsere Erwartungen wider. Erwartungsgemäß ist das Ergebnis, dass für effiziente Klassenführung keine Zusammenhänge mit den untersuchten Kompetenzfacetten nachgewiesen werden können. Dieses Ergebnis stützt die diskriminante Validität der Interpretation dieser Dimension als generisches Unterrichtsqualitätsmerkmal.

Konstruktive Unterstützung und fachdidaktische Strukturierung weisen schwach signifikant positive oder tendenziell positive Korrelationen zu den fachspezifischen Kompetenzfacetten auf ($0{,}15 < r \leq 0{,}24$). Dieses Ergebnis lässt darauf schließen, dass die unterrichtliche Implementation dieser Merkmalsbereiche durch Rückgriff auf fachspezifische Kompetenzfacetten erfolgt, dass dies aber Grenzen hat. Damit stellt sich die Frage, ob diese Qualitätsdimensionen als generisch oder fachspezifisch zu bezeichnen sind.

Widersprüchliche Ergebnisse zeigen sich in Bezug auf das Potential zur kognitiven Aktivierung. Zu dieser als fachspezifisch angenommenen Dimension lassen sich keine bedeutsamen Korrelationen mit MCK oder MPCK nachweisen. Damit fehlen konvergente Validitätsbelege in dieser Hinsicht. Im Gegensatz dazu zeigt sich der stärkste positive Zusammenhang ($r = 0{,}36$) zwischen kognitiver Aktivierung und der situationsbezogenen Kompetenz M_PID, was die Interpretation dieser Unterrichtsqualitätsdimension als fachspezifisch stützt.

6 Diskussion

Im Rahmen dieses Beitrags wurde die Erfassung der Qualität alltäglichen Mathematikunterrichts in der unteren Sekundarstufe beschrieben. Auf Grund der skizzierten Überlegungen zur fachbezogenen Erhebung der Unterrichtsqualität und den empirischen Befunden zu deren Wirksamkeit (im Überblick Charalambous und Praetorius 2018) bestand das Ziel der Instrumentenentwicklung darin, sowohl generische als

auch fachspezifische Merkmale der Qualität von Mathematikunterricht zu erfassen. Dazu wurden drei Forschungsfragen zur Struktur des Beobachtungsinstruments und zur Fachspezifität der erhobenen Merkmale bearbeitet.

6.1 Zur Faktorenstruktur des Beobachtungsinstruments

Die vorliegenden Ergebnisse einer Parallel- und einer explorativen Faktorenanalyse auf der Datenbasis von 156 Unterrichtsstunden deuten darauf hin, dass dem Beobachtungsinstrument eine vierdimensionale Struktur zugrunde liegt. Erwartungsgemäß waren nahezu alle beobachteten Interkorrelationen der Unterrichtsqualitätsdimensionen signifikant positiv, was die Vermutung zulässt, dass sie zu *einem* gemeinsamen, übergeordneten Konstrukt beitragen (Unterrichtsqualität). Die Zusammenhänge fallen allerdings eher moderat aus, was darauf hindeutet, dass jeweils eigenständige Teilkonstrukte abgebildet werden.

Eine der vier empirisch ermittelten Dimensionen lässt sich eindeutig als Basisdimension Klassenführung interpretieren. Die Dimension bildet die Merkmale Zeitmanagement und Disziplin, nicht aber die unterrichtliche Strukturierung ab (Klieme und Rakoczy 2008; Rakoczy und Pauli 2006). Die Ergebnisse sind damit in Einklang mit einem Begriffsverständnis von Klassenführung in Anlehnung an Kounin (1970), das auch in der COACTIV-Studie zum Ausdruck kommt (Baumert et al. 2010).

Konzeptionell gesehen kann auch die zweite Dimension als Basisdimension konstruktive Unterstützung interpretiert werden, die Merkmale unterrichtlicher Differenzierung und Individualisierung abbildet. Mit der empirisch ermittelten vierten Dimension (fachdidaktische Strukturierung) wurden dagegen Unterrichtsqualitätsmerkmale zusammengefasst, die sich auf stoffbezogene, diagnostische, evaluierende oder strukturierende Maßnahmen im Mathematikunterricht beziehen.

Mit Blick auf bestehende Operationalisierungen der Basisdimension konstruktive Unterstützung (u. a. Baumert et al. 2010; Rakoczy und Pauli 2006; Praetorius et al. 2018; zusammenfassend Kunter und Ewald 2016) lassen sich die vorliegenden Befunde so deuten, dass die empirisch ermittelten Dimensionen 2 und 4 unterschiedliche *motivationale* bzw. *kognitive* Konzeptualisierungen des Merkmals konstruktive Unterstützung abbilden, die nicht notwendig stark miteinander zusammenhängen. Während etwa in der Pythagoras-Studie ein integratives Verständnis konstruktiver Unterstützung zum Tragen kommt, bei dem sich nahezu alle der mit den Dimensionen 2 und 4 erfassten Unterrichtsmerkmale wiederfinden (Rakoczy und Pauli 2006), berücksichtigen Baumert et al. (2010) Maßnahmen zur Differenzierung und Individualisierung nicht.

Unsere Ergebnisse könnte man also derart interpretieren, dass die zuletzt genannten Unterrichtsmerkmale, die in der Dimension konstruktive Unterstützung zusammengefasst sind, stärker auf eine *motivationale* Unterstützung der Lernenden abzielen, während der Faktor fachdidaktische Strukturierung stärker *kognitiv-strukturierende* Unterstützungsmaßnahmen thematisiert. Ähnliche Unterscheidungen finden sich bei Praetorius und Charalambous (2018), bei Rakoczy et al. (2007), zur Konzeptualisierung des CLASS-Instruments (Pianta und Hamre 2009) und in einer Studie zur Qualität von Sachunterricht (Kleickmann 2012). Diese Unterscheidung würde auch erklären, warum die Interkorrelationen zwischen konstruktiver Unter-

stützung und den übrigen Qualitätsdimensionen vergleichsweise niedrig ausfallen. Eine Trennung von motivationalen und kognitiven Zieldimensionen hat sich in der Vergangenheit in vielen Studien gezeigt (z. B. Gruehn, 1995; Blömeke und Olsen, 2019).

Das Potential zur kognitiven Aktivierung im Mathematikunterricht lässt sich in generische (z. B. Maier et al. 2010) und stärker fachbezogene Merkmale unterteilen (z. B. Leuders und Holzäpfel 2011; Rakoczy und Pauli 2006). Dies wurde bei der Instrumentenentwicklung berücksichtigt, war konzeptionell aber zwei verschiedenen Dimensionen von Unterrichtsqualität zugeordnet worden. Die empirischen Befunde zeigen, dass die Ratingskalen zum generischen und fachbezogenen Potential zur kognitiven Aktivierung auf einem gemeinsamen Faktor laden. Eine konzeptionell-analytische Trennung dieser Merkmalsbereiche ist also vor dem Hintergrund der empirischen Ergebnisse zwar weiterhin möglich, im Unterrichtsalltag aber kaum durchzuführen. Dieses Ergebnis steht in Einklang mit frühen Überlegungen zur Konzeptualisierung der drei Basisdimensionen, in denen mehrfach auf die Fachbezogenheit dieser Dimension hingewiesen wurde (Klieme und Rakoczy 2008; Praetorius et al. 2014; Lipowsky et al. 2018).

6.2 Zur Fachspezifität des Beobachtungsinstruments

Die dritte Forschungsfrage befasste sich damit, die Interpretation der erfassten Unterrichtsmerkmale als generisch oder fachspezifisch zu validieren. In Anlehnung an Bromme (1995) haben wir mit dem Ziel der konvergenten und diskriminanten Validierung untersucht, inwieweit die erhobenen Qualitätsdimensionen mit fachlichen und fachdidaktischen Kompetenzfacetten der Mathematiklehrpersonen zusammenhängen.

Für Klassenführung fanden wir keine statistisch signifikanten Zusammenhänge. Damit stützen die Ergebnisse unsere Interpretation dieses Qualitätsmerkmals als Basisdimension im Sinne einer diskriminanten Validierung (Hartig et al. 2008). Dagegen fanden wir eine moderate Korrelation zwischen der professionellen Wahrnehmung von Mathematikunterricht (M_PID) und der kognitiven Aktivierung, eine schwach positive Korrelation zur konstruktiven Unterstützung und – in der Tendenz – einen Zusammenhang zur fachdidaktischen Strukturierung. Zudem fanden wir eine schwache Korrelation zwischen dem fachlichen Wissen (MCK) der Lehrpersonen und der konstruktiven Unterstützung im Mathematikunterricht sowie tendenziell Zusammenhänge zwischen MCK bzw. MPCK und fachdidaktischer Strukturierung. Unerwartet fanden wir keine statistisch signifikanten Korrelationen dieser Kompetenzfacetten mit kognitiver Aktivierung.

Auch wenn die Korrelationen niedriger ausfallen als in ähnlich angelegten Studien (z. B. Learning Mathematics for Teaching Project 2011), interpretieren wir die Befunde dahingehend, dass mit Ausnahme der Klassenführung jede der hier erfassten Dimensionen zumindest anteilig mit fachspezifischen Kompetenzfacetten der Lehrpersonen assoziiert ist (vgl. auch Praetorius und Charalambous 2018). Dies wirft etwa die Frage auf, inwieweit es gerechtfertigt ist, konstruktive Unterstützung und kognitiver Aktivierung als generischen Dimensionen zu verstehen.

Die Ergebnisse fallen jedoch keineswegs eindeutig aus und sollten in zukünftigen Untersuchungen auf ihre Generalisierbarkeit hin überprüft werden. So ist etwa der Befund, dass konstruktive Unterstützung mit MCK, nicht aber mit MPCK zusammenhängt, konzeptuell erwartungswidrig, da konstruktive Unterstützung vor allem im Kontext pädagogisch-psychologischer Theorien diskutiert wird (z. B. Rakoczy und Pauli 2006). Detaillierte Aufgabenanalysen (z. B. Jordan et al. 2008), die in diesem Beitrag aus Platzgründen nicht vorgenommen werden, könnten hierzu möglicherweise eine Erklärung leisten.

Die insgesamt eher schwachen Zusammenhänge zwischen der Unterrichtsqualität und den Kompetenzfacetten der Lehrpersonen könnten auch ein Hinweis darauf sein, dass das Professionswissen nicht handlungsnah genug erfasst wurde (Blömeke et al. 2015; Kaiser et al. 2015). Diese Annahme wird zwar nicht für alle Dimensionen, zumindest aber durch die differentielle Korrelation der kognitiven Aktivierung gestützt: Keine bzw. schwache Zusammenhänge zu wissensbezogenen Maßen, moderate Zusammenhänge zur als situationsspezifisch geltenden Kompetenzfacette M_PID (Blömeke et al. 2014).

Für die vierte Dimension (fachdidaktische Strukturierung) ist es daher bemerkenswert, dass die Zusammenhänge zu *allen* untersuchten Kompetenzfacetten schwach bzw. tendenziell positiv ausfallen. Dies deutet einerseits darauf hin, dass andere Merkmale der Lehrpersonen ihr unterrichtliches Handeln in dieser Hinsicht erklären könnten. Neben weiteren, wissensbezogenen und situationsspezifischen Kompetenzfacetten (pädagogisches Wissen und Können, z. B. König 2015) sei hier beispielhaft auf die (fachspezifischen) Überzeugungen und das (fachspezifische) Interesse der Mathematiklehrpersonen verwiesen. Andererseits lässt das Zusammenhangsmuster vermuten, dass diese vier fachbezogenen Konstrukte (MCK, MPCK, M_PID und fachdidaktische Strukturierung) anteilig durch einen gemeinsamen Faktor erklärt werden könnten (z. B. mathematikbezogenes Wissen und Können der Lehrpersonen).

Als Weiterführung unserer Überlegungen könnte in zukünftigen Studien bei geeigneter Stichprobengröße ein Modell im Strukturgleichungsansatz geprüft werden, das neben den drei Basisdimensionen der Unterrichtsqualität einen dazu orthogonalen „Fachspezifitätsfaktor" erfasst (z. B. Geiser 2010). Durch diese Modellierung wäre es möglich, für jede eingesetzte Ratingskala anzugeben, welcher Varianzanteil durch welche fachspezifische Faktoren zu erklären ist. Damit ließe sich dann auch eine Verortung der Ratingskalen auf einem „Fachspezifitätskontinuum" vornehmen (Charalambous und Praetorius 2018; z. B. Indikatoren für die Klassenführung mit geringem, Indikatoren zur kognitiven Aktivierung mit höherem fachspezifischen Varianzanteil).

6.3 Grenzen der Studie und Schlussfolgerungen

Als Limitation der Studie ist zunächst anzuführen, dass die Stichprobengröße von $n = 76$ Lehrpersonen eine geringe Teststärke mit sich bringt (Tab. 3; vgl. auch Schönbrodt und Perugini 2013). Daher sollten die Ergebnisse der vorliegenden Studie mit Daten einer größeren Stichprobe repliziert werden. Diese Stichprobe sollte zudem randomisiert gezogen werden, da die Ergebnisse nur geringe interindividuelle Un-

terschiede in der Unterrichtsqualität (z. B. Klassenführung) und damit eine große Homogenität der Stichprobe nahelegen, die – wie häufig bei freiwilliger Teilnahme – vermutlich eine Positivauswahl an Mathematiklehrpersonen darstellt.

Ein anderes Moment, das bedacht werden sollte, ist die Variation von Unterrichtsmerkmalen über die Zeit (z. B. Praetorius et al. 2014). Obwohl es sich um ein wichtiges psychometrisches Gütemaß der Erfassung von Qualitätsdimensionen handelt, wurde in diesem Beitrag die Stabilität bzw. Instabilität der Rater-Urteile nicht eigens modelliert. Die angesichts der Stichprobengröße stattdessen vorgenommene pragmatische Mittelwertbildung von Rater-Einschätzungen über Messzeitpunkte und Unterrichtsstunden sollte die Unterrichtsqualität im beobachteten Zeitraum aber zumindest angemessen abbilden. Eine ausführliche Bearbeitung dieser Frage ist an anderer Stelle erfolgt (Jentsch et al. 2019).

Abschließend möchten wir auf methodische Herausforderungen unseres Ratingverfahrens eingehen. Dass in dieser Studie die Korrelationen zwischen Kompetenzfacetten der Lehrpersonen und Unterrichtsqualität relativ niedrig ausfallen, könnte damit zusammenhängen, dass als Erhebungsmethode kognitiv anspruchsvolle Live-Ratings zum Einsatz kamen (u. a. König 2015). Zur Bearbeitung dieser Problematik wurde vor der Datenerhebung eine intensive Rater-Schulung vorgenommen, die die Belastung vermindern sollte. Denkbar wäre dennoch, dass eine weitere Komplexitätsreduktion durch Videoanalysen oder kürzere Rating-Intervalle andere Zusammenhangsmuster hervorbringen würde, wie Studien zum CLASS-Instrument vermuten lassen (vgl. Casabianca et al. 2013; Mashburn et al. 2014). Diskussionswürdig dürfte ferner sein, dass die Rater eine Nachbesprechung vorgenommen haben. Diese bot allerdings auch die Chance einer höheren Inhaltsvalidität, da hierdurch Beobachterfehler verringert werden dürften.

Trotz dieser Limitationen scheint die Operationalisierung von Unterrichtsqualität durch hoch-inferente Ratingskalen weiterhin ein gangbarer Weg zu sein, um die Komplexität unterrichtlicher Lehr-Lernprozesse angemessen abzubilden. Unsere Ergebnisse deuten an, dass die Ratingskalen psychometrisch günstige Eigenschaften aufweisen und sich zu vier Faktoren zusammenfassen lassen. In Bezug auf das Ziel unserer Studie deuten wir die vorliegenden Ergebnisse derart, dass durch das Beobachtungsinstrument sowohl generische als auch fachspezifische Qualitätsmerkmale von Mathematikunterricht abgebildet werden, wenngleich die Bündelung dieser Merkmale zu Dimensionen empirisch gesehen anders ausfiel als erwartet (weder nur fachspezifische Ausdifferenzierung der Basisdimensionen, noch additiv fachspezifische Unterrichtsqualität, vgl. Abschn. 1.1).

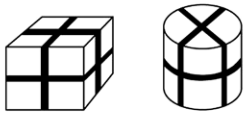
Funding Open Access funding provided by Projekt DEAL.

Anhang

Abb. 1 Beispiel-Item zur Erfassung des mathematischen Wissens (MCK)

3.

Zwei Geschenkschachteln sind mit Bändern verschnürt, wie unten dargestellt. Schachtel A ist ein Würfel mit der Seitenlänge 10 cm. Schachtel B ist ein Zylinder, dessen Höhe und Durchmesser ebenfalls 10 cm betragen.



A B

Für welche der beiden Schachteln wird mehr Band benötigt? _____

Erklären Sie, wie Sie zu der Antwort gekommen sind.

Abb. 2 Beispiel-Item zur Erfassung des mathematikdidaktischen Wissens (MPCK)

12.

Ein Mathematiklehrer möchte einigen Schüler(inne)n zeigen, wie die Formel zum Lösen quadratischer Gleichungen hergeleitet werden kann.

Entscheiden Sie, ob die folgenden Kenntnisse benötigt werden, um die Herleitung zu verstehen.

Kreuzen Sie ein Kästchen pro Zeile an.

		Nötig	Nicht nötig
A.	Wie man lineare Gleichungen löst.	☐	☐
B.	Wie man Gleichungen der Form $x^2 = k$ löst, wobei $k > 0$.	☐	☐
C.	Wie man das Quadrat eines Trinoms ergänzt.	☐	☐
D.	Wie man komplexe Zahlen addiert und subtrahiert.	☐	☐

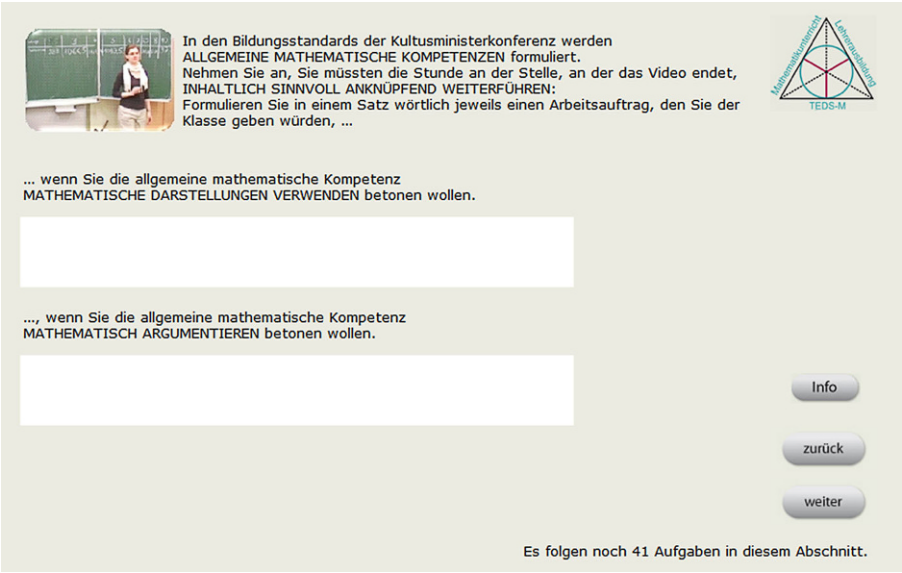


Abb. 3 Beispielitem zur mathematikdidaktischen Wahrnehmungsfähigkeit (M_PID)

Literatur

AERA et al. [American Educational Research Association, American Psychological Association & National Council on Measurement in Education] (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

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

Bell, C., Gitomer, D., McCaffrey, D. & Hamre, B. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2/3), 62–87.

Blömeke, S., & Olsen, R. V. (2019). Consistency of results regarding teacher effects across subjects, school levels, outcomes and countries. *Teaching and Teacher Education*, *77*, 170–182.

Blömeke, S., Kaiser, G., & Lehmann, R. (Hrsg.). (2010). *TEDS-M 2008 – Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster: Waxmann.

Blömeke, S., König, J., Busse, A., Suhl, U., Benthien, J., Döhrmann, M., & Kaiser, G. (2014). Von der Lehrerausbildung in den Beruf: Fachbezogenes Wissen als Voraussetzung einer genauen Wahrnehmung und Analyse von Unterricht. *Zeitschrift für Erziehungswissenschaft*, *17*(3), 509–542.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies. *Zeitschrift für Psychologie*, *223*(1), 3–13.

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

Borg, S. (2006). The distinctive characteristics of foreign language teachers. *Language Teaching Research*, *10*(1), 3–31.

Bortz., J. & Döring, N. (2006). *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler*. Heidelberg: Springer.

Bromme, R. (1995). Was ist „pedagogical content knowledge"? Kritische Anmerkungen zu einem fruchtbaren Forschungsprogramm. *Zeitschrift für Pädagogik, Beiheft*, *33*, 105–115.

Bruder, R., Hefendehl-Hebeker, L., Schmidt-Thieme, B., & Weigand, H.-G. (Hrsg.). (2015). *Handbuch der Mathematikdidaktik*. Berlin: Springer Spektrum.

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *Journal für Mathematikdidaktik*, *39*(2), 257–284.
Buchholtz, N., Kaiser, G., & Blömeke, S. (2014). Die Erhebung mathematikdidaktischen Wissens – Konzeptualisierung einer komplexen Domäne. *Journal für Mathematik-Didaktik*, *35*(1), 101–128.
Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.
Casabianca, J. M., Mccaffrey, D., Gitomer, D., & Bell, C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*(5), 757–783.
Charalambous, C., & Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: setting the ground for understanding instructional quality more comprehensively. *ZDM Mathematics Education*, *50*(3), 355–366.
Cohen, J. (1992) A power primer. *Psychological Bulletin*, *112*(1), 155–159.
Cramer, C. (2012). Empirische Befunde zur Religionslehrerbildung in Baden-Württemberg. *Zeitschrift für Pädagogik und Theologie*, *64*(4), 347–361.
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.
Deci, E. L., & Ryan, R. M. (1985). >*Intrinsic motivation and self-determination in human behavior. Perspectives in social psychology*. New York: Plenum.
Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit: fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht*. Münster: Waxmann.
Eichelmann, A., Narciss, S., Schnaubert, L., & Melis, E. (2012). Typische Fehler bei der Addition und Subtraktion von Brüchen – Ein Review zu empirischen Fehleranalysen. *Journal für Mathematik Didaktik*, *33*(1), 29–57.
Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9.
Geiser, O. (2010). *Datenanalyse mit MPlus: Eine anwendungsorientierte Einführung*. Wiesbaden: VS Verlag für Sozialwissenschaften.
Gruehn, S. (1995). Vereinbarkeit kognitiver und nichtkognitiver Ziele im Unterricht. *Zeitschrift für Pädagogik*, *41*(4), 531–553.
Hartig, J., Frey, A. & Jude, N. (2008). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 135–163). Berlin u.a.: Springer-Verlag.
Heinze, A. & Lindmeier, A. (2020). Die fachdidaktische Perspektive in der Unterrichtsqualitätsforschung: (bisher) ignoriert, implizit enthalten oder nicht relevant? In A.-K. Praetorius, J. Grünkorn & E. Klieme (Hrsg.), *66. Beiheft der Zeitschrift für Pädagogik* (S. 255–268). Weinheim: Beltz.
Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett.
Hiebert, J., Gallimore, R., Garnier, H., Stigler, J., et al. (2003). *Teaching mathematics in seven countries. Results from the TIMSS 1999 video study*. Washington: National Center for Education Statistics.
Hill, H. C., & Chin, M. (2018). Connections between teachers' knowledge of students, instruction, and achievement outcomes. *American Journal of Education*, *55*(5), 1076–1112.
Hill, H. C., Umland, K., Litke, E., & Kapitula, L. R. (2012). Teacher quality and quality teaching: examining the relationship of a teacher assessment to practice. *American Journal of Education*, *118*(4), 489–519.
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.
Jentsch, A., Casale, G., Schlesinger, L., Kaiser, G., König, J., & Blömeke, S. (2019). Variabilität und Generalisierbarkeit von Ratings zur Qualität von Mathematikunterricht zwischen und innerhalb von Unterrichtsstunden. *Unterrichtswissenschaft*. https://doi.org/10.1007/s42010-019-00061-8.
Jordan, A., Krauss, S., Löwen, K., Blum, W., Neubrand, M., Brunner, M., et al. (2008). Aufgaben im COACTIV-Projekt: Zeugnisse des kognitiven Aktivierungspotentials im deutschen Mathematikunterrichts. *Journal für Mathematikdidaktik*, *29*(2), 83–107.
Kaiser, G., Busse, A., Hoth, J., König, J., & Blömeke, S. (2015). About the complexities of video-based assessments: theoretical and methodological approaches to overcoming shortcomings of research on teachers' competence. *International Journal of Science and Mathematics Education*, *13*(3), 369–387.
Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: combining high-quality observations with student surveys and achievement gains*. Seattle: Bill & Melinda Gates Foundation.

Kane, M. T. (2013) Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73.
Kelcey, B., Hill, H. C. & Chin, M. (2019). Teacher mathematical knowledge, instructional quality, and student outcomes: a multilevel quantile mediation analysis. *School Effectiveness and School Improvement*, *30*(7), 1–35.
Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, *49*(3), 568–589.
Kleickmann, T. (2012). *Kognitiv aktivieren und inhaltlich strukturieren im naturwissenschaftlichen Sachunterricht*. Kiel: IPN Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik an der Universität Kiel.
Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, *54*, 222–237.
Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht: Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts „Pythagoras". In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (S. 127–146). Münster: Waxmann.
König, J. (2015). Measuring classroom management expertise (CME) of teachers: a video based assessment approach and statistical results. *Cogent Education*, *2*(1), 991178.
Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York: Holt, Rinehart & Winston.
Kunter, M., & Ewald, S. (2016). Bedingungen und Effekte von Unterricht: Aktuelle Forschungsperspektiven aus der pädagogischen Psychologie. In N. McElvany, W. Bos, H. G. Holtappels, M. M. Gebauer & F. Schwabe (Hrsg.), *Bedingungen und Effekte guten Unterrichts* (S. 9–31). Münster: Waxmann.
Kunter, M., & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften – Ergebnisse des Forschungsprogramms COACTIV* (S. 85–113). Münster: Waxmann.
Learning Mathematics for Teaching Project (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*, 25–47.
Leuders, T. (2001). *Qualität im Mathematikunterricht der Sekundarstufe I und II*. Berlin: Cornelsen Scriptor.
Leuders, T., & Holzäpfel, L. (2011). Kognitive Aktivierung im Mathematikunterricht. *Unterrichtswissenschaft*, *39*(3), 213–230.
Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, *19*(6), 527–537.
Lipowsky, F., Drollinger-Vetter, B., Klieme, E., Pauli, C., & Reusser, K. (2018). Generische und fachdidaktische Dimensionen von Unterrichtsqualität – Zwei Seiten einer Medaille? In M. Martens, K. Rabenstein, K. Bräu, M. Fetzer, H. Gresch, I. Hardy & C. Schelle (Hrsg.), *Konstruktionen von Fachlichkeit* (S. 183–202). Bad Heilbrunn: Klinkhardt.
Lotz, M. (2015). *Kognitive Aktivierung im Leseunterricht der Grundschule. Eine Videostudie zur Gestaltung und Qualität von Leseübungen im ersten Schuljahr*. Wiesbaden: Springer VS.
Maier, U., Kleinknecht, M., Metz, K., & Bohl, T. (2010). Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben. *Beiträge zur Lehrerbildung*, *28*(1), 84–96.
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.
Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist*, *59*(1), 14–19.
Muthén, B. O., & Muthén, L. K. (2010). *Mplus user's guide* (6. Aufl.). Los Angeles: Muthén & Muthén.
Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119.
Pietsch, M., & Tosana, S. (2008). Beurteilereffekte bei der Messung von Unterrichtsqualität. Das Multifacetten-Rasch-Modell und die Generalisierbarkeitstheorie als Methoden der Qualitätssicherung in der externen Evaluation von Schulen. *Zeitschrift für Erziehungswissenschaft*, *11*(3), 430–452.
Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*, 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., Vieluf, S., Saß, S., Bernholt, A., & Klieme, E. (2016). The same in German as in English? Investigating the subject-specifity of teaching quality. *Zeitschrift für Erziehungswissenschaft*, *19*(1), 191–210.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of three basic dimensions. *ZDM Mathematics Education*, *50*(3), 407–426.

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)* (S. 189–205). Frankfurt am Main: GFPF.

Rakoczy, K., Klieme, E., Drollinger-Vetter, B., Lipowsky, F., Pauli, C., & 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 I. M. Prenzel (Hrsg.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (S. 101–120). Münster: Waxmann.

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.7.2. Evanston, Illinois: Northwestern University.

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM Mathematics Education*, *48*(1), 29–40.

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *ZDM Mathematics Education*, *50*(3), 475–490.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2003). Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss. Beschluss vom 04.12.2003. http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2003/2003_12_04-Bildungsstandards-Mathe-Mittleren-SA.pdf. Zugegriffen: 9. Juni 2019.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educational Researcher*, *15*, 4–14.

Taut, S. & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction*, *46*, 45–60.

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.

Vollrath, H.-J., & Roth, J. (2012). *Grundlagen des Mathematikunterrichts in der Sekundarstufe*. Heidelberg: Spektrum.

Weinert, F. E. (1994). Lernen lernen und das eigene Verstehen lernen. In K. Reusser & M. Reusser-Weyeneth (Hrsg.), *Verstehen. Psychologischer Prozess und didaktische Aufgabe* (S. 183–205). Bern: Huber.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.