Unterrichtswiss (2020) 48:663–682
https://doi.org/10.1007/s42010-020-00084-6

ALLGEMEINER TEIL

Können Viertklässlerinnen und Viertklässler Unterrichtsqualität valide einschätzen? Ergebnisse zum Fach Deutsch

Ruven Stahns · Svenja Rieser · Anke Hußmann

Eingegangen: 8. Oktober 2018 / Überarbeitet: 31. Juli 2020 / Angenommen: 7. August 2020 / Online publiziert: 31. August 2020

Zusammenfassung Die Bedeutung der Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung für erfolgreiche Lernprozesse von Schüler(inne)n ist Gegenstand aktueller Studien zum Unterricht unterschiedlicher Fächer. In *Large-Scale*-Studien wird der Unterricht häufig aus der Perspektive von Schüler(inne)n eingeschätzt. Die Validität dieser Einschätzungen wird kritisch diskutiert. Im Rahmen der *Internationalen Grundschul-Lese-Untersuchung 2016* (IGLU) wurden Viertklässler(innen) gebeten, ihren Deutschunterricht hinsichtlich einiger Merkmale einzuschätzen, die diesen drei Dimensionen zugeordnet werden können. In der vorliegenden Studie wird zur Prüfung der Validität dieser Einschätzungen untersucht, ob die Kinder in ihren Urteilen zwischen den drei Dimensionen differenzieren und ob sich Zusammenhänge zwischen den Einschätzungen und der Leseleistung der Grundschüler(innen) zeigen. Die Datengrundlage der Untersuchung bilden Angaben der im Rahmen von IGLU 2016 befragten Viertklässler(innen) ($N = 3797$, 192 Klassen) und die Leistungswerte aus dem IGLU-Lesetest. Mittels konfirmatorischer Mehrebenen-Faktorenanalysen wird geprüft, ob sich eine Struktur mit drei Faktoren in den Daten aus den Fragebogenerhebungen findet. Die Zusammenhänge mit der Leistung werden mithilfe von doppelt-latenten Mehrebenen-Regressionsmodellen

Zusatzmaterial online Zusätzliche Informationen sind in der Online-Version dieses Artikels (https://doi.org/10.1007/s42010-020-00084-6) enthalten.

R. Stahns (✉)
FHNW University of Applied Sciences and Arts Northwestern Switzerland, Hofackerstrasse 30, 4132 Muttenz, Schweiz
E-Mail: Ruven.Stahns@fhnw.ch

S. Rieser
Institut für Bildungsforschung in der School of Education, Bergische Universität Wuppertal, Rainer-Gruenter-Straße 21, 42119 Wuppertal, Deutschland

A. Hußmann
Fakultät Rehabilitationswissenschaften, AG Unterrichtsforschung mit dem Schwerpunkt Inklusion, TU Dortmund, Emil-Figge-Straße 50, 44227 Dortmund, Deutschland

geschätzt. Die drei Dimensionen lassen sich mit den Daten abbilden. Zudem hängt die Einschätzung der Klassenführung auf individueller Ebene und auf Klassenebene mit der Leseleistung zusammen. Die Einschätzung der kognitiven Aktivierung hängt auf Klassenebene mit der Leistung zusammen. Für die konstruktive Unterstützung lässt sich auf individueller Ebene ein bedeutsamer Zusammenhang nachweisen. Werden die drei Dimensionen gemeinsam in einem Modell betrachtet, sind die Zusammenhänge mit der Leistung auf Klassenebene nicht mehr nachzuweisen.

Schlüsselwörter Unterrichtsqualität · Validität · Grundschule · Einschätzungen von Grundschüler(inne)n

Are fourth grade students able to rate instructional quality validly? Results from German Language classes

Abstract Recent studies on teaching quality in different subjects often focus on classroom management, constructive support and cognitive activation and their role in fostering successful learning processes. In large scale assessments teaching characteristics are usually assessed by student ratings. The validity of these ratings, however, is often discussed critically. With this study, we aim to contribute to the discussion about the validity of student ratings: First, we tested the ability of fourth graders to differentiate between the three dimensions. Secondly, we examined connections between each dimension and reading achievement as an indicator for predictive validity. In PIRLS 2016 (*Progress in International Reading Literacy Study*), fourth grade students in Germany (3797 students from 192 classes) were asked to rate the teaching quality in their German language classes according to the three dimensions. Using that data and the PIRLS reading achievement test we specified multilevel confirmatory factor analyses. A model with three factors on the individual and the classroom level fitted the data acceptably. In order to analyse predictive validity, doubly-latent multilevel models were also specified. The ratings of classroom management were positively related to reading achievement on both levels. Ratings of cognitive activation were positively associated with reading achievement on the classroom-level, while student support only showed a significant connection to reading achievement on the individual level. When all three dimensions were entered into the model together, no relation could be identified on the classroom level, whereas the connections on the individual level remained stable.

Keywords Instructional quality · Validity · Primary school · Student ratings

1 Einleitung

Ergebnisse von Schulleistungsstudien wie IGLU (*Internationale Grundschul-Lese-Untersuchung*), PISA (*Programme for International Student Assessment*) oder des IQB-Bildungstrends ermöglichen es, Lesekompetenzen von Schüler(inne)n zu vergleichen (vgl. Weis et al. 2016; Bremerich-Vos et al. 2017; Wittig und Weirich 2017). Im Rahmen von IGLU durchgeführte Kontextbefragungen erlauben es, einen

Eindruck von den unterrichtlichen Umständen zu gewinnen, unter denen Grundschüler(innen) Lesekompetenzen erwerben (vgl. Lankes und Carstensen 2007; Stahns et al. 2017). Ergebnisse einiger Unterrichtsstudien belegen, dass die Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung für die Lernergebnisse von Schüler(inne)n bedeutsame Größen sind. Ein diese sogenannten Basisdimensionen der Unterrichtsqualität umfassendes Modell gründet v. a. auf Forschungsbefunden zum Mathematikunterricht in der Sekundarstufe I (vgl. Klieme et al. 2001, 2006; Baumert und Kunter 2011). Auf dieses Modell wird auch in Studien Bezug genommen, in denen die Qualität des Deutschunterrichts in der Sekundarstufe I untersucht wird (vgl. Klieme et al. 2010; Praetorius et al. 2015). Allerdings ist die Relevanz aller drei Dimensionen für das Lernen im Deutschunterricht nicht nachgewiesen. Insbesondere der aus konstruktivistischen Vorstellungen des Lernens abgeleitete (vgl. Mayer 2004) und in Studien zum Mathematikunterricht z. T. belegte positive Zusammenhang zwischen der kognitiven Aktivierung und den Lernergebnissen von Schüler(inne)n kann in Studien zum Deutschunterricht nicht immer repliziert werden (vgl. Klieme et al. 2010). In der jüngeren Vergangenheit ist das skizzierte dreidimensionale Modell von Unterrichtsqualität auch in Studien berücksichtigt worden, in denen die Qualität des Deutschunterrichts in der Grundschule im Fokus steht (vgl. Lotz 2016; Hanisch 2018). Untersuchungen, in denen der Zusammenhang von Klassenführung, konstruktiver Unterstützung und kognitiver Aktivierung sowie Lernergebnissen im Deutschunterricht in der Grundschule untersucht wird, finden sich allerdings kaum. Insofern ist die Relevanz des Modells für den Deutschunterricht in der Grundschule nicht nachgewiesen. Zudem steht in den vorliegenden Studien zum Deutschunterricht in der Grundschule nicht die Einschätzung des Unterrichts aus der Perspektive der Lernenden im Mittelpunkt. Ob Grundschüler(innen) valide urteilen, wenn sie um eine Einschätzung ihres Unterrichts gebeten werden, ist Gegenstand einiger Studien jüngeren Datums, die sich nicht (im Speziellen) auf den Deutschunterricht beziehen (vgl. Fauth et al. 2014a, 2014b; Kloss 2014; Lenske 2016). Diskutiert werden dabei u. a. Auswirkungen eines Halo-Effekts (vgl. Lance et al. 1994). Es wäre möglich, dass Grundschulkinder Unterricht global einschätzen, ohne dass eine Unterscheidung unterschiedlicher Dimensionen von Unterrichtsqualität vorgenommen würde. Zudem sind die prädiktive bzw. kriteriale Validität der Einschätzungen von Grundschüler(inne)n Gegenstand dieser Studien (vgl. Fauth et al. 2014a, 2014b).

Es ist ungeklärt, ob sich die Struktur und die Bedeutung des skizzierten Modells von Unterrichtsqualität für den Deutschunterricht bestätigt, wenn Einschätzungen von Viertklässler(inne)n zugrunde gelegt werden. In die Fragebögen, die die Grundschüler(innen) im Rahmen von IGLU 2016 bearbeitet haben, wurden Items zur Einschätzung der Qualität des Deutschunterrichts aufgenommen. Auf Grundlage vorliegender Forschungsbefunde können diese Items den Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung zugeordnet werden. Im Folgenden wird überprüft, ob ein drei Faktoren umfassendes Modell in den Einschätzungen der Grundschüler(innen) nachzuweisen ist und ob sich zwischen den Einschätzungen und der Leseleistung der Kinder Zusammenhänge nachweisen lassen.

2 Dimensionen von Unterrichtsqualität: Klassenführung, konstruktive Unterstützung und kognitive Aktivierung

Lesekompetenzen sind das Ergebnis von Lernprozessen, die durch ein Zusammenspiel von Variablen auf der Individualebene sowie im familiären, schulischen und unterrichtlichen Kontext von Schüler(inne)n beeinflusst werden (vgl. Lotz 2016; Hußmann et al. 2017). In Studien zum Deutschunterricht, deren Ziel es ist, Unterrichtsmerkmale zu identifizieren, die den Erwerb von Lesekompetenzen unterstützen, wird zunehmend ein Modell von Unterrichtsqualität berücksichtigt, das die Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung umfasst (vgl. Gabriel 2014; Lotz 2016; Stahns et al. 2017). Dieses Modell hat sich in der Unterrichtsforschung im deutschsprachigen Raum in den letzten Jahren etabliert (vgl. Klieme et al. 2001, 2006; Kunter und Trautwein 2013; Praetorius et al. 2018).

Effiziente Klassenführung zeigt sich darin, dass Unterricht störungsfrei verläuft und die Unterrichtszeit mit der Auseinandersetzung mit Fachinhalten zugebracht wird. Operationalisiert wird effiziente Klassenführung z. B., indem eingeschätzt wird, ob im Unterricht Regeln und Routinen erkennbar sind und wie Lehrkräfte mit Störungen umgehen. Ergebnisse von Unterrichtsstudien belegen einen Zusammenhang zwischen effizienter Klassenführung und den Fachleistungen von Schüler(inne)n (vgl. Klieme et al. 2006; Fauth et al. 2014a, 2014b). Merkmale der Dimension konstruktive Unterstützung werden häufig vor dem Hintergrund der Selbstbestimmungstheorie der Motivation bestimmt (vgl. Deci und Ryan 1993). In der Operationalisierung der Dimension orientiert man sich an der Unterstützung von drei Grundbedürfnissen: dem Streben nach Autonomie, Kompetenz und sozialer Eingebundenheit. Ein Merkmal des Verhaltens von Lehrkräften, das dieser Dimension zugeordnet werden kann, ist z. B. der Verzicht auf eine soziale Bezugsnormorientierung im Umgang mit Fehlern. Forschungsergebnisse belegen v. a. einen Zusammenhang zwischen Merkmalen dieser Dimension sowie der Lernmotivation und dem fachlichen Interesse von Schüler(inne)n (vgl. Klieme et al. 2006; Fauth et al. 2014a). Neben der emotionalen Unterstützung ist auch eine Ausrichtung der Dimension auf die Unterstützung kognitiver Prozesse möglich. Bei einer entsprechenden Operationalisierung kann ein Zusammenhang zwischen der konstruktiven Unterstützung und den fachlichen Leistungen von Schüler(inne)n bestehen (vgl. Kunter und Trautwein 2013). Wird das Potenzial zur kognitiven Aktivierung der Lernenden operationalisiert, werden Aspekte des Unterrichts berücksichtigt, die eine eigenständige und verständnisorientierte Auseinandersetzung der Lernenden mit den Lerngegenständen anregen können (z. B. das kognitive Niveau der Fragen von Lehrkräften; vgl. Klieme et al. 2001, 2006). Im Forschungsfokus steht vor allem der Zusammenhang zwischen der kognitiven Aktivierung und den fachlichen Leistungen von Schüler(inne)n (vgl. Klieme et al. 2001, 2006, 2010; Baumert und Kunter 2011). Allerdings ist die Befundlage zum Zusammenhang von kognitiver Aktivierung und den Leistungen von Schüler(inne)n insgesamt uneinheitlich (vgl. Praetorius et al. 2018).

Ergebnisse von Studien zum Deutschunterricht in der Grundschule sowie weiterer Studien zum Grundschulunterricht (vgl. Fauth et al. 2014a, 2014b; Stahns et al.

2017), in denen das skizzierte Modell von Unterrichtsqualität berücksichtigt wurde, bestätigen die Erwartungen hinsichtlich der Zusammenhänge zwischen den drei Basisdimensionen und Lernergebnissen von Grundschüler(inne)n nur zum Teil – in dieser Hinsicht stellt sich die Forschungssituation ähnlich da wie für den Unterricht in der Sekundarstufe I (vgl. Praetorius et al. 2018). Auch zeigen sich diesbezüglich Unterschiede in Abhängigkeit davon, ob der Unterricht aus der Perspektive von Lernenden, Lehrenden oder Externen eingeschätzt wird (vgl. Fauth et al. 2014b). Im Folgenden stehen Einschätzungen von Grundschulkindern im Mittelpunkt.

3 Zur Einschätzung der Unterrichtsqualität durch Grundschüler(innen)

In Unterrichts- und Leistungsstudien werden Einschätzungen des Unterrichts von Schüler(inne)n genutzt, wenn Unterricht beschrieben und/oder ein Zusammenhang zwischen Prozessmerkmalen des Unterrichts und Lernergebnissen hergestellt werden soll (vgl. Prenzel und Lankes 2013). Ein Vorteil dieses Zugangs zur Unterrichtsqualität (vgl. Wagner et al. 2013; Lenske und Helmke 2015) sind die vergleichsweise geringen Kosten bei der Erhebung. Wenn das Studiendesign es erlaubt, können Einschätzungen ganzer Klasse erhoben werden. Darüber hinaus ermöglichen Befragungen von Schüler(inne)n es anders als z. B. die meisten Videostudien, Daten zu gewinnen, die auf Erfahrungen beruhen, die die Lernenden mit einer Lehrkraft über einen längeren Zeitraum gemacht haben (vgl. Wagner et al. 2016). Auch Studienergebnisse zur prädiktiven Validität der Einschätzungen tragen zur Wertschätzung der Befragung von Schüler(inne)n bei (vgl. Clausen 2002; Wagner et al. 2016). Die Ergebnisse dazu stammen vornehmlich aus Studien zum Unterricht in der Sekundarstufe. Hinsichtlich der Einschätzungen von Schüler(inne)n sind aber auch Einschränkungen zu beachten. Schwierigkeiten zeigen sich insbesondere, wenn die kognitive Aktivierung eingeschätzt werden soll und für die Bearbeitung von Items fachliches und/oder fachdidaktisches Wissen vorausgesetzt werden (vgl. Clausen 2002; Klieme et al. 2010). Zudem wird diskutiert, ob Schüler(innen) zwischen unterschiedlichen Dimensionen der Unterrichtsqualität unterscheiden können – befürchtet wird ein Halo-Effekt (vgl. Lance et al. 1994). Diese Einschränkungen sind auch bei der Befragung von Schüler(inne)n im Grundschulalter zu berücksichtigen (vgl. Lenske und Helmke 2015; Lenske 2016).

In der jüngeren Vergangenheit wurden Ergebnisse einiger Studien publiziert, die Aspekte der Validität von Unterrichtseinschätzungen durch Grundschüler(innen) fokussieren. Fauth et al. (2014a) untersuchen anhand von Daten zum Sachunterricht in dritten Klassen, ob sich ein Modell mit den drei oben beschriebenen Qualitätsdimensionen auf der Grundlage von Einschätzungen der Lernenden nachweisen lässt. Zudem prüfen sie, ob Zusammenhänge zwischen den Unterrichtseinschätzungen und der Entwicklung von Leistung sowie Fachinteresse nachzuweisen sind. Die drei Dimensionen lassen sich auf individueller Ebene und auf Klassenebene abbilden. Dieses Ergebnis deutet darauf hin, dass bereits Kinder im Grundschulalter differenzierte Urteile abgeben können, die keinem Halo-Effekt unterliegen. Unter Kontrolle der Beliebtheit der Lehrkräfte sind Zusammenhänge zwischen der kog-

nitiven Aktivierung sowie der konstruktiven Unterstützung und dem Fachinteresse festzustellen. Außerdem zeigt sich auf Klassenebene ein Zusammenhang zwischen der Klassenführung und der Leistungsentwicklung. Zwischen der kognitiven Aktivierung und der konstruktiven Unterstützung sowie der Leistungsentwicklung kann kein Zusammenhang nachgewiesen werden. Weitere Hinweise zur kriterialen Validität der Einschätzungen von Grundschüler(inne)n liefern Stahns et al. (2017). Sie setzen in ihren Analysen von Daten aus IGLU 2016 die aus der Forschung abgeleiteten Dimensionen Klassenführung, Strukturierung, Sozialklima und kognitive Aktivierung voraus und können für den Deutschunterricht auf der Individualebene einen positiven Zusammenhang zwischen Klassenführung, Strukturierung und Sozialklima sowie der Leseleistung der Viertklässler(innen) nachweisen. Zudem belegen sie einen Zusammenhang zwischen allen vier Dimensionen sowie der Lesemotivation der Kinder. Ein Nachweis der Faktorenstruktur erfolgt in der Untersuchung von Stahns et al. (2017) allerdings nicht. Zudem werden lediglich bivariate Korrelationen auf Individualebene berechnet, um den Zusammenhang zwischen den Unterrichtseinschätzungen sowie der Leseleistung und der Lesemotivation zu prüfen. Lenske (2016) untersucht die Validität von Items aus Befragungen von Grundschüler(inne)n, bezieht sich allerdings nicht auf das dreidimensionale Modell von Unterrichtsqualität. Sie zeigt u. a., dass einige Items von den Kindern anders verstanden werden, als das bei der Entwicklung der Items intendiert war. Insofern kommen Zweifel an der inhaltlichen Validität der Items auf, die zuvor von Expert(inn)en als valide eingeschätzt worden sind. Zudem prüft sie die Strukturvalidität der Einschätzungen von Grundschüler(inne)n am Beispiel der Qualitätsdimension Schülerorientierung und unterstützendes Lernklima. Einige Items, die sich zuvor als inhaltlich invalide erwiesen haben, erweisen sich hinsichtlich der Faktorenstruktur als valide. Der Nachweis einer bestimmten Faktorenstruktur genügt also nicht, um die Frage abschließend zu beantworten, ob Einschätzungen von Grundschüler(inne)n valide sind.

Hinsichtlich der Validität der Einschätzungen der Unterrichtsqualität durch Grundschüler(innen) ergibt sich kein eindeutiges Bild. Die Faktorenstruktur des drei Dimensionen umfassenden Modells von Unterrichtsqualität konnte in der Studie von Fauth et al. (2014a) für den Sachunterricht nachgewiesen werden. Auch die Befunde von Lenske (2016) weisen darauf hin, dass Grundschüler(innen) zwischen verschiedenen Konstrukten differenzieren. Lenske (ebd.) zeigt aber auch, dass der Nachweis von Strukturvalidität auch mit Items funktioniert, die inhaltlich invalide sind. Forschungsbefunde zur prädiktiven bzw. kriterialen Validität der Einschätzungen von Lernenden lassen unterschiedliche Ergebnisse für die drei Dimensionen von Unterrichtsqualität, unterschiedliche Zielkriterien und Analyseebenen erwarten (vgl. Praetorius et al. 2018). Zum Beispiel können Fauth et al. (2014a) keinen Zusammenhang zwischen der kognitiven Aktivierung aus Sicht der Grundschüler(innen) und der Leistungsentwicklung nachweisen. Ähnliches deutet sich auf Individualebene in den Daten aus IGLU 2016 an (vgl. Stahns et al. 2017). Bislang liegen keine Untersuchungen zum Deutschunterricht am Ende der Grundschulzeit vor, in denen die Struktur des skizzierten dreidimensionalen Modells unter Berücksichtigung der verschiedenen Analyseebenen auf Grundlage der Einschätzungen von Grundschüler(inne)n und die kriteriale Validität der Einschätzungen von Grundschüler(inne)n geprüft werden.

4 Forschungsfragen

Die Forschungsergebnisse, die zur Etablierung des skizzierten Modells von Unterrichtsqualität geführt haben, stammen v. a. aus Studien zum Mathematikunterricht in der Sekundarstufe I. Daher ist die Übertragbarkeit auf andere Fächer und Klassenstufen zu überprüfen.

Im Folgenden wird untersucht, ob sich ein Modell mit den Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung in Einschätzungen von Grundschüler(inne)n zum Deutschunterricht nachweisen lässt (Fragestellung 1). Erwartet wird, dass ein Modell mit drei Dimensionen sowohl auf der individuellen Ebene als auch auf der Klassenebene besser zu den Daten passt als ein Modell mit nur einer Dimension. Des Weiteren wird untersucht, ob die Einschätzungen des Deutschunterrichts mit den Leseleistungen der Grundschüler(innen) zusammenhängen (Fragestellung 2). Erwartet wird, dass die Einschätzung der Klassenführung mit der Leseleistung zusammenhängt. Ein Zusammenhang zwischen der Leseleistung und der kognitiven Aktivierung lässt sich theoretisch herleiten, allerdings fällt die empirische Befundlage in der bisher vorliegenden Forschung nicht eindeutig aus. Da die Operationalisierung der konstruktiven Unterstützung in der vorliegenden Studie nicht nur auf die Unterstützung emotionaler Bedürfnisse ausgerichtet ist (vgl. Abschn. 5.2), ist ein Zusammenhang mit der Leseleistung denkbar.

5 Methoden

5.1 Daten

Die im Folgenden vorgestellten Ergebnisse basieren auf Daten der Studie IGLU 2016. IGLU ist eine querschnittlich angelegte Schulleistungsstudie der *International Association for the Evaluation of Educational Achievement* (IEA). Die Studie wird in Deutschland in der vierten Klasse durchgeführt. Im Mittelpunkt von IGLU steht ein Lesetest, der es ermöglicht, Leseleistungen von Grundschüler(inne)n international zu vergleichen (vgl. Hußmann et al. 2017). Zur Beantwortung der Forschungsfragen werden in der vorliegenden Untersuchung neben Ergebnissen des Leistungstests Daten aus den Fragebögen herangezogen, die die Viertklässler(innen) und ihre Eltern bearbeiten. Der Original-Datensatz umfasst in Deutschland 3959 Grundschüler(innen) aus 221 Klassen (ebd.). Aus diesem Datensatz wurden für die vorliegende Untersuchung 29 Klassen (162 Kinder) ausgeschlossen, da aus diesen Klassen Angaben von weniger als zehn Kindern zur Verfügung stehen. Damit wäre die Aussagekraft auf Klassenebene stark eingeschränkt (vgl. Lüdtke et al. 2009). Die Stichprobe beläuft sich daher auf 3797 Grundschüler(innen) aus 192 vierten Klassen. Die Kinder sind im Mittel 10,34 Jahre alt ($SD = 0{,}51$), 50,2 % sind männlich, 32,1 % sprechen zuhause manchmal eine andere Sprache als Deutsch. Das lässt auf einen Migrationshintergrund schließen. Der HISEI der Stichprobe variiert zwischen 14,2 und 89,0 Punkten. Im Mittel liegt er bei $M = 54{,}0$ ($SD = 20{,}2$).

5.2 Messinstrumente

Die Unterrichtsqualität wird mittels 17 Items aus den Fragebögen erhoben, die die Viertklässler(innen) im Rahmen der Teilnahme an IGLU bearbeitet haben. Mit den Items werden Aspekte des Unterrichts erfasst, die sich (theoretisch) den drei Dimensionen des skizzierten Modells von Unterrichtsqualität zuordnen lassen. Die Items zur Klassenführung (sechs Items) wurden für IGLU 2016 von Baumert et al. (2009) adaptiert. Diese Items sind z. T. bereits in vorherigen IGLU-Zyklen zum Einsatz gekommen. Auf Grundlage der Skala lässt sich die Zeitnutzung bzw. die Störungsfreiheit des Unterrichts beurteilen. Die Items zur konstruktiven Unterstützung (sechs Items) und zur kognitiven Aktivierung (fünf Items) wurden von Fauth et al. (2014a) adaptiert und ergänzt. Die Items zur konstruktiven Unterstützung fokussieren die Beziehungsqualität. Erfasst werden Aspekte eines freundlichen und wertschätzenden Umgangs im Unterricht. Es werden aber auch Merkmale einbezogen, die sich auf den Umgang der Lehrkräfte mit Beiträgen der Lernenden (v. a. Fehlern) beziehen. Die Items, die zur Einschätzung des Potenzials zur kognitiven Aktivierung dienen, beziehen sich auf Verhalten der Lehrkräfte, das die Lernenden zu einer intensiven Auseinandersetzung mit den Unterrichtsinhalten anregen soll. Hierzu zählt z. B., dass Unterrichtsinhalte an Beispielen verdeutlicht werden oder dass die Lernenden ihre Antworten erklären. Die Viertklässler(innen) wurden gebeten, alle Items auf einer vierstufigen Antwortskala einzuschätzen. Die Items wurden so kodiert, dass höhere Werte für eine positivere Einschätzung der Unterrichtsqualität stehen.

Leseleistungen werden in IGLU mittels eines standardisierten Leistungstests erhoben. Die Viertklässler(innen) bearbeiten jeweils ein Testheft, das einen Sachtext, einen literarischen Text und Aufgaben zu den Texten enthält. Auf Grundlage aller über die Teilnehmer(innen) vorliegenden Daten werden für die Schätzung der Lesekompetenz *Plausible Values* berechnet. Die fünf für jedes Kind berechneten *Plausible Values* werden in dieser Arbeit genutzt. Die mittlere Lesekompetenz der deutschen Stichprobe liegt bei 537 Punkten ($SD = 78$ Punkte) (vgl. Bremerich-Vos et al. 2017). Rund 25 % der Varianz in der Leseleistung (ICC1 = 0,25) können der Klassenebene zugeordnet werden.

Als Kovariaten werden das Geschlecht, der sozioökonomische Status sowie die Mehrsprachigkeit der Kinder verwendet. Die Angaben zum Geschlecht stammen aus dem Fragebogen der Viertklässler(innen). Als Indikator für den sozioökonomischen Status der Familien wird aus den Angaben im Elternfragebogen der höchste *International Socio-Economic Index of Occupational Status* im Haushalt (HISEI) verwendet (vgl. Ganzeboom et al. 1992). Der HISEI kann Werte zwischen 10 und 90 annehmen, wobei höhere Werte auf einen höheren sozioökonomischen Status hinweisen. Da die Rücklaufquote des Elternfragebogens lediglich 72 % beträgt, fällt die Anzahl der fehlenden Werte für diese Variable hoch aus. Ob die Kinder in ihren Familien mehrsprachig leben, wurde auf Grundlage ihrer Angaben zur Nutzung der deutschen Sprache in der Familie ermittelt. Kinder, die angegeben haben, zuhause manchmal eine andere Sprache als Deutsch zu sprechen, werden der Gruppe der Mehrsprachigen zugeordnet.

5.3 Analyseverfahren

Die Daten weisen eine geschachtelte Struktur auf (Grundschüler(innen) in Klassen). Daher sind Mehrebenenanalysen für die Auswertung angemessen (vgl. Raudenbush und Bryk 2010).

Um die erste Forschungsfrage zu beantworten, werden mehrere konfirmatorische Mehrebenen-Faktorenanalysen durchgeführt. Im ersten Schritt wird ein Modell mit drei Faktoren (Klassenführung, konstruktive Unterstützung und kognitive Aktivierung) auf zwei Ebenen (individuelle Ebene und Klassenebene) geschätzt und die Passung zu den Daten bewertet. Zwischen zwei Items, die sich auf die Klassenführung beziehen, wird eine Korrelation der Residuen zugelassen, da diese eine inhaltliche Ähnlichkeit aufweisen.

Zur Beurteilung des Modell-Fits werden die Empfehlungen von Hu und Bentler (1999) herangezogen. Demnach ist der Modell-Fit als angemessen zu beurteilen, wenn der CFI und der TLI Werte von etwa 0,95 oder höher, der RMSEA Werte von etwa 0,06 oder geringer und der SRMR Werte von etwa 0,08 oder geringer aufweisen. Die Werte dienen der Orientierung. Die einzelnen Fit-Indizes werden von einer Vielzahl von Faktoren beeinflusst, sodass eine Entscheidung über die Annahme nicht auf Basis eines einzelnen Wertes getroffen werden sollte (vgl. Hu und Bentler 1999; Schermelleh-Engel et al. 2003). Analog zu Fauth et al. (2014a) wird in einem zweiten Schritt dieses 3/3-Faktoren-Modell mit einem Modell verglichen, in dem alle Items auf beiden Ebenen jeweils auf einen Faktor laden (1/1-Faktor-Modell), sowie einem zweiten Modell, bei dem auf individueller Ebene eine Struktur mit drei Faktoren angenommen und auf Klassenebene ein einzelner Faktor spezifiziert werden. Um zu entscheiden, welches Modell besser zu den Daten passt, werden relative Fit-Indizes verglichen. Fokussiert werden der AIC und der BIC, wobei niedrigere Werte eine bessere Passung implizieren (vgl. Raftery 1993). Außerdem werden die beiden alternativen Modelle mittels des Wald-Chi-Quadrat-Tests mit dem 3/3-Faktoren-Modell verglichen.

Zur Beantwortung der zweiten Fragestellung werden doppelt-latente Mehrebenen-Regressionsmodelle spezifiziert. Dieses Vorgehen ermöglicht es, gleichzeitig den potenziellen Mess- und den Stichprobenfehler zu berücksichtigen (vgl. Lüdtke et al. 2011). Für jede Dimension wird ein latenter Faktor gebildet. Diese Faktoren werden einzeln als Prädiktoren für die Leseleistung auf beiden Ebenen eingesetzt. Anschließend werden in einem Modell alle drei Faktoren gleichzeitig als Prädiktoren eingesetzt. Als manifeste Kovariaten werden auf individueller Ebene das Geschlecht, der sozioökonomische Status sowie die Mehrsprachigkeit der Grundschüler(innen) kontrolliert. Dieselben Variablen werden klassenweise aggregiert und als Prädiktoren auf Klassenebene genutzt. Da in der Stichprobe ein deutlicher Zusammenhang zwischen der Mehrsprachigkeit und dem sozioökonomischen Status besteht ($r_{\text{Individuell}} = -0{,}14^*$; $r_{\text{Klasse}} = -0{,}26^*$), wird eine Korrelation zwischen beiden Konstrukten auf beiden Ebenen zugelassen. Alle Variablen auf individueller Ebene werden am *Groupmean*, alle Variablen auf Klassenebene am *Grandmean* zentriert (vgl. Enders und Tofighi 2007). Aufgrund des querschnittlichen Designs von IGLU 2016 können keine Aussagen über kausale Zusammenhänge gemacht werden.

Die Auswertungen werden in MPlus 8 (vgl. Muthén und Muthén 2017) vorgenommen. Alle Variablen werden vorab standardisiert ($M = 0$; $SD = 1$). Fehlende Werte werden mithilfe des *Full Information Maximum Likelihood* Verfahrens (FIML, Arbuckle 1996) geschätzt.

In der Forschung zur Unterrichtsqualität wird häufig die Klassenebene fokussiert (vgl. Lüdtke et al. 2009). In der vorliegenden Untersuchung wird auch die individuelle Ebene berücksichtigt, weil den individuellen Einschätzungen der Grundschüler(innen) eine inhaltliche Bedeutung zugesprochen werden kann (vgl. Göllner et al. 2018).

6 Ergebnisse

Wie aus den deskriptiven Kennwerten in Tab. 1 hervorgeht, beurteilen die Viertklässler(innen) die konstruktive Unterstützung und die kognitive Aktivierung relativ positiv ($M \geq 2{,}91$; Tab. 1). Die Mittelwerte der Items zur Klassenführung ($M = 2{,}46$ bis 2,57) liegen um den theoretischen Mittelwert von 2,5. Alle Skalen weisen Varianz zwischen den Klassen auf. Diese fällt aber insbesondere bei den Items zur kognitiven Aktivierung für ein Klassenmerkmal sehr gering aus.

Tab. 1 Deskriptive Kennwerte der eingesetzten Fragebogenskalen sowie Beispielitems

Skala	Anzahl der Items	Beispielitem	Item	*M*	*SD*	ICC1	ICC2	Missing (%)
Klassenführung	6	*Im Deutschunterricht ist es laut und unruhig.*	KF1	2,52	0,84	0,20	0,83	14,2
			KF2	2,46	0,92	0,22	0,85	15,1
			KF3	2,72	0,95	0,08	0,65	16,8
			KF4	2,74	0,92	0,12	0,73	16,1
			KF5	3,12	0,97	0,12	0,73	16,0
			KF6	2,77	0,90	0,15	0,77	15,6
Konstruktive Unterstützung	6	*Unsere Deutschlehrerin/unser Deutschlehrer ist auch dann nett zu mir, wenn ich einen Fehler mache.*	KU1	3,57	0,75	0,07	0,58	15,6
			KU2	3,11	0,92	0,08	0,64	16,9
			KU3	3,32	0,84	0,07	0,61	19,7
			KU4	3,35	0,82	0,04	0,44	17,1
			KU5	3,44	0,80	0,07	0,59	17,3
			KU6	3,61	0,73	0,09	0,65	16,9
Kognitive Aktivierung	5	*Unsere Deutschlehrerin/unser Deutschlehrer möchte, dass wir unsere Antworten erklären.*	KA1	3,51	0,76	0,04	0,45	17,6
			KA2	3,28	0,81	0,04	0,47	18,0
			KA3	3,22	0,91	0,07	0,58	18,2
			KA4	3,40	0,76	0,05	0,51	18,7
			KA5	2,93	0,97	0,04	0,47	19,2

Anmerkung: Alle Items wurden so kodiert, dass höhere Werte als positivere Ausprägung des Merkmals zu deuten sind.

6.1 Dimensionalität der Einschätzungen

Zunächst wird geprüft, ob die Daten die Annahme unterstützen, dass die Viertklässler(innen) zwischen den drei Dimensionen unterscheiden können, wenn sie den Deutschunterricht einschätzen (Forschungsfrage 1). Um die Dimensionalität der Einschätzungen zu prüfen, wird zunächst eine konfirmatorische Mehrebenen-Faktorenanalyse mit jeweils drei Faktoren auf individueller Ebene und auf Klassenebene gerechnet. Die Ergebnisse sind in Abb. 1 sowie in Tab. 2 dargestellt.

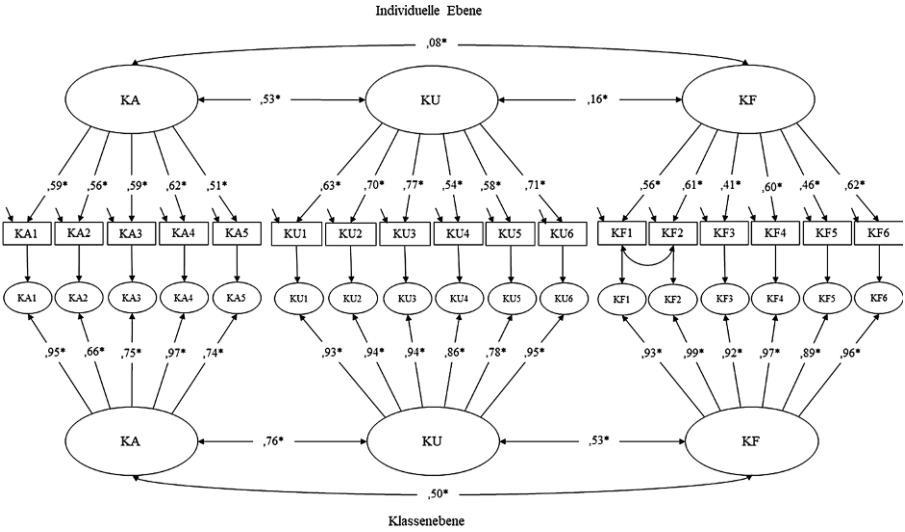

Abb. 1 Ergebnisse der konfirmatorischen Mehrebenen-Faktorenanalyse mit drei Faktoren auf individueller Ebene und auf Klassenebene. *KA* Kognitive Aktivierung; *KU* Konstruktive Unterstützung; *KF* Klassenführung; * $p<0,05$

Tab. 2 Fit-Indizes der konfirmatorischen Mehrebenen-Faktorenanalysen

	Modell 1: 3/3-Faktoren-Modell	Modell 2: 1/1-Faktor-Modell	Modell 3: 1/3-Faktor-Modell
χ^2 (*df*) $p(\chi^2)$	814,231* (231) <0,001	4030,042* (237) <0,001	912,309* (234) <0,001
RMSEA	0,03	0,07	0,03
CFI	0,95	0,68	0,94
TLI	0,94	0,63	0,93
SRMR (within)	0,03	0,10	0,033
SRMR (between)	0,09	0,24	0,214
AIC	137.323,687	140.747,718	137.409,306
BIC	137.592,994	140.999,462	177.669,832
Vergleich mit Modell 1	–	$\chi^2 = 35.994,523$ ($df = 6$) $p(\chi^2) < 0,001$	$\chi^2 = 14.264,972$ ($df = 3$) $p(\chi^2) < 0,001$

Die Betrachtung der Fit-Indizes in Tab. 2 zeigt eine akzeptable Passung der Daten zum Modell. Eine Ausnahme stellt der signifikante Chi²-Wert dar. Dieser kann durch verschiedene Merkmale des Modells (z. B. Stichprobengröße, Anzahl der Parameter im Modell) beeinflusst worden sein. Der Wert alleine führt nicht zur Ablehnung des Modells (vgl. Schermelleh-Engel et al. 2003). Die Faktorladungen variieren auf individueller Ebene zwischen 0,41 und 0,77 und auf Klassenebene zwischen 0,66 und 0,99. Alle Faktorladungen sind signifikant. Die latenten Faktoren auf jeder Ebene korrelieren signifikant miteinander. Die Korrelation ist jedoch nicht so hoch, dass die Trennbarkeit der Faktoren angezweifelt werden müsste. Sowohl auf individueller Ebene als auch auf Klassenebene ist die Korrelation zwischen den Faktoren kognitive Aktivierung und konstruktive Unterstützung am deutlichsten ausgeprägt (vgl. Abb. 1).

Im nächsten Schritt wird dieses Modell mit zwei weiteren Faktormodellen verglichen: einem Modell, in dem alle Items auf beiden Ebenen auf einen einzigen Faktor laden (Modell 2, Tab. 2), und einem Modell, in dem auf individueller Ebene eine Struktur mit drei Faktoren angenommen wird, während auf Klassenebene ein einzelner Faktor spezifiziert wird (Modell 3, Tab. 2). Der Vergleich von AIC und BIC zeigt für beide Modelle höhere Werte als für das 3/3-Faktoren-Modell. Das deutet auf eine bessere Passung dieses Modells hin. Auch der Wald-Chi-Quadrat-Test indiziert eine signifikant bessere Passung des 3/3-Faktoren-Modells (vgl. Tab. 2).

Die Dimensionen kognitive Aktivierung und konstruktive Unterstützung korrelieren mit $r_{Individuell} = 0{,}53$ und $r_{Klasse} = 0{,}76$ (vgl. Abb. 1). Eine Korrelation ähnlicher Stärke zwischen diesen beiden Dimensionen berichten auch Fauth et al. (2014a, S. 6: $r_{Individuell} = 0{,}67$; $r_{Klasse} = 0{,}70$). Um zu prüfen, ob es sich tatsächlich um zwei trennbare Konstrukte handelt, wurde in einer weiteren Analyse ein Modell mit zwei Faktoren auf beiden Ebenen spezifiziert. Dieses Modell passt deutlich schlechter zu den Daten ($\chi^2 = 2525{,}057^*$; df = 235; RMSEA = 0,05; CFI = 0,80; TLI = 0,77; SRMR (within) = 0,06; SRMR (between) = 0,15) als das 3/3-Faktoren-Modell (vgl. Tab. 2).

6.2 Zum Zusammenhang zwischen Unterrichtseinschätzungen und Leseleistung

Die Ergebnisse der Mehrebenen-Regressionsanalysen zur Klärung der Frage der kriterialen Validität der Einschätzungen sind in Tab. 3 dargestellt (Forschungsfrage 2).

Für die Leseleistung zeigen sich signifikante Zusammenhänge mit allen drei Kovariaten auf individueller Ebene. Auf Klassenebene zeigt sich kein Zusammenhang mit dem Anteil an Jungen in einer Klasse (Modell 1, Tab. 3). Für die kognitive Aktivierung findet sich auf Klassenebene ein signifikanter Zusammenhang mit der Leseleistung (Modell 2, Tab. 3). Die konstruktive Unterstützung hängt auf der individuellen Ebene signifikant mit der Leseleistung zusammen (Modell 3, Tab. 3). Für die Klassenführung ist auf beiden Ebenen ein signifikanter Zusammenhang mit der Leseleistung nachzuweisen (Modell 4, Tab. 3). Auf individueller Ebene bleiben die Zusammenhänge erhalten, wenn alle drei Dimensionen gemeinsam als Prädiktoren für die Leseleistung in ein Modell aufgenommen werden. Auf Klassenebene finden sich dann keine signifikanten Zusammenhänge mit der Leseleistung mehr (Modell 5, Tab. 3).

Tab. 3 Mehrebenen-Regression für die Leseleistung

		Leseleistung (AV)				
		Modell 1: Kovariaten	Modell 2: Kognitive Aktivierung	Modell 3: Konstruktive Unterstützung	Modell 4: Klassenführung	Modell 5: Unterrichtsqualität
		β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Individuelle Ebene	Mehrsprachigkeit	–0,07* (0,02)	–0,08* (0,02)	–0,07* (0,02)	–0,07* (0,02)	–0,07* (0,02)
	HISEI	0,31* (0,02)	0,31* (0,02)	0,30* (0,02)	0,30* (0,02)	0,30* (0,02)
	Geschlecht (1 = Junge)	–0,06* (0,02)	–0,06* (0,02)	–0,05* (0,02)	–0,06* (0,02)	–0,05* (0,02)
	Kognitive Aktivierung	–	0,03 (0,02)	–	–	–0,05 (0,02)
	Konstruktive Unterstützung	–	–	0,12* (0,02)	–	0,14* (0,02)
	Klassenführung	–	–	–	0,07* (0,02)	0,06* (0,03)
Klassenebene	Anteil Mehrsprachiger	–0,17* (0,07)	–0,18* (0,07)	–0,16* (0,07)	–0,11 (0,09)	–0,14 (0,07)
	Mittlerer HISEI	0,53* (0,07)	0,53* (0,07)	0,53* (0,07)	0,51* (0,07)	0,52* (0,07)
	Anteil Jungen	–0,15 (0,12)	–0,14 (0,11)	–0,14 (0,12)	–0,14 (0,13)	–0,14 (0,12)
	Kognitive Aktivierung	–	0,31* (0,13)	–	–	0,40 (0,29)
	Konstruktive Unterstützung	–	–	0,16 (0,13)	–	–0,27 (0,26)
	Klassenführung	–	–	–	0,29* (0,12)	0,22 (0,14)
Modellfit	Chi^2 (df)	7,206 (4)	224,486* (53)	475,945* (68)	191,015* (67)	1478,511* (366)
	RMSEA	0,02	0,03	0,04	0,02	0,03
	CFI	0,99	0,94	0,92	0,97	0,92
	TLI	0,97	0,91	0,90	0,96	0,90
	SRMR (within)	0,00	0,04	0,05	0,03	0,04
	SRMR (between)	0,05	0,09	0,08	0,12	0,11

* $p < 0,05$

7 Diskussion der Ergebnisse

Das Ziel der beschriebenen Untersuchung war es, die bisher vorliegenden Befunde zur Validität der Einschätzungen von Grundschüler(inne)n zu ergänzen. Zu diesem Zweck wurden die faktorielle und die kriteriale Validität von Einschätzungen der Unterrichtsqualität geprüft, die Viertklässler(innen) im Rahmen von IGLU 2016 vorgenommen haben. Theoretischer Ausgangspunkt der Analysen war ein dreidimensionales Modell von Unterrichtsqualität, das die Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung umfasst. Dieses Modell wurde aus der Forschung zum mathematisch-naturwissenschaftlichen Unterricht abgeleitet (vgl. Fauth et al. 2014a) und auf den Deutschunterricht in der Grundschule übertragen. Insofern wird mit den Befunden der vorliegenden Untersuchung der Forschungsstand zur Validität der Einschätzungen von Grundschüler(inne)n sowie zur Relevanz des dreidimensionalen Modells von Unterrichtsqualität für den Deutschunterricht in der Grundschule erweitert. Die faktorielle Validität der Einschätzungen wurde mithilfe von konfirmatorischen Mehrebenen-Faktorenanalysen geprüft. Es zeigt sich, dass auf individueller Ebene und auf Klassenebene ein drei Faktoren umfassendes Modell besser zu den Daten passt als ein Modell mit einer Dimension. Auch für den Deutschunterricht unterstützen die Ergebnisse aus IGLU 2016 die Annahme, dass Grundschüler(innen) zu differenzierten Urteilen über den Unterricht imstande sind. Ein Beleg für die inhaltliche Validität der Items ist das jedoch noch nicht (vgl. Lenske 2016). Um die kriteriale Validität der Einschätzungen der Viertklässler(innen) zu prüfen, wurden auf Grundlage von doppelt-latenten Mehrebenen-Regressionsmodellen Zusammenhänge zwischen den Unterrichtseinschätzungen und der Leseleistung berechnet. Auf individueller Ebene und auf Klassenebene ist ein Zusammenhang zwischen der Einschätzung der Klassenführung und der Leseleistung nachweisbar, wenn die Dimension alleine in die Analyse einbezogen wird. Für die Einschätzung der kognitiven Aktivierung zeigt sich nur auf Klassenebene ein positiver Zusammenhang mit der Leseleistung, während für die konstruktive Unterstützung nur auf individueller Ebene ein positiver Zusammenhang nachzuweisen ist. Bei gemeinsamer Betrachtung aller drei Dimensionen in einem Modell sind auf der Klassenebene keine Zusammenhänge mit der Leistung nachzuweisen, auf der individuellen Ebene bleiben die Zusammenhänge erhalten.

Die Befunde zum Zusammenhang von Klassenführung und den Leseleistungen der Viertklässler(innen) ergänzen das Bild, das sich durch andere Studien ergibt (vgl. Fauth et al. 2014a; Lipowsky und Bleck 2019): Auch die Ergebnisse von IGLU sprechen dafür, dass effiziente Klassenführung eine leistungsrelevante Dimension ist. Die Befunde zum Zusammenhang von kognitiver Aktivierung und den Leseleistungen auf Klassenebene entsprechen den Erwartungen, die sich aufgrund von konstruktivistischen Vorstellungen des Lernens ergeben, wenn die Dimension alleine in die Analyse einbezogen wird: Auf dieser Ebene ist ein Zusammenhang mit der Leseleistung nachzuweisen. Auf individueller Ebene zeigt sich kein Zusammenhang. Die Items zur Einschätzung der kognitiven Aktivierung fokussieren Handeln von Lehrkräften, das kognitive Aktivitäten aufseiten der Lernenden anregen sollte. Dieses Handeln muss nicht von allen Lernenden als anregend empfunden werden und zum individuellen Lernerfolg beitragen. Das Handeln sollte aber an den Voraussetzungen

der Klasse orientiert sein und positiv mit der Leistung auf Klassenebene zusammenhängen. Insofern sind die Ergebnisse erklärbar. Die Einschätzung der konstruktiven Unterstützung hängt auf individueller Ebene positiv mit der Leistung zusammen. Auf Klassenebene zeigt sich kein Zusammenhang, wenn nur diese Dimension in die Analyse einbezogen wird. Der fehlende Zusammenhang auf Klassenebene könnte mit der Operationalisierung der Dimension zusammenhängen. Zur Erfassung der konstruktiven Unterstützung wurden Items genutzt, die nach dem Erleben der Lernenden fragen. Das kann sich zwischen den Kindern einer Klasse unterscheiden. Daher sollte geprüft werden, ob sich die Items eignen, konstruktive Unterstützung auf Klassenebene zu erfassen. Der positive Zusammenhang auf individueller Ebene spricht dafür, dass eine positiv wahrgenommene Unterstützung bedeutsam für den individuellen Lernerfolg ist.

Werden die drei Dimensionen gemeinsam in ein Modell eingebracht, bleiben die Zusammenhänge auf individueller Ebene bestehen. Auf Klassenebene sind keine Zusammenhänge mehr nachzuweisen. Ein Grund dafür könnte in der hohen Interkorrelation zwischen den Dimensionen liegen – insbesondere der zwischen kognitiver Aktivierung und konstruktiver Unterstützung. Die Analysen zeigen, dass es sich um zwei voneinander abgrenzbare Konstrukte handelt. Die hohe Korrelation zwischen den beiden Dimensionen könnte durch eine konzeptionelle Nähe in der Operationalisierung begründet sein. Zum Beispiel wurde das Item „*Unsere Deutschlehrerin/unser Deutschlehrer glaubt, dass ich schwierige Aufgaben lösen kann*“ der konstruktiven Unterstützung zugeordnet. Damit entsteht eine inhaltliche Nähe zu den Items der kognitiven Aktivierung, die sich auf die kognitive Herausforderung im Unterricht beziehen. Der Überschneidungsbereich der beiden Dimensionen hätte verringert werden können, wenn hinsichtlich der konstruktiven Unterstützung nur die emotionale Unterstützung der Grundschüler(innen) erfasst worden wäre. Das ist allerdings mit der gängigen Fassung konstruktiver Unterstützung kaum zu vereinbaren (vgl. Lipowsky und Bleck 2019). Bis zu einem gewissen Grad aufzulösen wäre die Überschneidung von konstruktiver Unterstützung und kognitiver Aktivierung auch mit einer relativ engen Fassung beider Konstrukte: Minnameier et al. (2015) z. B. bestimmen kognitive Aktivierung als „Induktion eines Problems beim Lernenden“ (ebd., S. 842) und konstruktive Unterstützung als „Anleitung und Begleitung des dadurch in Gang gesetzten Problemlöseprozesses“ (ebd.). In zukünftigen Studien wäre bei der Item-Auswahl eine entsprechende Ausrichtung möglich, um die Konstrukte besser abzugrenzen. Damit ginge allerdings eine andere Ausrichtung dieser beiden Dimensionen einher, als sie für die vorliegende Studie leitend war.

Neben der hohen Korrelation zwischen kognitiver Aktivierung und konstruktiver Unterstützung ist die eher geringe Korrelation zwischen diesen beiden Konstrukten und der Klassenführung auffällig. Diese könnte dadurch zustande kommen, dass sich die Items zur Klassenführung von denen der beiden anderen Dimensionen hinsichtlich ihrer Polung unterscheiden. Die Items zur Klassenführung sind negativ formuliert, die Items zur kognitiven Aktivierung und zur konstruktiven Unterstützung positiv. Dadurch könnte die empirische Trennbarkeit der Klassenführung von den beiden anderen Konstrukten verbessert worden sein.

Bei der Interpretation der Ergebnisse zur kriterialen Validität ist zu beachten, dass es sich bei IGLU um eine Querschnittsstudie handelt. Somit können keine Aussagen

über kausale Zusammenhänge gemacht werden. Zudem liegen Hinweise vor, dass die Höhe von Zusammenhängen zwischen Unterrichtsmerkmalen und Leistungsmaßen in querschnittlichen Analysen tendenziell überschätzt wird (vgl. Kuger et al. 2017).

Zudem ist zu bedenken, dass insbesondere die Einschätzungen der kognitiven Aktivierung eine geringe Übereinstimmung innerhalb der Klassen aufweisen (ICC1 und ICC2). Diesem Problem wurde in den Analysen mit einer doppelt-latenten Modellierung begegnet, die sowohl den Stichproben- als auch den Messfehler berücksichtigt (vgl. Lüdtke et al. 2011). Trotzdem erscheint eine kritische Prüfung der Operationalisierung der kognitiven Aktivierung angebracht. Die Grundschüler(innen) mussten zur Bearbeitung der Items zur kognitiven Aktivierung die Häufigkeit, in der ein bestimmtes Handeln vorgekommen ist, über einen längeren Zeitraum aus dem Gedächtnis rekonstruieren. Nach dieser Rekonstruktion mussten sie eine Antwort auswählen. Das ist ein fehleranfälliger Prozess (vgl. Helmke und Lenske 2015). Es kann angenommen werden, dass eine Beschränkung des zu beurteilenden Zeitraums auf einige genau spezifizierte Unterrichtsstunden die Einschätzung der kognitiven Aktivierung verbessern könnte.

Die Ergebnisse der vorliegenden Untersuchung lassen weitere Fragen für folgende Forschungsprojekte offen: Der IGLU-Leistungstest bezieht sich nur auf einen von vier Kompetenzbereichen des Deutschunterrichts (vgl. KMK 2005, S. 7: „Lesen – mit Texten und Medien umgehen“, „Schreiben“, „Sprechen und Zuhören“ sowie „Sprache und Sprachgebrauch untersuchen“). In weiterführenden Studien wäre zu prüfen, ob sich das gefundene Zusammenhangsmuster auch für Leistungen in anderen Kompetenzbereichen zeigt (vgl. Naumann et al. 2019). Die Items zur Einschätzung der Unterrichtsqualität dagegen fokussieren den Deutschunterricht im Allgemeinen. Es wäre zu prüfen, ob es von Vorteil wäre, in den Items eine Beschränkung auf den Leseunterricht vorzunehmen. Allerdings ist fraglich, ob Viertklässler(innen) Deutsch- und Leseunterricht unterscheiden (können).

Auch die eingesetzten Items sind einer kritischen Prüfung zu unterziehen. Mit Prenzel und Lankes (2013) sind auf den Unterricht bezogene Items dahingehend zu unterscheiden, ob sie einer Beschreibung oder einer Einschätzung des Unterrichts zugrunde liegen. Beschreibende Items geben Auskunft über konkrete Aspekte der Unterrichtsgestaltung und liefern Hinweise auf mögliche Anpassungen durch Lehrkräfte. Items, die Einschätzungen erlauben, sind weniger konkret – mitunter sind sie überhaupt nicht auf bestimmte Merkmale des Unterrichts zu beziehen. Diese Items weisen häufig statistische Zusammenhänge mit Lernergebnisse auf. Allerdings lassen sie sich kaum dazu nutzen, Lehrkräften Hinweise auf Anpassungen im Unterricht zu geben. Die im Rahmen von IGLU 2016 verwendeten Items sind auf den Polen Beschreibung und Einschätzung an unterschiedlichen Stellen einzuordnen. So wird z. B. in dem Item „*Unsere Deutschlehrerin/unser Deutschlehrer ist auch dann nett zu mir, wenn ich einen Fehler mache*“ ein konkretes Ereignis beschrieben, aber hinsichtlich der Reaktion der Lehrkräfte ist eine Einschätzung durch die Kinder notwendig. Das Item „*Im Deutschunterricht ist es laut und unruhig*“ erfordert eine Einschätzung ohne eine Konkretisierung an einem bestimmten Ereignis. Das Item „*Unsere Deutschlehrerin/unser Deutschlehrer möchte, dass wir unsere Antworten erklären*“ bezieht sich auf konkrete Unterrichtsereignisse und ist (tendenziell) beschreibend. Kloss (2014) stellt die Validität der Einschätzungen von Grundschüler(inne)n auf der

Grundlage von eher unterrichtszentrierten, sachlichen und klassenbezogenen Items im Vergleich zu lehrkräfte- und selbstbezogenen sowie evaluativen Items heraus. Das Item „*Unsere Deutschlehrerin/unser Deutschlehrer ist auch dann nett zu mir, wenn ich einen Fehler mache*“ kann als lehrkräfte- und selbstbezogen sowie als evaluativ angesehen werden. Das Item „*Im Deutschunterricht ist es laut und unruhig*“ ist unterrichtszentriert, evaluativ und klassenbezogen. Das Item „*Unsere Deutschlehrerin/unser Deutschlehrer möchte, dass wir unsere Antworten erklären*“ ist sowohl lehrkräfte- als auch klassenbezogen und eher sachlich als evaluativ. Allerdings ist unklar, wann ein Kind von einer Erklärung sprechen würde. Entscheidungen hinsichtlich dieser Merkmale bei der Item-Auswahl sollten zukünftig vor dem Hintergrund der jeweiligen Studienziele getroffen und konsequent verfolgt werden.

Abschließend noch eine Anmerkung zur Frage, ob die Unterrichtsqualität in der vorliegenden Untersuchung fachspezifisch erfasst wurde: Das in der Untersuchung zugrunde gelegte Modell der Unterrichtsqualität wird häufig als generisch verstanden (vgl. Charalambous und Praetorius 2018; Praetorius et al. 2018). Zudem findet sich in der Literatur der Hinweis, dass bei der Operationalisierung der kognitiven Aktivierung fachspezifische Merkmale berücksichtigt werden sollten (vgl. Klieme et al. 2006; Lotz 2016). Bei der Einschätzung des Potenzials zur kognitiven Aktivierung im Deutschunterricht können nicht nur fachspezifische, sondern auch lernbereichs- und themenspezifische Aspekte berücksichtigt werden. Lotz (2016) z. B. operationalisiert kognitive Aktivierung für den Leseunterricht unter Berücksichtigung lernbereichsspezifischer Merkmale – u. a. zur Instruktion von Lesestrategien. Diese wiederum dürften vor allem von Bedeutung sein, wenn die Förderung des Textverstehens das Ziel ist. Weniger relevant ist die Vermittlung von Lesestrategien für die kognitive Aktivierung im Leseunterricht, in dem die Förderung der Leseflüssigkeit im Mittelpunkt steht. Das Potenzial dazu könnte z. B. daran festgemacht werden, dass lesestarke Schüler(innen) bei Lautleseübungen als Tutor(inn)en für schwächere Kinder fungieren (vgl. Rosebrock und Nix 2014). Darüber hinaus berücksichtigt Lotz (2016) zur Einschätzung der kognitiven Aktivierung Merkmale, die in anderen Fächern oder Lernbereichen des Deutschunterrichts für die Einschätzung des Potenzials zur kognitiven Aktivierung genutzt werden (u. a. anregende Fragen und Feedback der Lehrkraft). In Studien, in denen das Potenzial zur kognitiven Aktivierung der Schüler(innen) erhoben wird, ist demnach eine Orientierung an fachübergreifenden und fach-, lernbereichs- oder themenspezifischen Merkmalen möglich. Auch für die Einschätzung der anderen Qualitätsdimensionen könnten fachspezifische Merkmale berücksichtigt werden, sodass man in Abhängigkeit von der jeweiligen Operationalisierung der Dimensionen in Anlehnung an die Begrifflichkeit von Charalambous und Praetorius (2018, S. 357) von einem *hybriden Modell* von Unterrichtqualität sprechen könnte. In diesem Modell würden drei Dimensionen angesetzt, die jeweils fachspezifisch und/oder fachunspezifisch operationalisiert werden könnten. Die Unterrichtsqualität wurde für die vorliegende Untersuchung nicht unter Berücksichtigung fachspezifischer Aspekte operationalisiert. Es wäre problematisch vorauszusetzen, dass Grundschüler(innen) Unterrichtsaspekte einschätzen können, wenn das fachwissenschaftliches/fachdidaktisches Wissen erfordert (vgl. Clausen 2002; Lenske 2016). Daher schien der Verzicht auf eine fachspezifische Operationalisierung als gangbarer Weg für die Einschätzung des Deutschunterrichts

durch die Lernenden. Damit bleiben allerdings viele Unterrichtsaspekte unberücksichtigt, die für den Erwerb von Lesekompetenzen von Bedeutung sind.

Funding Open access funding provided by FHNW University of Applied Sciences and Arts Northwestern Switzerland

Literatur

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Hrsg.), *Advanced structural equation modelling. Issues and techniques* (S. 243–277). Mahwah, N.J.: Lawrence Erlbaum.

Baumert, J., & Kunter, M. (2011). Das mathematikspezifische Wissen von Lehrkräften, kognitive Aktivierung im Unterricht und Lernfortschritte von Schülerinnen und Schülern. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 163–192). Münster: Waxmann.

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., Krauss, S., Kunter, M., Löwen, K., Neubrand, M., & Tsai, Y. M. (2009). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV). Dokumentation der Erhebungsinstrumente*. Berlin: Max-Planck-Institut für Bildungsforschung.

Bremerich-Vos, A., Wendt, H., & Bos, W. (2017). Lesekompetenzen im internationalen Vergleich: Testkonzeption und Ergebnisse. In A. Hußmann, H. Wendt, W. Bos, A. Bremerich-Vos, D. Kasper, E.-M. Lankes, N. McElvany, T. C. Stubbe & R. Valtin (Hrsg.), *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 79–142). Münster: Waxmann.

Charalambous, C. Y., & Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: Setting the ground for understanding instructional quality more comprehensively. *ZDM*, *50*(3), 355–366.

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

Deci, E. L., & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik*, *39*(2), 224–238.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, *12*(2), 121–138.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014a). Students ratings of teaching quality in primary school: dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014b). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg. *Zeitschrift für Pädagogische Psychologie*, *28*(3), 127–137.

Gabriel, K. (2014). *Videobasierte Erfassung von Unterrichtsqualität im Anfangsunterricht der Grundschule – Klassenführung und Unterrichtsklima in Deutsch und Mathematik*. Kassel: University Press.

Ganzeboom, H. B. G., de Graaf, P. M., Treiman, D. J., & de Leeuw, J. (1992). *A standard international socio-economic index of occupational status*. Tilburg: WORC, Work and Organization Research Centre. WORC Reprint.

Göllner, R., Wagner, W., Eccles, J. S., & Trautwein, U. (2018). Students' idiosyncratic perceptions of teaching quality in mathematics: A result of rater tendency alone or an expression of dyadic effects between students and teachers*? Journal of Educational Psychology*, *110*(5), 709–725. https://doi.org/10.1037/edu0000236.

Hanisch, A. (2018). *Kognitive Aktivierung im Rechtschreibunterricht. Eine Interventionsstudie in der Grundschule*. Münster: Waxmann.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Hußmann, A., Wendt, H., Kasper, D., Bos, W., & Goy, M. (2017). Ziele, Anlage und Durchführung der Internationalen Grundschul-Lese-Untersuchung (IGLU 2016). In A. Hußmann, H. Wendt, W. Bos, A. Bremerich-Vos, D. Kasper, E.-M. Lankes, N. McElvany, T. C. Stubbe & R. Valtin (Hrsg.), *IGLU 2016 Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 29–78). Münster: Waxmann.

Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts „Pythagoras". In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms BiQua* (S. 127–146). Münster: Waxmann.

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

Klieme, E., Steinert, B., & Hochweber, J. (2010). Zur Bedeutung der Schulqualität für Unterricht und Lernergebnisse. In W. Bos, E. Klieme & Ö. Köller (Hrsg.), *Schulische Lerngelegenheiten und Kompetenzentwicklung. Festschrift für Jürgen Baumert* (S. 231–255). Münster: Waxmann.

Kloss, J. (2014). *Grundschüler als Experten für Unterricht. Empirische Überprüfung der Validität von Unterrichtsbeurteilungen durch Schüler der dritten und vierten Jahrgangsstufe*. Frankfurt a. M.: Peter Lang.

KMK – Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005). *Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4)*. München: Wolters Kluwer.

Kuger, S., Klieme, E., Lüdtke, O., Schiepe-Tiska, A., & Reiss, K. (2017). Mathematikunterricht und Schülerleistung in der Sekundarstufe: Zur Validität von Schülerbefragungen in Schulleistungsstudien. *Zeitschrift für Erziehungswissenschaft*, *20*(Suppl 2), 61–98.

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

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependence of three causal models of Halo Rater Error. *Journal of Applied Psychology*, *79*(3), 332–340.

Lankes, E.-M., & Carstensen, C. H. (2007). Der Leseunterricht aus der Sicht der Lehrkräfte. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert & R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 161–193). Münster: Waxmann.

Lenske, G. (2016). *Schülerfeedback in der Grundschule. Untersuchung zur Validität*. Münster: Waxmann.

Lenske, G., & Helmke, A. (2015). Child respondents—Do they really answer what scientific questionnaires ask for? In W. Schnotz, A. Kauertz, H. Ludwig, A. Müller & J. Pretsch (Hrsg.), *Multidisciplinary research on teaching and learning* (S. 146–166). Basingstoke: Palgrave Macmillan.

Lipowsky, F., & Bleck, V. (2019). Was wissen wir über guten Unterricht? – Ein Update. In U. Steffens & R. Messner (Hrsg.), *Konzepte und Bilanzen gelingenden Lehrens und Lernens – Grundlagen der Qualität von Schule* (Bd. 3, S. 219–249). Münster: Waxmann.

Lotz, M. (2016). *Kognitive Aktivierung im Leseunterricht in der Grundschule. Eine Videostudie zur Gestaltung und Qualität von Leseübungen im ersten Schuljahr*. Wiesbaden: Springer.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: how to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, *34*(2), 120–131.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 X 2 taxonomy of multilevel latent contextual models: accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*(4), 444–467.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, *59*(1), 14–19.

Minnameier, G., Hermkes, R., & Mach, H. (2015). Kognitive Aktivierung und Konstruktive Unterstützung als Prozessqualitäten des Lehrens und Lernens. *Zeitschrift für Pädagogik*, *61*(6), 837–856.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8. Aufl.). Los Angeles, CA: Muthén & Muthén.

Naumann, A., Rieser, S., Musow, S., Hochweber, J., & Hartig, J. (2019). Sensitivity of test items to teaching quality. *Learning and Instruction*, *60*, 41–53.

Praetorius, A.-K., Vieluf, S., Saß, S., Bernholt, A., & Klieme, E. (2015). The same in German as in English? Investigating the subject-specifity of teaching quality. *Zeitschrift für Erziehungswissenschaft*, *19*, 191–209.

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.

Prenzel, M., & Lankes, E.-M. (2013). Was können Schüler(innen) über ihren Unterricht sagen? Ein Blick in die Schülerfragebogen von internationalen Vergleichsstudien. In N. McElvany & H. G. Holtappels (Hrsg.), *Empirische Bildungsforschung. Theorien, Methoden, Befunde und Perspektiven. Festschrift für Wilfried Bos* (S. 93–107). Münster: Waxmann.

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen & J. S. Long (Hrsg.), *Testing structural equation models*. Sage focus editions, (Bd. 154, S. 163–180). Newbury Park: SAGE.

Raudenbush, S. W., & Bryk, A. S. (2010). *Hierarchical linear models: applications and data analysis methods* (2. Aufl.). Advanced quantitative techniques in the social sciences, Bd. 1. Thousand Oaks: SAGE.

Rosebrock, C., & Nix, D. (2014). *Grundlagen der Lesedidaktik und der systematischen schulischen Leseförderung* (7. Aufl.). Baltmannsweiler: Schneider.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Stahns, R., Rieser, S., & Lankes, E.-M. (2017). Unterrichtsführung, Sozialklima und kognitive Aktivierung im Deutschunterricht in vierten Klassen. In A. Hußmann, H. Wendt, W. Bos, A. Bremerich-Vos, D. Kasper, E.-M. Lankes, N. McElvany, T. C. Stubbe & R. Valtin (Hrsg.), *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 251–277). Münster: Waxmann.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: dimensionality and generalizability of domain-independet assessments. *Learning and Instruction*, *28*, 1–11.

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, *108*(5), 705–721.

Weis, M., Zehner, F., Sälzer, C., Strohmaier, A., Artelt, C., & Pfost, M. (2016). Lesekompetenz in PISA 2015: Ergebnisse, Veränderungen und Perspektiven. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme & O. Köller (Hrsg.), *PISA 2015. Eine Studie zwischen Kontinuität und Innovation* (S. 249–284). Münster: Waxmann.

Wittig, J., & Weirich, S. (2017). Mittelwerte und Streuungen der im Fach Deutsch erreichten Kompetenzen. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 153–167). Münster: Waxmann.