Fauth, Benjamin; Decristan, Jasmin; Rieser, Svenja; Klieme, Eckhard; Büttner, Gerhard

Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive. Zusammenhänge und Vorhersage von Lernerfolg

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Zeitschrift für pädagogische Psychologie 28 (2014) 3, S. 127-137

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg. *Zeitschrift für Pädagogische Psychologie*, *28* (3), 127-137.
doi: 10.1024/1010-0652/a000129

Originalartikel

Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg

B. Fauth et al.: Perspektiven auf Unterrichtsqualität

Benjamin Fauth[1,2*], Jasmin Decristan[1,3], Svenja Rieser[1,2], Eckhard Klieme[1,3] und Gerhard Büttner[1,2]

[1] IDeA-Forschungszentrum, Frankfurt am Main

[2] Goethe-Universität Frankfurt am Main

[3] Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt am Main

Zusammenfassung: Unterrichtsqualität kann mit drei Basisdimensionen beschrieben werden: strukturierte Klassenführung, kognitive Aktivierung und unterstützendes Klima. Untersuchungen aus dem Sekundarbereich finden oftmals nur geringe Übereinstimmungen zwischen Urteilen zur Unterrichtsqualität von Schülerinnen und Schülern, Lehrpersonen und externen Beobachtern. Es stellt sich damit die Frage, ob aus unterschiedlichen Perspektiven dieselben Konstrukte erfasst werden und wie diese mit dem Lernerfolg der Schülerinnen und Schüler zusammenhängen. Für die Grundschule werden in der vorliegenden Studie substanzielle Zusammenhänge zwischen Urteilen zur strukturierten Klassenführung aus allen drei Perspektiven gefunden. Zudem zeigen sich hier in Mehrebenen-Regressionsanalysen Effekte aller drei Urteilerperspektiven auf die Leistungsentwicklung. Keine Zusammenhänge zwischen den Perspektiven zeigen sich im Bereich kognitive Aktivierung. Beim unterstützenden Klima bestehen Zusammenhänge nur zwischen Schüler-

und Lehrerurteilen. Bei den Dimensionen kognitive Aktivierung und unterstützendes Klima waren nur die Urteile externer Beobachter entscheidend für die Leistungsentwicklung. Die Ergebnisse können mit der unterschiedlichen Beobachtbarkeit der drei Basisdimensionen und mit den Stärken und Grenzen der einzelnen Urteilerperspektiven erklärt werden.

Schlüsselwörter: Unterrichtsqualität, Beurteilungen, Validität

Teaching Quality in Primary School from the Perspectives of Students, Teachers, and External Observers: Relationships Between Perspectives and Prediction of Student Achievement

Abstract: The contribution examines three basic dimensions of teaching quality: cognitive activation, supportive climate, and classroom management. Previous studies in secondary schools show that ratings of teaching quality from students, teachers, and external observers are only slightly correlated. This leads to the question if taken from different perspectives, measures actually refer to the same construct and how these measures are related to student achievement. The answer seems to depend on the basic dimension concerned. Drawing on a primary school sample, we found substantial correlations between ratings of classroom management from each perspective but no correlations between the ratings of cognitive activation. In the dimension of supportive climate, only student and observer ratings were significantly related. Longitudinal multilevel regression analyses showed that from each perspective, ratings of classroom management had predictive power for student achievement. Regarding cognitive activation and supportive climate, only ratings of external observers were predictive for achievement. These results can be explained by taking into account that the basic dimensions cannot be observed in the same way and that each perspective has its strengths and weaknesses.

Keywords: teaching quality, ratings, validity

1 Theoretischer Hintergrund

1.1 Basisdimensionen von Unterrichtsqualität

Merkmale des Unterrichts sind von zentraler Bedeutung für den Lernerfolg von Schülerinnen und Schülern (Scheerens & Bosker, 1997; Seidel & Shavelson, 2007). In den letzten Jahren hat die Unterrichtsforschung große Fortschritte in der empirischen und theoretischen Bestimmung von Merkmalen „guten Unterrichts“ gemacht (Good, Wiley & Florez, 2009; Hattie, 2009). Ausgehend von hoch-inferenten Urteilen externer Videobeobachter haben Klieme, Pauli & Reusser (2009) ein theoretisches Rahmenmodell vorgestellt, das drei Basisdimensionen von Unterrichtsqualität beschreibt: unterstützendes Klima, kognitive Aktivierung und strukturierte Klassenführung. Die inhaltliche Bestimmung eines *unterstützenden Klimas* erfolgt im Rückgriff auf die Selbstbestimmungstheorie (Ryan & Deci, 2000). Zu Unterrichtsmerkmalen, die ein Erleben von Selbstbestimmung begünstigen, gehören positives Lehrerfeedback, ein konstruktiver Umgang mit Fehlern und ein insgesamt wertschätzender Umgang der Lehrperson mit den Schülerinnen und Schülern (Klieme & Rakoczy, 2008). Die Dimension der *kognitiven Aktivierung* zielt stärker auf fachdidaktisch relevante Aspekte des Unterrichts. Vor allem in Studien zum Mathematikunterricht wurden herausfordernde Aufgabenstellungen, die Exploration von Konzepten, Ideen und Lösungswegen durch die Lehrperson und eine diskursive Unterrichtspraxis mit dem Lernerfolg von Schülerinnen und Schülern in Verbindung gebracht (Baumert et al., 2010; Lipowsky et al., 2009; Pauli, Drollinger-Vetter, Hugener & Lipowsky, 2008). Im naturwissenschaftlichen Sachunterricht der Grundschule wird kognitive Aktivierung vor allem im Zusammenhang mit Theorien zum Konzeptwechsel

diskutiert (Einsiedler & Hardy, 2010). Im Sinne der Arbeiten von Kounin (1970) geht es bei einer *strukturierten Klassenführung* bzw. beim *Classroom Management* nicht nur um das Reagieren auf Unterrichtsstörungen, sondern um eine insgesamt störungspräventive Unterrichtsgestaltung, in der z.B. Übergänge zwischen Unterrichtsphasen durch Routinen geregelt sind. Ziel ist es, möglichst viel Unterrichtszeit auf die Auseinandersetzung mit dem Lernstoff verwenden zu können („time on task").

1.2 Urteile von Schülerinnen und Schülern, Lehrpersonen und externen Beobachtern

Die in der Unterrichtsforschung am häufigsten verwendeten Datenquellen sind Urteile von Schülerinnen und Schülern, Lehrpersonen oder externen Beobachtern, wobei jede dieser Perspektiven mit spezifischen Vor- und Nachteilen verbunden ist. Die Validität von Schülerurteilen wird häufig angezweifelt, weil Schülerinnen und Schüler aufgrund von fehlendem pädagogisch-didaktischen Urteilsvermögen nicht in der Lage seien, den Unterricht adäquat zu beurteilen (Baumert et al., 2004, S. 321; Waldis, Grob, Pauli & Reusser, 2010). Diese Probleme bestehen insbesondere bei Urteilen von Grundschülerinnen und Grundschülern (Kloss, 2012, S. 136). Ein Vorteil der Schülerperspektive ist, dass sich die Urteile auf eine breite Basis an Erfahrungen über viele Unterrichtstunden hinweg stützen (Baumert et al., 2004, S. 319). Lehrpersonen sollten aufgrund ihrer Ausbildung genügend pädagogisch-didaktische Expertise zur validen Einschätzung von Unterricht haben. Ihre Urteile können jedoch beispielsweise selbstwertdienlichen Verzerrungen unterliegen, was dazu führen würde, dass der eigene Unterricht in einem positiveren Licht erscheint (Wubbels, Brekelmans & Hooymayers, 1992). Dabei kann es sich bei Lehrerurteilen zum Unterricht je nach Zielkonstrukt in unterschiedlichem Ausmaß um Selbstbeurteilungen handeln (Clausen, 2002, S. 46). Urteile von externen (Video-)Beobachtern werden mitunter als „Königsweg zu Beschreibung und Bewertung des Unterrichts" gesehen (Helmke, 2009, S. 288), der sowohl eine hohe Objektivität als auch den geschulten methodisch-didaktischen Blick gewährleisten soll. Allerdings sind Video- und Beobachtungsstudien mit einem

immensen Aufwand verbunden, der dazu führt, dass sich die Urteile externer Beobachter häufig nur auf wenige (eine bis maximal fünf) Unterrichtsstunden pro Klasse beziehen, was die Validität der Ergebnisse in Frage stellen kann (Praetorius, Pauli, Reusser, Rakoczy & Klieme, 2013).

1.3 Übereinstimmung und Kombinierbarkeit unterschiedlicher Urteile zu den Basisdimensionen von Unterrichtsqualität

Ein Weg, mit den oben beispielhaft genannten Nachteilen der jeweiligen Datenquellen umzugehen, sind Multitrait-Multimethod-Designs, in denen unterschiedliche Datenquellen genutzt werden, um Konstrukte zu beschreiben (Campbell & Fiske, 1959). In der Unterrichtsforschung herrscht mittlerweile Konsens, dass es erstrebenswert ist, die Qualität des Unterrichts mit multiplen Indikatoren zu erfassen (Kane, McCaffrey, Miller & Staiger, 2013). In der US-amerikanischen MET-Study (Measures for Effective Teaching) wurden beispielsweise Schüler- und Beobachterurteile zusammen mit Value-Added Measures zu einem gemeinsamen Score für „effective teaching“ verrechnet (Mihaly, McCaffrey, Staiger & Lockwood, 2013). Dieses Vorgehen scheint aus theoretischer Sicht fragwürdig, da so Maße der Prozessqualität von Unterricht mit Schüleroutcome-Maßen vermischt werden. Zudem ist auch vor der Kombination von Indikatoren, die sich gleichermaßen auf die Prozessqualität von Unterricht beziehen, zu prüfen, ob aus unterschiedlichen Perspektiven überhaupt dasselbe Konstrukt erfasst wird.

Sehr geringe Übereinstimmungen zwischen unterschiedlichen Urteilerperspektiven führten schon bei Clausen (2002, S. 78) zu der Frage, ob Unterrichtsqualität als perspektivenspezifisches Konstrukt konzeptualisiert werden sollte. Die Antwort hierauf kann für die einzelnen Basisdimensionen von Unterrichtsqualität unterschiedlich ausfallen (Clausen, 2002; Desimone, Smith & Frisvold, 2010; Herman, Klein & Abedi, 2000; Kunter & Baumert, 2006; Mayer, 1999). Die deutlichsten Zusammenhänge zwischen den drei Perspektiven finden sich im Bereich der *strukturierten Klassenführung* (Kunter & Baumert,

2006; Waldis et al., 2010). Dies lässt sich nach Clausen (2002, S. 117) damit erklären, dass es sich bei der Klassenführung um ein relativ gut beobachtbares Geschehen handelt. Im Bereich *unterstützendes Klima* zeigen sich signifikante Korrelationen zwischen den Urteilen von Schülerinnen und Schülern und Lehrpersonen (Clausen, 2002, S. 129; Kunter & Baumert, 2006). Dagegen finden sich bei diesem Konstrukt keine Korrelationen mit der Perspektive der Videobeobachter (Clausen, 2002, S. 129; De Jong & Westerhof, 2001), was mit der schwierigen Beobachtbarkeit innerhalb der recht begrenzten Beobachtungsstichprobe einer Unterrichtsstunde erklärt werden kann. Im Bereich der *kognitiven Aktivierung* ist die Befundlage uneinheitlich. Studien finden hier Zusammenhänge zwischen bestimmten Aspekten kognitiver Aktivierung, z.B. beim Faktor „Repetitives Üben" (als Negativindikator; Zusammenhänge zwischen allen drei Perspektiven bei Clausen, 2002, S. 129) oder beim Anwenden eigener Lösungsstrategien (Zusammenhänge zwischen Schüler- und Lehrerurteilen bei Kunter & Baumert, 2006). Bei den Aspekten kognitiver Aktivierung, auf die in der vorliegenden Studie abgezielt wird (Herausfordernder Unterricht und Exploration von Schülerkonzepten), finden sich jedoch keinerlei Zusammenhänge zwischen den Perspektiven (Clausen, 2002, S. 129; Kunter & Baumert, 2006). Zieht man die unter 1.2 beschriebenen Vor- und Nachteile der einzelnen Urteilerperspektiven in Betracht, so können für den Grundschulbereich ähnliche Ergebnisse wie in der Sekundarstufe erwartet werden.

1.4 Prädiktive Validität der einzelnen Perspektiven für Schülerleistungen

Studien aus dem Sekundarbereich belegen die prädiktive Validität von klassenaggregierten Schülerurteilen zur strukturierten Klassenführung (Kunter & Baumert, 2006; Seidel & Shavelson, 2007) und zu Aspekten kognitiver Aktivierung (Dubberke, Kunter, McElvany, Brunner & Baumert, 2008; Klieme, Steinert & Hochweber, 2010). Für die Urteile von Grundschülern ist die Forschungslage weniger klar. Kane et al. (2013, S. 10) zeigen, dass Schülerurteile in der Primarstufe insgesamt eine größere prädiktive Kraft als Beobachterurteile haben können. In früheren Analysen fanden Fauth, Decristan, Rieser,

Klieme und Büttner (2014), dass die schülereingeschätzte strukturierte Klassenführung in der Grundschule mit dem Lernerfolg der Schülerinnen und Schüler zusammenhängt. Für das unterstützende Klima und die kognitive Aktivierung fanden sich hingegen keine Effekte.

Die Befundlage zur prädiktiven Kraft von Lehrerurteilen ist uneinheitlich. In der Sekundarstufe fand Wenglinsky (2002) Zusammenhänge zwischen Lehrerangaben zu Aspekten kognitiver Aktivierung und dem Lernerfolg der Schülerinnen und Schüler andererseits. Auch bei Clausen (2002, S. 171f.) finden sich Zusammenhänge zwischen Leistungsindikatoren und Lehrerurteilen zur kognitiven Aktivierung, nicht jedoch zu Urteilen zur strukturierten Klassenführung und zum unterstützenden Klima. In der Metaanalyse von Seidel und Shavelson (2007) finden sich kaum Zusammenhänge zwischen Lehrerurteilen und (motivationalen und kognitiven) Schüleroutcomes.

Seit der TIMSS 1995 Videostudie (Stigler & Hiebert, 1999) hat es einige groß angelegte Beobachtungsstudien gegeben, in denen deutliche Zusammenhänge zwischen den Urteilen externer Beobachter zur Qualität des Unterrichts und der Leistungsentwicklung gefunden wurden (Helmke et al., 2008; Kane et al., 2013; Pianta & Hamre, 2009; Seidel et al., 2006). Bei Seidel und Shavelson (2007) finden sich insgesamt kleine Effekte von Beobachterurteilen zur kognitiven Aktivierung auf Schüleroutcomes. In der deutsch-schweizerischen Videostudie „Pythagoras“ konnte jedoch gezeigt werden, dass insbesondere kognitive Aktivierung und strukturierte Klassenführung Leistungsentwicklungen erklären können (Lipowsky et al., 2009).

Für den Grundschulbereich besteht weiterhin Forschungsbedarf zu Fragen der prädiktiven Validität. Während für Lehrpersonen und Videobeobachter ähnliche Ergebnisse wie im Sekundarbereich erwartet werden können, ist dies im Hinblick auf Schülerurteile zu methodisch-didaktisch anspruchsvollen Konstrukten wie kognitiver Aktivierung eher fraglich.

2 Fragestellungen und Hypothesen

1. Welche Zusammenhänge gibt es zwischen den Urteilen von Schülerinnen und Schülern, Lehrpersonen und Beobachtern zu den Dimensionen kognitive Aktivierung, unterstützendes Klima und strukturierte Klassenführung? Wir erwarten einen positiven Zusammenhang zwischen Schüler- und Lehrerperspektive im Bereich unterstützendes Klima (Hypothese 1), jedoch keine Zusammenhänge im Bereich kognitive Aktivierung (Hypothese 2). Im Bereich strukturierte Klassenführung werden positive Zusammenhänge zwischen allen drei Perspektiven erwartet (Hypothese 3).
2. Wessen Urteile zu welcher Dimension sagen Schülerleistungen vorher? Für die Urteile von externen Beobachtern wird erwartet, dass sie zu den Dimensionen kognitive Aktivierung und strukturierte Klassenführung Leistungsentwicklungen vorhersagen (Hypothese 4). Es wird weiterhin erwartet, dass Lehrerurteile zur kognitiven Aktivierung und Schülerurteile zur strukturierten Klassenführung mit der Leistungsentwicklung der Schüler zusammenhängen (Hypothesen 5 und 6). Auf der Grundlage der vorliegenden Daten wurden in Fauth et al. (2014) bereits Effekte von Schülerurteilen geprüft, dort jedoch unter Kontrolle der Popularität der Lehrperson. Da die Zusammenhänge zwischen Schülerurteilen und Lernerfolg unabhängig von der Popularität der Lehrperson bestanden, wird diese Variable in der vorliegenden Arbeit nicht in die Analysen miteinbezogen. So können die Analysen zur Schülerperspektive parallel zu den Analysen zur Perspektive der Lehrpersonen und externen Beobachter gehalten werden und sind damit besser vergleichbar.
3. Lassen sich latente Faktoren identifizieren, die Urteile aus allen drei Perspektiven berücksichtigen? Welche Vorhersagekraft haben diese Faktoren für Schülerleistungen? Erwartet wird, dass insbesondere die Urteile im Bereich strukturierte Klassenführung konvergieren und sich hieraus ein latenter Faktor bilden lässt (Hypothese 7). Dieser sollte ebenfalls einen Effekt auf die Schülerleistungen haben (Hypothese 8).

3 Methoden

3.1 Stichprobe

Die vorliegende Studie basiert auf Daten von 1070 Schülerinnen und Schülern aus 54 Klassen der dritten Jahrgangsstufe. Das durchschnittliche Alter lag bei 8.8 Jahren (SD = 0.5). 49% der Kinder waren weiblich. Pro Klasse konnten durchschnittlich ca. 20 Kinder befragt werden (min. = 10, max. = 27). Es liegen weiterhin Daten von 54 Lehrpersonen vor, die im Erhebungszeitraum in den entsprechenden Klassen Sachunterricht unterrichtet haben. Die Lehrpersonen waren im Mittel 42.6 Jahre alt (SD = 9.3), waren zu 86% weiblich und hatten eine durchschnittliche Berufserfahrung von 16.4 Jahren (SD = 8.6). Für 53 Klassen liegen Daten von externen Beobachtern vor. Davon wurden 37 Fälle videobasiert beurteilt (Videoratings). Bei 16 Fällen lagen keine Einverständnisse der Lehrpersonen oder der Erziehungsberechtigten für Videographierungen vor. In diesem Fall wurden Beobachtungen vor Ort in der Klasse vorgenommen (Liveratings). Eine Klasse konnte aus organisatorischen Gründen nicht beobachtet werden.

Die Zielpopulation der Studie waren alle öffentlichen Grundschulen im Raum Mittel- und Südhessen. Lehrpersonen und Schulleiter wurden telefonisch kontaktiert und zu Informationsabenden eingeladen. Danach wurden an 84 interessierte Lehrpersonen Briefe mit näheren Informationen zur Studie verschickt. Von diesen Lehrpersonen erklärten sich 51 bereit, an der Studie teilzunehmen. Im Folgenden konnten noch drei weitere Lehrpersonen für die Teilnahme an der Studie rekrutiert werden. Um die Motivation zur Teilnahme zu erhöhen, bekamen die teilnehmenden Lehrpersonen Unterrichtsmaterialien sowie einen Gutschein für die Klassenkasse. Die Teilnahme an der Studie war für Lehrpersonen und Schülerinnen und Schüler (Einverständnis der Erziehungsberechtigten vorausgesetzt) freiwillig. Die Teilnahmequote innerhalb der Klassen lag durchschnittlich bei 96%.

3.2 Design

Die an der Studie teilnehmenden Lehrpersonen führten mit ihren Schülerinnen und Schülern zwei Unterrichtseinheiten zum Thema Schwimmen und Sinken durch, für die Ablaufpläne

und Materialien zur Verfügung gestellt wurden (vgl. Decristan et al., 2014; Hardy et al., 2011). Die Einheiten basieren auf den von Möller und Jonen (2005) konzipierten KiNT-Boxen, die in früheren Studien bereits empirisch evaluiert worden sind (Hardy, Jonen, Möller & Stern, 2006). In der ersten Einheit wurde das Dichtekonzept behandelt, in der zweiten Einheit ging es um die Konzepte von Verdrängung und Auftrieb. Die Einheiten umfassten jeweils 4.5 Doppelstunden. Der Komplexität des Stoffes für Grundschulkinder wurde mit der Aufteilung in zwei Unterrichtseinheiten Rechnung getragen. So war es auch möglich Prä-Tests, Unterrichtsqualitätsurteile und Post-Tests zu drei unterschiedlichen Zeitpunkten zu erfassen. Die standardisierten Leistungstests wurden vor und nach den Unterrichtseinheiten durchgeführt (Messzeitpunkte A/B und D, Abbildung 1). Eine Doppelstunde der ersten Unterrichtseinheit wurde von externen Beobachtern beurteilt. Die Urteile von Schülerinnen und Schülern und Lehrpersonen wurden nach Abschluss der ersten Unterrichtseinheit erhoben (Messzeitpunkt C, Abbildung 1).

3.3 Instrumente

Alle Items der Schüler-, Lehrer- und Beobachterinstrumente wurden auf einer vierstufigen Likert-Skala eingeschätzt (von 1 = *stimmt nicht* bis 4 = *stimmt genau*).

3.3.1 Schülerfragebogen

Der Schülerfragebogen umfasste 21 Items zu den Skalen kognitive Aktivierung (7 Items), unterstützendes Klima (9 Items) und strukturierte Klassenführung (5 Items), die auf die Unterrichtsqualität in den Einheiten zum Thema Schwimmen und Sinken fokussierten. Die Items wurden von bestehenden Instrumenten adaptiert (Diel & Höhner, 2008; Rakoczy, Buff & Lipowsky, 2005) und für den Gebrauch in der Grundschule überarbeitet. Bei der Konstruktion wurde besonders darauf geachtet, negative Formulierungen, Invertierungen und schwierige Begriffe zu vermeiden. Die Verständlichkeit der Items wurde in einer Pilotstudie mit Zweit- und Drittklässlern (N = 159 Schülerinnen und Schüler, 6 Klassen)

geprüft (Fauth et al., 2014). Die Reliabilität und faktorielle Validität des Schülerfragebogens konnte mittels Mehrebenen-Faktorenanalysen belegt werden (Fauth et al., 2014). Die Reliabilitäten der drei Skalen waren zufriedenstellend, sowohl hinsichtlich der internen Konsistenz (Cronbachs α), als auch hinsichtlich der Übereinstimmung der Urteile von Schülerinnen und Schülern innerhalb von Klassen (ICC2; Lüdtke, Trautwein, Kunter & Baumert, 2006, siehe Tabelle 1). In alle Analysen gingen die Schülerurteile als aggregierte Klassenmittelwerte ein.

3.3.2 Lehrerfragebogen

Der Lehrerfragebogen umfasste die Skalen kognitive Aktivierung (5 Items), unterstützendes Klima (8 Items) und strukturierte Klassenführung (4 Items), die inhaltlich parallel zu den Schülerskalen gehalten waren. Die Lehreritems wurden von bestehenden Instrumenten adaptiert (Baumert et al., 2004; Clausen, 2002; Kunter & Baumert, 2006; Rakoczy et al., 2005), die sich in der TIMS-Studie (Clausen, 2002) und in der Lehrerbefragung zum Unterricht im Rahmen von PISA 2003 (Baumert et al., 2004; Kunter & Baumert, 2006) bewährt haben. Die Skalen konnten mit einer akzeptablen Reliabilität erfasst werden (Tabelle 1). Anders als in früheren Studien (z.B. Kunter & Baumert, 2006) wurden hier nicht gleichlautende Items für Schülerinnen und Schüler sowie Lehrpersonen verwendet. Stattdessen wurde in den Formulierungen auf Angemessenheit für die jeweils befragte Personengruppe geachtet.

3.3.3 Beobachtungsinstrument

Die externen Beobachter gaben ihre Urteile zur Unterrichtsqualität auf hoch-inferenten Ratingitems ab. Aufgrund der begrenzten Aufmerksamkeitskapazitäten der Live-Beobachter war die Anzahl der Ratingitems begrenzt. Pro Basisdimension wurde ein Item eingeschätzt, das besonders repräsentativ für die jeweilige Basisdimension ist (vgl. Lipowsky et al., 2009) und inhaltlich mit den Schüler- und Lehrerinstrumenten korrespondiert: Herausfordernde

Unterrichtsgestaltung (kognitive Aktivierung), Anerkennung durch die Lehrperson (unterstützendes Klima) und Unterrichtsstörungen & präventives Lehrerhandeln (strukturierte Klassenführung; Tabelle 1). Die Items wurden für den Sachunterricht der Grundschule adaptiert aus Rakoczy und Pauli (2006). Für jedes Item wurden mehrere verhaltensnahe Indikatoren formuliert und in einem Ratingmanual festgehalten. Hinweise auf die Validität der Beobachtungsitems finden sich in den empirischen Ergebnissen der Pythagoras-Studie (Lipowsky et al., 2009). Alle Beobachter absolvierten eine Raterschulung mit einem Umfang von ca. 40 Stunden. Um die Objektivität der Urteile zu gewährleisten, wurden alle Videoratings und 50% der Liveratings von zwei unabhängigen Beobachtern vorgenommen. Diese gingen als Mittelwerte in die Analysen ein. Alle Items erreichten zufriedenstellende Beurteilerreliabilitäten (ICC > .70, vgl. Wirtz & Caspar, 2002; Tabelle 1).

3.3.4 Leistungstests

Der Lernerfolg wurde über das Wissen der Schülerinnen und Schüler zum Thema Schwimmen und Sinken operationalisiert und mittels standardisierter Leistungstests erfasst. Die Tests wurden aus der Studien von Hardy et al. (2006) und Hardy et al. (2010) adaptiert. Der Prä-Test umfasste 16 Items (EAP/PV-Reliabilität = .52) und der Post-Test nach der zweiten Unterrichtseinheit 13 Items (EAP/PV-Reliabilität = .76; Beispielitem: „Wie kommt es, dass ein riesiges, schweres Schiff aus Metall im Wasser nicht untergeht?“). Beide Tests wurden separat nach dem Partial Credit Model skaliert (Masters, 1982). Als Personenparameter wurden weighted likelihood estimates geschätzt (Warm, 1989). Als Intelligenzindikator wurde der CFT 20-R (Weiß, 2006) erhoben (56 Items, Cronbachs α = .72). Naturwissenschaftliche Kompetenz (Nawi-Kompetenz) wurde mit einem standardisierten Test in Anlehnung an den TIMSS 2007 Naturwissenschaftstest (Bos et al., 2007) gemessen. Dieser Test umfasste 12 Items (EAP/PV-Reliabilität = .70) und zeigte eine gute Passung zum 1PL-Rasch Modell.

3.4 Statistische Analysen und fehlende Werte

Die Zusammenhänge zwischen den Perspektiven wurden in einer Multitrait-Multimethod-Matrix (MTMM-Matrix; Campbell & Fiske, 1959) abgebildet. Zur Prüfung von Fragestellung 1 wurden die Validitätsdiagonalen (Monotrait-Heteromethod-Korrelationen) betrachtet. Zusätzlich können in der MTMM-Matrix Ergebnisse zur diskriminanten Validität der Urteile abgelesen werden. Hier ist die Frage, ob die Korrelationen der Validitätsdiagonalen größer sind als die übrigen Korrelationen in den jeweiligen Spalten und Zeilen der Heterotrait-Heteromethod-Dreiecke (Campbell & Fiske, 1959, S. 82; siehe Tabelle 1).

Die Analysen zur Vorhersage von Schülerleistungen wurden mittels Mehrebenen-Regressionsanalysen vorgenommen (Fragestellung 2). Die Urteile zur Unterrichtsqualität gingen als Prädiktoren auf Ebene 2 (Klassenebene) in die Analysen ein. Alle Kovariaten wurden als grand-mean-zentrierte Prädiktoren auf Ebene 1 (Individualebene) eingeführt (Lüdtke, Robitzsch, Trautwein & Kunter, 2009). Für die Beantwortung von Fragestellung 3 wurden Strukturgleichungsmodelle genutzt. So können latente Faktoren mit den Urteilen aus allen drei Perspektiven als Indikatoren modelliert und als Ebene-2-Prädiktoren in die Regressionsgleichung zur Leistungsvorhersage eingeführt werden. Um zu prüfen, ob es perspektivenspezifische oder gemeinsame Varianzanteile sind, die für die Leistungsentwicklung prädiktiv sind, wurden zusätzliche Modelle geschätzt, in denen die einzelnen Urteile zu einer Basisdimension gleichzeitig als Prädiktoren eingeführt wurden. In allen Mehrebenen-Regressionsmodellen wurden zur Vorhersage der Post-Test-Werte die folgenden Kovariaten einbezogen: der Prä-Test Wissen zum Thema Schwimmen und Sinken, naturwissenschaftliche Kompetenz und Intelligenz (CFT). Die Kovariaten sollten vorunterrichtliche individuelle Unterschiede zwischen den Schülerinnen und Schülern kontrollieren.

Signifikanztests wurden mit einem α-Niveau von 5% durchgeführt. Da bei Hypothese 2 die Forschungshypothese der Nullhypothese entspricht, wurde hier Bortz und Döring

(2006, S. 651) folgend ein α-Niveau von 10% zugrunde gelegt um das β-Fehler-Risiko zu senken.

Der Anteil an fehlenden Werten war in der vorliegenden Studie mit durchschnittlich 7.8% (Range: 6.8% – 8.8%) relativ klein. Fehlende Werte kamen zustande, wenn Kinder zu einem Messzeitpunkt aufgrund von Krankheit, Klassenwechsel oder Umzug nicht an den Erhebungen teilgenommen haben. Fälle mit fehlenden Werten auf den jeweiligen Prädiktorvariablen wurden in den Analysen nicht berücksichtigt. Auf der in der vorliegenden Untersuchung vor allem interessierenden Klassenebene (Ebene 2) gab es keine fehlenden Werte. Alle Analysen wurden mit MPlus 7 (Muthén & Muthén, 1998–2012) unter Verwendung eines Robust Maximum Likelihood Schätzers (MLR) durchgeführt.

4 Ergebnisse

4.1 Deskriptive Ergebnisse

Die Mittelwerte aller drei Beurteilerperspektiven liegen am oberen Ende der vierstufigen Skala. Einzig die Schülerurteile zur Klassenführung weisen einen Wert unter drei auf. Innerhalb der Perspektiven sind beinahe alle Skalen signifikant korreliert (Ausnahme: unterstützendes Klima und kognitive Aktivierung aus Beobachtersicht; Tabelle 1).

4.2 Perspektivenvergleiche

Wie erwartet findet sich beim unterstützenden Klima eine signifikante Korrelation zwischen den Urteilen der Schülerinnen und Schüler und der Lehrpersonen (Hypothese 1). Dagegen finden sich keine signifikanten Zusammenhänge zwischen den Urteilen zur kognitiven Aktivierung (Hypothese 2). Signifikante Zusammenhänge zwischen allen drei Perspektiven finden sich im Bereich der strukturierten Klassenführung (Hypothese 3). Die signifikanten Korrelationen zwischen unterschiedlichen Urteilen zum selben Konstrukt sind die höchsten in den jeweiligen Zeilen und Spalten der Heterotrait-Heteromethod-Dreiecke, was als ein Hinweis auf die diskriminante Validität der Urteile gewertet werden kann (Clausen, 2002, S.

130). Wo die Zusammenhänge der Urteile zum selben Konstrukt nicht signifikant sind (Urteile aller drei Perspektiven zur kognitiven Aktivierung und Beobachterurteile zum unterstützenden Klima), ist auch diese diskriminante Validität nicht gegeben (Tabelle 1).

4.3 Prädiktion von Schülerleistungen

Die einbezogenen Kovariaten waren in allen Modellen signifikante Prädiktoren für das Kriterium (Tabelle 2). Wie in Hypothese 4 angenommen, zeigten sich auf der Klassenebene signifikante Effekte der Urteile externer Beobachter zur kognitiven Aktivierung und zur strukturierten Klassenführung. Außerdem gab es einen signifikanten Effekt der Beobachterurteile zum unterstützenden Klima. Auch die Urteile der Lehrpersonen zur strukturierten Klassenführung hingen signifikant mit dem Lernerfolg zusammen. Allerdings fand sich entgegen der in Hypothese 5 formulierten Erwartungen kein Effekt der Lehrerurteile zur kognitiven Aktivierung. Für die Schülerurteile findet sich der in Hypothese 6 angenommene Effekt der Urteile zur strukturierten Klassenführung.

Auf die Ergebnisse des Perspektivenvergleichs aufbauend, wurde ein latenter Faktor strukturierte Klassenführung mit den Urteilen von Schülerinnen und Schülern, Lehrpersonen und Beobachtern als Indikatoren gebildet (standardisierte Faktorladungen: .71, .66 und .72). Dieses Modell zeigte eine sehr gute Passung zu den empirischen Daten ($\chi^2(2) = 1.39$, $p = .50$, CFI = 1.0, $SRMR_{between} = .02$; Hypothese 7). Wie in Hypothese 8 angenommen, hing auch dieser latente Faktor signifikant mit dem Lernerfolg der Schülerinnen und Schüler zusammen.

Zusätzlich zu den oben beschriebenen Analysen wurde geprüft, ob es perspektivenspezifische oder gemeinsame Varianzanteile sind, die für die Leistungsentwicklung prädiktiv sind. Im Bereich kognitive Aktivierung sind es wie zu erwarten perspektivenspezifische Varianzkomponenten der externen Beobachter, die prädiktiv sind. Im Bereich strukturierte Klassenführung haben nur die Schülerurteile eine

spezifische Vorhersagekraft, jenseits der mit den anderen Perspektiven geteilten Varianzanteile. Der Effekt der Beobachterurteile zum unterstützenden Klima ist trotz der geringen Korrelationen mit den anderen Perspektiven auf gemeinsame Varianzkomponenten zurückzuführen.

5 Diskussion

Der vorliegende Beitrag ergänzt bisherige Studien aus dem Sekundarschulbereich mit Befunden zu Konvergenzen und Divergenzen der Urteile von Schülerinnen und Schülern, Lehrpersonen und externen Beobachtern in der Grundschule. Ähnlich wie in der Sekundarstufe wird auch im Grundschulbereich das komplexe Konstrukt Unterrichtsqualität über die Perspektiven hinweg nicht einheitlich wahrgenommen. Darüber hinaus zeigen die Ergebnisse, dass die Vorhersagekraft der Urteile für Lernerfolg sowohl von der gewählten Perspektive als auch von der betrachteten Basisdimension abhängt. Die Studie liefert damit einen Beitrag zur präziseren Erfassung dessen, was sich hinter dem häufig verwendeten Label *Unterrichtsqualität* verbirgt. Das theoretische Modell der Basisdimensionen von Unterrichtsqualität erlaubt es – im Zusammenspiel mit einer Analyse der Stärken und Grenzen der einzelnen Perspektiven – theoretisch und empirisch fundierte Erwartungen an Zusammenhänge zwischen den Perspektiven und mit externen Kriterien zu formulieren und zu prüfen.

5.1 Kognitive Aktivierung

Die Zusammenhänge zwischen den Perspektiven im Bereich kognitive Aktivierung waren auch unter Anwendung eines erhöhten α-Fehler-Niveaus von 10% nicht bedeutsam. Obgleich auch dieses Niveau letztlich arbiträr ist (Bortz & Döring, 2006, S. 651), weisen die Ergebnisse nicht auf große Konvergenzen der Urteile hin. Sowohl dieses Ergebnis als auch die fehlende Diskriminanz der Urteile stimmen mit Befunden zu kognitiver Aktivierung im Sekundarbereich überein (Clausen, 2002, S. 130; Kunter & Baumert, 2006). Das Konstrukt

bringt es mit sich, dass in die Schülerurteile zum Teil Einschätzungen der eigenen kognitiven Aktiviertheit mit einfließen, die Lehrpersonen und Beobachtern von außen per se nicht zugänglich sind.

Umso bemerkenswerter ist der Befund, dass gerade die Urteile der externen Beobachter Leistungsentwicklungen vorhersagen und mithin als prädiktiv valide für Schülerleistungen gesehen werden können. Aus konstruktivistischer Sicht hätte man annehmen können, dass nicht so sehr der „objektive“ Anregungsgehalt einer Lernumgebung entscheidend ist, sondern in welchem Ausmaß dieser von den Schülerinnen und Schülern auch als aktivierend erlebt wird (Waldis et al., 2010, S. 172). In der dritten Jahrgangsstufe scheint es jedoch – wenn es um die Entwicklung von Leistungsunterschieden zwischen Klassen geht – nicht darauf anzukommen, wie kognitiv aktiviert sich die Schülerinnen und Schüler einer Klasse fühlen, sondern als wie kognitiv aktivierend ein Unterricht von geschulten Beobachtern eingeschätzt wird. Es mag sein, dass Grundschulkinder Schwierigkeiten bei der Beurteilung dieser methodisch-didaktisch anspruchsvollen Dimension haben. Es herrscht bei den Schülerinnen und Schülern zwar eine hinreichende Einigkeit in der Einschätzung der Skala kognitive Aktivierung (angezeigt durch den ICC2, siehe Tabelle 1), allerdings hat das Konstrukt, das hier erfasst wird, nur wenig mit der kognitiven Aktivierung zu tun, wie sie von Lehrpersonen und Videobeobachtern beurteilt wird. Dasselbe gilt für die Urteile der Lehrpersonen, die eigentlich genügend methodisch-didaktische Expertise zur Einschätzung von kognitiver Aktivierung mitbringen müssten. Der fehlende Effekt der Lehrerurteile könnte mit selbstwertdienlichen Verzerrungen zusammenhängen. Es ist plausibel, dass hiervon gerade Urteile zur kognitiven Aktivierung betroffen sein können, da hier das professionelle Selbstverständnis unmittelbar berührt ist. Waldis et al. (2010, S. 180) machen darauf aufmerksam, dass die empirischen Befunde zur Bedeutung der kognitiven Aktivierung für die Leistungsentwicklung weniger eindeutig sind, als dies von der theoretischen Konzeption des Konstrukts her zu erwarten ist. In künftigen Untersuchungen sollte daher geprüft werden, unter welchen Bedingungskonstellationen

Urteile zur kognitiven Aktivierung erhalten werden können, die prädiktiv valide für Lernerfolg sind. Dabei sollten auch Faktoren wie das Forschungsdesign und die Wahl der Leistungsmaße berücksichtigt werden (z.B. Fokus auf spezifisches Unterrichtsthema vs. Bezug zu einem Fach; kurzfristige Wissenszuwächse vs. Kompetenzzuwächse über ein Schuljahr).

5.2 Unterstützendes Klima

Wie in früheren Studien aus dem Sekundarbereich (Clausen, 2002; De Jong & Westerhof, 2001; Kunter & Baumert, 2006) fanden sich in der vorliegenden Studie im Bereich unterstützendes Klima signifikante Zusammenhänge zwischen der Lehrer- und Schülerperspektive, nicht jedoch mit der Beobachterperspektive. Das gefundene Korrelationsmuster kann mit Besonderheiten der Beobachtbarkeit des Konstrukts erklärt werden. Obgleich sich die Lehrer- und Schüleritems in der vorliegenden Studie nur auf den begrenzten Zeitraum der Unterrichtseinheiten zum Schwimmen und Sinken beziehen, ist davon auszugehen, dass auch vorher gemachte Erfahrungen in die Urteile mit einfließen. Dieser „Vorteil“ der Schüler-Lehrer-Übereinstimmung sollte vor allem bei schwierig zu beobachtenden Konstrukten wie dem unterstützenden Klima zum Tragen kommen (Waldis et al., 2010, S. 174). Hinzu kommt, dass externe Beobachter dem Unterrichtsgeschehen nur passiv folgen können, während Lehrpersonen und Schülerinnen und Schüler einmal gebildete Erwartungen an das Verhalten des jeweils anderen in der Interaktion überprüfen können (Clausen, 2002, S. 83f.).

Das unterstützende Klima wird in der Literatur insbesondere mit motivationalen Outcomes in Verbindung gebracht (Fauth et al., 2014; Kunter et al., 2013). In der vorliegenden Studie zeigt sich jedoch entgegen unserer Erwartungen auch ein kleiner Effekt der Beobachterurteile auf die Leistungsentwicklung. Es scheint leistungsförderliche Elemente eines unterstützenden Klimas zu geben, die allerdings nur in den Urteilen der externen Beobachter prononciert genug erfasst wurden, um in der Vorhersage von

Leistungsentwicklung sichtbar zu werden. Es mag zudem sein, dass die Anerkennung durch die Lehrperson für die Leistungsentwicklung in der Grundschule eine größere Rolle spielt als in der Sekundarstufe.

5.3 Strukturierte Klassenführung

Die substanziellen Übereinstimmungen und die diskriminanten Validitäten der Urteile zur strukturierten Klassenführung (vgl. auch Kunter & Baumert, 2006; Waldis et al., 2010) können mit der guten Beobachtbarkeit des Konstrukts anhand von verhaltensnahen Indikatoren erklärt werden. Zur Einschätzung der Klassenführung benötigen Beurteiler weder besonderes pädagogisch-didaktisches Verständnis, noch ist das Konstrukt besonders anfällig für selbstwertdienliche Verzerrungen seitens der Lehrperson. Die Ergebnisse legen nahe, dass im Bereich Klassenführung aus den unterschiedlichen Perspektiven tatsächlich in erheblichem Ausmaß dasselbe Konstrukt erfasst wird.

Anders als in früheren Studien (z.B. Kunter & Baumert, 2006) waren in der vorliegenden Studie auch die Urteile von Lehrpersonen bedeutsam für die Vorhersage des Lernerfolgs. Dies mag mit dem Design der Studie zusammenhängen, in dem nicht die Leistungsentwicklung über ein Schuljahr hinweg, sondern die spezifischen Effekte des Unterrichts auf den Lernerfolg in einem eng umgrenzten Themenfeld untersucht wurde. In Bezug auf die Urteile der Grundschulkinder ist bemerkenswert, dass diese in der vorliegenden Studie deskriptiv mehr Varianz des Kriteriums aufklären (21%) als die Urteile externer Beobachter (14%). Die Schülerurteile haben zudem eine spezifische Vorhersagekraft jenseits der durch die anderen Perspektiven erklärten Leistungsunterschiede (siehe Abschnitt 4.3). Auch mit Blick auf die deskriptiven Daten (Abschnitt 4.1) kann vermutet werden, dass insbesondere die Schülerurteile ein realistisches Bild der Klassenführung abgeben (vgl. Kane et al., 2013, S. 10).

Eine besondere Qualität des kombinierten Faktors strukturierte Klassenführung ist, dass in ihm die oben beschriebenen Nachteile der einzelnen Urteile, die sich in den einzelnen

Skalen als Messungenauigkeiten niederschlagen, ausgeglichen werden. Geht man also von einem perspektivenunspezifischen Konstrukt strukturierte Klassenführung aus, dann ist der kombinierte Faktor ein genauerer Schätzer dieses Konstrukts (Kunter et al., 2013).

5.4 Implikationen für Forschung und Praxis

Maße zur Qualität von Unterricht gewinnen sowohl in der empirischen pädagogisch-psychologischen Forschung als auch in der Praxis, z.B. im Rahmen von Schulevaluationen, zunehmend an Bedeutung (Kunter & Trautwein, 2013; Pietsch, 2010). Der Abgleich unterschiedlicher Perspektiven kann in der pädagogischen Praxis zudem wertvolle Anstöße zur Reflexion von Wahrnehmungsunterschieden geben, die wiederum Maßnahmen zur Unterrichtsentwicklung anstoßen können (Helmke & Helmke, 2004). Für die Verwendung von Unterrichtsmaßen in der Forschung kann festgehalten werden, dass sich der vergleichsweise große technische Aufwand von (Video-)Beobachtungen zu lohnen scheint, wenn es um die Erfassung von pädagogisch-didaktisch anspruchsvollen Konstrukten wie kognitive Aktivierung geht. Demgegenüber sind es gerade die Schülerurteile zur strukturierten Klassenführung, die einen besonders präzisen Einblick in diesen Aspekt des Unterrichts erlauben. Im Fall von Video- und Beobachtungsstudien ist es daher sinnvoll, die Schülerurteile mit zu erheben.

5.5 Stärken und Grenzen der Studie

Eine Besonderheit der Studie ist, dass sich die Items und Beobachtungsindikatoren auf den begrenzten Zeitraum von einer Unterrichtseinheit beziehen. Anders als in früheren Studien, in denen eher global z. B. nach „dem Mathematikunterricht“ gefragt wurde (z.B. Clausen, 2002, S. 112), ist der Gegenstand der Beobachtung so deutlicher umrissen. Durch den Fokus auf ein bestimmtes Thema wurde es auch möglich, die Unterrichtseinheiten für alle Klassen vergleichbar zu gestalten, was es leichter macht, die gefundenen Effekte in der Leistungsentwicklung auch tatsächlich auf den Unterricht zurückzuführen. Allerdings kann

durch die freiwillige Teilnahme nicht ausgeschlossen werden, dass es sich bei den untersuchten Lehrpersonen um eine selektive Stichprobe handelt, was die Generalisierbarkeit der Ergebnisse einschränken würde. Zudem bleibt es eine Schwäche des Vergleichs der unterschiedlichen Perspektiven, dass die externen Beobachter ihr Urteil auf der Basis von nur einer Doppelstunde trafen (Praetorius et al., 2013), während Schüler- und Lehrerurteile sich auf die gesamte erste Unterrichtseinheit beziehen. Hier sind schon designbedingt höhere Übereinstimmungen zwischen Schülern und Lehrpersonen zu erwarten. Eine weitere Grenze ist, dass von den Beobachtern pro Basisdimension nur ein Ratingitem eingeschätzt wurde. Obgleich diese Items jeweils auf der Grundlage mehrerer Indikatoren beurteilt wurden, kann es sein, dass die Beobachtungsitems die jeweiligen Konstrukte nicht in ihrer ganzen Breite erfassen konnten. Umso bemerkenswerter scheint es allerdings, dass die Urteile externer Beobachter insgesamt die vergleichsweise größte Erklärungskraft für den Lernerfolg der Schülerinnen und Schüler hatten.

Literatur

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

Baumert, J., Kunter, M., Brunner, M., Krauss, S., Blum, W. & Neubrand, M. (2004). Mathematikunterricht aus Sicht der PISA-Schülerinnen und -Schüler und ihrer Lehrkräfte. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. et al. (Hrsg.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland* (S. 314–354). Münster: Waxmann.

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

Bos, W., Bonsen, M., Baumert, J., Prenzel, M., Selter, C. & Walther, G. (Hrsg.). (2007). *TIMSS 2007 – Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster, New York, München, Bern: Waxmann.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminat Validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

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

De Jong, R. & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*, 51–85.

Decristan, J., Hondrich, A. L., Büttner, G., Hertel, S., Klieme, E., Kunter, M. et al. (in press). Impact of additional guidance in science education on primary students' conceptual understanding. *The Journal of Educational Research*.

Desimone, L. M., Smith, T. M. & Frisvold, D. E. (2010). Survey measures of classroom instruction comparing student and teacher reports. *Educational Policy, 24*, 267–329.

Diel, E. & Höhner, W. (2008). *Fragebögen zur Unterrichtqualität*. Wiesbaden: Institut für Qualitätsentwicklung.

Dubberke, T., Kunter, M., McElvany, N., Brunner, M. & Baumert, J. (2008). Lerntheoretische Überzeugungen von Mathematiklehrkräften. *Zeitschrift für Pädagogische Psychologie, 22*, 193–206.

Einsiedler, W. & Hardy, I. (2010). Kognitive Strukturierung im Unterricht. Einführung und Begriffsklärungen. *Unterrichtswissenschaft, 38*, 194–209.

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

Good, T. L., Wiley, C. & Florez, I. R. (2009). Effective teaching: An emerging synthesis. In G. Dworkin (Ed.), *International handbook of research on teachers and teaching* (pp. 803–816). New York: Springer.

Hardy, I., Hertel, S., Kunter, M., Klieme, E., Warwas, J., Büttner, G. et al. (2011). Adaptive Lerngelegenheiten in der Grundschule: Merkmale, methodisch-didaktische Schwerpunktsetzungen und erforderliche Lehrerkompetenzen. *Zeitschrift für Pädagogik, 57*, 819–833.

Hardy, I., Jonen, A., Möller, K. & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking". *Journal of Educational Psychology, 98*, 307–326.

Hardy, I., Kleickmann, T., Koerber, S., Mayer, D., Möller, K., Pollmeier, J. et al. (2010). Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung.* (S. 115–125). Weinheim: Beltz.

Helmke, A. & Helmke, T. (2004). Videobasierte Unterrichtsreflexion. *Seminar – Lehrerbildung und Schule, 10*, 48-66.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts*. Stuttgart: Kallmeyer.

Helmke, A., Helmke, T., Schrader, F. W., Wagner, W., Klieme, E., Nold, G. et al. (2008). Wirksamkeit des Englischunterrichts. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch* (S. 382–397). Weinheim: Beltz.

Herman, J., Klein, D. & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice, 19*, 16–24.

Kane, T. J., McCaffrey, D. F., Miller, T. & Staiger, D. O. (2013). *Have we identified effective teachers?* MET Project Research Paper, Bill & Melinda Gates Foundation.

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

Klieme, E., Pauli, C. & Reusser, K. (2009). The Pythagoras study. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster: Waxmann.

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

Kloss, J. (2012). *Grundschüler als Experten für Unterricht? Empirische Überprüfung der Validität von Unterrichtsbeurteilungen durch Schüler der dritten und vierten Jahrgangsstufe.* Unveröffentlichte Dissertation, Universität Erfurt.

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

Kunter, M. & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*, 231–251.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T. & Hachfeld, A. (2013). Professional competence of teachers: Effects on quality and student development. *Journal of Educational Psychology, 29*.

Kunter, M. & Trautwein, U. (2013). *Psychologie des Unterrichts*. Paderborn: Schöningh.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E. & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean theorem. *Learning and Instruction, 19*, 527–537.

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*, 120–131.

Lüdtke, O., Trautwein, U., Kunter, M. & Baumert, J. (2006). Analyse von Lernumwelten Ansätze zur Bestimmung der Reliabilität und Übereinstimmung von Schülerwahrnehmungen. *Zeitschrift für Pädagogische Psychologie, 20*, 85–96.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Mayer, D. P. (1999). Measuring instructional practice. *Educational Evaluation and Policy Analysis, 21*, 29–45.

Mihaly, K., McCaffrey, D. F., Douglas O., Staiger, D. O &. Lockwood, J. R. (2013). *A Composite Estimator of Effective Teaching*. MET Project Research Paper, Bill & Melinda Gates Foundation.

Möller, K. & Jonen, A. (2005). *Die KiNT-Boxen – Kinder lernen Naturwissenschaft und Technik. Klassenkisten für den Sachunterricht. Paket 1: Schwimmen und Sinken*. Essen: Spectra-Verlag.

Muthén, L. K. & Muthén, B. O. (1998–2012). *Mplus user's guide (7th ed.)*. Los Angeles, CA: Muthén & Muthén.

Pauli, C., Drollinger-Vetter, B., Hugener, I. & Lipowsky, F. (2008). Kognitive Aktivierung im Mathematikunterricht. *Zeitschrift für Pädagogische Psychologie, 22,* 127–133.

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

Pietsch, M. (2010). Evaluation von Unterrichtsstandards. *Zeitschrift für Erziehungswissenschaft, 13*, 121–148.

Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K. & Klieme, E. (in press). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Intsruction*.

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht*. Münster: Waxmann.

Rakoczy, K. & Pauli, C. (2006). Hoch inferentes Rating. Beurteilung der Qualität unterrichtlicher Prozesse. In I. Hugener, C. Pauli & K. Reusser (Hrsg.), *Videoanalysen. Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie* (S. 206–233). Frankfurt am Main: GFPF.

Rakoczy, K., Buff, A. & Lipowsky, F. (2005). *Befragungsinstrumente. Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie*. Frankfurt am Main: GFPF.

Ryan, R. M. & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55,* 68–78.

Scheerens, J. & Bosker, R. J. (1997). The Foundations of educational effectiveness. *International Review of Education, 45,* 113–120.

Seidel, T. & Shavelson, R. (2007). Teaching effectiveness research in the past decade. *Review of Educational Research, 77,* 454–499.

Stigler, J. W. & Hiebert, J. (1999). *The teaching gap*. NewYork: Free Press.

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* (S. 171–208). Münster: Waxmann.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Weiß, R. H. (2006). *CFT 20-R. Grundintelligenztest Skala 2 - Revision*. Göttingen: Hogrefe.

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10,* 1–30.

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

Wubbels, T., Brekelmans, M. & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education, 8,* 47–58.

Benjamin Fauth
IDeA-Forschungszentrum
Solmsstr. 73
60486 Frankfurt am Main
Deutschland
fauth@dipf.de

Tabelle 1

Beispielitems und -beobachtungsindikatoren, Reliabilitäten, deskriptive Statistiken und MTMM-Matrix (bivariate Korrelationen)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	α[a]	ICC[b]	M	SD	Instruktion und Beispielitems bzw. -indikatoren
Schüler														Im Unterricht zum Schwimmen und Untergehen…
(1) Kog. Aktivierung	-									.91	.73	3.19	.28	haben wir Aufgaben gemacht, über die ich ganz genau nachdenken musste.
(2) Unterst. Klima	.83*	-								.95	.78	3.30	.31	Unsere Lehrerin im Unterricht zum Schwimmen und Untergehen war freundlich zu mir.
(3) Klassenführung	.28*	.37*	-							.97	.86	2.56	.47	Im Unterricht zum Schwimmen und Untergehen hat keiner den Unterricht gestört.
Lehrpersonen														Im Unterricht zu Schwimmen und Sinken…
(4) Kog. Aktivierung	**.22**	*.22*	*.32**	-						.67	-	3.51	.39	habe ich Aufgaben gestellt, über die die Schüler genau nachdenken mussten.
(5) Unterst. Klima	*.25*	**.37***	*.34**	.56*	-					.79	-	3.34	.39	habe ich mich um ein wertschätzendes Verhältnis zu den Schülern bemüht.
(6) Klassenführung	*.13*	*.21*	**.45***	.31*	.50*	-				.86	-	3.11	.58	haben die Schüler den Unterricht nur selten gestört.
Beobachter														
(7) Kog. Aktivierung	**-.05**	*-.06*	*.13*	**.13**	*.08*	*.14*	-			-	.77	3.17	.84	Es werden offene Fragen gestellt, die zum Nachdenken anregen.
(8) Unterst. Klima	*.03*	**.12**	*.29**	*.10*	**.20**	*.47**	.23	-		-	.72	3.08	.74	Der Umgang mit den SchülerInnen ist wertschätzend.
(9) Klassenführung	*-.03*	*.04*	**.52***	*.11*	*.33**	**.54***	.46*	.55*	-	-	.81	3.34	.78	Der Unterricht wird nicht stark gestört.

Anmerkungen: [a]Bei Schülerskalen: Cronbachs α der Klassenmittelwerte. [b]Angegeben sind bei den Schülerskalen der ICC2 und bei den Beobachteritems der ICC1. Konvergente Validitäten (Monotrait-Heteromethod) sind fett gedruckt. Heterotrait-Heteromethod-Dreiecke sind kursiv.

* $p < .05$ (einseitige Testung).

Tabelle 2

Mehrebenen Regressionsanalysen. Urteile von Schülern, Lehrpersonen und Beobachtern zur Unterrichtqualität. Abhängige Variable: Wissen zum Thema Schwimmen und Sinken

Prädiktor	Schüler	Lehrpersonen	Beobachter	Kombiniert
Individualebene				
Prä-Test	.21* (.03)	.22* (.03)	.22* (.03)	.23* (.03)
CFT	.23* (.03)	.22* (.03)	.22* (.03)	.21* (.03)
Nawi-Kompetenz	.28*[1] (.03)	.27* (.03)	.26* (.03)	.27* (.03)
Klassenebene				
Kognitive Aktivierung	.03 (.12)	.22 (.15)	.43* (.13)	-
R^2 (between)	.00	.05	.19	-
Unterstützendes Klima	.12 (.14)	.18 (.16)	.24* (.14)	-
R^2 (between)	.02	.03	.06	-
Strukturierte Klassenführung	.46* (.13)	.30* (.14)	.37* (.14)	.57* (.16)
R^2 (between)	.21	.09	.14	.32

Anmerkungen: Die drei Basisdimensionen von Unterrichtsqualität wurden jeweils einzeln in die Analysen eingeführt (die Werte der Kovariaten gelten für die einzelnen Modelle gleichermaßen). Angegeben sind standardisierte Regressionsgewichte; in Klammern: Standardfehler.

* $p < .05$ (einseitige Testung).

[1] Im Modell mit Schülerurteilen zu strukturierter Klassenführung als Prädiktor beträgt das beta-Gewicht der Nawi-Kompetenz .29.

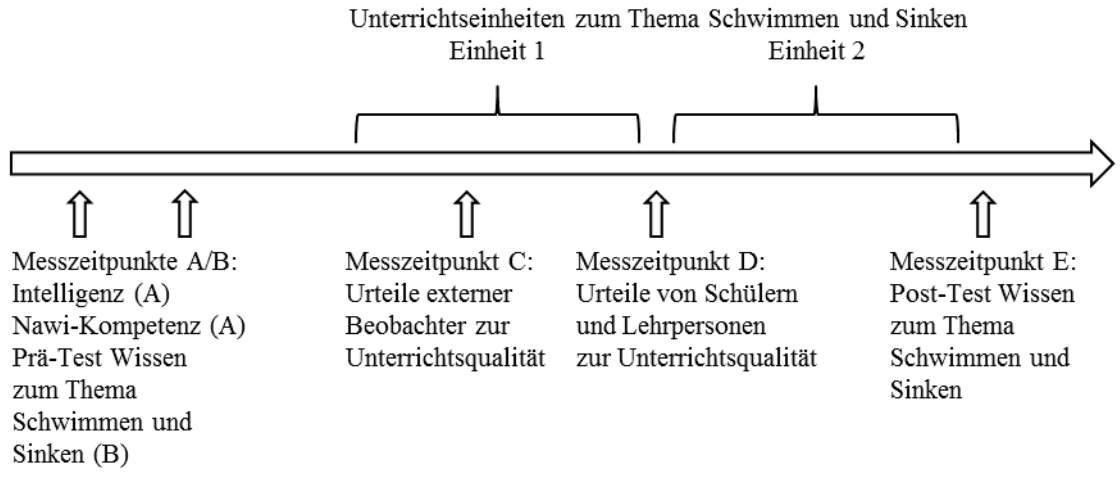

Abbildung 1. Design der Studie, Messzeitpunkte und erfasste Konstrukte.