Unterrichtswiss (2021) 49:137–163
https://doi.org/10.1007/s42010-020-00086-4

ALLGEMEINER TEIL

90 Minuten Mathematikunterricht bei gleichbleibender Unterrichtsqualität? – Analysen zur zeitlichen Stabilität und Generalisierbarkeit von Ratings zur Unterrichtsqualität im 2. Schuljahr

Katrin Gabriel-Busse · Frank Lipowsky

Eingegangen: 7. Januar 2020 / Überarbeitet: 24. September 2020 / Angenommen: 26. September 2020 / Online publiziert: 28. Oktober 2020

Zusammenfassung Studien für die Sekundarstufe zeigen, dass nicht zwangsläufig von der Qualität einer beobachteten Stunde einer Lehrperson auf die Qualität einer anderen Stunde geschlossen werden kann. Anhand sogenannter Generalisierbarkeitsstudien (G-Studien) konnte vor allem für die kognitive Aktivierung aufgezeigt werden, dass die Ausprägungen zwischen den einzelnen Unterrichtsstunden einer Lehrperson (bei gleicher Klasse) stark variieren (hoher stundenspezifischer Varianzanteil) und somit situationale bzw. kontextuelle Einflussfaktoren eine größere Rolle spielen als bislang angenommen, mit der Konsequenz, dass deutlich mehr Stunden für eine hinreichend zuverlässige Beobachtung der kognitiven Aktivierung notwendig sind. Für die Grundschule mangelt es bislang an entsprechenden Studien zur Analyse der zeitlichen Stabilität und Generalisierbarkeit der Unterrichtsqualität bzw. es werden lediglich Korrelationskoeffizienten als Hinweis zur Stabilität der Ausprägungen der einzelnen Basisdimensionen berichtet. Solche Korrelationskoeffizienten können jedoch nicht im Sinne einer zeitlichen Stabilität der Unterrichtsqualitätsmerkmale interpretiert werden, da sie mit anderen Faktoren konfundiert sind. Im vorliegenden Beitrag wird aus diesem Grund für die drei Basisdimensionen der Unterrichtsqualität untersucht, inwieweit sich die Unterrichtsqualität in einer Doppelstunde zur „Einführung in die Multiplikation“ (2. Schuljahr) in Abhängigkeit von dem Beobachtungszeitpunkt (1. Stunde vs. 2. Stunde einer 90-minütigen Unterrichtseinheit) unterscheidet. Grundlage bilden Videoaufzeichnungen von 36 Lehrpersonen. Die Ergebnisse zeigen – ähnlich wie für die Sekundarstufe – Unterschiede in der zeitlichen

K. Gabriel-Busse (✉)
Institut für Erziehungswissenschaft, Johannes Gutenberg-Universität Mainz, Jakob-Welder-Weg 12, 55128 Mainz, Deutschland
E-Mail: kgabriel@uni-mainz.de

F. Lipowsky
Fachgebiet Empirische Schul- und Unterrichtsforschung, Universität Kassel, Nora-Platiel-Str. 1, 34109 Kassel, Deutschland
E-Mail: lipowsky@uni-kassel.de

Stabilität je nach untersuchtem Unterrichtsqualitätsmerkmal. Im Unterschied zu den Ergebnissen aus der Sekundarstufe weist vor allem die Klassenführung eine höhere Variation in den Ausprägungen zwischen den zwei Teilstunden einer Lehrperson auf. Die Ergebnisse der explorativen D-Studien legen nahe, dass für eine hinreichend reliable Erfassung der Klassenführung sowie der kognitiven Aktivierung in der Grundschule mindestens drei und für das Unterrichtsklima zwei Unterrichtsstunden einer Lehrperson benötigt werden.

Schlüsselwörter Unterrichtsqualität · Grundschule · Mathematik · Generalisierbarkeitsstudie · Entscheidungsstudie

90 minutes of classroom teaching—same teaching quality? stability and generalizability of ratings of teaching quality across time in year two mathematics

Abstract Studies in secondary school education show that it is not invariably possible to draw conclusions from the quality of one observed lesson by a particular teacher about the quality of another lesson by the same teacher. Using generalizability studies (G studies), studies in secondary schools showed that the intensity of cognitive activation varies among individual lessons (given to the same class) by a given teacher (high lesson-specific proportion of variance). Thus situational and contextual factors play a greater role than has hitherto been assumed. Consequently, up to nine hours per teacher suffices to measure cognitive activation in secondary schools. Available studies for elementary schools do not analyze the stability and generalizability of ratings of teaching quality over time; rather, they confine themselves to reporting correlation coefficients as indicators of the stability of three dimensions of teaching quality. Since, however, such correlation coefficients are confounded with other factors, they cannot be used to draw conclusions about the temporal consistency of teaching quality. The present paper, therefore, studies the three basic dimensions of teaching quality to determine the extent to which teaching quality differs in a double period of "Introduction to Multiplication" (Year Two) depending on the time observed (first vs. second period of a 90-minute teaching unit), based on video recordings of 36 teachers. As in the case of secondary school, the results reveal differences in temporal consistency depending on the quality dimension that was studied. As opposed to the findings for those for secondary school in that it is chiefly classroom management that exhibits greater variation between the two lesson periods of the same teacher. The results of the exploratory D-studies suggest that threes lesson per teacher suffices to measure classroom management and cognitive activation in elementary school, whereas two lessons would be needed for classroom climate.

Keywords Teaching quality · Elementary School · Mathematics · Generalizability Theory · decision study

1 Theoretischer und empirischer Hintergrund

Als grundlegende Qualitätsmerkmale von Unterricht konnten in der empirischen Unterrichtsforschung drei sogenannte Basisdimensionen von Unterrichtsqualität (Klassenführung, unterstützendes Unterrichtsklima bzw. konstruktive Unterstützung und kognitive Aktivierung) identifiziert werden (z. B. Klieme et al. 2006; Klieme 2019; Praetorius et al. 2018). Im Rahmen von Videostudien werden diese Merkmale, unter Verwendung verschiedener Begrifflichkeiten und Schwerpunktsetzungen, häufig genutzt, um Unterricht zu beschreiben, mit Hilfe von Beobachterratings einzuschätzen und auf ihre Wirkungen zu untersuchen (vgl. Clausen et al. 2003; Gabriel 2014; Klieme 2002; Klieme et al. 2006; Kunter et al. 2006, 2007; Lipowsky et al. 2009). Inzwischen liegen zahlreiche Befunde zur Reliabilität und Validität von Beobachterratings vor (u. a. Fauth et al. 2014a, 2014b; Praetorius et al. 2012; Rakoczy 2008). Weniger gut erforscht ist bislang jedoch die Frage nach der zeitlichen Stabilität der Unterrichtsqualitätsmerkmale (Jentsch et al. 2019; Praetorius et al. 2014). Hier mangelt es vor allem an Studien im Grundschulbereich (vgl. Gabriel et al. 2015). Zudem untersuchten bisherige Studien die zeitliche Stabilität der Unterrichtsqualitätsmerkmale häufig über einen längeren Zeitraum (z. B. wenige Wochen oder Monate, siehe auch Jentsch et al. 2019), mit dem Ergebnis, dass die zeitliche Stabilität in Abhängigkeit vom untersuchten Unterrichtsqualitätsmerkmal differiert. Inwieweit die Merkmale der Unterrichtsqualität auch in kürzeren Abständen variieren (z. B. innerhalb einer Doppelstunde), wurde bislang kaum untersucht (Ausnahme z. B. Studie von Jentsch et al. 2019).

Die vorliegende Studie greift die genannten Desiderate im Rahmen einer Generalisierbarkeitsstudie (Cronbach et al. 1972) auf, indem die zeitliche Variabilität von Beobachterratings für Messwiederholungen innerhalb von Doppelstunden im Mathematikunterricht der Grundschule (2. Schuljahr) untersucht wird. Zusätzlich werden explorative Simulationsstudien (*decision studies*, vgl. Shavelson und Webb 1991) durchgeführt, um Aussagen darüber machen zu können, wie viele Unterrichtsstunden pro Lehrperson nötig sind, um die drei Basisdimensionen der Unterrichtsqualität im Mathematikunterricht im 2. Schuljahr hinreichend zuverlässig zu erfassen.

1.1 Zeitliche Stabilität bzw. Variabilität von Unterrichtsqualität in der Grundschule

Studien, die sich in den vergangenen Jahren mit der zeitlichen Stabilität der Unterrichtsqualität in der Grundschule beschäftigt haben, berechneten häufig Korrelationen zwischen verschiedenen Messzeitpunkten, wobei nicht nur die Anzahl der beobachteten Unterrichtsstunden variiert (vgl. im Überblick: Gabriel-Busse und Lipowsky in Vorbereitung; NICHD ECCRN 2002; Eckerth et al. 2012), sondern – in Abhängigkeit vom untersuchten Unterrichtsqualitätsmerkmal – auch die Höhe der Korrelation. So zeigen z. B. die Befunde aus der SCHOLASTIK-Studie (vgl. Helmke und Weinert 1997), dass die Klassenführung neben der Strukturiertheit am stabilsten über die Zeit (vom 3. auf das 4. Schuljahr, $r = 0{,}57^{**}$) ist, dicht gefolgt vom sozialen Klima mit $r = 0{,}46^{**}$. Analysen im Rahmen der Validierung des CLASS-Instruments (vgl. Pianta et al. 2008; Pianta und Hamre 2009) bestätigen, dass die

über einen Zeitraum von zwei Stunden beobachteten Unterrichtsqualitätsmerkmale im dritten Schuljahr relativ stabil über die vier einbezogenen Messzeitpunkte sind. Die Korrelationen für die Merkmale der Classroom Organization variieren zwischen 0,68 und 0,98, die für den Emotional Support zwischen 0,82 und 0,98 und die für den Instructional Support[1] zwischen 0,72 und 0,98.

Vor allem für das Merkmal kognitive Aktivierung ist die Befundlage für die Grundschule jedoch nicht eindeutig: In einer älteren Studie von Brophy et al. (1975) wurden 19 Lehrpersonen über zwei Schuljahre hinweg (2.–3. Schuljahr) mehrmals mit Hilfe des COS (Classroom Observation System) beobachtet. Für jedes Schuljahr wurde ein Mittelwert aus allen Beobachtungen gebildet. Die Korrelation der beiden Mittelwerte für das Merkmal „Teacher-Initiated Problem Solving" in Höhe von $r = 0{,}59$* zeigt beispielsweise, dass die beobachteten Lehrpersonen „frequently direct[...] questions or problems to the class, followed by elaborating questions to other students or probe[...] into students answers" (S. 874). Dies führt auch dazu, dass das Schülerverhalten bzw. das „level of cognitive pupil behavior" relativ stabil über beide Schuljahre hinweg ausgeprägt ist ($r = 0{,}64$*). Weitaus geringere Korrelationen zeigten sich jedoch in der Metaanalyse von Shavelson und Dempsey-Atwood (1976) für die Merkmale „higher cognitive level of student behavior" und „conceptual" (jeweils $r = 0{,}10$) sowie „problem solving" ($r = 0{,}27$), die inhaltlich der kognitiven Aktivierung zugeordnet werden können.

Für den Anfangsunterricht existieren bislang kaum Studien, die die zeitliche Stabilität der Basisdimensionen der Unterrichtsqualität untersucht haben. Im Rahmen des FiS-Projekts[2] fielen die Korrelationen zwischen verschiedenen Beobachtungszeitpunkten für die Merkmale Klassenführung, individuelle Unterstützung, Strukturierung und kognitive Aktivierung jedoch eher gering aus und waren nicht signifikant (min. $r = 0{,}18$; *ns*; max. $r = 0{,}24$; *ns*) (vgl. Eckerth et al. 2012). Die Autoren vermuten, dass die geringen Korrelationen mit der kurzen Schulbesuchszeit zu erklären sind. Nicht nur die Schülerinnen und Schüler finden sich mit zunehmender Verweildauer besser im Unterrichtsalltag zurecht (z. B. halten sich an Regeln, können Zusammenhänge alleine herausfinden), sondern auch die Lehrpersonen kennen die Bedürfnisse der Kinder besser und passen die Unterrichtsgestaltung entsprechend an, was wiederum Auswirkungen auf die Stabilität der Qualitätsdimensionen haben kann.

1.2 Korrelative Ansätze vs. Generalisierbarkeitsstudien zur Messung der zeitlichen Stabilität

Wie bereits dargestellt, liegen Ergebnisse zur zeitlichen Stabilität der Unterrichtsqualitätsmerkmale in der Grundschule häufig nur in Form von Korrelationen (= korrelative Ansätze) vor (vgl. Abschn. 1.1), die als Indikator für die Generalisierbarkeit bzw. zeitliche Stabilität des jeweiligen Unterrichtsqualitätsmerkmals angesehen

[1] Das Merkmal Instructional Support beinhaltet u. a. Aspekte der kognitiven Aktivierung (z. B. concept development).

[2] FiS steht als Abkürzung für „Förderung der Lern- und Bildungsprozesse von Kindern in der Schuleingangsphase".

werden (zusf. Gabriel et al. 2015). Das Problem bei Studien, die die Stabilität von Unterrichtsmerkmalen mithilfe von Korrelationen berechnen, ist, dass solche Korrelationen nach Shavelson und Dempsey-Atwood (1976) „nicht eindeutig im Sinne einer vorhandenen oder nicht vorhandenen Stabilität interpretiert werden können, da im Rahmen von Korrelationen die Stabilität mit weiteren Faktoren konfundiert ist" (Praetorius 2014, S. 238). Um die Stabilität von Unterrichtsmerkmalen von weiteren Einflussfaktoren zu trennen (z. B. Rater, Messzeitpunkte, Unterrichtsfächer) bietet sich die Durchführung sogenannter Generalisierbarkeitsstudien (G-Studien; vgl. Cronbach et al. 1972; Shavelson und Webb 1991) an, wie sie im Sekundarschulbereich in den letzten Jahren vermehrt durchgeführt wurden (z. B. Praetorius 2014; Praetorius et al. 2014; Schlesinger und Jentsch 2016; Jentsch et al. 2019). Im Kern geht es in solchen G-Studien darum, eine gemessene Variation auf verschiedene potenzielle Varianzquellen zurückzuführen (bzw. die Varianzquellen einer Messung aufzuschlüsseln) und deren relative Anteile zu bestimmen (vgl. Brennan 2001). Mithilfe von G-Studien konnte beispielsweise Praetorius (2014; vgl. auch Praetorius et al. 2012) für die Sekundarstufe feststellen, dass das Ausmaß an Varianz in den Ratings zur Unterrichtsqualität davon abhängig ist, welches Unterrichtsqualitätsmerkmal (Klassenführung, Schülerorientierung oder kognitive Aktivierung) untersucht wurde. Ihre Analysen zeigen sowohl für die Klassenführung als auch für die Schülerorientierung eine eher geringe Variation in den Ausprägungen der Dimensionen zwischen unterschiedlichen Stunden einer Lehrperson. Der stundenspezifische Varianzanteil lag bei 13 % für die Klassenführung und bei 5 % für die Schülerorientierung. Hypothesenkonform variieren hingegen die Ausprägungen der kognitiven Aktivierung in hohem Maße zwischen den unterschiedlichen Unterrichtsstunden der Lehrpersonen (stundenspezifische Varianzanteil = 46 %). Dies kann damit begründet werden, dass es sich bei der kognitiven Aktivierung um ein Merkmal handelt, das sowohl in Abhängigkeit von der Art (Einführungsstunde vs. Übungsstunde) als auch vom Inhalt der Stunde (Algebra, Problemlösen, Beweis des Satzes von Pythagoras usw.) zwischen den Unterrichtsstunden variieren kann (vgl. Praetorius 2014; Pauli und Reusser 2011), während die Klassenführung und Schülerorientierung eher als generische gegenstandsunabhängige Merkmale gelten (vgl. Rakoczy 2008; Lipowsky et al. 2018; Jentsch et al. 2019). Die Ergebnisse von Jentsch et al. (2019) zeigen zudem, dass die Beobachterratings für die Basisdimensionen Klassenführung und konstruktive Unterstützung auch innerhalb einer beobachteten Doppelstunde (vier Messzeitpunkte) zeitlich variiert. Als Erklärung fügen die Autoren an, dass vor allem zu Beginn einer Doppelstunde „Maßnahmen der Lehrperson zur Herstellung von Aufmerksamkeit und kognitiver Aktivierung [...] erfolgen oder zum Ende einer Unterrichtsstunde ausbleiben" (ebd. S. 16).

1.3 Beobachterzeiträume in bisherigen Studien

In bisherigen Studien, in denen die Unterrichtsqualität erfasst wird (mittels in-vivo- oder Videobeobachtung), werden eher wenige Stunden (1–3 Unterrichtsstunden) einer Lehrperson beobachtet (im Überblick z. B. Praetorius et al. 2014; Schlesinger und Jentsch 2016; für die Grundschule: Gabriel-Busse und Lipowsky in Vorbereitung) und die Ergebnisse generalisiert, unter der Annahme, dass eine

bis wenige Unterrichtsstunden ausreichen, um von einer stabilen Einschätzung von Unterrichtsqualität auszugehen. Dieses Vorgehen wird vor allem damit begründet, dass sich bestimmte Verhaltensweisen, die die Lehrperson bzw. die Schülerinnen und Schüler im Unterricht zeigen, nicht kurzfristig verändern und damit relativ zeitstabil, d. h. über die Zeit nicht variierend sind (z. B. Kunter 2005; Praetorius et al. 2016; Jentsch et al. 2019). Dieser Annahme folgend sollte ein bestimmtes Verhalten einer Lehrperson aber auch der Schülerinnen und Schüler in jeder Unterrichtsstunde zu verschiedenen Zeitpunkten beobachtbar sein (z. B. Klieme et al. 2001; Stigler et al. 1999), unabhängig davon, ob z. B. mehrere Unterrichtsstunden über verschiedene Tage oder zwei direkt aufeinander folgende Unterrichtsstunden (Doppelstunde) beobachtet werden. Inwieweit die Merkmale der Unterrichtsqualität jedoch in kurzen Zeitabständen (z. B. innerhalb einer Doppelstunde) variieren, wurde selten untersucht. Eine Ausnahme stellt u. a. die Studie von Jentsch et al. (2019) dar. Die Autoren konnten für die Sekundarstufe (7.–10. Klasse) zeigen, dass die zeitliche Stabilität innerhalb einer Doppelstunde – die Einschätzung der Unterrichtsqualität erfolgte in dieser Studie viermal in annähernd gleichen zeitlichen Abständen innerhalb der Doppelstunde – vor allem für die Klassenführung stärker variiert als bisherige Studien zeigen, und besonders zu Beginn der Stunde höher ausfällt.

Mashburn et al. (2014) führten eine Experimentalstudie zur zeitlichen Stabilität der Unterrichtsqualität in 47 Klassen (6. Klasse) durch. Zur Einschätzung der Unterrichtsqualität wurde das CLASS-Instrument (Classroom Assessment Scoring System, Pianta und Hamre 2009) verwendet. Die Autorengruppe variierte einerseits die Länge der einzuschätzenden Segmente im Unterricht (10, 20 oder 40 min) und andererseits untersuchten sie in einer weiteren Bedingung zwanzigminütige Segmente innerhalb einer Unterrichtsstunde, die den Ratern in zufälliger Reihenfolge dargeboten wurden. Im Vergleich zu den anderen überprüften Bedingungen fiel die zeitliche Variabilität in der Bedingung mit zufälliger Reihenfolge geringfügig höher aus. Für alle untersuchten Bedingungen ergaben sich moderate bis hohe Zusammenhänge, sowohl innerhalb als auch zwischen beobachteten Doppelstunden.

Inwieweit auch in der Grundschule die Unterrichtsqualität innerhalb einer Doppelstunde variiert, wurde unseres Wissens nach bislang nicht untersucht. Zwar wurden beispielsweise im Rahmen der IGEL-Studie (vgl. Fauth et al. 2014a, 2014b; Decristan et al. 2015) neben einer einführenden Unterrichtsstunde (45 min) zusätzlich zwei Doppelstunden (im Fach Sachunterricht, zwei verschiedene Themen) beobachtet, jedoch erfolgte die Beurteilung der Unterrichtsqualität nur einmal am Ende einer jeden Stunde sowie anhand von jeweils nur einem Item pro Basisdimension. Analysen zur zeitlichen Stabilität bzw. Variabilität der Unterrichtsqualität ließen sich hier lediglich auf Itemebene und über die zwei Doppelstunden hinweg, jedoch nicht innerhalb der Doppelstunden durchführen.

1.4 D-Studien zur Vorhersage der Anzahl notwendiger Unterrichtsstunden zur Erfassung der Unterrichtsqualität

Neben der Schätzung der verschiedenen Varianzkomponenten im Rahmen von G-Studien können die Varianzkomponenten in sogenannten *decision studies* (D-

Studien, vgl. Shavelson und Webb 1991) zur Bestimmung der Reliabilität einer Messung bei gegebener Anzahl von Unterrichtsstunden herangezogen werden. Im Rahmen dieser Entscheidungsstudien lässt sich so u. a. die Anzahl an Unterrichtsstunden ermitteln, die notwendig ist, um die Basisdimensionen der Unterrichtsqualität hinreichend reliabel zu messen. Wie viele Unterrichtsstunden einer Lehrperson beobachtet werden sollten, wird jedoch kaum untersucht (Ausnahme: Praetorius 2014; Praetorius et al. 2014). Praetorius (2014) konnte beispielsweise für den Mathematikunterricht in der Sekundarstufe zeigen, dass für die Erfassung der Unterrichtsqualität mit einer Reliabilität von mindestens 0,70 für die Unterrichtsqualitätsmerkmale Klassenführung und Schülerorientierung lediglich eine Unterrichtsstunde pro Lehrperson beobachtet werden muss. Für die kognitive Aktivierung sind hingegen neun Unterrichtsstunden pro Lehrperson notwendig, um eine Reliabilität von 0,70 und höher zu erreichen. Jentsch et al. (2019) kamen in ihrer Studie zu dem Ergebnis, dass auch für die konstruktive Unterstützung im Mathematikunterricht der Sekundarstufe bis zu sechs Doppelstunden für eine hinreichend reliable Messung notwendig sind. Inwieweit die Ergebnisse aus dem Sekundarschulbereich auch die Grundschule übertragbar sind, soll im Rahmen der vorliegenden Studie untersucht werden.

2 Forschungsfragen und Hypothesen

Der bisherige Forschungsstand zur zeitlichen Stabilität der Unterrichtsqualität lässt keine eindeutige Antwort auf die Frage zu, inwieweit die Ergebnisse bisheriger Generalisierbarkeitsstudien im Sekundarschulunterricht auch für den Grundschulunterricht generalisierbar sind. Im Rahmen der vorliegenden Studie soll untersucht werden, inwieweit sich in Abhängigkeit vom untersuchten Unterrichtsqualitätsmerkmal (Klassenführung, Schülerorientierung bzw. Unterrichtsklima, kognitive Aktivierung) auch für die Grundschule Unterschiede in den Ausprägungen der jeweiligen Dimension zwischen unterschiedlichen Stunden einer Lehrperson, im speziellen Fall zwischen zwei Teilstunden einer Doppelstunde (Mathematikunterricht, 2. Schuljahr) einer Lehrperson in ein und derselben Klasse, bestätigen lassen.

Forschungsfrage Nr. 1 In welchem Ausmaß variieren die Ausprägungen der drei Basisdimensionen von Unterrichtsqualität zwischen zwei Teilstunden einer Doppelstunde?

In Anlehnung an die Ergebnisse aus der Sekundarstufe können je nach Unterrichtsqualitätsmerkmal auch für die Grundschule unterschiedliche Annahmen getroffen werden, die anschließend näher erläutert werden.

2.1 Klassenführung

Da die Klassenführung einer Lehrperson eher auf allgemeinem pädagogischen Wissen basiert und als gegenstandsunabhängiges (vgl. Rakoczy 2008) bzw. generisches Merkmal (vgl. Jentsch et al. 2019) betrachtet wird, könnte auch in der vorliegenden Studie – ähnlich den Ergebnissen aus der Sekundarstufe (s. oben) – eine eher ge-

ringe Variation der Ausprägungen der Merkmale der Klassenführung zwischen den zwei Stunden einer 90-minütigen Unterrichtseinheit einer Lehrperson angenommen werden. Allerdings wurde in bisherigen Studien die zeitliche Stabilität über mehrere Wochen und Monate hinweg untersucht (z. B. Meyer et al. 2011). Erste Befunde zur zeitlichen Stabilität innerhalb einer Doppelstunde deuten jedoch darauf hin, dass die Beobachterratings zur Klassenführung stärker variieren können (vgl. Jentsch et al. 2019) und vor allem zu Beginn der Stunde höher ausfallen. Auch für den Grundschulunterricht kann eine größere Variation in den Beobachterratings angenommen werden. Der Unterricht in der Grundschule zeichnet sich im Vergleich zur Sekundarstufe durch eine höhere Methodenvielfalt (vgl. Götz et al. 2005), durch größere Schwankungen in den beobachtbaren Unterrichtsformen (vgl. Helmke und Weinert 1997) bzw. durch eine stärkere Abwechslung von eher lehrergelenkten mit schülerorientierten, selbstgesteuerten Phasen (vgl. Kammermeyer und Martschinke 2009) aus. Es werden demnach je nach Phase oder Sozialform andere Anforderungen an die Klassenführung der Lehrpersonen gestellt. In schülerorientierten und selbstgesteuerten Phasen kann die Lehrperson sogar einen Teil der Klassenführung (z. B. auf die Einhaltung von Regeln achten) an die Schülerinnen und Schüler abgeben (vgl. Bohl und Kucharz 2010). Der situationale Einfluss sollte somit höher und damit die Stabilität über verschiedene Unterrichtsstunden hinweg geringer sein, was sich in einem höheren stundenspezifischen Varianzanteil widerspiegeln sollte.

H1 Die Basisdimension Klassenführung variiert in der Grundschule in hohem Ausmaß zwischen den beiden Unterrichtsstunden (innerhalb einer Doppelstunde) einer Lehrperson.

2.2 Unterrichtsklima

Ähnlich wie bei der Klassenführung wird auch für das Unterrichtsklima davon ausgegangen, dass dies als ein eher generisches gegenstandsunabhängiges Merkmal über verschiedene Messzeitpunkte hinweg relativ stabil ausgeprägt ist (z. B. Satow 1999). Darauf verweisen auch verschiedene Definitionen, in denen das Unterrichtsklima als ein „relativ überdauerndes“ (z. B. von Saldern 1987, S. 17) Merkmal beschrieben wird, das jedoch nicht völlig stabil ist (vgl. von Saldern 2000). Als Klima werden demnach per Definition nur Merkmale z. B. der Schulumwelt, der Klasse oder des Unterrichts bezeichnet, wenn sie nicht nur episodischen und momentanen Charakter besitzen, sondern über die Zeit relativ stabil und somit charakteristisch für die jeweilige Klasse sind. Dass sich das Unterrichtsklima bzw. die Schülerorientierung (erfasst über Schülerbefragungen) jedoch auch bei gleicher Lehrperson und Klasse zwischen verschiedenen Fächern unterscheiden kann und demnach nicht gegenstandunabhängig ist, zeigen die Ergebnisse von Praetorius et al. (2016). In der Studie wurden Schülerdaten von 25 Klassen analysiert, die in den Fächern Deutsch und Englisch von derselben Lehrperson unterrichtet wurden. Zwar überwog im Rahmen der G-Studien der fachübergreifende Varianzanteil sowohl für die Klassenführung als auch für die Schülerorientierung, für die Schülerorientierung konnten jedoch auch zusätzlich substantielle fachspezifische Anteile bestätigt werden, was auf eine größere Bedeutung situationaler und kontextueller Bedingungen

für dieses Unterrichtsqualitätsmerkmal hinweist als bislang angenommen. Da im Rahmen der vorliegenden Studie jedoch die zeitliche Stabilität des Unterrichtsklimas in einer 90-minütigen Unterrichtseinheit in ein und demselben Fach bei ein und derselben Lehrperson in ein und derselben Klasse untersucht wird, sollte der unterrichtsstundenspezifische Varianzanteil relativ gering sein.

H2 Die Basisdimension Unterrichtsklima variiert in der Grundschule in geringem Ausmaß zwischen den beiden Unterrichtsstunden (innerhalb einer Doppelstunde) einer Lehrperson.

2.3 Kognitive Aktivierung

Ähnlich wie für die Sekundarstufe wird auch für die Grundschule angenommen, dass es sich bei der kognitiven Aktivierung eher um ein inhalts- bzw. gegenstandsabhängiges Merkmal handelt (vgl. Gabriel 2014; Praetorius 2014; Rakoczy 2008), das sowohl zwischen verschiedenen Fächern und Inhalten als auch vermutlich in Abhängigkeit von der Jahrgangsstufe unterschiedlich konkretisiert werden sollte (vgl. Minnameier et al. 2015; Klieme und Rakoczy 2008; Rakoczy 2008). So kann davon ausgegangen werden, dass beispielsweise eine kognitiv herausfordernde Aufgabe im Fach Mathematik anders operationalisiert werden muss als im Fach Deutsch oder Englisch (vgl. Klieme 2006). Ähnlich kann auch für Einführungs- und Übungsphasen argumentiert werden, in denen kognitive Aktivierung nicht dasselbe bedeuten muss (vgl. Praetorius 2014). Dass sich in Abhängigkeit der Unterrichtsphase beispielsweise das kognitive Niveau der Fragen der Lehrpersonen („Art der Fragen“), als ein Merkmal kognitiv aktivierenden Unterrichts, unterscheidet, konnten beispielsweise Hess und Lipowsky (2020) anhand von Daten aus der PERLE-Studie (Fach Deutsch) zeigen. So dienen die Fragen der Lehrpersonen im öffentlichen Unterricht stärker der Auseinandersetzung mit den Textinhalten. In Schülerarbeitsphasen wird ein Großteil der Fragen eher dazu genutzt, die Abläufe zu organisieren und aufrechtzuerhalten. Denkfragen – also Fragen, die die Schülerinnen und Schüler kognitiv herausfordern bzw. Wissen aufbauen und sichern sollen – kommen in beiden Unterrichtsphasen kaum vor, in Schülerarbeitsphasen allerdings noch seltener als im öffentlichen Unterricht. Für das Unterrichtsqualitätsmerkmal kognitive Aktivierung wird trotzdem – ausgehend auch von bisherigen Ergebnissen – angenommen, dass die Ausprägungen in einem hohen Ausmaß innerhalb einer 90-minütigen Unterrichtseinheit einer Lehrperson variieren, der stundenspezifische Varianzanteil also relativ hoch ausfällt.

H3 Die Basisdimension kognitive Aktivierung variiert in der Grundschule in hohem Ausmaß zwischen den beiden Unterrichtsstunden (innerhalb einer Doppelstunde) einer Lehrperson.

Die bisherigen Ergebnisse im Sekundarschulbereich lassen keine hinreichend zuverlässigen Aussagen darüber zu, inwieweit eine Doppelstunde bzw. zwei aufeinander aufbauende Teilstunden ausreichen, um von einer zuverlässigen Schätzung der Unterrichtsqualität in der Grundschule auszugehen. In einem weiteren Schritt sollen

im vorliegenden Beitrag deswegen explorative Entscheidungsstudien (sog. D-Studien, vgl. Shavelson und Webb 1991; Praetorius 2014; Praetorius et al. 2014; Jentsch et al. 2019) durchgeführt werden, um zu überprüfen, ob – ähnlich den Ergebnissen im Sekundarschulbereich (vgl. Praetorius 2014) – auch in der Grundschule die beiden Basisdimensionen Klassenführung und Unterrichtsklima mittels einer Unterrichtsstunde hinreichend reliabel erfasst werden können und ob für die kognitive Aktivierung deutlich mehr Unterrichtsstunden benötigt werden.

Forschungsfrage Nr. 2 Wie viele Unterrichtsstunden pro Lehrperson sind nötig, um die drei Basisdimensionen der Unterrichtsqualität im Mathematikunterricht im 2. Schuljahr hinreichend zuverlässig zu erfassen?

3 Methodisches Vorgehen

3.1 Stichprobe

Die Überprüfung der Forschungsfragen erfolgt anhand der Daten aus der Videostudie Mathematik der *PERLE-Studie* (*Persönlichkeits- und Lernentwicklung von Grundschulkindern*; vgl. Greb et al. 2007; Lipowsky et al. 2013). Die Videostudie Mathematik fand im zweiten Schuljahr (März 2008) statt (vgl. Greb et al. 2009; Lotz et al. 2011). Um die Vergleichbarkeit der einzelnen Lektionen in Mathematik (2. Schuljahr) zu gewährleisten, wurden im Vorfeld der Videoaufzeichnungen sowohl (1) die Aufnahmeprozeduren durch ein Kameraskript als auch (2) das Thema der Unterrichtsstunde „Einführung in die Multiplikation“ festgelegt (zusf. Gabriel 2014; Lotz et al. 2013b). Die Daten zur Unterrichtsqualität von insgesamt 36 Lehrpersonen[3] bilden die Grundlage für die nachfolgenden Analysen. Von jeder Lehrperson (♀ = 34; ♂ = 2) wurden 90 Minuten Mathematikunterricht (Doppelstunde) in ein und derselben Klasse beobachtet. 25 Lehrpersonen sind zwischen 35 und 55 Jahre alt, drei sind unter 35 und eine über 55 Jahre alt. Im Mittel haben die teilnehmenden Lehrpersonen 16,5 Jahre Berufserfahrung ($N = 26$; $Min = 0{,}5$; $Max = 35{,}0$; $SD = 10{,}6$). Die Angaben zu Alter und Berufserfahrung der Lehrpersonen stammen aus einem Lehrerfragebogen vom November 2006. Von sechs Lehrpersonen liegen keine Angaben zu Alter und Berufserfahrung vor. Von einer Lehrperson wurde der Mathematikunterricht in zwei unterschiedlichen zweiten Klassen gefilmt.

[3] In Gabriel et al. (2015) wurde eine kleinere Stichprobe (Zufallsstichprobe) verwendet, da es hier um die Frage ging, inwieweit sich Unterschiede in der Ausprägung der Klassenführung zwischen den Fächern Deutsch (Videostudie Deutsch, 1. Schuljahr) und Mathematik (Videostudie Mathematik, 2. Schuljahr) zeigen. Für die dort einbezogenen 21 Lehrpersonen liegen für beide Fächer Daten vor. Im vorliegenden Beitrag geht es im Unterschied zu Gabriel et al. (2015) darum, zu überprüfen, ob sich die Einschätzungen der Raterinnen und Rater über die drei Basisdimensionen der Unterrichtsqualität in Abhängigkeit von der jeweiligen Unterrichtsstunde im Fach Mathematik unterscheiden. Aus diesem Grund können die Daten von allen 36 Lehrpersonen genutzt werden.

3.2 Die hoch inferenten Ratingsysteme

Für jedes der in Tab. 1 aufgeführten Merkmale wurde die Grundidee des Merkmals beschrieben sowie diverse verhaltensnahe Indikatoren/Negativindikatoren formuliert (vgl. Gabriel 2014; Gabriel und Lipowsky 2013a, 2013b; Lauterbach et al. 2013). Da es sich bei den Basisdimensionen Klassenführung und Unterrichtsklima um generische Merkmale handelt (s. oben), wurden sie als fachunspezifische Unterrichtsmerkmale operationalisiert, d. h., weder nahmen die Indikatoren auf fachspezifische Besonderheiten Bezug noch erforderte das Rating spezifische fachwissenschaftliche oder fachdidaktische Kenntnisse von den Raterinnen und Ratern. Anders war dies bei der kognitiven Aktivierung, die von zwei studentischen Hilfskräften (höheres Studiensemester) mit entsprechend mathematikdidaktischem Hintergrund eingeschätzt wurde.

Die Entwicklung der Ratingsysteme erfolgte auf Basis theoretischer Grundlagen und Ergebnisse unterschiedlicher Studien (vgl. Gabriel 2014; Gabriel und Lipowsky

Tab. 1 Überblick über die hoch inferenten Merkmale, deskriptive Ergebnisse sowie Ergebnisse der Korrelationsstudien für die drei Basisdimensionen (Einzeldimensionen und Skalenebene)

	Hoch inferente Merkmale	Relativer G-Koeffizient	MW (SD) 1. Stunde	MW (SD) 2. Stunde	Bivariate Korrelation 1. Stunde/ 2. Stunde
Klassenführung[a]	(1) Allgegenwärtigkeit der Lehrperson	0,89	3,30 (0,58)	2,93 (0,71)	0,34**
	(2) Gruppenfokus	0,88	3,44 (0,50)	3,41 (0,58)	0,79***
	(3) Zeitmanagement/ effiziente Zeitnutzung	0,93	3,62 (0,62)	3,46 (0,60)	0,57***
	(4) Übergangsmanagement (Managing Transition)	0,91	3,29 (0,66)	3,03 (0,66)	0,66***
	(5) Regelklarheit bzw. -verwendung	0,89	3,40 (0,60)	3,26 (0,59)	0,64***
	(6) Störungsfreiheit	0,86	3,28 (0,66)	3,07 (0,71)	0,65***
	(7) Effektiver Umgang mit auftretenden Unterrichtsstörungen	0,88	3,36 (0,57)	3,08 (0,71)	0,63***
	(8) Gegenseitige Anerkennung der Schüler	0,87	3,44 (0,59)	3,25 (0,72)	0,62***
	Gesamtskala Klassenführung (8 Items)	–	3,41 (0,43)	3,19 (0,45)	0,73***
Unterrichtsklima	(1) Humorvolle Lernatmosphäre	0,83	1,83 (0,67)	1,58 (0,64)	0,74**
	(2) Anerkennung der Schüler durch die Lehrperson	0,87	3,56 (0,62)	3,50 (0,67)	0,68***
	(3) Fürsorglichkeit, Herzlichkeit und Wärme	0,89	2,61 (0,59)	2,64 (0,58)	0,81***
	Gesamtskala Unterrichtsklima (3 Items)	–	2,67 (0,51)	2,57 (0,48)	0,81***

Tab. 1 (Fortsetzung)

	Hoch inferente Merkmale	Relativer G-Koeffizient	MW (SD) 1. Stunde	MW (SD) 2. Stunde	Bivariate Korrelation 1. Stunde/ 2. Stunde
Kognitive Aktivierung	(1) Exploration von Vorwissen oder vorunterrichtlichen Vorstellungen	0,90	1,92 (0,69)	1,07 (0,24)	0,29 ns
	(2) Exploration der Denkweisen der Schüler	0,90	1,92 (0,74)	1,44 (0,52)	0,56***
	(3) Kognitiv herausfordernder Umgang mit Schülerbeiträgen	0,87	1,71 (0,85)	1,29 (0,44)	0,27 ns
	(4) Kognitiv aktivierende Aufgaben und Probleme	0,91	2,15 (0,71)	1,93 (0,51)	0,29 ns
	(5) Begründungspflicht/ Insistieren auf Erklärung und Begründung	0,92	1,92 (0,92)	1,43 (0,54)	0,49**
	(6) Unterstützung kognitiver Selbstständigkeit	0,84	2,15 (0,82)	1,94 (0,70)	0,60***
	Gesamtskala kognitive Aktivierung (6 Items)	–	1,96 (0,59)	1,52 (0,31)	0,51***

$N = 36$ Lehrpersonen
Antwortformat: vierstufig
MW Mittelwert, *SD* Standardabweichung
*** $p \leq 0{,}001$; ** $p \leq 0{,}01$; ns = nicht signifikant
[a]Grundlage bilden 34 Lehrpersonen aufgrund von Missings

2013a, 2013b; Lauterbach et al. 2013). Bei der Entwicklung des hoch inferenten Ratingsystems zur Analyse der Klassenführung wird ausgehend von den Studien von Kounin (2006) und in Anlehnung an Seidel (2009) bzw. Waldis et al. (2010) einem integrativen Ansatz gefolgt. Demnach wird die Klassenführung nicht eindimensional beispielsweise nur in Bezug auf den Umgang mit Störungen erfasst, weil sie damit nur auf ein Disziplinmanagement reduziert wäre, sondern es werden sowohl proaktive als auch reaktive Kriterien berücksichtigt (Gabriel und Lipowsky 2013a; zusf. Gabriel 2014; Gabriel et al. 2015). Unter dem Begriff Unterrichtsklima wird in Anlehnung an Bessoth (1989) primär ein lehrerabhängiges „soziales Klima" verstanden (S. 4). Demzufolge gelten in einer Klasse interpersonale Beziehungen der Schülerinnen und Schüler zur Lehrperson sowie die Schüler-Mitschüler-Beziehungen als zentrale Dimensionen bzw. Merkmale des Klimas (Zumhasch 2006). Dabei kann angenommen werden, dass sowohl Lehrpersonen als auch Schülerinnen und Schüler das Unterrichtsklima der Klasse in entscheidendem Ausmaß beeinflussen können (Lange et al. 1983). Neben Aspekten der Lehrer-Schüler-Beziehung und der Schüler-Schüler-Beziehung gelten allgemeine Unterrichtsmerkmale (z. B. Leistungsdruck) als bedeutsam für das Unterrichtsklima (von Saldern und Littig 1985). Als kognitiv aktivierend wird im Rahmen der Studie – in Anlehnung an theoretische Grundlagen von Piaget (1964), Vygotsky (1977, 1985) und Aebli (1983, 1987) – ein Unterricht verstanden, in dem die Lehrperson das (Vor-)Wissen der Schülerinnen und Schüler erkundet und einbezieht, anspruchsvolle Probleme und Aufgaben bear-

beiten lässt sowie zu einem vertieften Nachdenken über die Inhalte und zu einem gehaltvollen mathematischen Diskurs anregt (vgl. Lauterbach et al. 2013).

3.3 Die Datenerhebung

Die hoch inferenten Merkmale zur effektiven Klassenführung, zum Unterrichtsklima und zur kognitiven Aktivierung sind in Tab. 1 dargestellt. Jeweils zwei geschulte Raterinnen und Rater schätzten mithilfe hoch inferenter Ratingsysteme unabhängig voneinander insgesamt acht Merkmale der Klassenführung, drei Merkmale des Unterrichtsklimas und sechs Merkmale der kognitiven Aktivierung auf einer vierstufigen Skala ein. Die Einschätzungen erfolgten jeweils nach der 1. Stunde und der 2. Stunde einer 90-minütigen Unterrichtseinheit zur „Einführung in die Multiplikation“ (2. Schuljahr). Damit werden zwei aufeinander aufbauende Stunden einer Lehrperson in ein und derselben Klasse untersucht.

Zur Überprüfung der Qualität der hoch inferenten Daten wurde für jedes der 17 Unterrichtsmerkmale in Tab. 1 die Übereinstimmung zwischen den beiden Raterinnen und Ratern berechnet. Der relative Generalisierbarkeitskoeffizient (G-Koeffizient)[4] liegt für die Klassenführungsmerkmale zwischen 0,86 und 0,93, für die Merkmale des Unterrichtsklimas zwischen 0,83 und 0,89 und für die Merkmale der kognitiven Aktivierung zwischen 0,84 und 0,92 (siehe Tab. 1; vgl. auch Gabriel 2014; Lauterbach et al. 2013) und kann als zufriedenstellend angesehen werden (vgl. Lotz et al. 2013a).

3.4 Methode der G-Studien – Untersuchungsdesign

Um Aussagen über die Stabilität der drei Basisdimensionen der Unterrichtsqualität treffen zu können, wurden ebenfalls Generalisierbarkeitsstudien (G-Studien, vgl. Cronbach et al. 1972; Shavelson und Webb 1991) mithilfe der Software urGENOVA 2.1 (vgl. Brennan 2001) durchgeführt[5]. Die Software erlaubt den Umgang mit fehlenden Werten. Im Rahmen von urGENOVA wird mit fehlenden Werten so umgegangen, dass die jeweils vorhandene Anzahl an Items pro Person analysiert wird (vgl. Praetorius 2014, S. 80 f). Für die Basisdimensionen Unterrichtsklima und kognitive Aktivierung liegen keine fehlenden Werte vor. Für die Basisdimension Klassenführung liegt der Anteil fehlender Werte bei ca. 0,5 %. Auf eine Imputation der fehlenden Werte wurde verzichtet, da diese dazu führen würde, dass für

[4] Dieser gibt an, inwieweit die Einschätzungen der jeweils betrachteten hoch inferenten Merkmale der drei Basisdimensionen der Unterrichtsqualität durch die beiden Raterinnen und Rater ausreichende Generalisierbarkeit (Zuverlässigkeit) aufweisen (vgl. Lotz et al. 2013a). Analog zum Reliabilitätskoeffizienten der klassischen Testtheorie kann der G-Koeffizient einen Wert zwischen „Null“ und „Eins“ annehmen. Ein relativer G-Koeffizient von ≥0,70 gilt als Kriterium für eine zufriedenstellende Qualität der Daten.

[5] Die hier berechneten Modelle weichen von den Modellen zur Berechnung der relativen G-Koeffizienten (siehe Abschn. 3.3) ab, da unterschiedliche Ziele verfolgt wurden bzw. die Modelle verschiedene Funktionen erfüllen. Die Berechnung des relativen G-Koeffizienten diente ausschließlich der Analyse der Inter-Rater-Reliabilität, demnach stand die Analyse der Rater-Effekte im Vordergrund. Um einen größeren Anteil der beobachteten Varianz zu erklären, wurden in den Generalisierbarkeitsstudien zusätzlich die stundenspezifische Variation innerhalb der Doppelstunde und die Variationen in den Itemausprägungen innerhalb der drei Skalen berücksichtigt.

ein bestimmtes Merkmal , das aus Sicht der zwei Rater übereinstimmend nicht beobachtbar bzw. beurteilbar war, Werte für eine Lehrperson geschätzt werden (vgl. Praetorius 2014). Aus diesem Grund liegen für das Unterrichtsqualitätsmerkmal Klassenführung nur Daten von 34 Lehrpersonen vor.

Die Fragestellung Nr. 1 sowie die entsprechenden Hypothesen werden durch teilweise geschachtelte Generalisierbarkeitsstudien (G-Studien) untersucht. Das Ziel der G-Studie, die für jede der drei Basisdimensionen der Unterrichtsqualität separat berechnet wurde, stellt die Schätzung zeitstabiler Klassenunterschiede dar (= wahre Varianz), d. h. es wird untersucht, wieviel Varianz auf tatsächliche Unterschiede zwischen den Lehrpersonen bzw. Klassen zurückgeführt werden kann (*l*). Zusätzlich liefern die G-Studien Informationen darüber, inwieweit die Beobachterratings in Abhängigkeit von dem jeweiligen Messzeitpunkt innerhalb der Doppelstunde variieren, wobei die beiden Meszeitpunkte in ein und der selben Klasse geschachtelt in Lehrpersonen (*u:l*) in die Analysen eingehen. Die Facette Items (*i*) wurde aufgenommen, da die Analysen auf Itemebene durchgeführt wurden. Somit lassen sich zusätzlich Aussagen darüber treffen, inwieweit z. B. die Itemschwierigkeiten zwischen den Messzeitpunkten variieren oder Lehrpersonen Stärken und Schwächen z. B. in Bezug auf einzelne Unterrichtsmerkmale zeigen. Da ja nach Unterrichtsqualitätsmerkmal verschiedene Raterinnen bzw. Rater die Doppelstunden beurteilten,

Tab. 2 Ergebnisse G-Analysen für die drei Basisdimensionen in Mathematik (2. Schuljahr)

	Klassenführung		Unterrichtsklima		Kognitive Aktivierung	
Varianzkomponenten	VK	%[b]	VK	%[b]	VK	%[b]
Stabile Komponenten						
l	0,11	25	0,15	12	0,03	5
li	0,02	5	0,08	6	0,07	12
Stundenspezifische Komponenten						
u:l	0,05	11	0,01	1	0,18	31
(u:l)i	0,07	16	0,06	5	0,08	14
Raterspezifische Komponenten						
r	0	0	0	0	0[a]	0
rl	0,02	5	0[a]	0	0[a]	0
r(u:l)	0,02	5	0	0	0,02	3
ri	0[a]	0	0[a]	0	0,01	2
rli	0,03	7	0,07	6	0,03	5
Itemspezifische Komponenten						
i	0,02	5	0,81	64	0,05	8
Residuum						
(u:l)ri, e	0,10	23	0,08	6	0,12	20
Gesamtvarianz	0,44	–	1,26	–	0,59	–

$N = 36$ (die Analysen für die Klassenführung basieren auf $N = 34$ aufgrund von Missings)
l Lehrperson, *u* Teil der Unterrichtsstunde, *r* Rater, *i* Item, *e* Fehler, *VK* Varianzkomponente, *%* relativer Anteil der Varianzkomponente an der Gesamtvarianz
[a]Eine sehr kleine negative Varianz wurde auf null gesetzt (vgl. Brennan 2001)
[b]Die Varianzkomponenten sind an der Gesamtvarianz relativiert (mit 100 multipliziert, vgl. Shavelson und Webb 1991)

wurden neben den Items auch die Rater (*r*) als Facette aufgenommen. Durch die hier einbezogenen Varianzquellen ergibt sich ein Zwei-Facetten-Design ((*u:l*) × *r* × *i*-Design), mit dem es möglich ist, insgesamt jeweils elf Varianzquellen (vier Haupteffekte inkl. deren Interaktionen) für jede der drei Basisdimensionen voneinander zu unterscheiden (siehe Tab. 2). Beispielsweise zerfällt die Variation der Item-Ausprägungen einer Skala im Rahmen des vorliegenden Designs in fünf Varianzkomponenten: in den Item-Haupteffekt (*i*), die Interaktion zwischen Item und beobachteter Lehrperson innerhalb ein und derselben Klasse (*li*), sowie die stundenspezifische ((*u:l*)*i*) und raterspezifische Item-Variation (*rli*). Die letzte Interaktion (*u:l*)*ri, e* ist konfundiert mit der Residualvarianz konfundiert und kann deswegen inhaltlich nicht interpretiert werden (vgl. Shavelson und Webb 1991).

3.5 Methodisches Vorgehen bei den D-Studien

Zur Beantwortung der Frage, wie viele Unterrichtsstunden pro Lehrperson notwendig sind, um die drei Basisdimensionen der Unterrichtsqualität in der Grundschule hinreichend zuverlässig zu schätzen (Forschungsfrage Nr. 2), wurde in explorativen Simulationsstudien (decision study, Shavelson und Webb 1991) in EXCEL die Anzahl der beobachteten Unterrichtsstunden variiert und jeweils der relative G-Koeffizient geschätzt. Es wurden für jede Basisdimension der Unterrichtsqualität Analysen für ein bis zwölf Unterrichtsstunden (bzw. eine bis sechs Doppelstunden) pro Lehrperson durchgeführt, wobei die Anzahl der Rater ($N = 2$) sowie die jeweilige Anzahl an Items auf die in der Studie tatsächlich vorhandene Anzahl (siehe Tab. 2) fixiert wurde.

4 Ergebnisse

4.1 Deskriptive Ergebnisse und Ergebnisse der Korrelationsstudien

Tab. 1 stellt die Mittelwerte und Standardabweichungen für die beobachteten Unterrichtsqualitätsmerkmale (Item- und Skalenebene) für die beiden Teilstunden sowie – der Vollständigkeit halber – deren Korrelationen dar. Im Mittel sind die Merkmale der Klassenführung für beide Teilstunden der 90-minütigen Unterrichtseinheit deutlich positiv ausgeprägt, sowohl auf Ebene der einzelnen Unterrichtsmerkmale als auch auf Skalenebene ($MW_{Klassenführung\ \ 1.\ \ Stunde} \geq 3{,}41$; $MW_{Klassenführung\ \ 2.\ Stunde} \geq 3{,}19$), gefolgt vom Unterrichtsklima ($MW_{Unterrichtsklima\ 1.\ Stunde} \geq 2{,}67$; $MW_{Unterrichtsklima\ 2.\ Stunde} \geq 2{,}57$). Die kognitive Aktivierung ist im Vergleich dazu in beiden Teilstunden eher gering ausgeprägt ($MW_{kognitive\ Aktivierung\ \ 1.\ \ Stunde} \geq 1{,}96$; $MW_{kognitive\ Aktivierung\ \ 2.\ Stunde} \geq 1{,}52$). Für alle drei Basisdimensionen der Unterrichtsqualität wurde die 1. Stunde der 90-minütigen Unterrichtseinheit von den zwei Raterinnen und Ratern im Mittel positiver eingeschätzt. Die bivariaten Korrelationen zwischen den zwei Teilstunden der 90-minütigen Unterrichtseinheit liegen je nach Basisdimension zwischen $r = 0{,}51$ und $r = 0{,}81$; auf Ebene der einzelnen Unterrichtsmerkmale teilweise niedriger und nicht signifikant (z. B. kognitiv aktivierende Aufgaben und Probleme: $r = 0{,}29\,ns$).

4.2 Variation der Unterrichtsqualitätsmerkmale zwischen zwei Teilstunden einer 90-minütigen Unterrichtseinheit – Ergebnisse der G-Studien

Die Ergebnisse der G-Studien (vgl. Tab. 2) zeigen, dass sich die Unterschiede in den Raterurteilen für alle drei Basisdimensionen der Unterrichtsqualität zu einem gewissen Anteil auf Unterschiede zwischen den Lehrpersonen in den spezifischen Klassen zurückführen lassen (stabile Varianz: *l; li*). Für die Klassenführung liegt der Anteil der stabilen Varianz bei 30 %, für das Unterrichtsklima bei 18 % und für die kognitive Aktivierung bei 17 %. Das macht etwa ein Drittel der Gesamtvarianz in den Beobachterratings für das Merkmal Klassenführung bzw. etwa ein Fünftel für die Merkmale Unterrichtsklima und kognitive Aktivierung aus, was einer großen Effektstärke entspricht (vgl. Döring und Bortz 2016).

Für das Unterrichtsklima entfällt nur wenig Varianz auf stundenspezifische Unterschiede ($u{:}l + (u:l)i = 6\,\%$), während die stundenspezifischen Anteile für die Klassenführung (26 %) und die kognitive Aktivierung (45 %) vergleichsweise hoch sind. Je höher der stundenspezifische Anteil, desto deutlicher variieren die Ausprägungen der Unterrichtsqualitätsmerkmale zwischen den beiden Stunden der 90-minütigen Unterrichtseinheit einer Lehrperson. Das bedeutet, dass sich vor allem die Klassenführung und die kognitive Aktivierung in Abhängigkeit vom Beobachtungszeitpunkt (1. Stunde/2. Stunde der 90-minütigen Unterrichtseinheit) in einem hohen Maß unterscheiden, während das Unterrichtsklima eine relativ hohe Stabilität zwischen den beiden Stunden aufweist. Die Hypothesen können demnach bestätigt werden.

Auch graphisch bestätigt sich, dass die Ausprägung der kognitiven Aktivierung am stärksten zwischen den beiden Stunden der 90-minütigen Unterrichtseinheit variiert (vgl. Abb. 1), gefolgt von der Klassenführung (vgl. Abb. 2), während für das Unterrichtsklima (vgl. Abb. 3) die Variation eher gering ist. Für das Unterrichtsklima zeigt sich (mit Ausnahme einzelner Lehrpersonen, z. B. Nr. 11, 21 oder 26), dass die Rater die erste und zweite Stunde der 90-minütigen Unterrichtseinheit relativ ähnlich einschätzen. In Abb. 1, 2 und 3 sind jeweils die Skalenmittelwerte (Y-Achse) über beide Raterinnen und Rater, getrennt nach der jeweiligen Stunde (1. Stunde vs. 2. Stunde) der 90-minütigen Unterrichtseinheit abgebildet. Die Lehrpersonen (X-

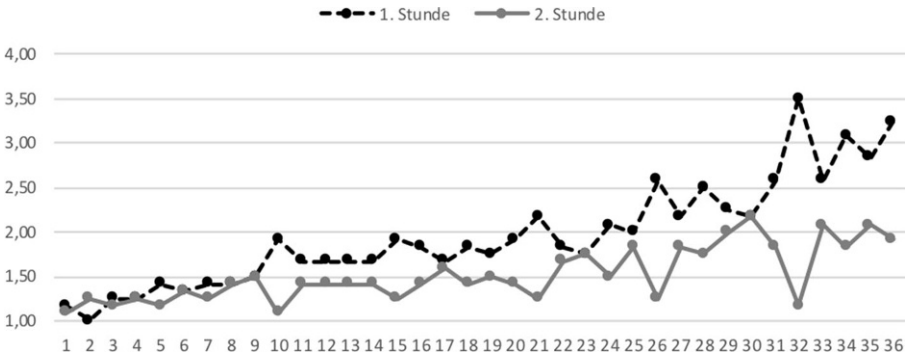

Abb. 1 Unterschiede in der Höhe der Ausprägung zwischen 1. und 2. Stunde für die kognitive Aktivierung (Skalenebene)

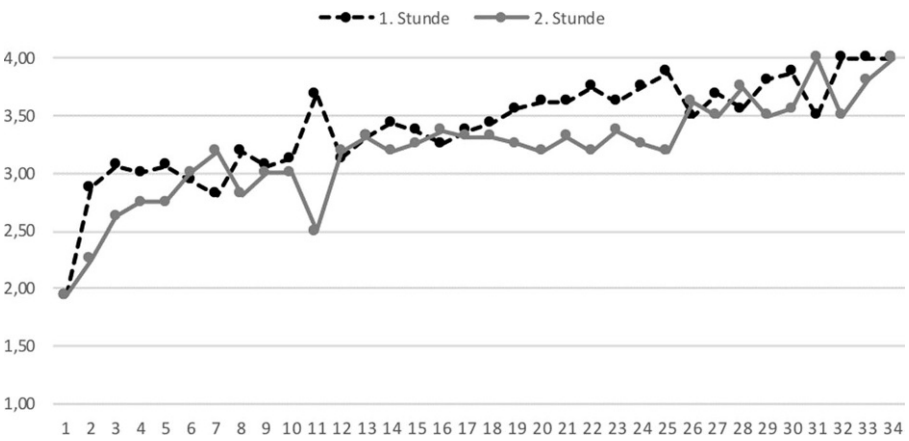

Abb. 2 Unterschiede in der Höhe der Ausprägung zwischen 1. und 2. Stunde für die Klassenführung (Skalenebene)

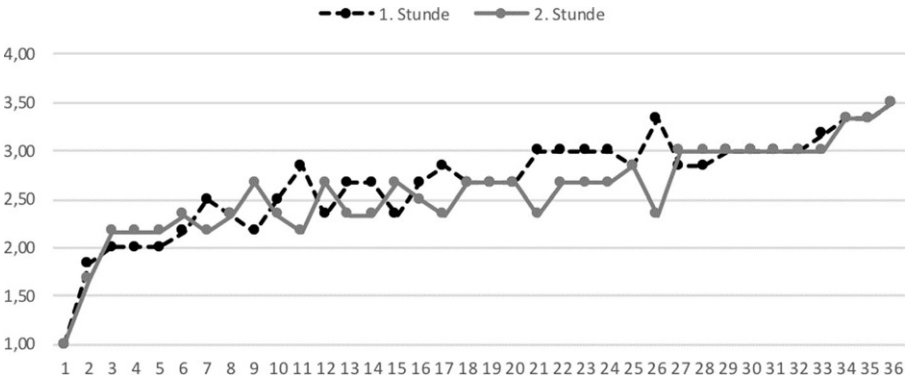

Abb. 3 Unterschiede in der Höhe der Ausprägung zwischen 1. und 2. Stunde für das Unterrichtsklima (Skalenebene)

Achse) sind je nach Abbildung aufsteigend nach der Ausprägung ihrer Klassenführung, des Unterrichtsklimas bzw. der kognitiven Aktivierung (jeweils gemittelt über alle Items) sortiert, die Nummerierung ist demnach nicht gleichbedeutend mit der ID einer Lehrperson.

Der Anteil, der auf raterspezifische Unterschiede zurückgeführt werden kann (*r, rl, r(u:l), ri, ril*), liegt je nach Unterrichtsqualitätsmerkmal zwischen 6 % und 17 %, der Anteil der residualen Varianz zwischen 6 % und 23 %. Auffällig ist der Haupteffekt der Items (*i*) von 64 % für das Unterrichtsklima, der die zeitstabilen Unterschiede in den Itemschwierigkeiten innerhalb einer Skala beschreibt. Relativ hohe Varianzanteile in Höhe von 23 % bzw. 20 % entfallen bei der Klassenführung und der kognitiven Aktivierung auf die Residuen, was wiederum darauf hinweist, dass zusätzliche potentielle Fehlerquellen in weiteren Studien berücksichtigt werden sollten.

4.3 Ergebnisse der explorativen Simulationsstudien (D-Studien)

Anschließend wird dargestellt, wie viele Unterrichtsstunden je nach Basisdimension der Unterrichtsqualität in Bezug auf die hier einbezogenen Daten notwendig sind, um Reliabilitäten (relative G-Koeffizienten) von 0,70, 0,80 oder 0,90 zu erreichen (vgl. Tab. 3, fett hervorgehoben). Während für das Unterrichtsklima im 2. Schuljahr Mathematikunterricht mindestens zwei Unterrichtsstunden (bzw. eine Doppelstunde) einer Lehrperson ausreichen, um einen relativen G-Koeffizienten von mindestens 0,70 zu erzielen, müssen für die Klassenführung und die kognitive Aktivierung mindestens drei Unterrichtsstunden (bzw. zwei Doppelstunden) beobachtet werden. Will man die Unterrichtsqualität mit einer Reliabilität von mindestens 0,80 und höher erfassen, so sind für das Unterrichtsklima drei Unterrichtsstunden (bzw. zwei Doppelstunden) und für die Klassenführung und die kognitive Aktivierung fünf Unterrichtsstunden (bzw. drei Doppelstunden) pro Lehrperson notwendig. Sofern man als Mindestreliabilität 0,90 festsetzt, lassen sich die Klassenführung und die kognitive Aktivierung erst mittels elf bzw. zehn Unterrichtsstunden einer Lehrperson (sechs bzw. fünf Doppelstunden) reliabel erfassen, das Unterrichtsklima lässt sich bereits mit drei Unterichtsstunden einer Lehrperson (bzw. zwei Doppelstunden) erfassen.

5 Zusammenfassung und Diskussion

5.1 Die Variation der drei Basisdimensionen der Unterrichtqualität innerhalb von Doppelstunden

In der vorliegenden Studie wurde die zeitliche Stabilität der drei Basisdimensionen der Unterrichtsqualität (Klassenführung, Unterrichtsklima und kognitive Aktivierung) innerhalb einer 90-minütigen Unterrichtseinheit zur „Einführung in die Multiplikation“ im 2. Schuljahr untersucht. Wie aus den Korrelationsanalysen ersichtlich, variiert die Stabilität innerhalb der 90-minütigen Unterrichtseinheit in Abhängigkeit des untersuchten Qualitätsmerkmals sowohl auf Item- als auch auf Skalenebene (vgl. Tab. 1). Die zusätzlich durchgeführten G-Studien bestätigen dieses Ergebnis: Vor allem die Ausprägungen der Unterrichtsqualitätsmerkmale Klassenführung und kognitive Aktivierung variieren zwischen den beiden Stunden der 90-minütigen Unterrichtseinheit einer Lehrperson, was aus den hohen stundenspezifischen Varianzanteilen hervorgeht (vgl. Tab. 2). Das Unterrichtsklima ist hingegen relativ stabil. Damit können die Befunde aus der Sekundarstufe zur zeitlichen Stabilität von Unterrichtsqualitätsmerkmalen für die Grundschule – im vorliegenden Fall für 90 Minuten Mathematikunterricht im 2. Schuljahr – nur teilweise bestätigt werden: der stundenspezifische Varianzanteil für die Basisdimension Klassenführung fällt deutlich höher aus, was zur Folge hat, dass für eine hinreichend reliable Erfassung der Klassenführung im 2. Schuljahr Mathematik mindestens drei Einzelstunden bzw. zwei Doppelstunden beobachtet werden sollten, wie die Ergebnisse der Entscheidungsstudie (Abschn. 4.3) zeigen. Anschließend werden die Befunde diskutiert.

Im Rahmen der G-Studien erwies sich der stundenspezifische Varianzanteil (*u:l, (u:l)i*) für das Unterrichtsqualitätsmerkmal *Klassenführung* als relativ hoch, d. h. ein

Tab. 3 Relative G-Koeffizienten für D-Studien mit 1–12 Unterrichtsstunden (bzw. 1–6 Doppelstunden) pro Lehrperson für die drei Basisdimensionen der Unterrichtsqualität

Anzahl Doppelstunden	Anzahl Stunden (45 Min)	Klassenführung	Unterrichtsklima	Kognitive Aktivierung
1	1	0,45	0,63	0,47
	2	0,62	**0,77**	0,64
2	3	**0,71**	**0,84**	**0,72**
	4	0,77	0,87	0,78
3	5	**0,80**	**0,90**	**0,81**
	6	0,83	0,91	0,84
4	7	0,85	0,92	0,86
	8	0,87	0,93	0,88
5	9	0,88	0,94	0,89
	10	0,89	0,94	**0,90**
6	11	**0,90**	0,95	0,91
	12	0,91	0,95	0,91

großer Teil der Varianz kann auf bestehende Unterschiede zwischen den beiden Unterrichtsstunden der Doppelstunde zurückgeführt werden. Die Rater schätzen die Klassenführung zum Ende der Doppelstunde niedriger ein als noch zu Beginn. Damit lassen sich die Ergebnisse aus der Sekundarstufe (vgl. Jentsch et al. 2019) auch für die zeitliche Variabilität der Klassenführung innerhalb einer Doppelstunde Mathematikunterricht in der Grundschule bestätigen. Ein Grund für die niedrigeren Mittelwerte für die Klassenführungsmerkmale im zweiten Teil der 90-minütigen Unterrichtseinheit könnte sein, dass die Lehrpersonen die erste Stunde und damit den Unterrichtseinstieg bzw. die Phase der Einführung in die Multiplikation sorgfältiger und genauer geplant haben als die zweite Stunde, in der es häufig um die Übung der neuen Rechenart z. B. im Rahmen von Stationenlernen geht. Dies erklärt auch die höheren Mittelwerte für die erste Teilstunde (vgl. Tab. 1) für einen Großteil der Lehrpersonen. Ähnlich argumentieren auch Jentsch et al. (2019), die vermuten, dass Maßnahmen zur Herstellung der Aufmerksamkeit vor allem zu Beginn einer Unterrichtsstunde erfolgen oder zum Ende einer Unterrichtsstunde ausbleiben. Andererseits kann vermutet werden, dass, in Abhängigkeit von der Phase im Unterricht bzw. der Sozialform (Frontalunterricht, Einzel-, Partner- oder Gruppenarbeit), die Klassenführung unterschiedlich gut beobachtbar ist. Eine mögliche Ursache liegt darin, dass sich bei der Entwicklung des hoch inferenten Ratingsystems für die Klassenführung an Studien aus der Sekundarstufe orientiert wurde und somit die einzelnen Merkmale der Klassenführung vermutlich eher für lehrergelenkten Unterricht geeignet sind. Da sich der Grundschulunterricht jedoch zusätzlich durch einen hohen Anteil an offenen bzw. geöffneten und selbstbestimmenden Phasen (z. B. teilweise in Form von Stationenarbeit) auszeichnet, kommt der Klassenführung in diesen Phasen möglicherweise eine andere Bedeutung zu (vgl. Bohl und Kucharz 2010). So kann es für Lehrpersonen und somit auch für den Beobachter in eher geöffneten Phasen schwieriger sein, z. B. Störungssituationen adäquat einzuschätzen, da Störungen, die die aktive Lernzeit verringern, angesichts der Ausdifferenzierung der Lernsituationen (z. B. verschiedene Orte, unterschiedliche Tätigkeiten oder Kooperationsformen)

weniger offensichtlich sein können (ebd.) bzw. in solchen Settings gar nicht groß auffallen. Dies könnte bei der Messung der Klassenführung durch Beobachtungen zu Problemen führen, womit sich auch der vergleichbar hohe Varianzanteil erklären lässt, der auf raterspezifische Unterschiede bzw. Rater-Bias (vgl. zusf. Praetorius 2014) zurückgeführt werden kann (*r, rl, r(u : l), ri, ril* = 17 %).

Ein größerer Varianzanteil entfällt schließlich auf die Residualterme (23 %), was wiederum bedeutet, dass in den bisherigen Analysen für die Klassenführung Variablen (z. B. Sozialformen, Unterrichtsmethoden) unberücksichtigt geblieben sind und sich eine weitere G-Studie mit zusätzlichen potenziellen Varianzquellen lohnen könnte.

Hypothesenkonform ist die Stabilität der Ausprägungen über die beiden Teilstunden der 90-minütigen Unterrichtseinheit für das *Unterrichtsklima* sehr hoch, d. h. der unterrichtsstundenspezifische Varianzanteil ist mit 6 % eher gering. Im Gegensatz zu den Unterrichtsqualitätsmerkmalen Klassenführung und kognitive Aktivierung geht jedoch ein nicht unerheblicher Varianzanteil (64 %) auf den Haupteffekt der Items (*i*) zurück. Dies bedeutet, dass die Abstände zwischen den Itemausprägungen – und damit das Ausmaß ihrer Ähnlichkeit – variieren, wohingegen sich die Ordnung der Itemausprägungen – und damit die relative Itemschwierigkeit – zwischen den zwei Teilstunden nicht unterscheidet (vgl. Abb. 4). Für einen Großteil der Lehrpersonen gilt, dass die Ausprägungen für das Item 1 (= humorvolle Lernatmosphäre) am niedrigsten sind, gefolgt von Item 3 (= Fürsorglichkeit, Herzlichkeit und Wärme) und Item 2 (= Anerkennung der Schüler durch die Lehrperson) (vgl. Abb. 4). Außerdem zeigen sich kaum Unterschiede in den Ausprägungen der einzelnen Items des Unterrichtsklimas in Abhängigkeit von der Höhe des Skalenmittelwerts: Sowohl Lehrpersonen mit einem eher niedrigen Skalenmittelwert für das Unterrichtsklima (linke Seite der Abb. 4) als auch Lehrpersonen mit einem eher hohen Skalenmittel-

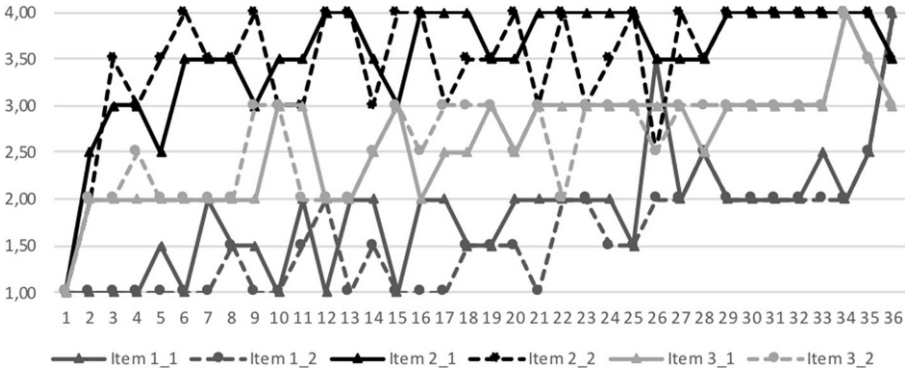

Abb. 4 Unterschiede in der Höhe der Itemausprägungen zwischen 1. und 2. Stunde für das Unterrichtsklima (Abgebildet sind die Itemausprägungen für die drei Items des Unterrichtsklimas getrennt nach der jeweiligen Stunde (1. Stunde vs. 2. Stunde) der 90-minütigen Unterrichtseinheit. Die Lehrpersonen (X-Achse) sind aufsteigend nach der Ausprägung des Unterrichtsklimas sortiert (gemittelt über alle Items des Unterrichtsklimas). *Item 1* = humorvolle Lernatmosphäre, *Item 2* = Anerkennung der Schüler durch die Lehrperson, *Item 3* = Fürsorglichkeit, Herzlichkeit und Wärme; *_1* = 1. Teil der Stunde; *_2* = 2. Teil der Stunde)

wert (rechte Seite der Abb. 4) weisen eine hohe Variation in den Itemausprägungen auf.

Mit 45 % erweist sich der stundenspezifische Anteil für das Unterrichtsqualitätsmerkmal *kognitive Aktivierung* als relativ hoch, deutlich höher als für die Klassenführung und das Unterrichtsklima. Damit ist der Varianzanteil ähnlich hoch ausgeprägt wie in den Studien in der Sekundarstufe (vgl. Praetorius 2014). Die Hypothese, dass sich die Ausprägungen der kognitiven Aktivierung auch im Rahmen der vorliegenden Studie zwischen den beiden Teilstunden (1. Teil/2. Teil der Unterrichtseinheit) unterscheiden, kann demnach beibehalten werden. Die Unterschiede in den Ausprägungen zwischen den zwei Teilstunden werden auch anhand der Abb. 1 deutlich. Bei einem Großteil der Lehrpersonen sind die Mittelwerte für die kognitive Aktivierung im 1. Teil der 90-minütigen Unterrichtseinheit (Phase zur Einführung in die Multiplikation) höher ausgeprägt als im 2. Teil (Übungsphase, z. B. in Form von Stationenarbeit). Hier stellt sich die Frage, ob bestimmte Unterrichtsformen eher als kognitiv aktivierend eingeschätzt werden können als andere oder ob die kognitive Aktivierung in beiden Teilstunden unterschiedlich gut beobachtbar war (vgl. Hess und Lipowsky 2020; Pauli und Reusser 2011).

In der vorliegenden Studie wurde die kognitive Aktivierung – wie auch die Klassenführung und das Unterrichtsklima – zu beiden Messzeitpunkten (1. Teil/2. Teil der Stunde) mit ein und demselben hoch inferenten Ratinginstrument eingeschätzt. Ausgehend von den hier berichteten Ergebnissen kann jedoch angenommen werden, dass kognitive Aktivierung in eher lehrergelenkten Phasen etwas anderes bedeutet als in individuellen Arbeitsphasen und somit die kognitive Aktivierung auch in der Grundschule je nach Phase unterschiedlich operationalisiert werden sollte. Dies könnte auch den relativ hohen Varianzanteil von 14 % erklären, der auf die Interaktion zwischen Unterrichtsstunde (geschachtelt in Lehrpersonen) und den einzelnen Items ((*u:l*)*i*) entfällt. Während für lehrergelenkte Phasen relativ klar ist, dass Lehrpersonen z. B. durch das Stellen von zum Denken anregenden Aufgaben oder herausfordernden Problemen die Schülerinnen und Schüler kognitiv aktivieren können, stellt sich für Übungsphasen die Frage, wie diese gestaltet sein müssen, damit die Schülerinnen und Schüler über das Automatisieren ihrer neu erlernten Fertigkeiten (z. B. zur Multiplikation) hinaus kognitiv aktiviert werden können.

Eine weitere Frage, die sich in diesem Zusammenhang stellt, ist die, inwieweit für das Merkmal der kognitiven Aktivierung überhaupt eine Stabilität in den Ausprägungen zwischen den einzelnen Phasen anzustreben ist (vgl. auch Praetorius, 2014). Minnameier et al. (2015) verstehen die kognitive Aktivierung beispielsweise als „Prozess der Induktion eines Problems“ bei Schülerinnen und Schüler (ebd., S. 842). In Phasen des Problemlöseprozesses steht – nach Meinung von Minnameier et al. (2015) – vielmehr die konstruktive Unterstützung im Sinne des Anleitens und Begleitens der Schülerinnen und Schüler durch die Lehrperson im Vordergrund. Eine solche Engführung kann jedoch kritisch betrachtet werden, da Schülerinnen und Schüler beispielsweise auch in Phasen der Reflexion oder in Übungssituationen (z. B. Leseübung im Deutschunterricht der Grundschule, vgl. Lotz 2016) kognitiv aktiviert werden können/sollten, indem sie z. B. Rückmeldungen bzw. Lehrerfeedback im Sinne prozessorientierter Lernbegleitung wahrnehmen, interpretieren und bestenfalls nutzen.

5.2 Die notwendige Anzahl an Unterrichtsstunden zur Erfassung der Unterrichtsqualität in der Grundschule

In der vorliegenden Studie wurde untersucht, wie viele Unterrichtsstunden (bzw. Doppelstunden) zur Erfassung der drei Basisdimensionen der Unterrichtsqualität in der Grundschule notwendig sind, um eine hinreichende Reliabilität zu erreichen (Forschungsfrage Nr. 2). In Ergänzung zu bisherigen Studien für die Sekundarstufe, deuten die Ergebnisse der vorliegenden Studie darauf hin, dass auch für die Grundschule davon ausgegangen werden kann, dass sich einzelne Unterrichtsstunden einer Lehrperson (sogar innerhalb von Doppelstunden) in ihrer Qualität so unterscheiden, dass vor allem in Bezug auf die Klassenführung und die kognitive Aktivierung mehr als eine Unterrichtsstunde pro Lehrperson herangezogen werden sollte, um zuverlässige Aussagen treffen zu können. Sofern eine Mindestreliabilität von 0,70 gesetzt wird, lässt sich das Unterrichtsklima in der vorliegenden Untersuchung bereits mit zwei Unterrichtsstunden (einer Doppelstunde) hinreichend reliabel erfassen, während für die Klassenführung und die kognitive Aktivierung mindestens drei Unterrichtsstunden (zwei Doppelstunden) beobachtet werden sollten. Dass sich die zeitliche Stabilität in Abhängigkeit von der jeweiligen Basisdimension unterscheidet, steht im Einklang mit bisherigen Studien aus der Sekundarstufe (vgl. Praetorius 2014; Jentsch et al. 2019; Kunter 2005; Rakoczy 2008), allerdings weicht die Anzahl an notwendigen Unterrichtsstunden pro Lehrperson für eine stabile Einschätzung der Basisdimensionen von den bisherigen Ergebnissen für die Sekundarstufe ab. „One Lesson is all you need“ (Praetorius et al. 2014) trifft in der Grundschule für keine der drei Basisdimensionen der Unterrichtsqualität zu.

6 Limitationen

Im Rahmen der vorliegenden Studie wurden keine zufälligen Unterrichtsstunden einer Lehrperson in die Analyse miteinbezogen, sondern es wurde von jeder Lehrperson eine Doppelstunde Mathematikunterricht videografiert. Als Thema der Doppelstunde wurde die „Einführung in die Multiplikation“ festgelegt, d. h. der Inhalt wurde standardisiert. Wie in den Videos ersichtlich, bauen die beiden Teilstunden größtenteils inhaltlich und methodisch aufeinander auf (1. Teil: Einführung in die Multiplikation, 2. Teil: Übung z. B. in Form von Stationenlernen). Demnach weisen die beiden Teilstunden keine stochastische Unabhängigkeit auf, womit auch Designeffekte nicht ausgeschlossen werden können. Vor diesem Hintergrund wäre es für nachfolgende Analysen (1) sinnvoll, die Doppelstunde in mehrere Segmente (ähnlich Mashburn et al. 2014 bzw. Jentsch et al. 2019) zu unterteilen und die Unterrichtsqualität beispielsweise für verschiedene Oberflächenmerkmale (Lehrmethoden sowie konkrete Organisations- und Sozialformen) des Unterrichts zu erfassen. Durch die Erfassung von Zusammenhängen mit Oberflächenmerkmalen können zusätzliche Varianzanteile von Beobachterratings erklärt werden (vgl. Jentsch et al. 2019; Patrick und Montzicopoulos 2016). (2) Es sollte überprüft werden, inwieweit die Ergebnisse auch auf Grundschulstudien übertragbar sind, die eine zufällige Anzahl an Unterrichtsstunden pro Lehrperson in ihre Analysen mit einbeziehen. (3) An-

dererseits sollten auch weitere Fächer mit einbezogen werden, da der Großteil der Studien häufig den Mathematikunterricht fokussiert (Ausnahme: Praetorius et al. 2016; Gabriel-Busse und Lipowsky in Vorbereitung). (4) Da die Ergebnisse der explorativen D-Studien nahelegen, dass mindestens drei Unterrichtsstunden bzw. zwei Doppelstunden für eine adäquate Einschätzung der Klassenführung und der kognitiven Aktivierung im Mathematikunterricht der Grundschule notwendig sind, wäre in zukünftigen Studien die Beobachtung einer größeren Anzahl an Unterrichtsstunden pro Lehrperson wünschenswert.

7 Schlussfolgerungen

Aus den hier berichteten Ergebnissen lassen sich folgende Schlussfolgerungen für die Grundschule ziehen: (1) Da sich deutliche Unterschiede in der zeitlichen Stabilität zwischen den drei Basisdimensionen der Unterrichtsqualität auch in der Grundschule zeigen, scheint es auch hier sinnvoll die zeitliche Stabilität für verschiedene Unterrichtsqualitätsmerkmale separat zu untersuchen. (2) Die Ergebnisse der vorliegenden Studie legen nahe, dass im Unterschied zur Sekundarstufe für eine genauere Messung der Klassenführung im Mathematikunterricht die Anzahl an Unterrichtsstunden pro Lehrperson erhöht werden sollte, da der Varianzanteil, der auf Unterrichtsstunden zurückzuführen ist, deutlich höher ist, als beispielsweise in der Studie von Praetorius (2014). Für eine adäquate Abbildung der kognitiven Aktivierung im Mathematikunterricht scheinen hingegen deutlich weniger Unterrichtsstunden pro Lehrperson notwendig zu sein.

Danksagung An dieser Stelle gilt unser Dank Frau Prof.' in Dr. Anna-Katharina Praetorius für die Unterstützung bei der Berechnung der Generalisierbarkeitsstudien und Herrn Dr. Armin Jentsch für die Hilfe bei der Berechnung der explorativen D-Studien.

Förderung Die PERLE-Studie wurde gemeinsam von den Universitäten Kassel und Bamberg sowie dem Deutschen Institut für Internationale Pädagogische Forschung (DIPF) unter Leitung von Prof. Dr. Frank Lipowsky und Prof.' in Dr. Gabriele Faust (†) durchgeführt. Das Projekt wurde vom Bundesministerium für Bildung und Forschung (BMBF) gefördert.

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

Aebli, H. (1983). *Zwölf Grundformen des Lehrens: Eine Allgemeine Didaktik auf psychologischer Grundlage*. Stuttgart: Klett.

Aebli, H. (1987). *Grundlagen des Lehrens*. Stuttgart: Klett.

Bessoth, R. (1989). *Verbesserung des Unterrichtsklimas: Grundlagen, Aufbau und Einsatz von Instrumenten*. Neuwied: Luchterhand.

Bohl, T., & Kucharz, D. (2010). *Offener Unterricht heute. Konzeptionelle und didaktische Weiterentwicklung*. Weinheim: Beltz.

Brennan, R. L. (2001). *Manual for urGenova (Version 2.1)*. Iowa City: Iowa Testing Programs. University of Iowa.

Brophy, J. E., Coulter, C. L., Crawford, W. J., Evertson, C. M., & King, C. E. (1975). Classroom Observation Scales: stability across time and context and relationships with student learning gains. *Journal of Educational Psychology*, *67*(6), 873–881.

Clausen, M., Reusser, K., & Klieme, E. (2003). Unterrichtsqualität auf der Basis hoch-inferenter Unterrichtsbeurteilungen. Ein Vergleich zwischen Deutschland und der deutschsprachigen Schweiz. *Unterrichtswissenschaft*, *31*(2), 122–141.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley.

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., & Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal*, *52*(6), 1133–1159.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Berlin, Heidelberg: Springer.

Eckerth, M., Hanke, P., & Hein, A. K. (2012). Schulische Bedingungen des Lehrens und Lernens im Anfangsunterricht der Grundschule – ausgewählte Ergebnisse aus dem FiS-Projekt. In F. Hellmich, S. Förster & F. Hoya (Hrsg.), *Bedingungen des Lehrens und Lernens in der Grundschule. Bilanz und Perspektiven* (S. 65–68). Wiesbaden: Springer.

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

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

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

Gabriel, K., & Lipowsky, F. (2013a). Hoch inferentes Rating: Klassenführung in drei Fächern. In M. Lotz, F. Lipowsky & G. Faust (Hrsg.). Technischer Bericht zu den PERLE-Videostudien (Materialien zur Bildungsforschung, Band 23/3, S. 145–168). Frankfurt am Main: Gesellschaft zur Förderung Pädagogischer Forschung (GFPF).

Gabriel, K., & Lipowsky, F. (2013b). Hoch inferentes Rating: Unterrichtsklima in drei Fächern. In M. Lotz, F. Lipowsky & G. Faust (Hrsg.), *Technischer Bericht zu den PERLE-Videostudien* (Materialien zur Bildungsforschung, Band 23/3,S. 169–190). Frankfurt am Main: Gesellschaft zur Förderung Pädagogischer Forschung (GFPF).

Gabriel, K., Praetorius, A.-K., & Lipowsky, F. (2015). Wie in der einen, so auch in der anderen Stunde? – Analysen zur Stabilität der Klassenführung im Anfangsunterricht. Jahrbuch für Allgemeine Didaktik. Themenheft: Klassenmanagement/Klassenführung. *Perspektiven, Befunde, Kontroversen*, *1*, 85–95.

Gabriel-Busse, K., & Lipowsky, F. (in Vorb.). Reicht eine Doppelstunde im Anfangsunterricht aus? – Ergebnisse aus den PERLE-Videostudien Deutsch und Mathematik zur zeitlichen Stabilität und Generalisierbarkeit von Ratings zur Klassenführung. In M. Hess, A. Denn, C. Theurer & F. Lipowsky (Hrsg.). *Determinanten und Effekte der Persönlichkeits- und Lernentwicklung in der Grundschule – Ergebnisse der PERLE-Studie*. Münster: Waxmann.

Götz, T., Lohrmann, K., Ganser, B., & Haag, L. (2005). Einsatz von Unterrichtsmethoden – Konstanz oder Wandel? *Empirische Pädagogik*, *19*, 342–360.

Greb, K., Faust, G., & Lipowsky, F. (2007). Projekt PERLE. Persönlichkeits- und Lernentwicklung von Grundschulkindern. *Diskurs Kindheits- und Jugendforschung*, *2*, 100–104.

Greb, K., Lipowsky, F., & Faust, G. (2009). Nina und Michael, Miró und ein Nussknacker! Persönlichkeits- und Lernentwicklung von Grundschulkindern. *Die Grundschulzeitschrift*, *23*, 18–21.

Helmke, A., & Weinert, F.E. (1997). Unterrichtsqualität und Leistungsentwicklung: Ergebnisse aus dem SCHOLASTIK-Projekt. In F.E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 241–251). Weinheim: Beltz.

Hess, M., & Lipowsky, F. (2020). Zur (Un-)Abhängigkeit von Oberflächen- und Tiefenstrukturen im Grundschulunterricht – Fragen von Lehrpersonen im öffentlichen Unterricht und in Schülerarbeitsphasen im Vergleich. *Zeitschrift für Pädagogik*. https://doi.org/10.3262/ZPB2001117.

Jentsch, A., Casale, G., Schlesinger, L., Kaiser, G., König, J., & Blömeke, S. (2019). Variabilität und Generalisierbarkeit von Ratings zur Qualität von Mathematikunterricht zwischen und innerhalb von Unterrichtsstunden. *Unterrichtswissenschaft*. https://doi.org/10.1007/s42010-019-00061-8.

Kammermeyer, G., & Martschinke, S. (2009). Qualität im Anfangsunterricht – Ergebnisse der KILIA-Studie. *Unterrichtswissenschaft*, *37*(1), 35–54.

Klieme, E. (2002). Was ist guter Unterricht? Ergebnisse der TIMSS-Videostudie im Fach Mathematik. In W. Bergsdorf, J. Court, M. Eckert & H. Hoffmeister (Hrsg.), *Herausforderungen der Bildungsgesellschaft*. 4. Ringvorlesung der Universität Erfurt. (S. 89–113). Weimar: Rhino.

Klieme, E. (2006). Empirische Unterrichtsforschung: Aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde. *Zeitschrift für Pädagogik*, *52*(6), 765–773.

Klieme, E. (2019). Unterrichtsqualität. In M. Harring, C. Rohlfs & M. Gläser-Zikuda (Hrsg.), *Handbuch Schulpädagogik* (S. 393–408). Münster: Waxmann.

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. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 43–58). Bonn: BMBF.

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.

Kounin, J.S. (2006). *Techniken der Klassenführung*. Stuttgart: Klett.

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht*. Münster: Waxmann.

Kunter, M., Dubberke, T., Baumert, J., Blum, W., Brunner, M., & Jordan, A. (2006). Mathematikunterricht in den PISA-Klassen 2004: Rahmenbedingungen, Formen und Lehr-Lernprozesse. In PISA-Konsortium (Hrsg.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (S. 161–194). Münster: Waxmann.

Kunter, M., Klusmann, U., Dubberke, T., Baumert, J., Blum, W., & Brunner, M. (2007). Linking aspects of teacher competence to their instruction. Results from the COACTIV project. In M. Prenzel (Hrsg.), *Studies on the educational quality of schools. The final report on the DFG priority programm* (S. 39–59). Münster: Waxmann.

Lange, B., Kuffner, H., & Schwarzer, R. (1983). *Schulangst und Schulverdrossenheit: Eine Längsschnittanalyse von schulischen Sozialisationseffekten*. Opladen: Westdeutscher Verlag.

Lauterbach, C., Gabriel, K., & Lipowsky, F. (2013). Hoch inferentes Rating: Kognitive Aktivierung im Mathematikunterricht. In M. Lotz, F. Lipowsky & G. Faust (Hrsg.). *Technischer Bericht zu den PERLE-Videostudien* (Materialien zur Bildungsforschung, Band 23/3, S. 405-421). Frankfurt am Main: Gesellschaft zur Förderung Pädagogischer Forschung (GFPF).

Lipowsky, F., Faust, G., Kastens, C., & Post, S. (2013). Die PERLE-Studie: Überblick und Hintergründe. In F. Lipowsky, G. Faust & C. Kastens (Hrsg.), *Persönlichkeits- und Lernentwicklung an staatlichen und privaten Grundschulen. Ergebnisse der PERLE-Studie zu den ersten beiden Schuljahren* (S. 9–28). Münster: Waxmann.

Lipowsky, F., Drollinger-Vetter, B., Klieme, E., Pauli, C., & Reusser, K. (2018). Generische und fachdidaktische Dimensionen von Unterrichtsqualität – Zwei Seiten einer Medaille? In M. Martens, K. Rabenstein, K. Bräu, M. Fetzer, H. Gresch, I. Hardy & C. Schelle (Hrsg.), *Konstruktionen von Fachlichkeit: Ansätze, Erträge und Diskussionen in der empirischen Unterrichtsforschung* (S. 183–202). Bad Heilbrunn: Klinkhardt-Verlag.

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

Lotz, M., Berner, N. E., Gabriel, K., Post, S., Faust, G., & Lipowsky, F. (2011). Unterrichtsbeobachtung im Projekt PERLE. In D. Kucharz, T. Irion & B. Reinhoffer (Hrsg.), *Grundlegende Bildung ohne Brüche* (S. 183–194). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lotz, M., Berner, N. E., & Gabriel, K. (2013a). Auswertung der PERLE-Videostudien und Überblick über die Beobachtungsinstrumente. In M. Lotz, F. Lipowsky & G. Faust (Hrsg.), Technischer Bericht zu den PERLE-Videostudien. In F. Lipowsky & G. Faust (Hrsg.), *Dokumentation der Erhebungsinstrumente des Projekts „Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE)* (Materialien zur Bildungsforschung, Band 23/3, S. 83–103). Frankfurt am Main: Gesellschaft zur Förderung Pädagogischer Forschung (GFPF).

Lotz, M., Lipowsky, F., & Faust, G. (2013b) (Hrsg.), Technischer Bericht zu den PERLE – Videostudien. In F. Lipowsky & G. Faust (Hrsg.), *Dokumentation der Erhebungsinstrumente des Projekts „Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE)* (Materialien zur Bildungsforschung, Band 23/3). Frankfurt am Main: Gesellschaft zur Förderung Pädagogischer Forschung (GFPF).

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

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

Minnameier, G., Hermkes, R., & Mach, H. (2015). Kognitive Aktivierung und Konstruktive Unterstützung als Prozessqualitäten des Lehrens und Lernens. *Zeitschrift für Pädagogik*, *61*(6), 837–854.

National Institute of Child Health and Human Development. Early Child Care Research Network (NICHD ECCRN) (2002). The relation of global first-grade classroom environment to structural classroom features and teacher and students behaviors. *The Elementary School Journal*, *102*(5), 367–387.

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

Pauli, C., & Reusser, K. (2010). Selbst- und Unterrichtswahrnehmung der Lehrpersonen. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 143–170). Münster: Waxmann.

Pauli, C., & Reusser, K. (2011). Expertise in Swiss mathematics instruction. In Y. Li & G. Kaiser (Hrsg.), *Expertise in mathematics instruction: An international perspective* (S. 85–107). New York: Springer.

Piaget, J. (1964). Development and learning. In R. Ripple & V. Rockcastle (Hrsg.), *Piaget rediscovered* (S. 7–20). Ithaca: Cornell University Press.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system. Manual*. Baltimore: Brookes.

Praetorius, A.-K. (2014). *Messung von Unterrichtsqualität durch Ratings*. Münster: Waxmann.

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: do they fulfill what they promise? *Learning and Instruction*, *22*, 387–400.

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

Praetorius, A.-K., Vieluf, S., Saß, S., Bernholt, A., & Klieme, E. (2016). The same in German as in English? Investigating the subject-specificity of teaching quality. *Zeitschrift für Erziehungswissenschaft*, *19*(1), 191–209.

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

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

v. Saldern, M. (1987). *Sozialklima von Schulklassen. Überlegungen und mehrebenenanalytische Untersuchungen zur subjektiven Wahrnehmung von Lernumwelten*. Frankfurt am Main: Peter Lang.

v. Saldern, M. (2000). Unterrichtsklima, Partizipation und soziale Interaktion. In M. K. W. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion. Pädagogisch-psychologische Aspekte des Lehrens und Lernens in der Schule* (S. 565–581). Opladen: Leske & Budrich.

v. Saldern, M., & Littig, K. E. (1985). Die Konstruktion der Landauer Skalen zum Sozialklima. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *17*(2), 138–149.

Satow, L. (1999). Zur Bedeutung des Unterrichtsklimas für die Entwicklung schulbezogener Selbstwirksamkeitserwartungen. Eine Mehrebenenanalyse mit latenten Variablen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *31*, 171–179.

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

Seidel, T. (2009). Klassenführung. In E. Wild & J. Möller (Hrsg.). *Pädagogische Psychologie* (S. 135–148). Heidelberg: Springer Verlag.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, *46*, 553–611.

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

Stigler, J., Gonzales, P., Kawanka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS-videotape classroom study (technical report)*. Los Angeles: University of California.

Vygotsky, L. (1977). *Denken und Sprechen*. Frankfurt: Fischer.

Vygotsky, L. (1985). *Die psychischen Systeme: Ausgewählte Schriften*. Köln: Pahl Rugenstein.

Waldis, M., Grob, U., Pauli, C., & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 171–208). Münster: Waxmann.

Zumhasch, C. (2006). Das Unterrichtsklima. In K. H. Arnold, U. Sandfuchs & J. Wiechmann (Hrsg.), *Handbuch Unterricht* (S. 144–147). Bad Heilbrunn: Klinkhardt.