Herbert, Benjamin; Schweig, Jonathan

Erfassung des Potenzials zur kognitiven Aktivierung über Unterrichtsmaterialien im Mathematikunterricht

Zeitschrift für Erziehungswissenschaft 24 (2021) 4, S. 955-983

Z Erziehungswiss (2021) 24:955–983
https://doi.org/10.1007/s11618-021-01020-9

ZfE

ALLGEMEINER TEIL

Erfassung des Potenzials zur kognitiven Aktivierung über Unterrichtsmaterialien im Mathematikunterricht

Benjamin Herbert · Jonathan Schweig

Eingegangen: 26. Januar 2020 / Überarbeitet: 10. Dezember 2020 / Angenommen: 21. April 2021 /
Online publiziert: 20. Mai 2021

Zusammenfassung Eine zentrale Voraussetzung der Unterrichtsforschung besteht darin, Unterrichtsmerkmale angemessen zu erfassen. Die Einschätzung des (gefilmten) Unterrichts durch geschulte Rater*innen gilt dabei als Königsweg, ist jedoch mit einem hohen organisatorischen und zeitlichen Aufwand verbunden. Im Rahmen dieses Artikels wird ein neu entwickeltes Erhebungsinstrument für das Unterrichtsqualitätsmerkmal des Potenzials zur kognitiven Aktivierung (PKA) vorgestellt. Das Instrument wurde für Mathematikunterricht zum Thema Quadratische Gleichungen entwickelt, basiert auf der gemeinsamen Auswertung aller Unterrichtsmaterialien einer Stunde durch geschulte Rater*innen und erfasst die von der Lehrperson schriftlich in den Unterricht getragenen Potenziale für kognitive Aktivierung. Die Validität der intendierten Interpretation als Indikator für das schriftlich in den Unterricht eingebrachte PKA einer Unterrichtstunde wird über einen argumentationsbasierten Ansatz untersucht und kann über verschiedene Evidenzen gestützt werden: Beispielsweise zeigt eine D-Studie, dass das Instrument bereits von einer einzigen Rater*in zuverlässig erfasst werden kann. Zudem korreliert es substanziell mit einer auf Videoratings basierenden Messung des PKA.

Disclaimer The TALIS Video Study is an OECD project. The development of the Study's instrumentation and data analyses and drafting of international reports were contracted by the OECD to RAND, ETS and DIPF. The authors of this work are solely responsible for its content. The opinions expressed and arguments employed in this work do not necessarily represent the official views of the OECD or its member countries.

B. Herbert (✉)
Lehr- und Lernqualität in Bildungseinrichtungen, DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Rostocker Str. 6, 60323 Frankfurt am Main, Deutschland
E-Mail: herbert@dipf.de

Dr. J. Schweig
RAND Corporation, 1776 Main Street, Santa Monica, 90403, USA
E-Mail: jschweig@rand.org

Schlüsselwörter Artefakt · Messung · Potenzial zur kognitiven Aktivierung · Unterrichtsmaterialien · Unterrichtsqualität

Measuring the potential for cognitive activation via teaching materials in mathematics lessons

Abstract A central challenge of instructional research is to measure teaching characteristics appropriately. The assessment of (filmed) instruction by trained raters is regarded as the best way to achieve this, but is associated with high organizational and time costs. In this paper, we present a newly developed instrument for assessing the teaching quality characteristic potential for cognitive activation (PCA). The instrument was developed for mathematics lessons on the topic of quadratic equations. It is based on the joint evaluation of all instructional artifacts of a lesson (lesson plans, student assignments, assessments, etc.) by trained raters and measures the potential for cognitive activation brought into the classroom by the teacher in writing. The validity of the intended interpretation as an indicator for the written PCA of a lesson is examined by an argument-based approach and can be supported by various pieces of evidence: For example, a decision study shows that the instrument can already be reliably assessed by a single rater. Furthermore, it correlates substantially with a measurement of the PCA based on video ratings.

Keywords Artifact · Measurement · Potential for Cognitive Activation · Teaching Materials · Teaching Quality

1 Einleitung

Das Potenzial zur kognitiven Aktivierung (PKA) einer Unterrichtstunde ist als Unterrichtsqualitätsmerkmal aus der deutschen Unterrichtsforschung nicht mehr wegzudenken. Das Konstrukt bildet eine der drei Qualitätsdimensionen des im deutschsprachigen Raum verbreitetsten Modells für Unterrichtsqualität (vgl. Klieme et al. 2001) und hat sich als Prädiktor für den Lernzuwachs von Schüler*innen erwiesen (Baumert et al. 2010; Klieme et al. 2001; Kunter et al. 2005; Lipowsky et al. 2009). Ungeachtet dessen stellt es die empirische Forschung noch immer vor Herausforderungen, das Konstrukt angemessen zu erfassen. Kunter und Voss (2011) empfehlen, das PKA von Beobachter*innen einschätzen zu lassen. Diese können vorab gezielt geschult werden und gewährleisten dadurch einen methodisch-didaktisch geschulten Blick auf Unterricht (Helmke 2009).

Auswertungen durch Beobachter*innen basieren in der Regel auf gefilmten Unterrichtsstunden, was aufwendige und invasive Erhebungen voraussetzt. Eine wenig genutzte Alternative stellt die Analyse von Unterrichtsmaterialien wie Lehrbuchseiten, Aufgabenblättern, Präsentationen oder Ablaufplänen dar. Diese lassen sich am Ende einer Unterrichtsstunde von der Lehrperson ohne großen Aufwand zusammentragen und können von geschulten Beobachter*innen mit geringerem zeitlichen Aufwand als Unterrichtsvideos ausgewertet werden. Inhaltlich spiegeln sie einen wichtigen Teil des Unterrichtsangebots wider und eignen sich dadurch als Indikator

für das schriftlich in den Unterricht eingebrachte PKA (Kunter et al. 2013; Kunter und Voss 2011; Lipowsky 2015). Ein auf der Auswertung von Unterrichtsmaterialien basierendes Messinstrument für das PKA wurde in Deutschland für den Mathematikunterricht bislang nur in der COACTIV-Studie eingesetzt, deren Erhebungen 2003/2004 stattfanden (Jordan et al. 2008). Mit der *TALIS Video Study* (TVS), wurde kürzlich eine weitere Studie durchgeführt, die entsprechende Auswertungen erlaubt (Opfer et al. 2020)[1]. In dieser wurde u. a. ein innovatives Ratingverfahren durchgeführt: Alle Materialien einer Unterrichtsstunde wurden gebündelt und gemeinsam ausgewertet. Das international eingesetzte Verfahren dient der Beschreibung verschiedener Facetten von *Instruction* (Schweig und Stecher 2020a). Auf Grundlage der um nationale Items erweiterten Erhebung der Studie in Deutschland wird im vorliegenden Aufsatz ein gezielt auf die Messung von PKA ausgerichtetes Instrument vorgestellt und theoriegeleitet validiert.

Dieser Beitrag befasst sich mit der Forschungsfrage, inwieweit sich das PKA einer Unterrichtsstunde im Fach Mathematik auf der Basis des neuen Messansatzes für Unterrichtsmaterialien, der ein breites Spektrum an Merkmalen des PKA abdeckt, erfassen lässt. Das Messinstrument wurde für den Einsatz in der Unterrichtsforschung konzipiert, soll die von der Lehrperson schriftlich in den Unterricht getragenen Potenziale für kognitive Aktivierung erfassen und als Indikator für das geplante PKA einer Unterrichtstunde interpretiert werden. Die Validität dieser intendierten Interpretation und Nutzung wird über einen argumentationsbasierten Ansatz untersucht (Kane 2006, 2013). Dazu werden überprüfbare, die intendierte Interpretation und Nutzung des Instruments stützende Grundannahmen aufgestellt, die sich den Inferenzbereichen Bewertung, Verallgemeinerung und Extrapolation zuordnen lassen. Die einzelnen Grundannahmen werden anhand empirischer Evidenzen und theoretischer Argumente überprüft und anschließend in ihrer Gesamtheit zusammenfassend bewertet.

Auf methodischer Ebene werden vier Schwerpunkte gesetzt: Die interne Struktur des Instruments wird statistisch über eine Faktorenanalyse beurteilt. Anhand einer D-Studie wird untersucht, wie viele Rater*innen benötigt werden, um Unterrichtsmaterialien mit dem entwickelten Vorgehen zuverlässig auszuwerten. Die inhaltliche Breite des Instruments wird über einen Vergleich mit gängigen fragebogen- und videobasierten Operationalisierungen eingeschätzt und die Auswertungsergebnisse des Instruments mit einem videobasierten Messverfahren des PKA korreliert.

2 Theoretischer Rahmen

Insbesondere im deutschsprachigen Raum hat sich zur Analyse der Unterrichtsqualität das *Modell der drei Basisdimensionen guten Unterrichts* durchgesetzt (Praetorius et al. 2020). Das Modell wurde im Jahr 2001 von Klieme, Schümer und Knoll aus den Daten der TIMSS-Videostudie auf Basis von hoch-inferenten Urteilen externer

[1] Die Studie wurde unter dem Namen *Teaching and Learning International Survey (TALIS) Video Study*, kurz TVS, durchgeführt. Im Zuge der Berichtlegung fand eine Umbenennung in *Global Teaching InSights Video Study* statt.

Videobeobachter*innen entwickelt. Es stimmt mit allgemeinen Unterrichtstheorien und etablierten Forschungstraditionen der Unterrichtspsychologie überein und setzt sich aus den drei Dimensionen *strukturierte Klassenführung, unterstützendes Klassenklima* und *kognitive Aktivierung* zusammen. Den Dimensionen werden positive Wirkungen auf die Leistung, aber auch auf motivationale und emotionale Merkmale von Schüler*innen zugeschrieben (Klieme und Rakoczy 2008). Die prognostizierten Wirkungen wurden bereits in vielen empirischen Studien untersucht (z. B. Baumert et al. 2010; Fauth et al. 2014; Lipowsky et al. 2009; zur Übersicht siehe Praetorius et al. 2018).

2.1 Kognitive Aktivierung und das Potenzial zur kognitiven Aktivierung

Das Konstrukt der kognitiven Aktivierung (KA) basiert auf kognitiv-konstruktivistischen Lerntheorien und zielt auf die Förderung eines vertieften Verständnisses der unterrichtlichen Inhalte (Lipowsky 2015). Zur theoretischen Fundierung des explorativ empirisch entstandenen Konstrukts wird in der Unterrichtsforschung weitestgehend einheitlich Bezug auf die konstruktivistischen Lerntheorien nach Vygotsky (1978) und Piaget (1985) genommen (vgl. Reusser 2006). Eine zentrale Gemeinsamkeit der beiden Theorien ist, dass kognitive Aktivität der Lernenden als Erfolgsmerkmal von Unterricht angesehen wird. Diese sorgt bei den Schüler*innen für ein tiefergehendes konzeptuelles Verständnis der unterrichtlichen Inhalte (Hardy et al. 2006; Mayer 2004). Das konzeptuelle Verständnis zeigt sich an der Fähigkeit, gedankliche Verbindungen zwischen Fakten, Prozeduren und Ideen herzustellen (Hiebert und Carpenter 1992). Der Bezug auf kognitiv-konstruktivistische Lerntheorien und das Ziel, ein konzeptuelles Verständnis der mathematischen Inhalte zu fördern, stellen nach dem Verständnis der Autoren die konstituierenden Eigenschaften der KA dar.

Operationalisierungen für empirischen Studien zu entwickeln, die als valide Indikatoren des Konstrukts interpretiert werden können, stellt noch immer eine Herausforderung der Unterrichtsforschung dar: Zum einen können verschiedenste Impulse zu kognitiver Aktivität anregen, weshalb sich das Konstrukt aus vielen Merkmalen zusammensetzt und sich dadurch schwierig erfassen lässt (Lipowsky 2015). Zum anderen lässt sich die tatsächliche kognitive Aktiviertheit der Schüler*innen nur schwer an deren Verhalten erkennen (Lipowsky 2015; Mayer 2004; Renkel 2011). Anstatt Versuche zu unternehmen, sinnbildlich in die Köpfe der Schüler*innen zu schauen, wird daher häufig stellvertretend das „Potenzial der Lerngelegenheit, zielgerichtete kognitive Tätigkeiten der Lernenden anzuregen“ oder kurz das *Potenzial zur kognitiven Aktivierung* erfasst (Kunter und Voss 2011, S. 88; vgl. Kunter und Trautwein 2013; Lipowsky und Bleck 2019). Das PKA ist eine notwendige Voraussetzung und Teilmenge der KA und bietet für empirische Studien den Vorteil, dass es ohne Vermutungen über die kognitive Aktiviertheit der Lernenden eingeschätzt werden kann.

Unterricht kann vielfältige Potenziale beinhalten, Schüler*innen zu kognitiver Aktivität anzuregen. Bereits die Auswahl der Unterrichtsinhalte ist von Bedeutung. Diese sollten so gewählt werden, dass sie den Lernvoraussetzungen der Schüler*innen entsprechen (Baumert et al. 2010) und auf deren Vorwissen aufbauen

(Greeno 2006). Hiebert und Grouws (2007) betonen, dass Schüler*innen ein besseres konzeptuelles Verständnis des Lerngegenstands erlangen, wenn sie Anstrengungen auf sich nehmen müssen, um den Sinn und die inhaltlichen Zusammenhänge von Inhalten zu erarbeiten und zu verstehen. Aufgabenstellungen sollten daher herausfordernd sein. Indem sie kognitive Konflikte provozieren, regen Aufgaben Schüler*innen dazu an, ihr Vorwissen zu reaktivieren und in Frage zu stellen und vertiefend über die Inhalte nachzudenken (Baumert et al. 2010; Lipowsky et al. 2009). Die Gestaltung des Unterrichts sollte die Schüler*innen dazu anregen, die Inhalte zu verarbeiten, zu reflektieren und zu diskutieren; hierzu zählt, Beziehungen zwischen Kernideen sowie deren Implikationen selbst zu erkennen und diese dazu zu nutzen, Lösungsstrategien zu entwickeln, zu vergleichen und Nicht-Routine-Probleme zu lösen (Brophy 2000). Auch Maßnahmen zur metakognitiven Förderung können zur kognitiven Aktivierung beitragen, indem diese zur Selbstreflexion anregen und die Fähigkeit der Schüler*innen zum selbstgesteuerten Lernen unterstützen (Lipowsky und Bleck 2019).

2.2 Unterrichtsmaterialien als Datengrundlage

Um das PKA zu erfassen, werden in der Regel Schüler*innen- oder Lehrer*innen-Fragebögen oder Beobachtungsinstrumente eingesetzt, in denen Merkmale des Unterrichtsangebots und/oder dessen Nutzung durch die Schüler*innen erhoben werden. Eine systematische Zusammenstellung bisheriger Operationalisierungen findet sich bei Praetorius et al. (2018). Ein alternatives Vorgehen kann in der Analyse von Unterrichtsmaterialien, auch *Artefakte* genannt, bestehen.

Die Auswahl der Unterrichtsmaterialien einer Stunde ergibt sich aus der Vorbereitung des Unterrichts durch die Lehrperson, die Materialien erstellt und selektiert. Da Unterrichtsmaterialien den Ablauf und Inhalt des Unterrichts stark beeinflussen, ist es möglich, über sie Rückschlüsse auf den Unterricht zu ziehen. Dies ist jedoch mit Einschränkungen verbunden. Für die Analyse des PKA äußert sich dies darin, dass Potenziale unberücksichtigt bleiben, die sich erst aus dem Unterrichtsgeschehen heraus entwickeln, bspw. durch Diskussionen oder Feedback, und nicht schriftlich festgehalten werden. Darüber hinaus resultiert erst aus der dynamischen Interaktion zwischen den Materialien und der Lehrperson, wie Materialien im Unterricht verwendet werden (Remillard 2005). Dadurch können Diskrepanzen auftreten zwischen den schriftlich von der Lehrperson in den Unterricht eingebrachten Potenzialen und der Nutzung dieser Potenziale. Empirisch zeigte sich dies in der TIMSS-Videostudie: In deutschem Mathematikunterricht wurden komplexe Aufgabenstellungen von Lehrpersonen häufig so kleinschrittig implementiert, dass die resultierenden Teilaufgaben die Schüler*innen nicht länger zur kognitiven Aktivität anregten (Klieme et al. 2001). Wie Unterrichtsmaterialien letztlich verwendet werden, hängt mit den Einstellungen und Fähigkeiten der Lehrperson zusammen (Brown 2009; Charalambous und Hill 2012). Aus einer festgelegten Auswahl an Unterrichtsmaterialien können daher eine Vielzahl verschiedener Unterrichtsabläufe resultieren (Brown 2009; Remillard 2005; Stein et al. 2007). Dabei stellt es für Lehrpersonen eine größere Herausforderung dar, anspruchsvolle Aufgaben im Unterricht umzusetzen als diese zu erstellen (Stein et al. 2007).

Zu erwarten ist daher, dass artefaktbasierte Messungen des PKA mit videobasierten Messungen zusammenhängen, allerdings mit begrenzter Effektstärke. Es kann nicht angenommen werden, dass sich sämtliche Potenziale einer Stunde in den Unterrichtsmaterialien widerspiegeln; abgebildet werden nur von der Lehrperson intendierte und schriftlich vorbereitete oder während der Stunde schriftlich festgehaltene Potenziale. Andererseits können sich in den Materialien Aspekte der Unterrichtsplanung niederschlagen, die im Unterrichtsgeschehen selbst nicht sichtbar sind. Insofern ist eine Überlappung, aber keine Identität zwischen artefakt- und videobasierter Messung zu erwarten. Jedenfalls sind Materialien eine wichtige Voraussetzung dafür, dass sich Schüler*innen vertieft mit den mathematischen Inhalten auseinandersetzen (Hill und Charalambous 2012).

2.3 Messung des PKA über Unterrichtsmaterialien

Von den verschiedenen Typen an Unterrichtsmaterialien, die im Unterricht anfallen, wurden in deutschsprachigen Studien bislang nur Aufgaben als Datengrundlage für die Auswertung des PKA oder einzelne Merkmale des Konstrukts herangezogen. Hervorzuheben sind zwei Publikationen zum Biologieunterricht in der neunten Jahrgangsstufe, bei denen über die Art des kognitiven Prozesses, der für das Lösen einer Aufgabe erforderlich ist, auf das kognitive Anforderungsniveau einer Aufgabe geschlossen wird (Förtsch et al. 2018; Jatzwauk et al. 2008). Das häufige Auftreten kognitiv anspruchsvoller Aufgaben im Unterricht zeigt dabei einen positiven Effekt auf das konzeptuelle Verständnis der Schüler*innen (Förtsch et al. 2018). Die Aufgaben wurden allerdings über Videographien analysiert und nicht als Artefakte erhoben. Artefakte wurden nach dem Wissen der Autoren in Deutschland bislang nur in der COACTIV-Studie systematisch ausgewertet. Als Alternative zu Videos wurden in der Studie zum Mathematikunterricht in der neunten Jahrgangsstufe Aufgaben aus Klassenarbeiten als Indikator für das PKA des Unterrichts herangezogen (vgl. Baumert et al. 2010; Jordan et al. 2006). Über drei Indikatoren (Aufgabentyp, Niveau der mathematischen Argumentation und innermathematische Übersetzung) wurde das kognitive Anforderungsniveau der Aufgaben ausgewertet und in weiteren Analysen als Indikator für das PKA des Unterrichts verwendet. COACTIV ist nach dem Wissen der Autoren bislang die einzige Studie, in der aus artefaktbasierten Items eine Skala zum PKA gebildet wurde.

Weitere Forschungsarbeiten zur Analyse des PKA über Unterrichtsmaterialien stammen aus den USA, wo in den vergangenen 20 Jahren verschiedene Ansätze entwickelt und getestet wurden, um Unterrichtsqualität über Unterrichtsmaterialien zu erfassen. In allen identifizierten Studien wurden dabei auch Merkmale des PKA untersucht.

Das zeitlich gesehen erste Projekt wurde zum Englischunterricht in der Grundschule durchgeführt. In diesem wurden die Aufgaben einer Unterrichtsstunde in Verbindung mit einigen Lösungen der Schüler*innen sowie einer kurzen, von der Lehrperson erstellten, leitfragengestützten Erläuterung zu den Zielen, der Einbettung und der Nutzung der Aufgaben untersucht (Aschbacher 1999; Clare 2000; Clare und Aschbacher 2001). Eingeschätzt wurden Items zum kognitiven Anspruchsniveau der Aufgaben sowie zur Klarheit und Passung zwischen Lernzielen und Aufgaben.

Dabei zeigen sich signifikante Zusammenhänge zwischen der Unterrichtsqualität, wenn sie über Artefakte erfasst wird, und der Unterrichtsqualität, wenn sie über Beobachtungen erfasst wird (Clare et al. 2001); zudem kann ein Teil der Leistung der Schüler*innen über die Qualität der Aufgaben erklärt werden (Matsumura et al. 2002). Vergleichbare Ergebnisse finden sich auch für den Mathematik- und Englischunterricht in der Mittelstufe. In einem zweiten Projekt wurde das Instrument des *Instructional Quality Assessment* (IQA) entwickelt (Junker et al. 2006; zu Mathematik vgl. Boston und Wolf 2006; zu Englisch vgl. Matsumura et al. 2006). Das IQA erfasst unter anderem das kognitive Potenzial der analysierten Aufgaben. Für beide Fächer zeigen sich gute Raterübereinstimmungen und signifikante Zusammenhänge mit dem Leistungszuwachs der Schüler*innen (Matsumura et al. 2006, 2008).

Zwei weitere Studien deuten darauf hin, dass auch andere, im regulären Unterrichtsgeschehen anfallende Unterrichtsmaterialien sinnvolle Datenquellen für das PKA sind. In dem sogenannten *Scoop Notebook Verfahren*, das für die Fächer Mathematik und Naturwissenschaften entwickelt wurde, werden neben Aufgaben, Lösungen der Schüler*innen und Erläuterungen durch die Lehrperson auch sämtliche weiteren Unterrichtsmaterialien wie beispielsweise Ablaufpläne oder Lehrmaterialien analysiert (Stecher et al. 2003, 2005). Die erhobenen Materialien werden unter anderem über Items eingeschätzt, die Merkmale des PKA darstellen (z. B. Erklärungen und Begründungen, Verknüpfungen und Anwendungen sowie kognitive Tiefe). In einer Stichprobe zum Mathematikunterricht der Mittelstufe zeigen sich für die einzelnen Items akzeptable Raterübereinstimmungen und hohe Korrelationen mit Items eines inhaltlich identischen Beobachtungsinstruments (Stecher et al. 2005, 2007). Der Ansatz des Scoop Notebook Verfahrens mündete in der Entwicklung eines Messinstruments mit vergleichbarer Vorgehensweise namens *Quality Assessment in Science Notebook* (QAS), für das in zwei Validierungsstudien von positiven Ergebnissen bezüglich Reliabilität und Zusammenhängen mit einem videobasierten Instrument berichtet wird (Martínez et al. 2012).

Im Anschluss an die Auswertung der durch das Scoop und das QAS Notebook Verfahren erhobenen Daten wurden die Rater*innen jeweils befragt, wie gut die unterschiedlichen Typen von Unterrichtsmaterialien dazu geeignet waren, die einzelnen Qualitätsmerkmale von Unterricht zu beurteilen. Materialien wie Ablaufpläne, Handouts und Arbeitsblätter wurden dabei als gute Indikatoren für Qualitätsbereiche genannt, die Merkmale des PKA darstellen (Martínez et al. 2012; Stecher et al. 2007).

2.4 Entwicklung der Forschungsfrage

Der dargelegte Forschungsstand deutet einheitlich darauf hin, dass Unterrichtsmaterialien Informationen über die Qualität einzelner Merkmale des PKA beinhalten und diese Informationen erfasst werden können (Jordan et al. 2008; Kunter und Voss 2011; Matsumura et al. 2008; Resnick et al. 2006; Stecher et al. 2005, 2007). Auch die Realisierbarkeit einer Skalenbildung zum PKA aus Artefaktratings wurde bereits durch die COACTIV-Studie demonstriert (Baumert et al. 2010; Kunter und Voss 2011). Die angeführten Studien offenbaren jedoch auch einige Schwächen

der bisherigen Vorgehensweisen. So setzt sich die einzige auf Artefaktratings basierende Skala zum PKA aus drei Items zusammen und deckt dadurch nur wenige Merkmale des Konstruktes ab. Zudem wurden in den vorgestellten Studien entweder sehr umfangreiche Datengrundlagen verwendet, die über natürlich auftretende Artefakte hinausgehen und zusätzliche Arbeit seitens der Lehrpersonen erfordern, oder die Datengrundlage beschränkte sich auf Aufgaben, wodurch weitere potenziell vorhandene Artefakte unberücksichtigt blieben. Der Beitrag knüpft an diesen Problemen an und befasst sich mit der Forschungsfrage, inwieweit sich das PKA einer Unterrichtsstunde im Fach Mathematik auf der Basis eines neuen Messansatzes für Unterrichtsmaterialien, der ein breites Spektrum an Merkmalen des PKA abdeckt, erfassen lässt.

Dazu wurde ein Messinstrument entwickelt, dem sämtliche natürlich auftretenden Artefakte einer Stunde als Datenquelle zugrunde liegen und das sich durch eine gemeinsame Bewertung aller Artefakte effizient auswerten lässt. Es soll die von der Lehrperson schriftlich in den Unterricht getragenen Potenziale für kognitive Aktivierung erfassen und in seiner intendierten Verwendung in der Unterrichtsforschung als Indikator für das PKA einer Unterrichtstunde interpretiert werden.

2.5 Validierungsansatz

Die Validität dieser geplanten Interpretation und Nutzung wird über einen argumentationsbasierten Ansatz evaluiert (Kane 2006, 2013). Dass sich dieser ursprünglich für Tests entwickelte Ansatz erfolgreich auf beobachtungsbasierte Instrumente übertragen lässt, zeigen Bell et al. (2012). Die ursprünglich auf Tests ausgerichtete Definition von Validität lässt sich unmittelbar auf andere Messungen übertragen, indem das Wort Test durch Messung ersetzt wird: Validität beschreibt demnach das Ausmaß, in dem empirische Befunde und theoretische Argumente die Interpretationen von Messwerten für die beabsichtigten Verwendungen von Messungen unterstützen (AERA et al. 2014, S. 11; Übersetzung nach Hartig et al. 2020, S. 530). Die Evaluation der Validität wird in zwei Schritten durchgeführt: Zunächst wird ein Interpretationsargument aufgestellt. Dieses besteht aus der bereits in Abschnitt 2.4 dargelegten Beschreibung der geplanten Interpretation und Nutzung des Instruments

Tab. 1 Grundannahmen des Interpretationsarguments

1. Bewertung
1.1 Die Regeln der Bewertung sind angemessen
1.2 Das Verständnis der Rater*innen über die einzelnen Items ist präzise
1.3 Alle Items des Instruments bilden das gleiche Konstrukt ab
1.4 Das PKA lässt sich über das Instrument intersubjektiv nachvollziehbar einschätzen
2. Verallgemeinerung
2.1 Die Stichprobe repräsentiert das Spektrum möglicher Mathematikstunden zum Thema Quadratische Gleichungen
2.2 Das PKA lässt sich über das Instrument intersubjektiv nachvollziehbar einschätzen
3. Extrapolation
3.1 Das Instrument erfasst relevante Inhaltsbereiche des Konstrukts
3.2 Das Instrument erfasst im Kern, wenn auch nicht vollständig, die gleichen Inhaltsbereiche des Konstrukts PKA wie videobasierte Messungen

sowie überprüfbaren Grundannahmen, die der Interpretation und Nutzung inhärent sind und diese stützen (siehe Tab. 1). Laut Kane (2013) lassen sich gängige Schlussfolgerungen aus dem Einsatz eines Instruments fünf Inferenzbereichen zuordnen: Bewertung, Verallgemeinerung, Extrapolation, Implementation und Entscheidungsfindung. Um Transparenz über die Validität möglicher Schlussfolgerungen zu erhalten, empfiehlt Kane, alle für die geplante Interpretation und Nutzung relevanten Inferenzbereiche in den Grundannahmen abzudecken. Da für das entwickelte Instrument weder eine feste Implementation noch die Verwendung als Entscheidungskriterium geplant sind, bleiben diese Bereiche unberücksichtigt. Die einzelnen Grundannahmen werden über empirische Evidenzen und theoretische Argumente überprüft und anschließend im Rahmen des sogenannten Validitätsarguments in ihrer Gesamtheit zusammenfassend bewertet. Die geplante Interpretation und Nutzung eines Instruments kann nur dann als valide betrachtet werden, wenn das Interpretationsargument klar, kohärent, vollständig und plausibel ist (Kane 2013).

Der Inferenzbereich Bewertung befasst sich damit, ob die auszuwertenden Daten angemessen in Zahlenwerte überführt werden. Hierzu werden vier Annahmen untersucht: (1.1) *Die Regeln der Bewertung sind angemessen*. Ausgehend von der bei Bell et al. (2012) genutzten Annahme, dass bei angemessenen Bewertungsregeln die gesamte Breite an Item-Ausprägungen Verwendung findet, werden die deskriptiven Statistiken der Einzelitems überprüft. (1.2) *Das Verständnis der Rater*innen über die einzelnen Items ist präzise*. Evidenzen für diese Annahme bilden das Schulungsdesign und die Zertifizierungsergebnisse der Rater*innen sowie das eingesetzte Ratingverfahren. (1.3) *Alle Items des Instruments bilden das gleiche Konstrukt ab*. Im Kontext der Skalenbildung werden statistische Evidenzen für die interne Struktur der Skala angeführt, indem die Korrelationen aller Items miteinander berechnet und die Eindimensionalität der Skala über eine konfirmatorische Faktorenanalyse getestet werden. (1.4) *Das PKA lässt sich über das Instrument intersubjektiv nachvollziehbar einschätzen*. Auf Item-Ebene wird die Übereinstimmung der zwei Ratings pro Artefakt-Set untersucht. Zudem wird eine D-Studie durchgeführt, um die Aussagekraft der entwickelten Skala bei gegebener Anzahl von Ratern zu testen.

Der Inferenzbereich Verallgemeinerung befasst sich damit, ob Ergebnisse des Instruments auf andere, vergleichbare Anwendungskontexte übertragen werden können. Hierzu werden zwei Annahmen untersucht: (2.1) *Die Stichprobe repräsentiert das Spektrum möglicher Mathematikstunden zum Thema Quadratische Gleichungen*. Evidenzen bilden das Erhebungsdesign sowie die Zusammensetzung der Stichprobe bezüglich der Verteilung der Unterrichtsstunden auf den Verlauf der Unterrichteinheit. (2.2) *Das PKA lässt sich über das Instrument intersubjektiv nachvollziehbar einschätzen.* Die auch für den Inferenzbereich Bewertung aufgestellte Grundannahme und die dazu untersuchten empirischen Evidenzen lassen sich auch hier einordnen, da sie eine Einschätzung dazu erlauben, ob eine Anwendung des Instruments über Rater hinweg verallgemeinerbar ist.

Der Inferenzbereich Extrapolation befasst sich damit, ob von einer Messung mit dem Instrument auf das gewünschte Konstrukt geschlossen werden kann. Hierzu werden ebenfalls zwei Annahmen untersucht: (3.1) *Das Instrument erfasst relevante Inhaltsbereiche des Konstrukts*. Evidenzen hierfür stellen die Auswahl Einzelitems vor dem Hintergrund der theoretischen Aufarbeitung des PKA sowie eine

Gegenüberstellung der letztlichen Item-Zusammensetzung des entwickelten Instruments mit den erfassten Inhaltsbereichen des PKA in gängigen Operationalisierungen. (3.2) *Das Instrument erfasst im Kern, wenn auch nicht vollständig, die gleichen Inhaltsbereiche des Konstrukts PKA wie videobasierte Messungen*. Es konnte wiederholt gezeigt werden, dass Messverfahren für das PKA, die auf verschiedenen Perspektiven basieren (Schüler*innen-, Lehrer*innen- oder Beobachter*innen-Perspektive) – anders als bei Messungen zur Klassenführung – nicht signifikant miteinander korrelieren (Fauth et al. 2020). Zur Analyse des Zusammenhangs zwischen der Skala und einem weiteren Messinstrument kommt daher nur ein Vergleich mit Videoratings in Frage, da diese ebenfalls auf der Perspektive von externen Beobachter*innen basieren. Ein entsprechendes Vorgehen für einzelne Items resultierte bei anderen Studien in signifikanten Zusammenhängen unterschiedlicher Stärke (Martínez et al. 2012; Stecher et al. 2007). Aufgrund des beschriebenen Fokus von Unterrichtsmaterialien auf die Vorbereitung von Unterricht sowie von Videos auf die Durchführung werden mit beiden Instrumenten zu Teilen unterschiedliche Potenziale erfasst und zugleich erfasste Potenziale aus unterschiedlichen Perspektiven auswertet. Daher wird ein schwacher bis mittlerer signifikanter Zusammenhang als positive Evidenz angesehen.

3 Methoden

3.1 Stichprobenbeschreibung

Als Datengrundlage dienen die Erhebungen der TVS in Deutschland. In dieser wurden 50 Klassen der 8. oder 9. Jahrgangsstufe aus verschiedenen Schulformen (84 % Gymnasien, 10 % Gesamt-, 4 % Real- und 2 % Berufsbildende Schulen) zur Unterrichtseinheit *Quadratische Gleichungen* untersucht. Aus der ersten und zweiten Hälfte der Unterrichtseinheit wurde jeweils eine von der Lehrperson ausgewählte Stunde videographiert. Wie Abb. 1 zeigt, verteilen sich die einzelnen Stunden annähernd gleichmäßig auf die Unterrichtseinheit, was die Annahme stützt, dass die Stichprobe das Spektrum möglicher Unterrichtsstunden zum Thema Quadratische Gleichungen repräsentiert.

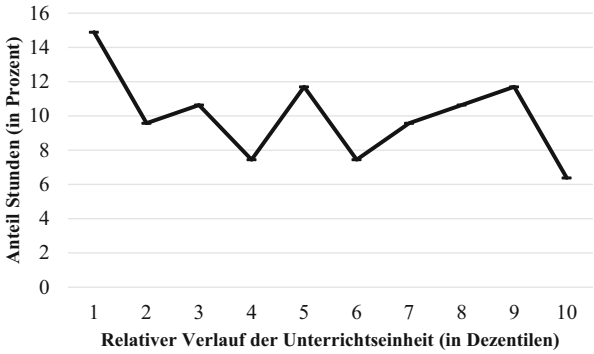

Abb. 1 Relative Position der videographierten Stunden innerhalb der Unterrichtseinheit Quadratische Gleichungen (Lesart: Von den videographierten Unterrichtsstunden lassen sich 14,89 % dem zeitlich gesehen ersten Zehntel der Unterrichtseinheit zum Thema Quadratische Gleichungen zuordnen)

Zudem wurden die Unterrichtsmaterialien der videographierten sowie der jeweiligen Folgestunde von der Lehrperson zu einer Informationseinheit, einem sogenannten Artefakt-Set, gebündelt und den Forschenden übergeben. Es wurde nicht überprüft, ob die Materialien im Unterricht vollständig Verwendung gefunden haben. Die Sets der videographierten Stunden wurden um Screenshots der Tafelanschriebe ergänzt, wodurch für jede der 100 Stunden ein Artefakt-Set vorliegt. Für die Folgestunden liegen nur 88 Artefakt-Sets vor, da in einigen Stunden entweder keine Materialien verwendet oder diese von der Lehrperson nicht eingereicht wurden.

Die Artefakt-Sets setzen sich aus unterschiedlichen Materialien zusammen: In 65 % der Sets befinden sich Aufgabenblätter. Visuelle Materialien wie beispielsweise Tafelanschriebe, PowerPoint-Präsentationen oder Overhead-Folien sind in 62 % der Artefakt-Sets enthalten. Für 43 % der Unterrichtsstunden wurden Ablaufpläne und für 42 % Lehrbuchseiten erfasst und für 4 % kurze formative Tests. In 10 % der Sets fanden sich Materialien, die sich keiner der Kategorien zuordnen lassen. In einem Artefakt-Set können auch mehrere Materialien des gleichen Typs vorhanden sein (z. B. mehrere Aufgabenblätter oder Lehrbuchseiten); die Prozentangaben beziehen sich auf das generelle Vorhandensein eines Material-Typs. Im Durchschnitt enthält ein Artefakt-Set zweieinhalb verschiedene Material-Typen.

3.1.1 Ratingverfahren

Jedes Artefakt-Set wurde als Ganzes ausgewertet. Die Auswertung fand über ein eigens für die TVS entwickeltes und im Rahmen dieser Arbeit ergänztes Kodiersystem statt und wurde von Studierenden der Fachrichtungen Erziehungswissenschaft, Psychologie, Lehramt und Wirtschaftspädagogik durchgeführt. Die sechs Rater*innen erhielten einen eintägigen Workshop zum Thema quadratische Gleichungen und wurden über drei Tage zu dem Kodiersystem geschult. Die Schulung beinhaltete eine umfangreiche Erläuterung der Codes, Beispiele zu allen Ausprägungen, wiederholte Anwendungs- und Übungsmöglichkeiten sowie eine abschließende Prüfung. Im internationalen Vergleich zeigten die deutschen Rater*innen bei der Zertifizierung und den sogenannten Validierungsratings, die im Verlauf der Ratingphase zweimal verdeckt durchgeführt wurden, gute Übereinstimmungswerte mit den Masterratings, die als Musterlösung für die Bewertung galten (Stecher und Schweig im Druck). Darüber hinaus wurde der Ratingprozess von wöchentlichen einstündigen Besprechungen begleitet, in denen Abweichungen in den Ratings und Verständnisschwierigkeiten hinsichtlich der Codes besprochen wurden (Schweig und Stecher 2020b). Dieses Vorgehen stützt die Annahme, dass Rater*innen ein präzises Verständnis über die einzelnen Items aufweisen.

Die Zuteilung der Rater*innen auf Artefakt-Sets wurde zufällig durchgeführt, wobei jede*r Rater*in maximal zwei Sets von einer Lehrperson auswertete. Es fand eine Doppelkodierung statt, sodass zwei unabhängige Ratings für jede Unterrichtsstunde entstanden. Die Anzahl der ausgewerteten Sets war für alle Rater*innen mit 61 bis 64 Sets ähnlich. Das beschriebene Design minimiert Rater-Effekte und sorgt dafür, dass die Codes wie intendiert angewendet werden. Dies stellt eine erste Evidenz für die Intersubjektivität der Ratings dar.

Die Items des Kodiersystems erfassen mathematische Inhaltsbereiche, Struktur- und Unterrichtsqualitätsmerkmale. Aus den letzteren wurden theoriegeleitet Items für die Analyse des PKA ausgewählt. Sie sind ähnlich aufgebaut und werden nach einem einheitlichen Schema ausgewertet. So wird bspw. über das Item V1 erfasst, ob Verknüpfungen zwischen verschiedenen Repräsentationsformen (z. B. zwischen einer Gleichung und ihrer grafischen Darstellung) hergestellt werden. In der Einschätzung des Items wird danach unterschieden, ob keine Verknüpfungen vorliegen und auch nicht hergestellt werden sollen (Rating = 1), eine Verknüpfung bereits auf den Unterrichtsmaterialien vorgegeben ist (Rating = 2) oder Schüler*innen dazu aufgefordert werden, diese selbst herzustellen (Rating = 3). Artefakt-Sets erhalten jeweils das höchste Rating, für das sich ein Beispiel in dem Set finden lässt. Es spielt dabei keine Rolle, wie häufig niedrigere Ausprägungen vorliegen oder auf welchem Material-Typ innerhalb des Sets das Merkmal vorliegt.

3.1.2 Beschreibung der Items

Aus dem Kodiersystem der TVS, das insgesamt breiter und eher deskriptiv angelegt ist (Schweig und Stecher 2020a), beziehen sich sieben Items auf Merkmale des PKA. Drei Items erfassen, ob das konzeptuelle Verständnis der mathematischen Inhalte gefördert wird. Auf innermathematischer Ebene wird eingeschätzt, ob Verknüpfungen zwischen mathematischen Repräsentationsformen vorliegen oder hergestellt werden sollen (V1) und ob von Beispielen auf generelle Eigenschaften des mathematischen Gegenstands geschlossen wird (V2). Auf außermathematischer Ebene wird zudem eingeschätzt, ob mathematische Gegebenheiten mit Echtwelt-Kontexten verknüpft werden (V3). Weitere Items zielen darauf ab, ob eine vertiefte Auseinandersetzung mit den mathematischen Inhalten gefördert wird. Es wird erfasst, ob Schüler*innen dazu aufgefordert werden, ihre Lösungen und Vorgehensweisen zu erklären und zu begründen (V4), und ob es ihnen ermöglicht wird oder sie sogar dazu angehalten werden, verschiedene mathematische Verfahren einzusetzen oder zu vergleichen (V5). Es wird weiter erhoben, ob Technologien genutzt werden, die ein konzeptuelles Verständnis fördern und über die sich Schüler*innen selbstständig vertiefend mit mathematischen Inhalten auseinandersetzen können (V6). Auch der Bereich der Metakognition wurde über ein Item abgedeckt. Dieses erfasst, ob Schüler*innen zur Selbstreflexion angeregt werden (V7).

Da eines der wichtigsten Merkmale des PKA, das kognitive Anspruchsniveau, über die international entwickelten Items der Studie nicht hinreichend abgedeckt wird, wurde das Kodiersystem auf nationaler Ebene um zwei Items ergänzt. Beide Items wurden in ähnlicher Form bereits in der COACTIV-Studie verwendet. Das Item V8 erfasst die Komplexität der Sprache von Aufgabentexten (Sprachlogische Komplexität; Cohors-Fresenborg 1996; Cohors-Fresenborg et al. 2004). Darüber hinaus wird über drei dichotome Items eingeschätzt, ob verschiedene Typen mathematischen Arbeitens in den Artefakt-Sets auftreten (Neubrand 2004; Neubrand et al. 2001). Die Informationen werden zu einer Variable zusammengefasst (V9), wobei der komplexeste vorhandene Typ mathematischen Arbeitens den Wert für das Set bestimmt. Die Ausprägungen geben an, ob (1) keine oder nur technische Aufgaben, (2) rechnerische Modellierungsaufgaben oder (3) begriffliche Modellierungsaufga-

ben vorliegen. Die beschriebene Auswahl der Items ist eine erste Evidenz dafür, dass relevante Inhaltsbereiche des PKA in der Skalenbildung berücksichtigt werden.

3.2 Datengrundlage

Die Datengrundlage bilden je zwei Ratings für 188 Artefakt-Sets, wobei für eine Unterrichtsstunde nur ein Rating vorliegt. Die deskriptiven Statistiken der Items sowie Angaben zur Raterübereinstimmung finden sich in Tab. 2. Da die Daten ordinal skaliert sind, werden die deskriptiven Statistiken auf Ratingebene berichtet.

Die Artefakt-Sets wurden für sieben der neun Items am häufigsten mit der niedrigsten der drei Item-Ausprägungen eingeschätzt; gleichwohl sind die in diesen Items beschriebenen Potenziale zur kognitiven Aktivierung in unterschiedlichen Qualitätsstufen (Ratings von 2 oder 3) in 8 bis 48,5 % der Artefakt-Sets zu finden. Ausnahmen bilden die Items V1 und V5, bei denen die höchste bzw. mittlere Ausprägung überwiegt. Insgesamt finden alle Item-Ausprägungen Verwendung, was eine Evidenz für die angemessene Bewertung der Artefakt-Sets darstellt.

Die exakte prozentuale Übereinstimmung zwischen den beiden Ratings jedes Artefakt-Sets liegt für alle Items zwischen 61,5 % und 90,4 %. Als zusätzliches Maß der Inter-Rater-Reliabilität wurde das gewichtete Kappa berechnet. Die Werte liegen im moderaten bis exzellenten Bereich, wobei der Wert für V2 am niedrigsten ausfällt (Fleiss et al. 2003). Die Raterübereinstimmungen zeigen, dass das Kodiersystem zuverlässig angewendet wurde und liefern damit eine weitere Evidenz für die Intersubjektivität der Auswertung.

Der für korrelationsbasierte Analysen klassisch genutzte Pearson-Korrelationskoeffizient setzt metrisch-skalierte Daten voraus und geht mit Nachteilen für ordinal-

Tab. 2 Deskriptive Statistiken, prozentualer Anteil der Kategorienhäufigkeiten und Angaben zur Raterübereinstimmung der Artefakt-Ratings

	Deskriptive Statistiken		Kategorienhäufigkeiten			Raterübereinstimmung	
Item	*M*	*SD*	1	2	3	%	K_w
V1 – Verknüpfen math. Repräsentationsformen	2,42	0,81	20,5	17,3	62,1	78,6	0,77
V2 – Explizite Muster und Generalisierungen	1,45	0,73	69,3	16,8	13,9	62,0	0,45
V3 – Echtweltbezüge	1,73	0,91	59,2	9,1	31,7	85,0	0,92
V4 – Fragen nach Erklärungen	1,56	0,67	53,6	36,5	9,9	71,7	0,78
V5 – Anwenden mehrerer math. Methoden	1,89	0,63	26,1	58,7	15,2	61,5	0,61
V6 – Verständnisfördernde Technologienutzung	1,29	0,60	78,1	14,4	7,5	83,4	0,76
V7 – Anregen zur Selbstevaluation	1,11	0,40	92,0	4,8	3,2	90,4	0,76
V8 – Sprachlogische Komplexität	1,49	0,62	57,1	36,5	6,4	69,0	0,72
V9 – Aufgabenklassen	1,63	0,72	51,5	34,1	14,4	68,4	0,77

Angaben zu Kategorienhäufigkeiten in Prozent (%). Die Raterübereinstimmung in Prozent (%) gibt den Anteil der exakten Übereinstimmungen zwischen den zwei Ratings der Artefakt-Sets an
K_w Quadratisch gewichtetes Kappa

Tab. 3 Polychorische Korrelationen

Item	V1	V2	V3	V4	V5	V6	V7	V8
V1 – Verknüpfen math. Repräsentationsformen	1,00	–	–	–	–	–	–	–
V2 – Explizite Muster und Generalisierungen	0,02	1,00	–	–	–	–	–	–
V3 – Echtweltbezüge	0,51	–0,10	1,00	–	–	–	–	–
V4 – Fragen nach Erklärungen	0,39	0,37	0,12	1,00	–	–	–	–
V5 – Anwenden mehrerer math. Methoden	0,50	0,01	0,39	0,31	1,00	–	–	–
V6 – Verständnisfördernde Technologienutzung	0,26	0,21	0,21	0,34	0,23	1,00	–	–
V7 – Anregen zur Selbstevaluation	0,02	0,06	0,33	0,28	0,19	0,02	1,00	–
V8 – Sprachlogische Komplexität	0,48	0,02	**0,69**	0,20	0,27	0,10	0,25	1,00
V9 – Aufgabenklassen	0,54	–0,04	**0,92**	0,23	0,34	0,21	0,28	**0,79**

Berechnet auf Rating-Ebene ($N = 375$). Korrelationen >0,6 in fett

skalierte Variablen einher (Bernstein und Teng 1989; Olsson 1979). Die durchgeführten Faktorenanalysen werden daher basierend auf polychorischen Korrelationen berechnet (Holgado-Tello et al. 2008). Die polychorische Korrelationsmatrix der Items ist in Tab. 3 abgebildet.

Einige Korrelationen fallen sehr niedrig oder negativ aus. Sie lassen sich weitestgehend V2 zuordnen, was darauf hindeutet, dass das Item ein anderes Konstrukt abbildet. Da auch die Raterübereinstimmung von V2 nur im moderaten Bereich liegt, wird das Item aus der Skalenbildung ausgeschlossen. Implikationen für die intendierte Interpretation des entwickelten Instruments werden im Zuge der Limitationen der Studie erörtert. Drei Korrelationen fallen zudem besonders hoch aus: Der stärkste Zusammenhang findet sich zwischen den Items V3 und V9. Beide Items korrelieren zudem hoch mit dem Item V8. Inhaltlich lassen sich die Zusammenhänge so erklären, dass es sich bei Aufgaben mit Echtweltbezug meist um Modellierungsaufgaben handelt, deren Aufgabentexte eine höhere sprachlogische Komplexität aufweisen. Auf statistischer Ebene können hohe Korrelationen (Werte >0,8) auf Multikollinearität hindeuten, was dazu führen kann, dass sich die Parameter der im Anschluss berechneten Faktorenanalysen nicht korrekt interpretieren lassen (Field 2009; Tabachnick und Fidell 2007). Als Kennwert für die Diagnose von Multikollinearität wird der SMC (Squared Multiple Correlation) verwendet, weil dieser über polychorische Korrelationen berechnet werden kann. Werte gegen 1 deuten auf Multikollinearität hin. Für die Variablen V3 und V9 fällt der SMC mit 0,88 und 0,90 sehr hoch aus, weshalb von Multikollinearität ausgegangen werden muss.

Auch der Kaiser-Meyer-Olkin-Koeffizient (KMO) und der Bartlett-Test auf Sphärizität wurden auf Basis polychorischer Korrelationen berechnet. Der KMO liegt bei 0,69 und deutet damit auf ein akzeptables Ausmaß an Interkorrelationen zwischen allen acht Items hin (ein häufig genannter Mindestwert für eine Faktorenanalyse liegt bei 0,60; Tabachnick und Fidell 2007). Der Bartlett-Test wird signifikant

($\chi^2(28) = 1615{,}1$, $p < 0{,}001$) und gibt damit an, dass die Items nicht vollständig unkorreliert sind und sich für eine Faktorenanalyse eignen (Field 2009).

3.3 Vorgehen beim Durchführen der Faktorenanalyse

Zur Skalenbildung und dem Überprüfen der internen Struktur des Instruments wird eine hierarchische konfirmatorische Faktorenanalyse durchgeführt. Diese erlaubt es, die Struktur der Daten nach aktuellen methodischen Standards angemessen zu berücksichtigen (McCaffrey et al. 2015). Die ordinale Skalierung der Items und die Auswertung der Artefakt-Sets durch zwei unabhängige Rater*innen werden in die Analysen einbezogen. Die erste Ebene des Modells bilden die zwei manifesten Ratings pro Item. Diese werden auf der zweiten Ebene zu latenten Item-Werten zusammengeführt. Die dritte Ebene bildet die Skala zum PKA, die sich aus den latenten Item-Werten zusammensetzt. Als Schätzer wurde DWLS (Diagonal Weighted Least Squares) mit robuster Schätzung der Standardfehler gewählt. Das robuste DWLS-Verfahren hat sich insbesondere bei ordinal skalierten und nicht normalverteilten Daten sowie kleinen Stichprobengrößen als geeignetes Verfahren herausgestellt (Flora und Curran 2004; Li 2016).

3.4 Vorgehen beim Durchführen der D-Studie

Um zu beurteilen, wie intersubjektiv das vollständige Instrument angewendet werden kann, wird das Framework der Generalisierbarkeitstheorie (G-Studie) genutzt und eine Abhängigkeitsstudie (D-Studie) durchgeführt (Shavelson und Webb 1991). In dieser werden hypothetische Szenarien für die Anzahl an Ausprägungen auf den einzelnen Facetten (in diesem Fall die Anzahl der Rater*innen) sowie die daraus resultierenden Auswirkungen auf die Zuverlässigkeit der Messung geschätzt. Da die interessierenden Merkmale einer D-Studie mehr als zwei Ausprägungen aufweisen sollten (Briesch et al. 2014), wurde eine Teilstichprobe der Artefakt-Sets von allen sechs Rater*innen ausgewertet. Die Stichprobengröße liegt entsprechend der Empfehlung von Shavelson et al. (1989) bei 20 Sets. Um weitere Ursachen für Varianz in den Daten zu reduzieren, handelt es sich jeweils um die Unterrichtsmaterialien der ersten videographierten Unterrichtsstunde. Die D-Studie wird mit einem vollständig gekreuzten Ein-Facetten-Random-Design durchgeführt. Der Skalenwert des PKA bildet das Messobjekt und wird als arithmetisches Mittel berechnet. Rater*innen stellen die Facette des Modells dar. Als Maß der Zuverlässigkeit der Messung wird der G-Koeffizient (ρ^2) angegeben.

3.5 Vorgehen beim Überprüfen erfasster Inhaltsbereiche des Instruments

Auf inhaltlicher Ebene wird überprüft, ob das Konstrukt umfassend erhoben wird. Zum einen basiert die Auswahl aller Einzelitems auf der theoretischen Ausarbeitung des PKA (siehe Abschnitt 2.1). Aufgrund des geplanten Einsatzes des Instruments in der Bildungsforschung wird darüber hinaus untersucht, ob die letztliche Item-Zusammensetzung ähnliche Inhaltsbereiche repräsentiert wie bisher eingesetzte Messinstrumente, die auf Befragungen von Schüler*innen und Lehrpersonen oder

Videobeobachtungen basieren. Für einen Vergleich wird ein Übersichtsbeitrag von Praetorius et al. (2018) herangezogen, in dem bisherige Operationalisierungen der drei Basisdimension zusammengefasst und anhand von Subdimensionen strukturiert werden. Für das PKA wurden sieben Subdimensionen herausgearbeitet, die unterschiedliche Inhaltsbereiche des Konstrukts repräsentierten. Die Items des entwickelten Instruments werden mit diesen verglichen.

3.6 Vorgehen beim Überprüfen des Zusammenhangs mit einer videobasierten Messung

Um zu überprüfen, ob die Skalenwerte des entwickelten Instruments signifikant mit videobasierten Messungen des PKA zusammenhängen, wird eine Korrelationsanalyse durchgeführt. Eine Voraussetzung für die Interpretierbarkeit des Zusammenhangs zwischen den verschiedenen Messverfahren ist, dass sich beide Messungen auf möglichst identische Situationen beziehen. Da das PKA zwischen den einzelnen Stunden einer Lehrperson stark variieren kann (vgl. Praetorius et al. 2014), wird als Vergleichsebene die Unterrichtsstunde gewählt. Datengrundlage sind die 100 Unterrichtsstunden der TVS Deutschland, für die Unterrichtsmaterialien und zugleich Videos vorliegen.

Die Auswertung der Unterrichtsvideos wurde mit dem Kodiersystem der TVS durchgeführt, das Items zu verschiedenen Merkmalen des PKA enthält. Aus sechs Items, die sich inhaltlich stark mit den Items zur Auswertung der Unterrichtsmaterialien überschneiden, wurde eine Skala gebildet. Eine inhaltliche Erläuterung der Items sowie des Ratingverfahrens findet sich im Technical Report der TVS (Bell 2020a, 2020b), die Skalenbildung wird von Köhler et al. (in Vorbereitung) beschrieben[2]. Über die Skala wird erhoben, ob Schüler*innen dazu aufgefordert werden, ihre mathematischen Vorgehens- und Denkweisen zu erläutern und zu begründen. Es wird erfasst, ob das aktive Verhalten der Schüler*innen, beispielsweise Antworten, Kommentare und Rückfragen, darauf hindeutet, dass sich diese vertiefend mit den mathematischen Inhalten auseinandersetzen und ein konzeptuelles Verständnis der Inhalte erlangen. Die Skala deckt darüber hinaus ab, ob mehrere mathematische Vorgehensweisen verwendet und Verknüpfungen zwischen mathematischen Inhalten hergestellt werden. Hervorzuheben ist, dass die beschriebenen Merkmale in beiden Ratingsystemen aus unterschiedlichen Perspektiven beurteilt werden. Im Fokus der Videoratings stehen das Verhalten der Schüler*innen und der Lehrperson und deren Umgang mit den mathematischen Inhalten. Über die Artefakte werden hingegen schriftliche, von der Lehrperson bereitgestellte Potenziale erfasst.

Beide Skalen werden zunächst separat in latenten Modellen geschätzt und anschließend in einem gemeinsamen Modell korreliert, wobei jeweils das robuste DWLS-Schätzverfahren eingesetzt wird. Aufgrund der reduzierten Stichprobengröße auf 100 Stunden ist eine hierarchische Modellierung der Artefakt-Skala nicht möglich. Für jede Unterrichtsstunde wird daher eines der beiden Ratings zufällig ausgewählt.

[2] Eine Übersicht der Items einschließlich deskriptiver Statistiken ist als Online-Anhang verfügbar.

3.7 Verwendete Software

Die Berechnungen wurden mit der Software *R* in der Version 3.6.2 durchgeführt (R Core Team 2014). Polychorische Korrelationen und Strukturgleichungsmodelle wurden mit dem Paket *lavaan* in der Version 0.6-5 geschätzt (Rosseel 2012). Für die G-Studie wurde das Paket *Hemp* in der Version 0.1.0 verwendet (Desjardins und Bulut 2018).

4 Ergebnisse

4.1 Faktorenanalyse

Die standardisierten Faktorladungen und robusten Fit-Werte der durchgeführten Faktorenanalysen sind in Tab. 4 abgebildet. Das als Modell 1 bezeichnete Ausgangsmodell wird mit den acht verbleibenden Items berechnet. Zwei der Faktorladungen liegen im problematischen Bereich: Das Item V6 unterschreitet mit einem Wert von 0,28 selbst liberale Angaben zu Cut-Off-Werten (0,32; Comrey und Lee 1992; Tabachnick und Fidell 2007). Zudem liegt die Faktorladung von Item V9 mit einem Wert von 1,12 außerhalb des inhaltlich sinnvoll interpretierbaren Wertebereichs, was verdeutlicht, dass die Multikollinearität der Items V3 und V9 ein Problem darstellt. Da V3 hinsichtlich der Raterübereinstimmung bessere Werte aufweist als V9, wurde ein zweites Modell ohne V9 berechnet. Durch dieses Vorgehen wurde auch sicher-

Tab. 4 Ergebnisse der hierarchischen Faktorenanalysen

Item/Gütekriterium	Modell 1	Modell 2
V1 – Verknüpfen math. Repräsentationsformen	0,70	0,81
V3 – Echtweltbezüge	0,94	0,83
V4 – Fragen nach Erklärungen	0,33	0,42
V5 – Anwenden mehrerer math. Methoden	0,59	0,69
V6 – Verständnisfördernde Technologienutzung	**0,28**	0,32
V7 – Anregen zur Selbstevaluation	0,32	0,31
V8 – Sprachlogische Komplexität	0,90	0,86
V9 – Aufgabenklassen	**1,12**	–
χ^2	243,99	152,08
Df	96	70
P	<0,001	<0,001
RMSEA	0,091	0,079
TLI	0,969	0,965
CFI	0,975	0,973

Die Modelle wurden als hierarchische konfirmatorische Faktorenanalysen berechnet. Abgetragen sind die standardisierten Faktorladungen der zweiten Ebene und robuste Fit-Indizes
$N = 188$. Faktorladungen $< 0{,}3$ und $> 1{,}0$ in fett
RMSEA Root-Mean-Square-Error-of-Approximation, *TLI* Tucker-Lewis-Index, *CFI* Comparative-Fit-Index

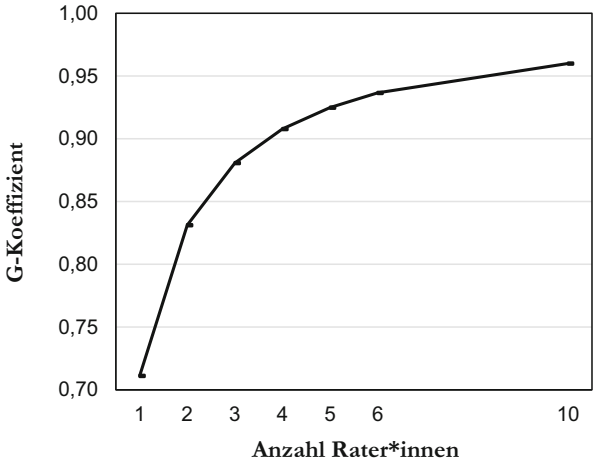

Abb. 2 Veränderung des G-Koeffizienten für die auf Artefaktratings basierende Skala zum PKA in Abhängigkeit der Anzahl an Rater*innen

gestellt, dass die niedrige Faktorladung von V6 kein Resultat der Multikollinearität ist.

Modell 2 besteht aus sieben Items und weist gute bis sehr gute Fit-Werte auf. Dies deutet auf die Passung zwischen den Daten und dem entwickelten Messmodell hin. Als Evidenz dafür, dass alle Items des Instruments das gleiche Konstrukt abbilden, kann angeführt werden, dass die standardisierten Faktorladungen zwischen 0,31 und 0,86 liegen.

4.2 D-Studie

Die durchgeführte D-Studie zeigt, dass der G-Koeffizient für die Einschätzung der Skala durch eine*n Rater*in bei 0,71 liegt (Abb. 2). Er übersteigt damit den Referenzwert von 0,70 aus anderen Studien der Unterrichtsforschung für eine gute Reliabilität (Praetorius et al. 2014), was darauf hindeutet, dass Artefakt-Sets mit dem entwickelten Messverfahren bereits durch eine*r Rater*in zuverlässig ausgewertet werden können. Dies stellt eine weitere positive Evidenz für die Intersubjektivität des Messinstruments dar. Um einen G-Koeffizienten von 0,80 zu überschreiten, muss die Auswertung von zwei Rater*innen durchgeführt werden; für einen Wert über 0,90 von vier Rater*innen.

4.3 Erfasste Inhaltsbereiche

Von den sieben Teilbereichen des PKA, die von Praetorius et al. (2018) herausgearbeitet wurden, werden Inhalte von vieren über die Items des Instruments abgedeckt (vgl. Tab. 5). Was unter dem Teilbereich *Herausfordernde Aufgaben und Fragen* zu verstehen ist, hängt immer vom fachlichen Kontext ab. Items zu diesem Bereich befassen sich mit fachdidaktisch einschlägigen Aspekten von Mathematikunterricht; sie bilden den Kern des Instruments und sind deshalb detailliert abgebildet. Alle weiteren Inhaltsbereiche lassen sich fachunabhängig erfassen und werden durch jeweils ein Items abgebildet. Die beiden ausgeschlossenen Items, V2 und V9, sind dem ers-

Tab. 5 Zuordnung der Items des Instruments zu Inhaltsbereichen des PKA

Inhaltsbereiche des PKA	Items des entwickelten Instruments
Herausfordernde Aufgaben und Fragen	V1 – Verknüpfen math. Repräsentationsformen
	V2 – Explizite Muster und Generalisierungen
	V3 – Echtweltbezüge
	V5 – Anwenden mehrerer math. Methoden
	V8 – Sprachlogische Komplexität
	V9 – Aufgabenklassen
Vorwissen explorieren und aktivieren	–
Denkweisen der Schüler*innen ergründen und sichtbar machen	V4 – Fragen nach Erklärungen
Rezeptives/transmissives Verständnis von Lernen der Lehrperson (negativer Indikator)	V6 – Verständnisfördernde Technologienutzung
Diskursives/ko-konstruierendes Lernen	–
Genetisch-Sokratisches Unterrichten	–
Unterstützen von Metakognition	V7 – Anregen zur Selbstevaluation

Inhaltsbereiche des PKA nach Praetorius et al. (2018, S. 414f.), übersetzt durch die Autoren. Aus dem Instrument ausgeschlossene Items in kursiv

ten Inhaltsbereich zugeordnet, gemeinsam mit vier weiteren Items, weshalb durch deren Ausschluss keine Einschränkungen in der inhaltlichen Breite des Instruments entsteht.

Nicht erfasst werden die Inhaltsbereiche *Vorwissen explorieren und aktivieren, Diskursives/ko-konstruierendes Lernen* und *Genetisch-Sokratisches Unterrichten*. Diese haben gemeinsam, dass sie stark auf Interaktionen zwischen Schüler*innen und der Lehrperson oder Schüler*innen untereinander basieren. Zu erwarten ist, dass sich Potenziale in entsprechenden Kontexten verbal äußern und nicht verschriftlicht werden. Dies trifft insbesondere auf die interaktionsbasierte Methode des genetisch-sokratischen Unterrichtens zu. Für die anderen zwei Bereiche sind hingegen auch schriftliche Impulse denkbar, so könnte bspw. die schriftliche Aufforderung, in Gruppen zu arbeiten, ein Hinweis auf ko-konstruierendes Lernen darstellen und ein an der Tafel festgehaltener Rückblick auf ein Aktivieren des Vorwissens hindeuten.

4.4 Zusammenhang mit einer videobasierten Messung

Um die Grundannahme zu überprüfen, dass das Messinstrument im Kern das gleiche Konstrukt wie eine videobasierte Messung des PKA erfasst, wurde eine Korrelationsanalyse durchgeführt. Die Modelle der Einzelskalen weisen akzeptable bis gute Fit-Werte auf (Artefakt-Skala mit korrelierten Residuen der Items V3 und V8: $\chi^2(13) = 14{,}75$, $p = 0{,}323$, RMSEA = 0,037, TLI = 0,977, CFI = 0,986; Video-Skala: $\chi^2(9) = 11{,}62$, $p = 0{,}235$, RMSEA = 0,054, TLI = 0,956, CFI = 0,973). In dem gemeinsamen Modell korrelieren die Skalen signifikant in einer mittleren Stärke miteinander ($r = 0{,}42$, $p = 0{,}002$), was bedeutet, dass die untersuchten Unterrichtsstunden mit beiden Messverfahren ähnlich eingeschätzt werden. Das Ergebnis entspricht der Erwartung, dass ein schwacher bis mittlerer signifikanter Zusammenhang vorliegt und wird deshalb als Evidenz für die beschriebene Annahme gesehen.

5 Diskussion

5.1 Interpretation der Ergebnisse

Der Beitrag befasst sich mit der Forschungsfrage, inwieweit sich das PKA einer Unterrichtsstunde im Fach Mathematik auf der Basis des vorgestellten Messinstruments erfassen lässt. Das Instrument soll die von der Lehrperson schriftlich in den Unterricht getragenen Potenziale für kognitive Aktivierung erfassen und als Indikator für das PKA einer Unterrichtstunde interpretiert werden. Dadurch soll es sich für den Einsatz in der Unterrichtsforschung eignen. Um die Validität dieser intendierten Interpretation und Nutzung zu evaluieren, wurden Grundannahmen formuliert, die diese stützen und sich den Inferenzbereichen Bewertung, Verallgemeinerung und Extrapolation zuordnen lassen (siehe Tab. 1). Die untersuchen empirischen Evidenzen und theoretischen Argumente für die verschiedenen Grundannahmen werden nachfolgend diskutiert und zusammenfassend bewertet.

5.1.1 Inferenzbereich Bewertung

Für den Inferenzbereich der Bewertung wurden vier Annahmen evaluiert: (1.1) *Die Regeln der Bewertung sind angemessen*. Die deskriptiven Statistiken der Einzelitems zeigen, dass sich für alle Item-Ausprägungen Beispiele in den Daten finden. Dass niedrige Werte tendenziell häufiger vertreten sind, entspricht den umfangreich validierten Ergebnissen der COACTIV Studie (Jordan et al. 2008). Ausgehend von diesem Befund wird die vorliegende Werteverteilungen als Evidenz für angemessene Bewertungsregeln interpretiert. (1.2) *Das Verständnis der Rater*innen über die einzelnen Items ist präzise*. Die umfangreiche Schulung der Rater*innen sowie die anschließenden Zertifizierungsergebnisse deutet auf ein hohes Verständnis der Rater*innen für das eingesetzte Kodiersystem hin. Aufgrund der fortlaufenden wöchentlichen Kalibrierungssitzungen während der Ratingphase ist nicht mit einer Abnahme dieses Verständnisses zu rechnen (Wendler et al. 2019). (1.3) *Alle Items des Instruments bilden das gleiche Konstrukt ab*. Über die Skalenbildung konnte gezeigt werden, dass sich sieben Items mit aktuellen statistischen Methoden sinnvoll zu einer Skala zusammenfassen lassen. Ergebnisse der hierarchischen konfirmatorischen Faktorenanalyse bestätigen die Passung der Daten zum Messmodell sowie die Eindimensionalität des Instruments. (1.4) *Das PKA lässt sich über das Instrument intersubjektiv nachvollziehbar einschätzen*. Die hohen Qualitätsstandards bei der Auswertung der Unterrichtsmaterialien spiegeln sich in den Ergebnissen zur Raterübereinstimmung und der D-Studie: Ausgehend von dem gewählten Grenzwert des G-Koeffizienten ist nur ein*e Rater*in erforderlich, um das PKA einer Unterrichtsstunde mit dem entwickelten Instrument zuverlässig auszuwerten. Dies steht im Einklang mit einem Teil der vorgestellten Studien, laut derer sich verschiedene Merkmale des PKA zuverlässig über Unterrichtsmaterialien auswerten lassen (Jordan et al. 2008; Matsumura et al. 2008). Zudem bestätigt das Ergebnis die intersubjektive Anwendbarkeit des entwickelten Messinstruments.

5.1.2 Inferenzbereich Verallgemeinerung

Der Inferenzbereich der Verallgemeinerung wurde über zwei Annahmen evaluiert: (2.1) *Die Stichprobe repräsentiert das Spektrum möglicher Artefakt-Sets zum Thema Quadratische Gleichungen*. Durch die weitestgehend lehrpersonengelenkte Auswahl der erfassten Unterrichtsstunden, wurden Artefakt-Sets aus allen Phasen der Unterrichtseinheit Quadratische Gleichungen erhoben und ausgewertet. Zudem stammen diese aus unterschiedlichen Schulformen. Obgleich Gymnasien überdurchschnittlich häufig vertreten sind, stellt die untersuchte Stichprobe eine gute Approximation an die unterschiedlichen Artefakt-Sets dar, die aus Stunden zum Thema Quadratische Gleichung resultieren können. Dies wird als Evidenz dafür gesehen, dass sich das Instrument generell auf Unterrichtsmaterialien zu diesem Thema anwenden lässt. Da die eingesetzten Items mit Fachbezug nicht inhaltsspezifisch sind und ein Bezug zum Thema Quadratische Gleichungen nur in der Schulung der Rater*innen hergestellt wurde, werden keine Einschränkungen in der Übertragbarkeit auf andere mathematische Inhalte erwartet; erforderlich wären Anpassungen bei der Schulung. Die Validität einer Nutzung des Instruments für andere Inhalte kann mit den untersuchten Daten jedoch nicht beurteilt werden. Als weitere Evidenzen für die Verallgemeinerbarkeit des Instruments können erneut die Befunde zu Annahme 1.4 angeführt werden. Das Rating-Design ist mit seiner zufälligen Zuordnung von Rater*innen zu Artefakt-Sets und wöchentlichen Kalibrierungssitzungen darauf ausgerichtet, Rater-Effekte zu minimieren. Zudem zeigen die statistischen Kennwerte zur Raterübereinstimmung und der D-Studie, dass Messungen mit dem Instruments über Rater*innen hinweg generalisiert werden können.

5.1.3 Inferenzbereich Extrapolation

Der Inferenzbereich der Extrapolation wurde ebenfalls über zwei Annahmen evaluiert: (3.1) *Das Instrument erfasst relevante Inhaltsbereiche des Konstrukts*. Basierend auf einer sorgfältigen konzeptionellen Ausarbeitung des Konstrukts wurden neun Items zusammengestellt, die verschiedene Merkmale des PKA abbilden. Eines der Items musste wegen geringer Korrelationen mit den anderen Items und ein weiteres wegen Multikollinearität ausgeschlossen werden. Ein Vergleich mit bisherigen Operationalisierungen des PKA entlang der Systematik von Praetorius et al. (2018) zeigt, dass vier der sieben identifizierten Inhaltsbereiche des PKA über das Instrument erfasst werden. Einschränkungen werden insbesondere in Hinblick auf stärker interaktionsbasierte Inhaltsbereiche deutlich, denen häufig kein schriftlicher Impuls zugrunde liegt, der über Unterrichtsmaterialien ausgewertet werden könnte, z. B. im Falle genetisch-sokratischen Unterrichtens. Dies illustriert deutlich den Fokus des Instruments auf schriftliche Potenziale von kognitiver Aktivierung, wie er in Abschnitt 2.2. herausgearbeitet wurde, sowie die Grenzen der Auswertung von Unterrichtsmaterialien. Obgleich potenziell noch weitere Inhaltsbereiche, wie bspw. die Aktivierung des Vorwissens, über Unterrichtsmaterialien eingeschätzt werden könnten, deckt das Instrument bereits mehr Merkmale des PKA ab als das bislang einzige andere artefaktbasierte Instrument, das im Rahmen der COACTIV-Studie entwickelt wurde (Baumert et al. 2010; Jordan et al. 2006). Hinzu kommt, dass bei

der Auswertung des PKA sämtliche natürlich auftretenden Artefakte einer Unterrichtsstunde berücksichtigt wurden. Dies stellt eine umfangreichere Datengrundlage dar als bislang in deutschsprachigen Studien üblich (Förtsch et al. 2018; Jatzwauk et al. 2008). (3.2) *Das Instrument erfasst im Kern, wenn auch nicht vollständig, die gleichen Inhaltsbereiche des Konstrukts PKA wie videobasierte Messungen*. Um diese Grundannahme zu überprüfen, wurde die Korrelation mit einem videobasierten Messverfahren berechnet. Zwischen den beiden Skalen zeigt sich ein signifikanter mittelstarker Zusammenhang. Das Ergebnis passt zu den bisherigen Befunden auf der Ebene einzelner Items, für die sich durchgängig signifikante, in ihrer Stärke jedoch sehr unterschiedliche Korrelationen mit videobasierten Messungen zeigen (Martínez et al. 2012; Stecher et al. 2007). Die Stärke des hier identifizierten Zusammenhangs deutet darauf hin, dass beide Messverfahren sowohl geteilte Aspekte des PKA abbilden als auch Anteile, die nur über eines der beiden Instrumente erfasst werden können (Martínez et al. 2012). Unterschiede ergeben sich daraus, dass Unterrichtsmaterialien für sich genommen nur die von der Lehrperson vorbereiteten und schriftlich in den Unterricht getragenen Potenziale abbilden. Dadurch wird die Unterrichtsplanung stärker einbezogen, während Potenziale, die erst im Unterrichtsverlauf entstehen, unberücksichtigt bleiben. Zudem liefern Artefakte keine Anhaltspunkte dafür, wie Potenziale umgesetzt wurden. Die jeweilige Umsetzung hängt mit den Einstellungen und Fähigkeiten der Lehrperson zusammen, wodurch aus einer festgelegten Auswahl an Unterrichtsmaterialien eine Vielzahl verschiedener Unterrichtsabläufe resultieren können (Brown 2009; Stein et al. 2007). Videobasierte Instrumente fokussieren diese Aspekte, lassen dafür aber häufig den Bereich vorbereiteter Inhalte außer Acht, die im Zentrum einer artefaktbasierten Auswertung stehen. Der signifikante Zusammenhang deutet darauf hin, dass beide Messverfahren trotzdem wesentliche inhaltliche Gemeinsamkeiten teilen und zu ähnlichen Ergebnissen führen. Dies steht im Einklang mit der Erkenntnis von Hill und Charalambous (2012), dass Unterrichtsmaterialien häufig eine Voraussetzung dafür sind, dass sich Schüler*innen vertieft mit den mathematischen Inhalten auseinandersetzen. Der signifikante Zusammenhang wird deshalb als Evidenz dafür gesehen, dass beide Verfahren im Kern das gleiche Konstrukt abbilden.

5.1.4 Zusammenfassung des Validitätsarguments

In Hinblick auf die Validität der geplanten Interpretation und Nutzung des entwickelten Instruments zeichnet sich ein vielversprechendes Bild: Die untersuchten Evidenzen deuten einheitlich darauf hin, dass in den Artefakt-Sets enthaltene Informationen angemessen in Zahlenwerte überführt werden. Angewendet wurde das Instrument bislang nur auf das Thema Quadratische Gleichungen im Mathematikunterricht. Für dieses sind keine Einschränkungen in der Generalisierbarkeit zu erwarten; die Übertragbarkeit auf andere Themen und vor allem Unterrichtsfächer gilt es jedoch zunächst zu überprüfen. Eine Generalisierbarkeit über Rater*innen hinweg ist gegeben. Weitere Evidenzen deuten darauf hin, dass von den Messergebnissen des Instruments angemessen auf das Konstrukt geschlossen werden kann. Trotz des Fokus auf schriftliche Potenziale zur kognitiven Aktivierung werden viele Inhaltsbereiche des Konstrukts erfasst, die auch mit fragebogen- und videobasierten

Messinstrumenten erhoben werden. Obwohl noch Spielraum für inhaltliche Erweiterungen des Instruments besteht, deuten die untersuchten Evidenzen darauf hin, dass die schriftlich in den Unterricht eingebrachten Potenziale zur kognitiven Aktivierung gut abgebildet werden und Hinweise auf das PKA einer Unterrichtsstunde liefern. Zusammenfassend folgt aus dem dargelegten Argumenten, dass die Auswertung der Unterrichtsmaterialien einer Unterrichtsstunde mit dem entwickelten Instrument valide als Indikator für das schriftliche PKA einer Unterrichtsstunde interpretiert werden kann und das Messinstruments für die Verwendung in der empirischen Unterrichtsforschung im Fach Mathematik zum Thema Quadratische Gleichungen geeignet ist.

5.2 Limitationen der Studie

Es konnte gezeigt werden, welches Potenzial in der Auswertung von Unterrichtsmaterialien steckt. Es kristallisieren sich aber auch Limitationen der gewählten Vorgehensweise heraus. Der Ausschluss des Items V2 zeigt, dass selbst innerhalb dessen, was über Unterrichtsmaterialien erfassbar wäre, einzelne theoretisch wichtige Merkmale (z. B. das Ableiten von Mustern und Generalisierungen) nicht angemessen erhoben werden können. Obwohl die verbleibenden sieben Items noch immer vielfältige Potenziale erfassen, die ein konzeptionelles Verständnis der mathematischen Inhalte fördern und Schüler*innen zu kognitiven Tätigkeiten anregen, bleibt ein aus Sicht der Fachdidaktik zentrales Merkmal von kognitiv aktivierendem Mathematikunterricht unberücksichtigt. Qualitative Analysen deuten darauf hin, dass das Identifizieren von Mustern und Generalisierungen einen Ausgangspunkt für Abweichungen zwischen den intendierten und den umgesetzten Potenzialen zur kognitiven Aktivierung darstellen kann (vgl. Klieme et al. 2001; sowie anhand der TVS Deutschland, Schreyer in Vorbereitung): Es konnten verschiedene Vorgehensweisen identifiziert werden, wie Lehrpersonen auf eine Identifikation von Mustern und Generalisierungen in ihrem Unterricht hinführten. Bei einigen zeigt sich, dass das schriftlich oder verbal kommunizierte vorhandene Potenzial zur kognitiven Aktivierung nicht genutzt wird, z. B. im Falle einer sehr kleinschrittigen Bearbeitung oder dem Beantworten der Fragestellung durch die Lehrperson selbst.

Aus der Zusammensetzung der Artefakt-Sets ergibt sich eine weitere Limitation. Es ist anzunehmen, dass Ratings von Artefakt-Sets, die Schulbuchseiten beinhalten, systematisch verzerrt sind. Schulbuchseiten beinhalten meist sehr viele Aufgaben, deren vollständige Bearbeitung von den Schüler*innen innerhalb einer Unterrichtsstunde schon aufgrund der großen Menge unwahrscheinlich ist. Trotzdem wurde in diesen Fällen die gesamte Seite ausgewertet. Für künftige Auswertungen wäre es sinnvoll, zu erfassen, welche Abschnitte oder Aufgaben der einzelnen Materialien in der Unterrichtsstunde tatsächlich verwendet bzw. bearbeitet wurden, um nur diese auszuwerten. Hierbei gilt es, den resultierenden höheren Aufwand für Lehrpersonen mit einer verbesserten Präzision des Instruments abzuwägen.

Eine weitere Problematik der gewählten Vorgehensweise besteht darin, dass die Artefakt-Sets videographierter Unterrichtsstunden um Tafelbilder ergänzt wurden, was für Sets der anderen Stunden nicht möglich war. Es zeigen sich jedoch keine signifikanten Unterschiede der Skalenmittelwerte zwischen den beiden Gruppen

($t(373) = -1{,}66$, $p = 0{,}098$). Darüber hinaus sind Tafelbilder keine Voraussetzung für hohe Skalenwerte.

Über die Problematik von Schulbuchseiten und Tafelbildern hinaus wäre es wünschenswert, Daten darüber zu haben, wie stark das PKA zwischen einzelnen Materialien innerhalb einer Unterrichtsstunde bzw. zwischen verschiedenen Arten von Materialien variiert. Der hier gewählte Ansatz, das gesamte Set an Materialien ganzheitlich zu bewerten, ist zwar effizient, aber Feinanalysen in noch aufwändigeren Validierungsstudien wären hilfreich.

Eine generelle Limitation des vorgestellten Messverfahrens ist dessen Generalisierbarkeit auf andere Unterrichtsthemen und -fächer. Das Anregen zu komplexen Denkprozessen und die vertiefende Auseinandersetzung mit dem Unterrichtsgegenstand – Kernelemente des PKA – sind immer an den Unterrichtsgegenstand und damit an Inhalte geknüpft (Klieme und Rakoczy 2008). Entsprechend bezogen sich einige der eingesetzten Items gezielt auf Mathematikunterricht und die Schulung der Rater*innen fokussierte das Thema Quadratische Gleichungen. Die Übertragbarkeit auf andere Fächer bringt größere Herausforderungen mit sich, da die Bedeutung dessen, was unter potenziell kognitiv aktivierendem Unterricht zu verstehen ist, vom jeweiligen Fach und Unterrichtsthema abhängt. Zudem basiert Unterricht in verschiedenen Fächern unterschiedlich stark auf Unterrichtsmaterialien, weshalb diese voraussichtlich nicht in allen Fächern eine geeignete Datengrundlage für die Auswertung des PKA darstellen.

5.3 Fazit

Es konnte gezeigt werden, dass sich das schriftlich in den Unterricht eingebrachte PKA einer Unterrichtsstunde auf der Basis des vorgestellten Messinstruments erfassen lässt. Dabei wird die hohe Objektivität, die mit der Auswertung durch externe Beobachter*innen einhergeht, kombiniert mit dem geringeren Aufwand für Lehrpersonen und Forschende im Vergleich zu videobasierten Verfahren. Da Unterrichtsmaterialien als Vorbereitung einer Unterrichtsstunde erstellt und ausgewählt werden, ermöglicht es das Instrument, Rückschlüsse auf Unterricht zu ziehen, ohne in diesen einzugreifen. Dies könnte die Reaktivität der Erhebung reduzieren und die Hemmung von Lehrpersonen und Schüler*innen senken, an einer Studie teilzunehmen.

Die Verwendung des entwickelten Instruments als alleiniger Indikator für das PKA ist vor allem dann zu empfehlen, wenn der geplante Inhalt einer Unterrichtsstunde und damit das von der Lehrperson intendierte PKA erfasst werden soll. In Nachfolgestudien gilt es diesbezüglich zu untersuchen, ob Messungen des PKA über das entwickelte Instrument wie zu erwarten mit anderen Konstrukten der Unterrichtsforschung zusammenhängen, z. B. mit der Leistung und Motivation von Schüler*innen (Klieme und Rakoczy 2008) sowie den Einstellungen und Kompetenzen von Lehrpersonen.

Darüber hinaus entstehen mehrere Potenziale für die Unterrichtsqualitätsforschung aus der Kombination des Instruments mit einem videobasierten Messverfahren. Das artefaktbasierte Instrument fokussiert Potenziale zur kognitiven Aktivierung, die über Videos kaum erfasst werden können. Deshalb ermöglicht es

die Kombination beider Messverfahren, das PKA umfassender zu erheben als in empirischen Studien bislang üblich. Ferner könnte untersucht werden, in welchen Fällen Potenziale zur kognitiven Aktivierung, die sich in Unterrichtsmaterialien zeigen, von Lehrpersonen auch so umgesetzt werden, dass sie die kognitive Aktivität der Schüler*innen fördern (vgl. Klieme et al. 2001). Erste Anhaltspunkte liefert der Zusammenhang zwischen den Einstellungen und Fähigkeiten einer Lehrperson und ihrer Vorgehensweise, Unterrichtsmaterialien einzusetzen (Brown 2009; Charalambous und Hill 2012).

Schließlich könnten mit Hilfe des Instruments Rückschlüsse auf einen wichtigen Aspekt der Planungskompetenz einer Lehrperson gezogen werden, nämlich der Kompetenz, für ihren Unterricht potenziell kognitiv aktivierende Materialien bereitzustellen (vgl. Reflexive Kompetenz; Lindmeier 2011). Die Validität dieser Interpretation gilt es jedoch zunächst in Folgestudien zu untersuchen, u. a. im Hinblick auf die Konvergenz der Ergebnisse über verschiedene Unterrichtsstunden derselben Lehrperson hinweg. Im Falle positiver Befunde könnte das Messinstrument auch in der Lehrerforschung und -fortbildung Anwendung finden.

Zusatzmaterial online Zusätzliche Informationen sind in der Online-Version dieses Artikels (https://doi.org/10.1007/s11618-021-01020-9) enthalten.

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt B. Herbert und J. Schweig geben an, dass kein Interessenkonflikt besteht.

Literatur

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform* (CSE Technical Report). Los Angeles: CRESST/University of California.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Yi-Miau, T. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180. https://doi.org/10.3102/0002831209345157.

Bell, C. A. (2020a). The development of the study observation coding system. In OECD (Hrsg.), *Global teaching insights technical report*. Paris: OECD Publishing.

Bell, C. A. (2020b). Rating teaching components and indicators of video observations. In OECD (Hrsg.), *Global teaching insights technical report*. Paris: OECD Publishing.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2/3), 62–87. https://doi.org/10.1080/10627197.2012.715014.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin*, *105*(3), 467–477. https://doi.org/10.1037/0033-2909.105.3.467.

Boston, M., & Wolf, M. K. (2006). *Assessing academic rigor in mathematics instruction: the development of the instructional quality assessment toolkit* (CSE technical report). Los Angeles: CRESST/University of California. https://doi.org/10.1037/e644922011-001.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: a practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008.

Brophy, J. (2000). *Teaching* (Educational practices series, Bd. 1).

Brown, M. W. (2009). The teacher-tool relationship: theorizing the design and use of curriculum materials. In J. T. Remillard, B. A. Herbel-Eisenmann & G. M. Lloyd (Hrsg.), *Mathematics teachers at work* (S. 17–36). New York: Routledge.

Charalambous, C. Y., & Hill, H. C. (2012). Teacher knowledge, curriculum materials, and quality of instruction: unpacking a complex relationship. *Journal of Curriculum Studies*, *44*(4), 443–466. https://doi.org/10.1080/00220272.2011.650215.

Clare, L. (2000). *Using teachers' assignments as an indicator of classroom practice* (CSE technical report, Bd. 532). Los Angeles: CRESST/University of California.

Clare, L., & Aschbacher, P. R. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment*, *7*(1), 39–59. https://doi.org/10.1207/S15326977EA0701_5.

Clare, L., Valdés, R., Pascal, J., & Steinberg, J. R. (2001). *Teachers' assignments as indicators of instructional qualitity in elementary school* (CSE technical report, Bd. 545). Los Angeles: CRESST/University of California.

Cohors-Fresenborg, E. (1996). Mathematik als Werkzeug der Wissensrepräsentation. In G. Kadunz, H. Kautschitsch & G. Ossimitz (Hrsg.), *Trends und Perspektiven. Beiträge zum 7. Internationalen Kärntner Symposium zur „Didaktik der Mathematik" in Klagenfurt* (S. 85–90). Wien: Hölder-Pichler-Tempsky.

Cohors-Fresenborg, E., Sjuts, J., & Sommer, N. (2004). Komplexität von Denkvorgängen und Formalisierung von Wissen. In M. Neubrand (Hrsg.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000* (S. 109–144). Wiesbaden: VS.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2. Aufl.). Hillsdale: Erlbaum.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. London: Chapman and Hall/CRC.

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. https://doi.org/10.1024/1010-0652/a000129.

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). *Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives* (Zeitschrift für Pädagogik: Bd. 66, S. 138–155). Weinheim: Beltz. Beiheft

Field, A. P. (2009). *Discovering statistics using SPSS: and sex, drugs and rock „n" roll* (3. Aufl.). Los Angeles: SAGE.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions*. Hoboken: John Wiley & Sons. https://doi.org/10.1002/0471445428.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466.

Förtsch, C., Werner, S., von Kotzebue, L., & Neuhaus, B. J. (2018). Effects of high-complexity and high-cognitive-level instructional tasks in biology lessons on students' factual and conceptual knowledge. *Research in Science & Technological Education*, *36*(3), 353–374. https://doi.org/10.1080/02635143.2017.1394286.

Greeno, J. G. (2006). Theoretical and practical advances through research on learning. In Y. L. Green, G. Camilli, P. Elmore, A. Skukauskaite & E. Grace (Hrsg.), *Handbook of complementary methods in education research* (S. 795–822). Washington DC: American Educational Research Association.

Grünkorn, J., Klieme, E., Praetorius, A.-K., & Schreyer, P. (Hrsg.). (2020). *Mathematikunterricht im internationalen Vergleich. Ergebnisse aus der TALIS-Videostudie Deutschland*. Frankfurt a. M.: DIPF.

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of „floating and sinking.". *Journal of Educational Psychology*, *98*(2), 307–326. https://doi.org/10.1037/0022-0663.98.2.307.

Hartig, J., Andreas, F., & Jude, N. (2020). Validität von Testwertinterpretationen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 529–545). Berlin: Springer.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (2. Aufl.). Seelze-Velber: Klett.

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Hrsg.), *Handbook of research on mathematics teaching and learning* (S. 65–97). New York: Macmillan.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Hrsg.), *Second handbook of research on mathematics teaching and learning* (1. Aufl., S. 371–404). Charlotte: IAP.

Hill, H. C., & Charalambous, C. Y. (2012). Teacher knowledge, curriculum materials, and quality of instruction: lessons learned and open issues. *Journal of Curriculum Studies*, *44*(4), 559–576. https://doi.org/10.1080/00220272.2012.716978.

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2008). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity*, *44*(1), 153–166. https://doi.org/10.1007/s11135-008-9190-y.

Jatzwauk, P., Rumann, S., & Sandmann, A. (2008). Der Einfluss des Aufgabeneinsatzes im Biologieunterricht auf die Lernleistung der Schüler – Ergebnisse einer Videostudie. *Didaktik der Naturwissenschaften*, *14*, 263–283.

Jordan, A., Ross, N., Krauss, S., Baumert, J., Blum, W., Neubrand, M., Löwen, K., Brunner, M., & Kunter, M. (Hrsg.). (2006). *Klassifikationsschema für Mathematikaufgaben: Dokumentation der Aufgabenkategorisierung im COACTIV-Projekt*. Berlin: Max-Planck-Inst. für Bildungsforschung.

Jordan, A., Krauss, S., Löwen, K., Blum, W., Neubrand, M., Brunner, M., Kunter, M., & Baumert, J. (2008). Aufgaben im COACTIV-Projekt: Zeugnisse des kognitiven Aktivierungspotentials im deutschen Mathematikunterricht. *Journal für Mathematik-Didaktik*, *29*(2), 83–107. https://doi.org/10.1007/BF03339055.

Junker, B., Weisberg, Y., Matsumura, L. C., Crosson, A., Wolf, M. K., Levison, A., & Resnick, L. (2006). *Overview of the instructional quality assessment* (CSE technical report). Los Angeles: CRESST/University of California. https://doi.org/10.1037/e644942011-001.

Kane, M. (2006). Validation. In R. Brennan (Hrsg.), *Educational measurement* (4. Aufl., S. 17–64). Westport: American Council on Education and Praeger.

Kane, M. (2013). The argument-based approach to validation. *School Psycholoy Review*, *42*(4), 448–457. https://doi.org/10.1080/02796015.2013.12087465.

Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, *54*(2), 222–237.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur" und Unterrichtsgestaltung. In Bundesministerium für Bildung und Forschung (Hrsg.), *TIMSS-Impulse für Schule und Unterricht* (S. 43–57). Bonn: Bundesministerium für Bildung und Forschung.

Köhler, C., Herbert, B., & Praetorius, A.-K. (in Vorb.). *Statistical Decisions in Modeling Effects of Teaching: An Example from the TALIS Video Study Germany.*

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

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.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Hrsg.). (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers: results from the COACTIV project*. Boston: Springer US. https://doi.org/10.1007/978-1-4614-5149-5.

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W., Jordan, A., & Neubrand, M. (2005). Der Mathematikunterricht der PISA-Schülerinnen und -Schüler: Schulformunterschiede in der Unterrichtsqualität. *Zeitschrift für Erziehungswissenschaft*, *8*(4), 502–520. https://doi.org/10.1007/s11618-005-0156-8.

Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, *21*(3), 369–387. https://doi.org/10.1037/met0000093.

Lindmeier, A. (2011). *Modeling and measuring knowledge and competencies of teachers. A threefold domain-specific structure model for mathematics*. Münster: Waxmann.
Lipowsky, F. (2015). Unterricht. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 69–105). Berlin: Springer. https://doi.org/10.1007/978-3-642-41291-2_4.
Lipowsky, F., & Bleck, V. (2019). Was wissen wir über guten Unterricht? – Ein Update. In U. Steffens & R. Messner (Hrsg.), *Unterrichtsqualität: Konzepte und Bilanzen gelingenden Lehrens und Lernens* (Bd. 3). Münster: Waxmann.
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. https://doi.org/10.1016/j.learninstruc.2008.11.001.
Martínez, J. F., Borko, H., & Stecher, B. M. (2012). Measuring instructional practice in science using classroom artifacts: lessons learned from two validation studies. *Journal of Research in Science Teaching*, *49*(1), 38–67. https://doi.org/10.1002/tea.20447.
Matsumura, L. C., Patthey-Chavez, G. G., Valdés, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower-and higher-achieving urban schools. *The Elementary School Journal*, *103*(1), 3–25.
Matsumura, L. C., Slater, S. C., Wolf, M. K., Crosson, A., Levison, A., & Peterson, M. (2006). *Using the instructional quality assessment toolkit to investigate the quality of reading comprehension assignments and student work* (CSE Report, Bd. 669). Los Angeles: CRESST/University of California.
Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment*, *13*(4), 267–300. https://doi.org/10.1080/10627190802602541.
Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, *59*(1), 14–19. https://doi.org/10.1037/0003-066X.59.1.14.
McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice*, *34*(2), 34–46. https://doi.org/10.1111/emip.12061.
Neubrand, M. (Hrsg.). (2004). *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland*. Wiesbaden: VS. https://doi.org/10.1007/978-3-322-80661-1.
Neubrand, M., Biehler, R., Blum, W., & Cohors-Fresenborg, E. (2001). Grundlagen der Ergänzung des internationalen PISA-Mathematik-Tests in der deutschen Zusatzerhebung. *ZDM*, *33*(2), 45–59. https://doi.org/10.1007/BF02652739.
Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.
Opfer, V. D., Bell, C. A., Klieme, E., McCaffrey, D. F., Schweig, J. D., & Stecher, B. M. (2020). Understanding and measuring mathematics teaching practice. In OECD (Hrsg.), *OECD global teaching insights: a video study of teaching*. Paris: OECD Publishing.
Piaget, J. (1985). *The equilibration of cognitive structures. The central problem of intellectual development*. Chicago: University of Chicago Press.
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. https://doi.org/10.1016/j.learninstruc.2013.12.002.
Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of three basic dimensions. *ZDM*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4.
Praetorius, A.-K., Klieme, E., Kleickmann, T., Brunner, E., Lindmeier, A., Taut, S., & Charalambous, C. Y. (2020). Towards developing a theory of generic teaching quality: origin, current status, and necessary next steps regarding the three basic dimensions model. *Zeitschrift für Pädagogik*. https://doi.org/10.3262/ZPB2001015.
R Core Team (2014). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing. http://www.R-project.org/
Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research*, *75*(2), 211–246. https://doi.org/10.3102/00346543075002211.
Renkel, A. (2011). Aktives Lernen = gutes Lernen? Reflektion zu einer (zu) einfachen Gleichung. *Unterrichtswissenschaft*, *39*(3), 194–196.
Resnick, L., Matsumura, L. C., & Junker, B. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: a pilot study of the instructional quality assessment* (CSE Technical Report, Bd. 681). Los Angeles: CRESST/University of California.

Reusser, K. (2006). Konstruktivismus – vom epistemologischen Leitbegriff zur Erneuerung der didaktischen Kultur. In M. Baer, M. Fuchs, P. Füglister, K. Reusser & H. Wyss (Hrsg.), *Didaktik auf psychologischer Grundlage. Von Hans Aeblis kognitionspsychologischer Didaktik zur modernen Lehr-und Lernforschung* (S. 151–168). Bern: h.e.p.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v048.i02.

Schreyer, P. (in Vorb.). *Kognitive Aktivierung als Interaktionsmerkmal*. Frankfurt a. M.: Goethe Universität.

Schweig, J. D., & Stecher, B. M. (2020a). Construct and code development for artefacts. In OECD (Hrsg.), *Global teaching insights technical report*. Paris: OECD Publishing.

Schweig, J. D., & Stecher, B. M. (2020b). Rating artefacts. In OECD (Hrsg.), *Global teaching insights technical report*. Paris: OECD Publishing.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: a primer*. Newbury Park: SAGE.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922–932. https://doi.org/10.1037/0003-066X.44.6.922.

Stecher, B. M., & Schweig, J. D. (2021). Artefact score characteristics. In OECD (Hrsg.), *Global teaching insights technical report*. Paris: OECD Publishing.

Stecher, B. M., Alonzo, A., Borko, H., Moncure, S., & McClam, S. (2003). *Artifact packages for measuring instructional practice: a pilot study* (CSE Report, Bd. 615). Los Angeles: CRESST/University of California.

Stecher, B. M., Wood, A. C., Gilbert, M. L., Borko, H., Kuffner, K. L., Arnold, S. C., & Dorman, E. H. (2005). *Using classroom artifacts to measure instructional practices in middle school mathematics: a two-state field test* (CSE report, Bd. 662). Los Angeles: CRESST/University of California.

Stecher, B. M., Borko, H., Kuffner, K. L., Martinez, F., Arnold, S. C., & Barnes, D. (2007). *Using artifacts to describe instruction: lessons learned from studying reform-oriented instruction in middle school mathematics and science* (CSE Technical Report, Bd. 705). Los Angeles: CRESST/University of California.

Stein, M. K., Remillard, J. T., & Smith, M. S. (2007). How curriculum influences student learning. In F. K. Lester (Hrsg.), *Second handbook of research on mathematics teaching and learning* (Bd. 1, S. 319–370). Charlotte: IAP.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5. Aufl.). Boston: Pearson/Allyn & Bacon.

Vygotsky, L. S. (1978). *Mind in society. The development of higher psychological processes*. Cambridge: Harvard University Press.

Wendler, C., Glazer, N., & Cline, F. (2019). *Examining the calibration process for raters of the GRE® general test* (ETS research report series, Bd. 1).