Die Nutzung von Schülerurteilen zur Erfassung von Unterrichtsqualität in Forschung und Schulpraxis

Dissertation

zur Erlangung des Doktorgrades

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Ann-Kathrin Jaekel
aus Ulm

Tübingen

2021

1. Betreuer:	Prof. Dr. Richard Göllner
2. Betreuer:	Prof. Dr. Benjamin Fauth
3. Betreuer:	Prof. Dr. Ulrich Trautwein
Tag der mündlichen Prüfung:	09. Juli 2021
Dekan:	Prof. Dr. Josef Schmid
1. Gutachter:	Prof. Dr. Richard Göllner
2. Gutachter:	Prof. Dr. Benjamin Fauth

DANKSAGUNG

Ich möchte mich ganz besonders bei Prof. Dr. Richard Göllner, Prof. Dr. Benjamin Fauth, Prof. Dr. Ulrich Trautwein und Dr. Wolfgang Wagner bedanken. Vielen Dank für eure Mühen, Ideen, Ratschläge und die vielen hilfreichen Gespräche, die zu dieser Arbeit beigetragen haben. Ich habe unglaublich viel von euch gelernt und bin euch dafür wirklich dankbar. Durch eure Unterstützung habe ich den Spaß an der Arbeit nie verloren. Richard, danke für alles!

Danke auch an meine Kolleginnen und Kollegen, die mittlerweile vor allem auch Freunde geworden sind: Lisa Z., Johanna, Till, Kathi, Anna, Fabienne, Lisa H., Alfi. Vielen Dank für die vielen lustigen und schönen Momente, die mir sehr durch die letzten Jahre geholfen haben. Ich freue mich auf viele weitere!

Schließlich möchte ich mich bei Personen bedanken, die mich schon seit vielen Jahren begleiten. Tini, Andi, Anton und Michel: Danke für unermüdliches Nachfragen, was ich da nochmal mache, euer Vertrauen, das mit nichts aufzuwiegen ist, interessante Diskussionen sowie Bücher, Sandkästen und Duplo-Bauwerke, die mich alle Anstrengungen für einen Moment haben vergessen lassen. Lisa K., Annika und Angy: Ihr habt mich sehr geprägt. Zuletzt und ganz besonders möchte ich meiner Familie danken. Danke, dass ihr immer für mich da seid und mich unterstützt!

ZUSAMMENFASSUNG

Unterricht wird in der empirischen Unterrichtsforschung als ein komplexes Wirkungsgefüge verstanden, das von unterschiedlichen Faktoren wie der Expertise und Persönlichkeit der Lehrkraft, von Kontextmerkmalen, aber auch von individuellen Lernvoraussetzungen der Schülerinnen und Schüler beeinflusst wird (Helmke, 2017). Damit ist Unterricht ein interaktives Geschehen, das von der Lehrkraft gesteuert und gelenkt wird, jedoch durch die Schülerinnen und Schüler aktiv mitgestaltet wird (Klieme, 2019). Ein wesentliches Merkmal für das Erreichen von Lernzielen der Schülerinnen und Schüler stellt die Unterrichtsqualität dar. In den vergangenen Jahren hat sich besonders in der deutschsprachigen Forschungslandschaft das „Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität“ (Klieme et al., 2001) etabliert, das Unterrichtsqualität durch die drei Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung abbildet. Diese Dimensionen sind mittlerweile Grundlage vieler Unterrichtsstudien und haben sich vielfach als relevant für verschiedene Zielkriterien des Unterrichts gezeigt (Göllner et al., 2018; Kunter et al., 2013; Wagner et al., 2013, 2016). Eine zentrale Herausforderung bei der Identifikation von lernförderlichen Merkmalen des Unterrichts ist jedoch eine reliable und valide Erfassung der Unterrichtsqualität. Hierfür sind Schülerurteile eine kostengünstige, effektive und niederschwellige Möglichkeit, um die Unterrichtsqualität aus der Perspektive der Schülerinnen und Schüler zu erfassen. Die Eignung von Schülerurteilen zur Erfassung von Unterrichtsqualität konnte in der bisherigen Forschung in vielerlei Hinsicht belegt werden (Fauth et al., 2014; Kuhfeld, 2017; Wagner et al., 2013). Dennoch sind im Hinblick auf die Nutzung von Schülerurteilen in Forschung und Praxis noch wichtige Fragen offen. Drei dieser Fragen wurden in der vorliegenden Dissertation in empirischen Studien adressiert, um so einen Beitrag zur verlässlichen Nutzung von Schülerurteilen zu leisten.

Studie 1 (*Digital teaching during the COVID-19 crisis: Social connectedness matters most for teaching quality and students' learning*) beschäftigte sich erstmals mit der Frage, ob Schülerurteile auch für den Distanzunterricht eine geeignete Methode zur Erfassung von Unterrichtsqualität sind. Während die Nutzung von Schülerurteilen im Präsenzunterricht in der Vergangenheit vielfach erforscht wurde, konnte dieser Frage durch die Schulschließungen aufgrund der COVID-19-Pandemie erstmals auch für den Distanzunterricht nachgegangen werden. Hierfür wurden Daten einer Onlinestudie genutzt, in welcher im Frühsommer 2020 die konkrete Umsetzung des Distanzunterrichts, Subdimensionen der Unterrichtsqualität

sowie verschiedene Zielkriterien des Unterrichts aus Schülersicht erfasst wurden. Die Ergebnisse ermöglichten nicht nur wertvolle Einblicke in das Unterrichtsgeschehen während der Schulschließungen, sondern zeigten zudem, dass Schülerurteile auch für den Distanzunterricht eine geeignete Methode zur Erfassung von Unterrichtsqualität sind. Die Urteile von Schülerinnen und Schülern zur Unterrichtsqualität waren zudem mit Zielkriterien des Unterrichts assoziiert.

In Studie 2 („*The teacher motivates us – or me?" – The role of the addressee in student ratings of teacher support*) wurde die Rolle der Itemformulierung zur Erfassung von Unterrichtsqualität näher betrachtet. Items unterscheiden sich häufig darin, ob sie sich auf die individuelle Sicht der Schülerinnen und Schüler beziehen (Ich-Adressat) oder auf die Sicht der gesamten Klasse (Wir-Adressat). Möglichen Unterschieden in den psychometrischen Eigenschaften beider Versionen wurde jedoch in der bisherigen Forschung kaum Beachtung geschenkt. Eine experimentelle Variation des Item-Adressaten in Subdimensionen der konstruktiven Unterstützung ermöglichte es zu untersuchten, ob sich die theoretischen Unterschiede zwischen einem Ich-Adressaten und einem Wir-Adressaten in den psychometrischen Eigenschaften der Schülerurteile niederschlagen. Die Ergebnisse zeigten, dass Items mit einem Wir-Adressaten zu leicht höheren Mittelwerten, ICCs und Interkorrelationen unterschiedlicher Subdimensionen bei gleichem Adressaten auf Klassenebene führten. Jedoch zeigten sich höhere Zusammenhänge für Zielkriterien des Unterrichts auf Schülerebene, wenn Items mit einem Ich-Adressaten verwendet wurden. Auch wenn die gefundenen Unterschiede relativ klein sind, verdeutlicht diese Studie die Relevanz des Item-Adressaten für die Erfassung von Unterrichtsqualität durch Schülerurteile.

Studie 3 (*How students' perceptions of teaching quality in one subject are impacted by the grades they receive in another subject – Dimensional comparisons in student evaluations of teaching quality*) ging schließlich der Frage nach, welche Bedeutung die Note für Schülerurteile zur Erfassung der Unterrichtsqualität hat. Speziell wurde untersucht, ob Schülerurteile zur Erfassung von Unterrichtsqualität eines Faches (z. B. Deutsch) durch die Note eines anderen Faches (z. B. Mathematik) beeinflusst sind. Die Ergebnisse zeigten auf Schüler- und auf Klassenebene positive Zusammenhänge zwischen der Note und den Schülerurteilen zur Unterrichtsqualität innerhalb eines Faches sowie negative Zusammenhänge zwischen den Fächern. Dies bedeutet, dass Schülerinnen und Schüler die Unterrichtsqualität desjenigen Faches, in dem sie die bessere (schlechtere) Note erhalten haben, relativ betrachtet aufwerten (abwerten). Schülerurteile eines Faches können somit durch Merkmale beeinflusst sein, die unabhängig von der zu beurteilenden Unterrichtsqualität

des jeweiligen Faches sind.

Die Ergebnisse der Studien dieser Dissertation tragen zur verlässlichen Nutzung von Schülerurteilen in Forschung und Praxis bei. Es zeigte sich, dass Schülerurteile sowohl für den Präsenzunterricht als auch für den Distanzunterricht eine geeignete Methode zur Erfassung von Unterrichtsqualität sind. Zudem wurde die Relevanz des Item-Adressaten deutlich, der sich in kleinen, aber systematischen Unterschieden in den psychometrischen Eigenschaften der Schülerurteile widerspiegelte. Schließlich zeigte sich, dass die Nutzung von Schülerurteilen auch mit Grenzen einhergeht, insbesondere dann, wenn Schülerinnen und Schüler zur Unterrichtsqualität in zwei Fächern befragt werden.

INHALT

1 EINLEITUNG UND THEORETISCHER HINTERGRUND

Die Bedeutung von schulischem Unterricht für den Lernerfolg von Schülerinnen und Schülern beschäftigt seit vielen Jahren Forscherinnen und Forscher unterschiedlicher Disziplinen (Klieme & Rakoczy, 2008). Während Studien aus den Anfängen der Unterrichtsforschung der Schule und dem Unterricht eine nur nebensächliche Rolle zusprachen (Coleman et al., 1966; Jencks et al., 1972), zeigt die umfassendere Forschung der letzten Jahrzehnte, dass Unterricht und dessen Qualität einen bedeutsamen Einfluss auf das schulische Lernen von Schülerinnen und Schülern haben (Hattie, 2009; Seidel & Shavelson, 2007; Wang et al., 1993). In der heutigen empirischen Unterrichtsforschung wird die Qualität von Unterricht im Hinblick auf diejenigen Merkmale von Unterricht betrachtet, die in Zusammenhang mit der Lernentwicklung der Schülerinnen und Schüler sowie dem Erreichen von Bildungs- und Erziehungszielen stehen (Klieme, 2019). Für die konkrete Betrachtung und Operationalisierung von Unterrichtsqualität hat sich in den letzten Jahren insbesondere in der deutschsprachigen Forschungslandschaft das „Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität“ (Klieme et al., 2001) etabliert. Dieses Modell beschreibt Unterrichtsqualität anhand der drei Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung und konnte in mehreren Studien bestätigt werden (Baumert et al., 2010; Lipowsky et al., 2018).

Um fundierte Aussagen über lernförderliche Merkmale des Unterrichts treffen zu können, ist die Nutzung verlässlicher Methoden unabdingbar. Da Schülerinnen und Schüler nach heutigem Verständnis von Unterricht Ko-Produzenten und nicht nur reine Rezipienten des Unterrichts sind (Helmke, 2017), liegt es nahe, ihre Wahrnehmung des Unterrichtsgeschehens zu berücksichtigen. Schülerurteile haben viele Vorteile: Sie sind kostengünstig, niederschwellig einsetzbar und es können in kurzer Zeit eine große Anzahl von Befragten erreicht werden (Fauth, Göllner et al., 2020; Göllner et al., 2016; Wagner, 2008; Wallace et al., 2016). Zudem erleben Schülerinnen und Schüler in ihrer schulischen Laufbahn viele verschiedene Lehrkräfte in unterschiedlichen Fächern, teilweise auch in unterschiedlichen Klassenkonstellationen. Somit greifen sie bei der Beurteilung von Unterrichtsqualität auf einen breiten Erfahrungsschatz zurück. In der Vergangenheit haben sich die Urteile von Schülerinnen und Schülern als reliable und in vielerlei Hinsicht valide Methode zur Erfassung von Unterrichtsqualität erwiesen (Kane et al., 2013; Wagner et al.,

2013; Wallace et al., 2016). So zeigte sich beispielsweise für die Erfassung von Unterrichtsqualität auf Grundlage der drei Basisdimensionen, dass Schülerurteile zur Klassenführung besonders bedeutsam für den Lernzuwachs, ihre Urteile zur konstruktiven Unterstützung insbesondere relevant für die Lernmotivation der Schülerinnen und Schüler sind (Brophy, 2006; Cornelius-White, 2007; Fauth et al., 2014; Göllner et al., 2018; Wagner et al., 2016). Kritiker von Schülerurteilen äußerten jedoch auch Zweifel hinsichtlich der Nutzung von Schülerurteilen. Diese betrafen beispielsweise die Fähigkeit von Schülerinnen und Schülern, Unterricht aufgrund ihrer Involviertheit objektiv beurteilen können (Fisicaro & Lance, 1990). Zudem konnte in Studien gezeigt werden, dass Schülerurteile von Merkmalen unabhängig der zu beurteilenden Unterrichtsqualität, wie beispielsweise der Klassenkomposition, beeinflusst sein können (Fauth et al., 2021; Fauth, Göllner et al., 2020; Göllner et al., 2020).

Trotz des häufigen Einsatzes von Schülerurteilen sind aktuell noch wichtige Fragen unbeantwortet, die in der vorliegenden Dissertation adressiert werden sollen. So konzentrierte sich die bisherige Forschung zur Nutzung von Schülerurteilen primär auf den Präsenzunterricht. Aufgrund der Schulschließungen im Frühjahr 2020 zur Eindämmung der COVID 19-Pandemie fand Unterricht jedoch fast ausschließlich auf Distanz statt. Studie 1 untersuchte deshalb, ob sich Schülerurteile auch in diesem Lehr-Lernsetting eignen, um Unterrichtsqualität zu erfassen und die Urteile in Zusammenhang mit Zielkriterien des Unterrichts stehen. Hinsichtlich der Itemformulierungen in Instrumenten zur Erfassung von Unterrichtsqualität wird zudem schnell ersichtlich, dass diese bezüglich des verwendeten Adressaten (Ich-Adressat oder Wir-Adressat) häufig variieren. Der Rolle des Adressaten wurde in Studie 2 nachgegangen, um die Frage zu beantworten: Macht es für die Urteile von Schülerinnen und Schüler einen Unterschied, ob sie zu ihrer individuellen Wahrnehmung des Unterrichts (Ich-Adressat) oder zur Wahrnehmung der gesamten Klasse (Wir-Adressat) befragt werden? Studie 3 befasst sich schließlich mit der Frage, ob die Urteile von Schülerinnen und Schülern durch Informationen eines anderen Faches beeinflusst sein könnten. Speziell wird geprüft, ob Schülerurteile zur Unterrichtsqualität in einem Fach durch die Note eines anderen Faches beeinflusst sein können, wenn Schülerinnen und Schüler zur Unterrichtsqualität in zwei Fächern befragt werden.

Die vorliegende Dissertation beginnt zunächst mit einer Einleitung und dem theoretischen Hintergrund, der dieser Arbeit zugrunde liegt (Kapitel 1). Dabei werden ein Überblick über die Entwicklung des heutigen Verständnisses von Unterricht und

Unterrichtsqualität (1.1) sowie Befunde zur Nutzung von Schülerurteilen zur Erfassung von Unterrichtsqualität gegeben (1.2). Schließlich werden offene Fragen zur Nutzung von Schülerurteilen in Forschung und Praxis aufgezeigt (1.3). In Kapitel 2 werden die Fragestellungen, die in der vorliegenden Arbeit adressiert werden, dargelegt. Es folgen in Kapitel 3, Kapitel 4 und Kapitel 5 die drei empirischen Studien, die zur Beantwortung der Fragestellungen beitragen sollen. Die Arbeit schließt mit einer Gesamtdiskussion (Kapitel 6). In diesem Kapitel werden die Ergebnisse der drei Studien zusammengefasst (6.1), Stärken und Grenzen aufgezeigt (6.2), Implikationen für Forschung und Praxis gegeben (6.3) und abschließend ein Fazit gezogen (6.4).

1.1 UNTERRICHT UND UNTERRICHTSQUALITÄT

Eine jede und ein jeder von uns hat eine individuelle Schullaufbahn durchlebt. Auf unterschiedlichen Wegen haben wir schrittweise die Stufen des Bildungssystems durchlaufen, haben verschiedene Abschlüsse erreicht, manchmal Ehrenrunden gedreht oder sind auch mal gescheitert. Eins jedoch haben wir alle gemeinsam: Wir haben den schulischen Unterricht in einer Vielfalt erlebt, die durch unterschiedliche Fächer, Lehrkräfte und auch Mitschülerinnen und Mitschüler geprägt war – und Stunden, die wir heute als *guten* oder *schlechten* Unterricht bezeichnen würden. Der Blick auf Unterricht hat sich in den vergangenen Jahrzehnten immer wieder verändert. Im folgenden Kapitel werden nach einer kurzen Geschichte der Unterrichtsforschung und einer Darstellung der Entwicklung des Verständnisses von Unterrichtsqualität aufgezeigt, was wir heute unter einem qualitätsvollen Unterricht verstehen und anhand welcher Merkmale dieser beschrieben werden kann.

1.1.1 Eine kurze Geschichte der Unterrichtsforschung

Der Ursprung der empirischen Perspektive zur Frage, was *guter* oder auch effektiver Unterricht ist und welche Rolle der Unterrichtsqualität für das Lernen der Schülerinnen und Schüler zukommt, wird häufig mit dem sogenannten Coleman-Report (Coleman et al., 1966) in Verbindung gebracht. Von der US-amerikanischen Regierung im Jahr 1964 beauftragt, untersuchten der Soziologe James Samuel Coleman und Kollegen in der bis dahin größten empirischen Studie die Chancenungleichheit von Schülerinnen und Schülern. Die Autoren kamen in ihrem Bericht, der unter dem Titel *Equality of Educational Opportunity* veröffentlicht wurde, zu dem Schluss, dass Leistungsunterschiede von Schülerinnen und Schülern hauptsächlich auf schulunabhängige Faktoren wie die Familie oder die soziale Zusammensetzung der Schülerschaft zurückzuführen sind, die Schulzugehörigkeit hingegen nur einen marginalen Anteil der Leistungsunterschiede erklärt (Coleman et al., 1966). Die Autoren konstatierten: „The social composition of the student body is more highly related to achievement, independent of the student's own social background, than is any school factor" (Coleman et al. 1966, S. 325). Christopher Jencks und Kollegen (1972) kamen in ihrer anschließenden Studie zu ähnlichen, die Annahmen Colemans bekräftigenden Ergebnissen, welche besagten, dass der Lernerfolg von Schülerinnen und Schülern der Sekundarstufe in erster Linie durch deren kognitive und soziale Voraussetzungen erklärt werden kann. Schule und Unterricht schienen für den Lernerfolg der Schülerinnen und Schüler also nur eine nebensächliche Rolle innezuhaben. Beide Studien wurden in der Bildungspolitik und -for-

schung intensiv diskutiert, auch vielfach kritisiert und gaben den Anstoß zu einer großen Anzahl weiterer Untersuchungen über die Bedeutsamkeit von Unterricht und Schule. Adressiert wurden in den nachfolgenden Studien unter anderem Schwächen in der Methodik und im Design der Studien von Coleman und Jencks wie beispielsweise die fehlende Berücksichtigung von Unterrichts- und Lehrkraftmerkmalen in der Studie von Coleman und Kollegen (1966) (Gruehn, 2000). Die Studien kamen zu Ergebnissen, die der Schule und dem Unterricht eine weit bedeutendere Rolle zusprachen (Babu & Mendro, 2003; Scheerens & Bosker, 1997; Wayne & Youngs, 2003). In ihrer umfangreichen Metaanalyse zeigten beispielsweise Wang, Haertel und Walberg (1993), dass Unterrichtsqualität, genauer die Komponenten Klassenführung und die Intensität und Qualität von Fragen und Antworten der Lehrkräfte, etwa in gleicher Weise zum schulischen Lernerfolg der Schülerinnen und Schüler beitragen wie kognitive und metakognitive Fähigkeiten sowie der familiäre Hintergrund. Damit konnte gezeigt werden, dass schulische Faktoren wie die im Unterricht stattfindenden Lehr-Lernprozesse einen bedeutenden Anteil am Lernerfolg von Schülerinnen und Schülern haben.

Eine der einschlägigsten Arbeiten der vergangenen Jahre zur Rolle verschiedener Einflussfaktoren auf den Lernerfolg von Schülerinnen und Schülern ist die umfassende Synthese von Metaanalysen – also gewissermaßen eine Meta-Metaanalyse – des neuseeländischen Bildungsforschers John Hattie (2009). Hatties Werk basiert auf über 800 Metaanalysen, die wiederum auf rund 50.000 Einzelstudien und den Daten von insgesamt etwa 250 Millionen Lernenden beruhen. Zum Vergleich: Im Jahr 2015 besuchten in ganz Europa ca. 49,3 Millionen Schülerinnen und Schüler Schulen des Primar- und Sekundarbereichs (Statistisches Amt der Europäischen Union [Eurostat], 2017a, 2017b). Der viel zitierte und bekannte Satz „teachers make the difference“ (Fischer & Platzbecker, 2018; Hattie, 2003; Lipowsky, 2006), indiziert, dass nach Hatties Studie insbesondere Merkmale aufseiten der Lehrkraft für den Lernerfolg der Schülerinnen und Schüler von Relevanz sind. Diese Aussage basiert jedoch nicht allein auf Merkmalen der Lehrkraft als Person, sondern auch auf durch die Lehrkraft steuerbare und beeinflussbare Merkmale des Unterrichts. Fasst man die für den Lernerfolg der Schülerinnen und Schüler relevanten Merkmale aufseiten der Lehrperson zusammen, z. B. Lehrer-Schüler-Beziehung [1], Klarheit der Lehrperson, Merkmale des Unterrichtsgeschehens wie beispielsweise Feedback und Lehrstrategien der Lehrkraft oder lautes Denken durch die Schülerinnen und Schüler, lassen sich dadurch bis zu 30 % der

[1] Die folgenden Begriffe „Lehrer-Schüler-Beziehung“, „Lehrerselbstberichte“, „Schülerurteile“, „Schülerebene“ etc. schließen alle Geschlechter ein.

Unterschiede in den Schülerleistungen erklären (Lotz & Lipowsky, 2015). Gemessen an einem Anteil von knapp 50 % der Varianzaufklärung durch individuelle Schülermerkmale, wie beispielsweise die kognitive Entwicklung oder das vorangegangene Leistungsniveau, zeigen diese Befunde, welche bedeutende Rolle der Unterrichtsqualität sowie dem Verhalten der Lehrkraft für den Lernerfolg von Schülerinnen und Schülern zukommt. Auch wenn es Kritik an der Studie von Hattie gibt und mit einer derartigen Zusammenfassung zahlreicher Studien Nachteile einhergehen, z. B. bezüglich der Qualität einzelner Studien oder möglicher Verzerrungen der Ergebnisse (Lotz & Lipowsky, 2015), bietet die Studie dennoch eine wichtige empirische Grundlage für das Verständnis schulischen Lernens.

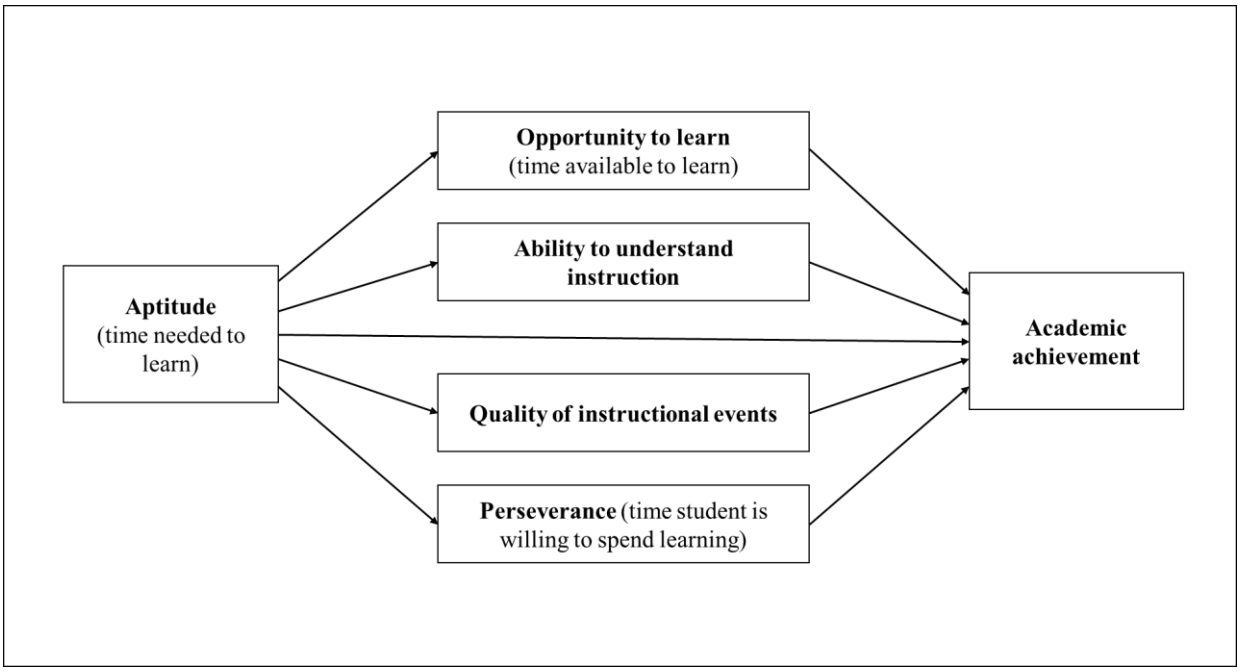
Heute ist die vermehrte Beachtung von Unterrichtsqualität auch daran ersichtlich, dass in nationalen und internationalen Leistungsvergleichsstudien wie PISA, TIMSS, VERA oder dem Ländervergleich Dimensionen der Unterrichtsqualität ein fester Bestandteil sind und sich eine Vielzahl von Studien mit den Zusammenhängen von Unterrichtsqualität und unterschiedlichen Zielkriterien von Unterricht befassen (Graf et al., 2016; Lenski et al., 2016; Mullis et al., 2009; Reiss et al., 2016; Stahns & Rieser, 2018). Welche Merkmale betrachtet werden, hängt maßgeblich vom vorherrschenden Forschungsparadigma ab, das heißt, von welchen Annahmen über die Wirkungsweisen von Unterricht ausgegangen wird.

1.1.2 Entwicklung des Verständnisses von Unterrichtsqualität

John B. Carroll stellte im Jahr 1963 erstmals ein Modell auf, das explizit die Qualität des Unterrichts als schulische Einflussgröße auf den Lernzuwachs von Schülerinnen und Schülern in den Blick nahm (Carroll, 1963). Zentral in Carrolls „Model of School Learning“ (Abb. 1) ist die Annahme, dass Lernerfolg auf dem Verhältnis der aufgewendeten Lernzeit zur tatsächlich benötigen Lernzeit beruht. Die aufgewendete Lernzeit setzt sich zusammen aus der zur Verfügung stehenden Lernzeit und der Ausdauer der Schülerinnen und Schüler, sich einen Inhalt anzueignen. Die benötigte Lernzeit resultiert aus den fachspezifischen Fähigkeiten der Schülerinnen und Schüler und der Unterrichtsqualität. Verfügen also Schülerinnen und Schüler über günstigere Lernvoraussetzungen, wie höheres Vorwissen oder höhere kognitive Fähigkeiten, reduziert sich die benötigte Lernzeit und die Schülerinnen und Schüler können in Summe einen größeren Lernzuwachs verzeichnen (Gräsel & Göbel, 2015).

Abbildung 1

Model of School Learning (Carroll, 1963)



Auch der Unterrichtsqualität – insbesondere der Klarheit und Verständlichkeit, der Berücksichtigung der Bedürfnisse der Schülerinnen und Schüler und einer angemessenen Lernschrittfolge – kommt Carrolls Modell zufolge eine für den Lernerfolg der Schülerinnen und Schüler wesentliche Bedeutung zu. Eine hoch ausgeprägte Unterrichtsqualität kann geringere Voraussetzungen aufseiten der Schülerinnen und Schüler kompensieren. Eine niedrig ausgeprägte Unterrichtsqualität, beispielsweise durch unklare Aufgabenstellungen oder Überforderung der Schülerinnen und Schüler, kann jedoch dazu führen, dass primär Schülerinnen und Schüler mit größerem Vorwissen und höheren kognitiven Fähigkeiten profitieren, da diese im Unterricht Unklarheiten eigenständiger auflösen oder Zusammenhänge konstruieren können (Gruehn, 2000; Harnischfeger & Wiley, 1977).

Carrolls Modell wurde in den darauffolgenden Jahren vielfach weiterentwickelt und modifiziert (z. B. Carroll, 1973, 1989; Harnischfeger & Wiley, 1977). Es diente als Grundlage für Modelle, die andere Schwerpunkte zur Erfassung relevanter Merkmale für den Lernerfolg von Schülerinnen und Schülern setzten. Beispielsweise betont Bloom (1976) in seinem Modell weniger als Carroll die zeitliche Komponente für den Lernprozess. Vielmehr steht in seinem Modell die Frage im Vordergrund, welche Bedeutung individuelle kognitive und motivationale Voraussetzungen der Schülerinnen und Schüler und auch die Qualität des Unterrichtsangebots zur Überwindung von Lernschwierigkeiten haben (Bloom, 1976).

Sowohl Carrolls als auch Blooms Modell sind dem Prozess-Produkt-Paradigma der Unterrichtsforschung zuzuordnen (Gage & Needels, 1989). Dieses Paradigma löste in den 1970er-Jahren das Persönlichkeits-Paradigma ab, welches Persönlichkeitseigenschaften der Lehrperson als zentrales Element für den schulischen Lernerfolg von Schülerinnen und Schülern betrachtete (Helmke, 2017). Das Prozess-Produkt-Paradigma fokussierte stärker auf Verhaltensweisen der Lehrkraft im Unterricht und basiert auf der Überlegung, dass ein qualitätsvoller Unterricht (Prozess) positive Zusammenhänge mit Zielkriterien des Unterrichts aufseiten der Schülerinnen und Schüler (Produkt) zur Folge hat (Terhart, 2006). Tatsächlich konnten für diese Annahmen einige empirische, meist korrelative Befunde gefunden werden, die zeigten, dass Unterschiede in der Unterrichtsdurchführung – beispielsweise in der Klarheit der Lehrkraft – ein für die Leistungsentwicklung von Schülerinnen und Schülern relevantes Merkmal sind (Bromme et al., 2006; Brophy & Good, 1986, 1990; Ditton, 2002). Jedoch stieß das Prozess-Produkt-Paradigma auch auf vielfache Kritik (Ditton, 2002; Dubs, 2009; Helmke, 2010; Neuweg, 2011). Die Betrachtung von isolierten Bestandteilen des Lehrens und Lernens sowie der angenommene lineare Zusammenhang von Unterrichtsmerkmalen und Outputs wie dem Leistungszuwachs der Schülerinnen und Schüler schien zu kurz gegriffen (Gage & Needels, 1989). Beispielsweise wurden Kontextfaktoren wie der sozialen Herkunft der Schülerinnen und Schüler oder der Zusammensetzung der Schulklasse keine Beachtung geschenkt (Helme & Weinert, 1997). Als Antwort auf die Kritik wurden seit den 1990er-Jahren vermehrt Mediationsvariablen wie die kognitiven Fähigkeiten der Schülerinnen und Schüler einbezogen (Borich, 1986; Doyle, 1977), wodurch versucht wurde, der Komplexität von Unterricht und schulischem Lernen gerecht zu werden. Auf dieses erweiterte Prozess-*Mediations*-Produkt-Paradigma folgend wichen die Annahmen immer weiter von einem linearen und direkten Zusammenhang (von der Lehrkraft zum Schüler) ab und orientierten sich stärker an interaktionistischen Merkmalen.

Im Zentrum der Forschungsparadigmen der Unterrichtsforschung steht primär die Frage, welche Rolle die Lehrkraft und andere Merkmale wie beispielsweise die Unterrichtsqualität für den Lernerfolg von Schülerinnen und Schülern spielen (Drechsel & Schindler, 2019). Ein für die Einordnung und Erfassung von Unterrichtsqualität relevantes Rahmenmodell ist das „Angebot-Nutzungs-Modell der Wirkungsweise des Unterrichts“ (Helmke, 2017), dessen vermehrte Beachtung auch für die Weiterentwicklung des Prozess-Produkt-Paradigmas zum Prozess-Mediations-Produkt-Paradigma verantwortlich ist.

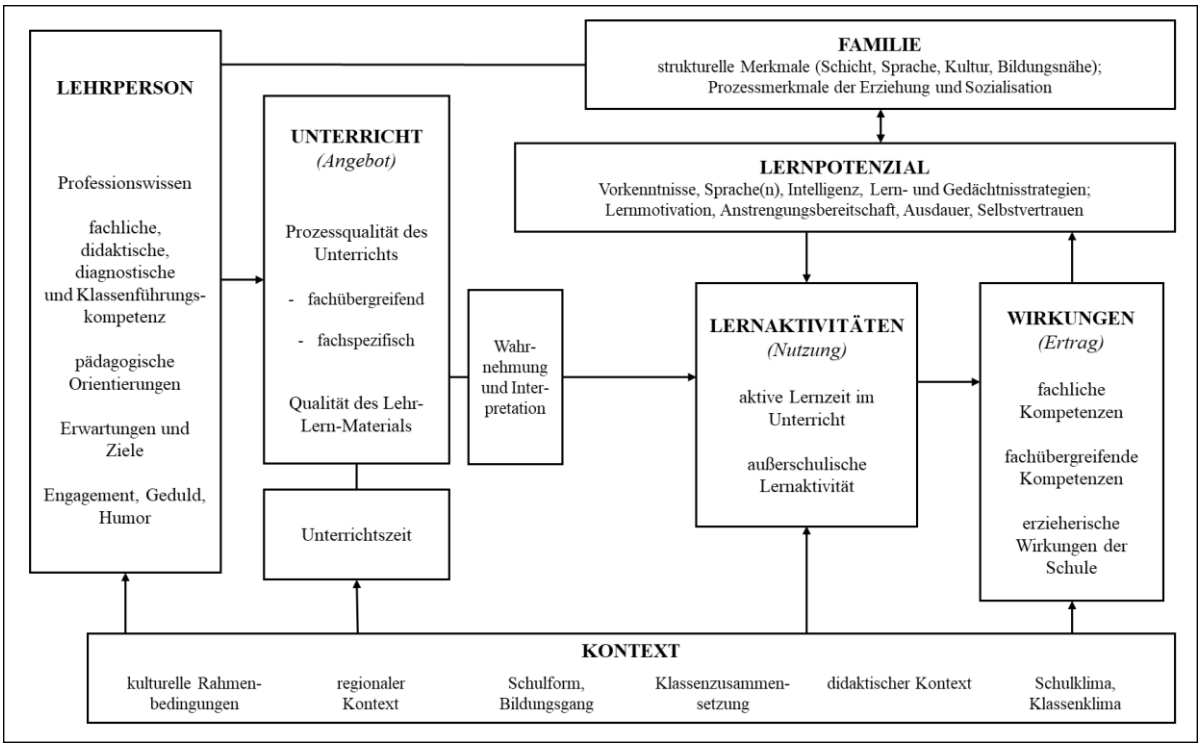
1.1.3 Das Angebot-Nutzungs-Modell unterrichtlicher Wirkungen

Das „Angebot-Nutzungs-Modell“ ist ein Rahmenmodell, das die komplexen Zusammenhänge und Wirkungsweisen von unterschiedlichen Einflussfaktoren auf den Schulerfolg von Schülerinnen und Schülern veranschaulicht. Es kann als deutschsprachige Weiterentwicklung des vor allem im US-amerikanischen Raum präsenten Prozess-Produkt-Paradigmas verstanden werden (Vieluf et al., 2020). Erstmalig publizierte der Pädagoge Helmut Fend das „Angebot-Nutzungs-Modell“ 1982 im Rahmen einer Studie zur Evaluation von Gesamtschulen im Vergleich zum gegliederten Schulsystem (Fend, 1982). Das Modell geht davon aus, dass für den schulischen Lernerfolg von Schülerinnen und Schülern zwei Ebenen von Bedeutung sind: eine Angebotsseite der Schule und eine Nutzungsseite der Schülerinnen und Schüler (Fend, 2019). Die Angebotsseite bezieht sich im Ursprungsmodell auf von der Schule vorgegebene systemische Kontextmerkmale wie Unterrichtsinhalte und Unterrichtsstandards oder die Lehrzeit, jedoch auch die Qualität des Unterrichts. Die Nutzungsseite bezieht sich auf die je individuellen Lernvoraussetzungen der Schülerinnen und Schüler, wie beispielsweise deren kognitive und affektive Kompetenzen, sowie auf die Annahme, dass Lernende „Produzent der eigenen Entwicklung“ (Fend, 2008a, S. 132) sind. Im Gegensatz zum vorherrschenden Prozess-Produkt-Paradigma wurden somit weitere Variablen zur Erklärung von schulischem Lernerfolg einbezogen.

In den folgenden Jahren erfuhr das Angebot-Nutzungs-Modell eine Vielzahl von Erweiterungen und Anpassungen (z. B. Fend, 2008a, 2008b). Für die Einordnung der Wirkung von Unterricht und dessen Qualität ist das von Helmke erweiterte „Angebot-Nutzungs-Modell der Wirkungsweise des Unterrichts“ (Helmke, 2017) primär in der heutigen deutschsprachigen Bildungsforschung von zentraler Bedeutung. Basierend auf den Überlegungen von Fend (1981) und Helmke und Weinert (1997) erweiterte Helmke das Rahmenmodell um weitere Einflussfaktoren auf den schulischen Lernerfolg wie die Rolle der Lehrperson sowie motivationale und emotionale Voraussetzungen der Schülerinnen und Schüler. Abbildung 2 zeigt das „Angebot-Nutzungs-Modell der Wirkungsweise des Unterrichts“ in seiner aktuellen Version (Helmke, 2017).

Abbildung 2

Angebot-Nutzungs-Modell der Wirkungsweise des Unterrichts (Helmke, 2017)



Dieses Rahmenmodell basiert auf der Grundannahme, dass Unterricht ein *Angebot* durch die Lehrkraft ist, welches nicht notwendigerweise auf direktem Wege einen gewünschten *Ertrag* wie etwa einen Zuwachs an fachlichen oder überfachlichen Kompetenzen bewirkt. Vielmehr sind auch schulische und außerschulische Merkmale auf Ebene der Schülerinnen und Schüler (z. B. familiärer Hintergrund, individuelles Lernpotenzial) sowie Merkmale auf Klassenebene (z. B. Qualität des Unterrichts) bedeutend für den Lernerfolg (Helmke, 2017; Lipowsky, 2015). Damit lässt sich das Modell in konstruktivistischen Lerntheorien verorten, nach welchen Wissenserwerb auch abhängig von den jeweiligen Kontextbedingungen wie dem Vorwissen oder der Lerngruppe stattfindet (Reusser, 2006).

Helmke nahm zudem auch diejenigen Merkmale der Lehrperson als Einflussvariable in das Modell auf, welche sich auf deren professionelle Kompetenz sowie auf Persönlichkeitseigenschaften wie Geduld oder Humor beziehen (Helmke, 2017). Dieses Modell wurde in den folgenden Jahren von Helmke selbst immer wieder modifiziert (Helmke, 2007, 2010, 2012), jedoch auch von anderen Forscherinnen und Forschern aufgegriffen und hinsichtlich verschiedener Schwerpunkte angepasst (z. B. Klieme et al., 2006; Lipowsky, 2006; Seidel, 2014). Heute stellt es eine zentrale Grundlage der Unterrichtsforschung dar.

1.1.4 Zum heutigen Verständnis von Unterricht und Unterrichtsqualität

Unterricht wird im „Angebot-Nutzungs-Modell der Wirkungsweise des Unterrichts“ als ein durch die Lehrkraft steuerbares und veränderbares Angebot verstanden. Ob und in welchem Ausmaß dieses Angebot zum gewünschten Ertrag führt, hängt davon ab, inwiefern es von den Schülerinnen und Schülern genutzt wird (Helmke, 2012; Kunter & Trautwein, 2013). Aber was genau kann unter Unterricht verstanden werden?

Es existiert eine Vielzahl an Definitionen des Begriffs „Unterricht“, die sich je nach theoretischem Schwerpunkt unterscheiden (für einen Überblick siehe Lüders, 2012 oder Merkens, 2010). Jedoch ist den aktuellen erziehungswissenschaftlichen Definitionen gemein, dass schulischer Unterricht als ein interaktiver, sozialer Prozess unter Anleitung einer Lehrperson verstanden wird, verbunden mit Ergebniserwartungen wie einem Zuwachs an fachlichen oder fächerübergreifenden Kompetenzen (Doyle, 1986; Fend, 1998; Klieme, 2019; Reusser, 2009). Klieme (2019, S. 393) definiert schulischen Unterricht folgendermaßen:

> Unterricht ist eine Form systematischen pädagogischen Handelns, die darauf abzielt, Lernenden ein Verständnis von Lerninhalten (‚Gegenständen‘) zu vermitteln, damit zugleich in unterschiedliche (fachliche) Modi des Denkens und Handelns einzuführen, den Erwerb fachlicher und fächerübergreifender Kompetenzen zu fördern und Bildung – als Aneignung von Kultur und als Entfaltung einer mündigen Persönlichkeit – zu ermöglichen.

Eng verbunden mit dem zugrunde liegenden Konstruktverständnis von Unterricht ist die Frage, was unter *gutem* Unterricht verstanden wird und, damit einhergehend, welche Merkmale einen qualitativ hochwertigen Unterricht ausmachen, sodass die oben genannten Ziele des Unterrichts erreicht werden.

Mit der Frage, durch welche Merkmale Bildungs- und Erziehungsziele im schulischen Unterricht erreicht werden, beschäftigen sich seit vielen Jahren ganz unterschiedliche Disziplinen, etwa die Schulpädagogik, die Fachdidaktiken, die Soziologie, die pädagogische Psychologie oder die empirische Bildungsforschung (Klieme & Rakoczy, 2008; Praetorius, Grünkorn et al., 2020). Die quantitative empirische Unterrichtsforschung als Teil der interdisziplinären empirischen Bildungsforschung hat zum Ziel, auf Basis empirischer

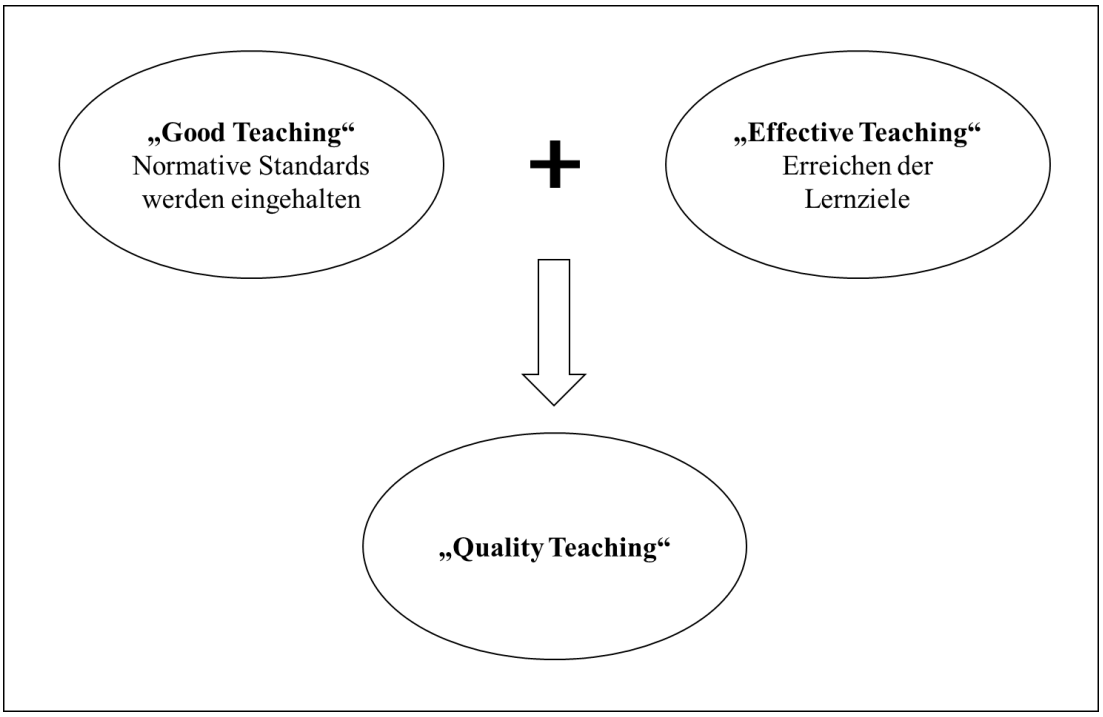
Befunde Merkmale zu identifizieren, die im Zusammenhang mit der Erreichung bestimmter Lernziele stehen (Gräsel & Göbel, 201). Als Indikatoren für erfolgreichen Unterricht stehen zumeist die fachliche Leistung oder die affektiv-emotionale Entwicklung der Schülerinnen und Schüler im Vordergrund (Lipowsky, 2015).

David Berliner (2005) unternimmt den Versuch eines übergreifenden Verständnisses von qualitätsvollem Unterricht. Er unterscheidet diesen in zwei Aspekte: „gutes" Unterrichten und „effektives" Unterrichten. „Gutes" Unterrichten bezieht sich dabei auf die Frage, ob im Unterricht normative Standards eingehalten werden, also ob gewisse Wertvorstellungen erfüllt werden. Beispielsweise könnte dies daran festgemacht werden, ob die Lehrkraft unterschiedliche gängige Unterrichtsmethoden durchführt, um einen abwechslungsreichen Unterricht zu halten, oder ob sie Wert auf ein warmherziges Verhältnis zu ihrer Schülerschaft legt. „Guter" Unterricht ist also gebunden an die im jeweiligen Kontext herrschenden Standards. Dies bedeutet auch, dass in anderen Kontexten, beispielsweise in ländlich gelegenen Schulen in Indien, andere Normen und Standards gelten als an deutschen Schulen und damit eine andere Form des Unterrichts als „guter" Unterricht gilt (Berliner, 2005; Klieme, 2019).

Da die Tatsache, dass – im Sinne des „guten" Unterrichtens – die Lehrkraft verschiedene Methoden anwendet oder ein gutes Verhältnis zu ihrer Schülerschaft pflegt, nicht zwangsläufig zu einer erfolgreichen Vermittlung des Lernstoffs führt, bedarf es eines zweiten Aspektes qualitätsvollen Unterrichts: dem „effektiven" Unterrichten. Unterricht ist dann „effektiv", wenn Lernziele, wie die Aneignung von Wissen oder die Erweiterung von sozialen und emotionalen Kompetenzen, erreicht werden (Berliner, 2005). Greifen schließlich beide Aspekte des „guten" und „effektiven" Unterrichtens ineinander, führt dies zu qualitätsvollem Unterricht (Abb. 3).

Abbildung 3

Modell des qualitätsvollen Unterrichts nach Berliner (2005)



Basierend auf diesem Verständnis von Unterricht lässt sich die empirische Unterrichtsforschung eher dem Aspekt des „effektiven“ Unterrichtens zuordnen. Jedoch geht es nicht allein darum, Zielkriterien wie die fachliche Leistung als isolierten Indikator für eine hohe Unterrichtsqualität zu verstehen. Vielmehr ist das Ziel, die zugrundeliegenden Prozesse des Unterrichtens zu verstehen, die zur Erreichung multipler Lernziele führen. Dies beinhaltet auch, dass je nach Lernziel Unterricht unterschiedlich gestaltet sein kann und demnach ganz vielfältige Formen von Unterricht als qualitätsvoller Unterricht bezeichnet werden können (Helmke, 2017). Übergreifend betrachtet kann Unterrichtsqualität also beschrieben werden als „Gesamtheit der empirisch beobachtbaren Merkmale des Unterrichtsgeschehens, die nachweislich mit einer Entwicklung der Lernenden im Sinne der Realisierung von Bildungs- und Erziehungszielen einhergehen“ (Klieme, 2019, S. 397).

Diesem Verständnis und den Annahmen des „Angebot-Nutzungs-Modells der Wirkungsweise des Unterrichts“ zufolge ist Unterricht ein komplexes und multidimensionales Gefüge, das von unterschiedlichen Kontextfaktoren beeinflusst wird. Welche Merkmale für eine hohe Unterrichtsqualität entscheidend sind, ist eines der prominentesten Forschungsfelder der empirischen Bildungsforschung.

1.1.5 Zur Systematisierung von Unterrichtsqualität

Ein prominenter Zugang zur Systematisierung und Erfassung von Unterrichtsqualität ist die Beschreibung von Unterricht durch zwei Ebenen: die Sichtstruktur und die Tiefenstruktur. Diese Unterscheidung basiert auf der Annahme, dass Unterricht sowohl hinsichtlich organisatorischer Merkmale als auch in Bezug auf die stattfindenden Lehr-Lernprozesse betrachtet werden kann (Aebli, 1961; Kunter & Trautwein, 2013; Oser & Baeriswyl, 2001). Oser und Baeriswyl (2001, S. 1032) bezeichnen die Verbindung der beiden Ebenen „sight structure“ und „basis-model“ als „bridging instruction and learning“. Sichtstrukturen beziehen sich dabei auf alle von außen sichtbaren Unterrichtsmerkmale wie die Organisation des Unterrichts, Sozialformen (z. B. Frontalunterricht oder Gruppenarbeit) und Unterrichtsmethoden. Sie bilden damit einen strukturellen Rahmen von Unterricht. Mit Tiefenstrukturen sind nicht direkt beobachtbare Lehr-Lernprozesse gemeint, also in welchem Ausmaß sich die Lernenden mit den Unterrichtsinhalten auseinandersetzen sowie die Interaktion zwischen den Lernenden und mit der Lehrkraft (Kunter & Voss, 2011). Die empirische Unterrichtsforschung hat gezeigt, dass für den Lernzuwachs von Schülerinnen und Schülern weniger die strukturellen Rahmenbedingungen, also die Sichtstrukturen, sondern primär die Tiefenstrukturen von Relevanz sind (Hattie 2009; Seidel & Shavelson, 2007). Zudem zeigte sich, dass Sicht- und Tiefenstrukturen voneinander unabhängig sind. Dies bedeutet, dass innerhalb desselben strukturellen Rahmens von Unterricht ein unterschiedliches Ausmaß an qualitätsvollen Lehr-Lernprozessen stattfinden kann (Kunter & Voss, 2011).

In den vergangenen zwei Jahrzehnten hat in der Unterrichtsforschung besonders im deutschsprachigen Raum der Fokus auf Tiefenstrukturen Einzug gehalten (Lipowsky, 2015). Ziel war und ist es, Antworten zu finden auf die Frage, welche Merkmale von Unterricht Schülerinnen und Schüler in ihrem Lernprozess unterstützen (Decristan et al., 2020). Hierbei hat das „Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität“ besondere Aufmerksamkeit erfahren.

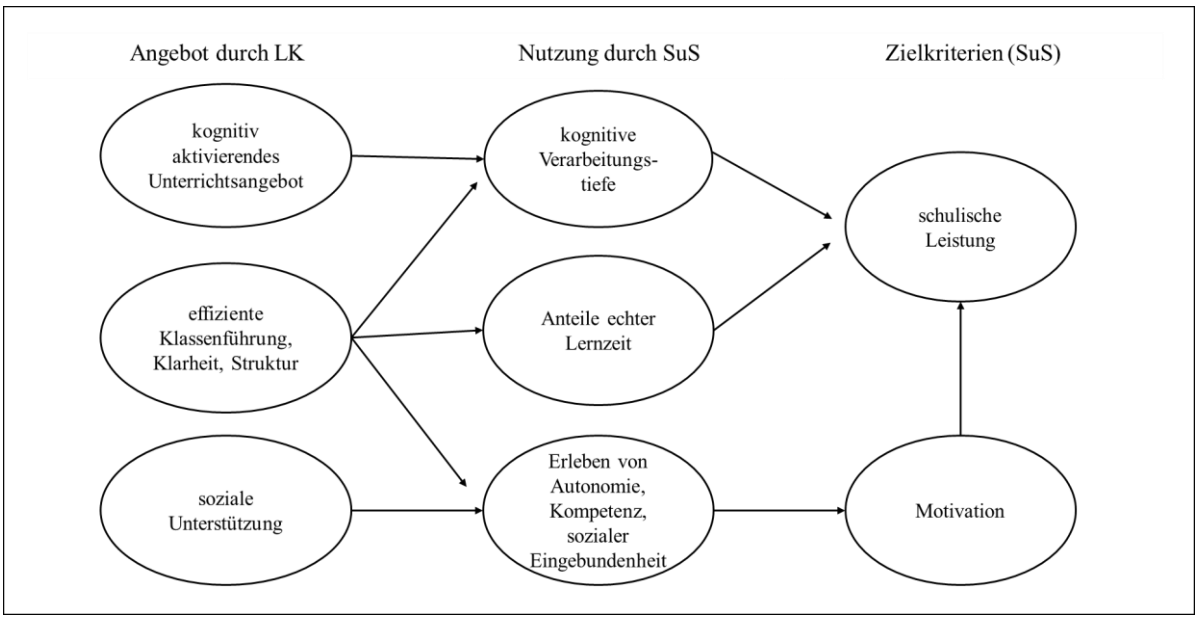
1.1.6 Das Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität

Das „Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität“ lässt sich in den Tiefenstrukturen des Unterrichts verorten und umfasst die Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung (Klieme et al., 2001). Seinen Ursprung hat diese generische, das heißt fächerübergreifende Systematisierung von Unterrichtsqualität in der TIMS-Studie (Baumert et al., 1997; Stigler & Hiebert, 1999), im Rahmen derer

Videoaufnahmen von Mathematikstunden analysiert wurden. Eine explorative Datenanalyse der hoch inferenten Beobachterratings dieser Videos brachte die Struktur der drei Dimensionen als die am besten passende hervor (Praetorius, Klieme et al., 2020). Zusätzlich konnten die Autoren zeigen, dass die Dimension kognitive Aktivierung in positivem Zusammenhang mit der Mathematikleistung und die Dimension konstruktive Unterstützung in positivem Zusammenhang mit dem Mathematikinteresse der Schülerinnen und Schüler stehen (Klieme et al., 2001). Als Bestandteil weiterer Studien wie der PISA-Studie im Jahre 2003 (Prenzel et al., 2013) sowie der Pythagoras-Studie (Klieme et al., 2006) fanden die Dimensionen, die wiederum unterschiedliche Subdimensionen enthalten können, erneute Anwendung. Auch in diesen Studien konnte die Annahme der drei Basisdimensionen der Unterrichtsqualität bestätigt werden (Baumert et al., 2010; Lipowsky et al., 2018; Rakoczy et al., 2010). Abbildung 4 zeigt die angenommenen Beziehungen zwischen dem durch die Lehrkraft bereitgestellten Unterrichtssetting, der Nutzung der Schülerinnen und Schüler dieses Angebotes sowie den Zielkriterien des Unterrichts (Klieme et al., 2009).

Abbildung 4

Mediationsmodell der Basisdimensionen der Instruktionsqualität und ihre Wirkungen auf Schülerlernen und -motivation (in Anlehnung an Klieme et al., 2009)



Dieses Modell, in welchem sich die Annahmen des „Angebot-Nutzungs-Modells der Wirkungsweise des Unterricht“ (Helmke, 2017) wiederfinden, wurde insbesondere in Bezug auf die Wirkungsweisen der drei Basisdimensionen untersucht (Praetorius, Rogh et al., 2020).

Im Folgenden werden die drei Basisdimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung genauer beschrieben.

Klassenführung

Klassenführung (oder auch „Classroom Management“) wird als Gesamtheit aller Maßnahmen verstanden, durch welche ein ungestörter und geordneter Unterrichtsablauf ermöglicht wird (Seidel, 2015). Als wegweisende Arbeiten zum Thema Klassenführung sind die Veröffentlichungen von Jacob Kounin zu nennen (1976, 2006). Der US-Amerikaner untersuchte in den 1970er-Jahren, wie Unterrichtsstörungen unterbunden werden können. Kounin identifizierte dabei fünf Merkmalsbereiche effektiver Klassenführung:
a) Disziplinierung, b) Allgegenwärtigkeit und Überlappung, c) Reibungslosigkeit und Schwung, d) Gruppenmobilisierung, e) Abwechslung und Herausforderung (Kounin, 1976, 2006; Seidel, 2015). Betrachtet man diese Merkmalsbereiche genauer, so wird deutlich, dass nicht allein die Fähigkeit der Lehrkraft zur Disziplinierung ihrer Schülerschaft dazu beiträgt, einen störungsfreien und geordneten Unterricht herzustellen. Ein großer Teil der Merkmale bezieht sich darüber hinaus auf die Organisation und die Inhalte des Unterrichts sowie auf das Ineinandergreifen verschiedener Unterrichtsphasen. Somit soll ein möglichst großes Maß an aktiver Lernzeit, der „time on task“, hergestellt werden (Brophy, 2000, S. 11), wodurch ein höheres Ausmaß schulischer Leistung erreicht werden kann (siehe Abb. 3). Heutige Konzeptualisierungen von Klassenführung bauen auf den theoretischen Grundlagen Kounins auf und werden beispielsweise durch Subdimensionen wie der Prävention von und der Reaktion auf Störungen, der Regelklarheit oder dem Monitoring der Lehrkraft operationalisiert (Kuhfeld, 2017; Kunter et al., 2007; Seidel, 2015). Dabei wird auch deutlich, dass sich die Subdimensionen nicht allein auf das Verhalten der Lehrkraft, sondern auch auf das Verhalten der Schülerinnen und Schüler beziehen (z. B. Ausmaß an Störungen im Unterricht). Hierdurch wird dem Unterricht als einem interaktiven Lehr-Lernsetting, zu dem die Lehrkraft *und* die Schülerschaft beitragen, Rechnung getragen (Helmke, 2017; Klieme, 2019; Reusser, 2009).

Konstruktive Unterstützung

Die Dimension der konstruktiven Unterstützung umfasst Aspekte einer wertschätzenden Lehrer-Schüler-Beziehung sowie eines positiven Unterrichtsklimas (Clausen 2002; Klieme, 2019). Konstruktive Unterstützung bezieht sich somit auf die Begleitung der Schülerinnen und Schüler in ihrem Lernprozess durch die Lehrperson in emotionaler und

motivationaler Hinsicht. Die Grundzüge dieser Dimension lassen sich auf die Selbstbestimmungstheorie der Motivation von Deci und Ryan (1985) zurückführen. Dieser Theorie zufolge haben Personen drei psychologische Grundbedürfnisse: das Bedürfnis nach Kompetenz, nach sozialer Eingebundenheit und nach Autonomie. Dies soll bei der Erfassung von konstruktiver Unterstützung durch unterschiedliche Subdimensionen abgebildet werden und zu einer höheren Motivation der Schülerinnen und Schüler führen (siehe Abb. 3). Beispiele hierfür sind der Grad der Autonomie der Schülerinnen und Schüler in ihrem Lernprozess, die Unterstützung und Ansprechbarkeit der Lehrkraft oder regelmäßiges Feedback zum Lernstand der Schülerinnen und Schüler (Fauth et al., 2014; Göllner et al., 2016; Hamre & Pianta, 2010; Wagner et al., 2013).

Kognitive Aktivierung

Die Dimension der kognitiven Aktivierung beschreibt das Potenzial, das Unterricht bietet, damit sich Schülerinnen und Schüler vertieft mit den Unterrichtsinhalten auseinandersetzen (Lipowsky & Hess, 2019). Basierend auf konstruktivistischen Lerntheorien (Staub & Stern, 2002) ist Unterricht dann besonders lernförderlich, wenn kognitive Lernprozesse angeregt werden und sich Schülerinnen und Schüler aktiv und nachhaltig mit den Lerninhalten auseinandersetzen. Dies führt den Modellannahmen zufolge zu einer höheren schulischen Leistung (siehe Abb. 3). Häufig wird auf das *Potenzial* zur kognitiven Aktivierung verwiesen, da kognitive Prozesse von der Lehrkraft lediglich angeregt werden können, die Lehrkraft jedoch die Denkprozesse der Schülerinnen und Schüler letztlich nicht kennt (Klieme, 2019; Lipowsky & Hess, 2019). Das Potenzial zur kognitiven Aktivierung kann beispielsweise durch Subdimensionen – wie das Herstellen von Bezügen zu bereits Gelerntem, die Begründung eigener Lösungswege oder das Erkennen von Regelmäßigkeiten beim Einüben von Inhalten – erfasst werden (Klieme & Rakoczy, 2008; Kuger et al., 2017; Kunter & Voss, 2011). Während die Dimensionen Klassenführung und konstruktive Unterstützung als relativ unabhängig vom Unterrichtsfach betrachtet werden können, wird der fachliche Bezug aufgrund der inhaltlichen Orientierung bei der kognitiven Aktivierung deutlich. Da die ersten Studien, welche auf diesem Modell basierten, Mathematik oder Fächer der Naturwissenschaften erfassten (Stigler & Hiebert, 1999; Jäger & Helmke, 2008) und diese aufgrund des inhaltlichen und fachlichen Bezugs nicht immer auf andere Fächer übertragbar sind, gibt es hier größere Variationen der Subdimensionen und Items (Maier et al., 2010; Praetorius et al., 2018). Beispielsweise ist das Item „Wir üben regelmäßig, um Aufgaben immer schneller und sicherer bearbeiten zu können“ auf Mathematik anwendbar, jedoch für

den Geschichts- oder Erdkundeunterricht weniger passend.

Die drei Basisdimensionen bilden die Annahmen des „Angebot-Nutzungs-Modells der Wirkungsweise des Unterrichts“ (Helmke, 2017) ab, indem Unterricht einerseits als Angebot durch die Lehrkraft geschaffen werden muss (z. B. durch eine klare Struktur des Unterrichts oder regelmäßiges Feedback), jedoch von den Schülerinnen und Schüler auch aktiv genutzt werden muss, wie beispielsweise durch die Bearbeitung der bereitgestellten Aufgaben. Die Intention hinter diesem Rahmenmodell ist der Versuch, das komplexe und vielfältige Unterrichtsgeschehen durch eine begrenzte Anzahl an Dimensionen sparsam, aber gleichzeitig möglichst umfänglich zu beschreiben (Klieme et al., 2001). Zwar wird in diesem Rahmenmodell keine hierarchische Beziehung zwischen den Dimensionen angenommen, jedoch wird in manchen Publikationen Klassenführung als Voraussetzung für die beiden anderen Dimensionen angesehen (Brunner, 2018; Praetorius, Klieme et al., 2020). Das dargestellte „Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität“ versteht sich nicht als ein starres, feststehendes Modell, sondern eher als ein Ausgangspunkt zur Verständigung über relevante Merkmale von Unterrichtsqualität. Dementsprechend gibt es Ansätze zur Erweiterung des Rahmenmodells, beispielsweise um eine vierte Dimension generischer Inhalte wie kognitive Unterstützung (Kleickmann et al., 2020), Klarheit (Nilsen & Gustafsson, 2016) oder individuelle Förderung (Paulicke et al., 2019), aber auch fachspezifischer Merkmale wie die fachliche Korrektheit (Brunner, 2018). Auch findet das Rahmenmodell eher in der deutschsprachigen Forschungslandschaft Anwendung (Praetorius, Klieme et al., 2020). Im US-amerikanischen Raum bekannte Instrumente zur Erfassung von Unterrichtsqualität, die auch inhaltliche Überschneidungen mit dem „Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität“ aufweisen, sind das Instrument „Classroom Assessment Scoring System“ (CLASS; Pianta et al., 2008) zur Beurteilung von Unterricht durch externe Beobachter, das Instrument zur Erfassung von Unterricht aus Sicht der Lehrkraft „Classroom Strategies Scales-Teacher Form“ (CSS-T; Reddy et al., 2015) oder der „Tripod 7Cs Student Perceptions Survey“ zur Erfassung von Unterrichtsqualität aus Schülersicht (Kuhfeld, 2017). Diese Instrumente machen deutlich, dass es unterschiedliche Wege zur Erfassung von Unterrichtsqualität gibt. Einen prominenten Weg hierfür bieten Befragungen von Schülerinnen und Schülern als den unmittelbaren Rezipientinnen und Rezipienten sowie Beteiligten des Unterrichtsgeschehens. Diese viel genutzte Methode zur Erfassung von Unterrichtsqualität soll im Folgenden beschrieben werden.

1.2 Nutzung von Schülerurteilen zur Erfassung von Unterrichtsqualität

Unterrichtsqualität kann durch drei unterschiedliche Perspektiven erfasst werden. In Lehrerselbstberichten machen Lehrkräfte Angaben zu ihrer eigenen Unterrichtspraxis, beispielsweise in Fragebögen, in welchen unterschiedliche Qualitätsdimensionen abgefragt werden. Eine zweite Möglichkeit sind die Bewertungen externer, geschulter Beobachter. Anhand vorgegebener Manuale raten sie beispielsweise videografierte Unterrichtssequenzen oder nehmen direkt am Unterricht teil und beurteilen diesen.[2] Die Beurteilung durch Schülerinnen und Schüler stellt eine dritte zentrale und häufig verwendete Methode zur Erfassung von Unterrichtsqualität dar. Die folgende Abbildung stammt aus der Studie „Unterricht aus Schülersicht“ (UNITAS) und zeigt, wie eine typische Befragung aussehen kann (Jaekel et al., 2020).

Abbildung 5

Ausschnitt eines Schülerfragebogens zur Erfassung von Unterrichtsqualität

		stimmt gar nicht	stimmt eher nicht	stimmt eher	stimmt genau
Störungen	Im Unterricht ist es manchmal laut und alles geht durcheinander.	☐	☐	☐	☐
	Der Unterricht wird oft gestört.	☐	☐	☐	☐
	Manchmal muss unsere Lehrkraft zu Beginn der Stunde lange warten, bis Ruhe eintritt.	☐	☐	☐	☐
	Im Unterricht wird fortwährend laut gequatscht.	☐	☐	☐	☐
Feedback	Von unserer Lehrkraft erfahren wir immer wieder, wo wir mit unseren Leistungen stehen.	☐	☐	☐	☐
	Die Lehrkraft gibt mir regelmäßig Rückmeldung, was ich schon kann.	☐	☐	☐	☐
	Unsere Lehrkraft gibt uns regelmäßig Rückmeldung, was wir noch nicht so gut können.	☐	☐	☐	☐
	Die Lehrkraft zeigt immer wieder, wie ich mich verbessern kann.	☐	☐	☐	☐

[2] Für eine ausführliche Darstellung der Erhebungsmethoden aus Perspektive der Lehrkräfte und Beobachter, den jeweiligen Vor- und Nachteilen sowie eine Gegenüberstellung aller drei Perspektiven siehe z. B. Clausen, 2002; Desimone et al., 2010; Fauth, Göllner et al., 2020; Kunter & Baumert, 2006; Praetorius et al., 2014; Wagner et al., 2016.

Meist werden in einem Fragebogen mehrere theoretisch distinkte Subdimensionen von Unterrichtsqualität erfragt, wie hier das Ausmaß der Störungen und das Feedback durch die Lehrkraft (siehe Kapitel 1.1.6). Wie auch in diesem Beispiel zu sehen ist, werden Schülerinnen und Schüler über einen längeren Zeitraum häufig zu ihren allgemeinen Eindrücken befragt, beispielsweise wie oft das im Item formulierte Geschehen im Laufe des vergangenen Schuljahres vorkam (siehe jedoch Ogrin et al., 2017). Dies gilt auch für die Inhalte: Meist werden eher generalisierte Angaben über die Lehrkraft oder die Schülerschaft erfragt als ganz konkrete Handlungen, das heißt eher hoch inferente als niedrig inferente Items verwendet (Wagner, 2008).

Die Urteile der Schülerinnen und Schüler können auf zwei Ebenen betrachtet und analysiert werden: Zum einen bieten sie Informationen zur individuellen Wahrnehmung des Unterrichts, also wie jede einzelne Schülerin und jeder einzelne Schüler die Lernumgebung wahrnimmt (Schülerebene). Zum anderen werden Schülerurteile häufig aggregiert, indem Mittelwerte der Individualdaten der Schülerinnen und Schüler einer Klasse für das jeweilige Merkmal gebildet werden (Klassenebene). Die aggregierten Werte bilden demnach das geteilte Urteil der Klasse zur jeweiligen Dimension ab (Lüdtke et al., 2006; Marsh et al., 2012).

Die Nutzung von Schülerurteilen geht mit unterschiedlichen Vor- und Nachteilen einher. Diese sowie damit verbundene empirische Befunde zur Genauigkeit und Verlässlichkeit von Schülerurteilen zur Erfassung von Unterrichtsqualität werden im Folgenden näher beschrieben.

1.2.1 Vor- und Nachteile in der Nutzung von Schülerurteilen zur Erfassung von Unterrichtsqualität

Schülerurteile bieten eine Reihe von Vorteilen: Sie sind kostengünstig, und mit wenig Aufwand können viele Schülerinnen und Schüler zu einer großen Anzahl unterschiedlicher Merkmale der Unterrichtsqualität befragt werden. Schülerinnen und Schüler haben meist viele unterschiedliche Lehrkräfte in verschiedenen Fächern erlebt und greifen somit auf einen großen Erfahrungsschatz zurück (Kunter et al., 2008; Wagner et al., 2013). Zudem liegt es nahe, Schülerinnen und Schüler als aktiv Beteiligte des täglichen Unterrichtsgeschehens nach ihren Eindrücken zu fragen und die Möglichkeit zu nutzen, als Lehrkraft die Qualität des eigenen Unterrichts weiterzuentwickeln. Schülerinnen und Schüler werden mit solchen Befragungen im Sinne des „Angebot-Nutzungs-Modells der Wirkungsweise des Unterrichts“ als *Ko-Produzenten*, und nicht nur als reine Konsumenten des Unterrichts wahrgenommen

(Helmke, 2017). Schließlich folgt aus der Aggregation einer häufig großen Zahl einzelner Schülerurteile auf Klassenebene meist eine hohe Reliabilität der zusammengefassten Werte (Lüdtke et al., 2006; Wagner, 2008).

In der Vergangenheit wurden jedoch auch immer wieder Vorbehalte hinsichtlich der Nutzung von Schülerurteilen geäußert (Abrami, 1989; Abrami et al., 2007; Aleamoni, 1981). Auch wenn Schülerinnen und Schüler Unterricht in vielfältiger Weise erleben, verfügen sie dennoch nicht über eine professionelle pädagogisch-didaktische Expertise. Daher ist es möglich, dass sie Sinn und Zweck einer bestimmten Unterrichtsdurchführung nicht adäquat einordnen können. Zudem liegen die Steuerung des Unterrichts und auch die Leistungsbeurteilung der Schülerinnen und Schüler in der Hand der Lehrkraft (siehe Kapitel 1.1.4). Aus diesem Grund besteht eine gewisse Abhängigkeit der Schülerschaft von der Lehrkraft, sodass eine neutrale Perspektive auf den Unterricht mitunter schwer einzunehmen sein könnte. Schließlich haben Studien gezeigt, dass Schülerurteile durch unterschiedliche Faktoren, wie beispielsweise Klassenkompositionsmerkmale, beeinflusst sein können (z. B. Benton & Cashin, 2012; Fauth et al., 2014; Göllner et al., 2018, 2020), was mit der Validität der Nutzung von Schülerurteilen einhergeht.

Grundsätzlich setzen belastbare Analysen von Schülerurteilen voraus, dass die Daten reliabel und valide sind (Eid et al., 2017). Im Folgenden werden die beiden zentralen Gütekriterien der Reliabilität und Validität in Bezug auf die Nutzung von Schülerurteilen sowie empirische Befunde hierzu vorgestellt.

1.2.2 Reliabilität von Schülerurteilen

Reliabilität ist definiert als Genauigkeit, mit der ein Merkmal gemessen wird (Rammstedt, 2010). Im Hinblick auf die Reliabilität von Skalen wird die Genauigkeit, mit der diese das intendierte Merkmal der Unterrichtsqualität erfassen (z. B. der Grad der Motivierung durch die Lehrkraft), häufig durch das Reliabilitätsmaß Cronbachs Alpha angegeben (Lüdtke et al., 2009). Cronbachs Alpha wird auch als „interne Konsistenz“ bezeichnet, da es angibt, wie stark die einzelnen Items einer Skala miteinander zusammenhängen (Eid et al., 2017, S. 863). Die Bestimmung der Reliabilität von Schülerurteilen hat den Vorteil, dass ihr klare Indizes und Grenzwerte zugrunde liegen. So kann Cronbachs Alpha Werte zwischen 0 und 1 annehmen. Beispielsweise besagt ein Wert von Cronbachs Alpha = 0.82, dass für diese Skala eine hohe Reliabilität angenommen werden kann. Bei hierarchisch strukturierten Daten (z. B. Schülerinnen und Schüler innerhalb von Klassen) kann die Reliabilität der aggregierten Individualdaten zu einem Konstrukt auf

Klassenebene durch die Intraklassenkorrelation (ICC) angegeben werden. Die ICC(1) ist damit ein Maß für die relative Übereinstimmung von Schülerinnen und Schülern einer Klasse hinsichtlich eines Merkmals und basiert auf dem Verhältnis der Varianz innerhalb von Klassen zur Varianz zwischen Klassen (Snijders & Bosker, 1999). Beispielsweise besagt ein Wert von ICC(1) = 0.26 der Skala „Motivierung“, dass 26 % der Varianz in den Urteilen zur wahrgenommenen Motivierung auf Unterschiede zwischen Klassen zurückzuführen ist. Die ICC(2) beschreibt die Reliabilität des auf Klassenebene aggregierten Urteils und bezieht zusätzlich die (durchschnittliche) Anzahl der Urteiler pro Klasse ein (Lüdtke et al., 2006). So werden Werte von ICC(2) > .70 als zufriedenstellende Reliabilität angenommen (Lüdtke et al., 2006).

Legt man die beiden gängigen Maße zur Erfassung der Reliabilität von Schülerurteilen zur Unterrichtsqualität interne Konsistenz und ICC zugrunde, so hat eine große Zahl an Studien gezeigt, dass Schülerinnen und Schüler sowohl der Grundschule als auch der Sekundarstufe Unterricht anhand unterschiedlicher Qualitätsdimensionen reliabel einschätzen können (De Jong & Westerhof, 2001; Fauth et al., 2014, 2019; Kane et al., 2013; Kunter et al., 2008; Lüdtke et al., 2006, 2009; van der Scheer et al., 2019; Wagner et al., 2013). Beispielsweise zeigen Praetorius et al. (2018) in einer breiten Übersicht unterschiedlicher Instrumente, dass sich die Reliabilitätsmaße der Schülerurteile zur Unterrichtsqualität größtenteils in zufriedenstellenden Bereichen befinden (Cronbachs Alpha: 0.73–0.95; ICC(1): 0.12–0.31; ICC(2): 0.59–0.98).

1.2.3 Validität von Schülerurteilen

Die Validität von Schülerurteilen lässt sich, anders als die Reliabilität, nicht direkt an einzelnen Kennzahlen mit festgelegten Schwellenwerten festmachen. Vielmehr wird sich der Frage, ob und inwiefern Schülerinnen und Schüler Unterrichtsqualität valide einschätzen können und tatsächlich das intendierte Merkmal gemessen wurde, aus mehreren Perspektiven genähert. Sind beispielsweise einzelne Subdimensionen (z. B. Strukturierung, Motivierung, anspruchsvolle Aufgaben) auch psychometrisch voneinander abgrenzbar? Sind Schülerurteile zur Unterrichtsqualität auch über unterschiedliche Kontexte wie Klassen oder Fächer hinweg vergleichbar und in welchem Zusammenhang stehen sie mit Zielkriterien des Unterrichts? In den vergangenen Jahren wurde eine Reihe von empirischen Studien zur Frage, wie valide sich Unterrichtsqualität mit Schülerurteilen erfassen lässt, durchgeführt. Im Folgenden werden einige zentrale Befunde zur Konstruktvalidität und zur prädiktiven Validität von Schülerurteilen vorgestellt.

Dimensionalität von Schülerurteilen

Instrumente zur Erfassung von Unterrichtsqualität bestehen meist aus einer theoriegeleiteten Zusammenstellung verschiedener Subdimensionen. Ein bedeutsamer Indikator für die Konstruktvalidität von Schülerurteilen ist der Nachweis der Multidimensionalität, also ob sich die theoretische Abgrenzung unterschiedlicher Subdimensionen auch psychometrisch nachweisen lässt und Schülerinnen und Schüler somit zwischen den Subdimensionen differenzieren (Greenwald, 1997). Um die im schulischen Kontext vorgegebene Datenstruktur (z. B. Schülerinnen und Schüler innerhalb von Klassen) abzubilden, werden häufig exploratorische oder konfirmatorische Mehrebenen-Strukturgleichungsmodelle durchgeführt, die erlauben, die faktorielle Struktur der Subdimensionen simultan auf Schüler- und auf Klassenebene zu überprüfen (Göllner et al., 2020; Lüdtke et al., 2007).

Obwohl Schülerurteile häufig genutzt werden, wurden diese im Hinblick auf ihre theoretisch angenommene multidimensionale Struktur vergleichsweise wenig untersucht. Kunter et al. (2008) fanden in ihrer Studie, dass Schülerinnen und Schüler für das Fach Mathematik die Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung differenziert beurteilen können. Die Autorinnen und Autoren nutzten Daten von fast 4.000 Neuntklässlerinnen und Neuntklässlern aus 323 Klassen. Sowohl auf Schüler- als auch auf Klassenebene ließ sich zeigen, dass ein dreifaktorielles Modell, das die drei Basisdimensionen der Unterrichtsqualität widerspiegelte, der Datenstruktur am besten entsprach. Die Ergebnisse lassen daher annehmen, dass Schülerinnen und Schüler Unterrichtsqualität differenziert beurteilen können, und stützen zudem die Annahmen des Modells der drei Basisdimensionen der Unterrichtsqualität (Klieme et al., 2001).

Wagner und Kollegen (2013) gingen in ihrer Studie explizit der Frage nach der Dimensionalität von Schülerurteilen nach. Sie untersuchten die Angaben von fast 7.000 Schülerinnen und Schülern ebenfalls der neunten Jahrgangsstufe aus 280 Klassen für die Fächer Deutsch und Englisch. Anhand konfirmatorischer Faktorenanalysen wurde überprüft, ob sich in den Fächern Englisch und Deutsch sowohl auf Schüler- als auch Klassenebene die gleiche Faktorenstruktur für die Dimensionen Motivation, Verständlichkeit, Einbezug der Schülerschaft, Strukturierung und Klassenführung abbildet. Die Ergebnisse bestätigten für beide Ebenen und für beide Fächer die multidimensionale Struktur der fünf untersuchten Subdimensionen. Dies bedeutet, dass die theoretisch angenommenen Subdimensionen empirisch durch die Schülerurteile trennbar sind und somit Schülerinnen und Schüler die

Subdimensionen differenziert beurteilen können. Diese Ergebnisse können als starker Hinweis für die Konstruktvalidität von Schülerurteilen gewertet werden.

Auch für jüngere Schülerinnen und Schüler konnten Belege für deren Fähigkeit, Unterricht differenziell zu beurteilen, gefunden werden. Fauth et al. (2014) prüften für 1.556 Drittklässler aus 89 Klassen, ob deren Urteile für den Sachunterricht den Annahmen des „Rahmenmodells der drei Basisdimensionen der Unterrichtsqualität“ entsprechen. Die Ergebnisse zeigten auf Schüler- und auf Klassenebene, dass bereits Grundschülerinnen und Grundschüler in der Lage sind, Unterricht anhand der drei Dimensionen zu beurteilen. Ebenfalls für Grundschülerinnen und Grundschüler konnten van der Scheer et al. (2019) anhand konfirmatorischer Faktorenanalysen zeigen, dass diese fünf Subdimensionen der Unterrichtsqualität differenziert beurteilen können.

Zusammengefasst zeigen diese und weitere Studien (z.B. De Jong & Westerhof, 2001; Kuhfeld, 2017; Wallace et al., 2016), dass sich die theoretisch angenommene multidimensionale Struktur auch psychometrisch nachweisen lässt. Dies zeigt, dass sowohl Schülerinnen und Schüler der Grundschule als auch der Sekundarstufe in der Lage sind, unterschiedliche Dimensionen und Subdimensionen der Unterrichtsqualität differenziert zu beurteilen.

Generalisierbarkeit von Schülerurteilen

Ein weiterer wichtiger Aspekt der Konstruktvalidität von Schülerurteilen zur Unterrichtsqualität ist die Generalisierbarkeit über Kontexte hinweg: Sind Schülerurteile zu Merkmalen der Unterrichtsqualität über verschiedene Klassen oder Schulfächer hinweg vergleichbar und können die Werte tatsächlich gleich interpretiert werden? Die damit einhergehende Frage nach einem äquivalenten Messmodell bzw. der faktoriellen Struktur für unterschiedliche Kontextbedingungen, nämlich verschiedene Klassen und die zwei Schulfächer Deutsch und Englisch, überprüften Wagner und Kollegen (2013) in ihrer Studie. Für die Vergleiche über Klassen hinweg fanden die Autoren, dass die Annahme äquivalenter Messmodelle für die Subdimensionen Strukturierung und Klassenführung gerechtfertigt ist. Für die Subdimensionen Motivierung, Verständlichkeit und Einbezug der Schülerschaft zeigte sich jedoch, dass diese durch Merkmale der jeweiligen Klasse beeinflusst waren, sodass ein direkter Vergleich der Schülerurteile für diese Subdimensionen nur bedingt angestellt werden kann. Unter Hinzunahme der Schulfächer zeigte sich ein ähnliches Muster: Eine vollständige Messäquivalenz über die Fächer hinweg konnte weiterhin nur für die Subdimensionen Strukturierung und Klassenführung angenommen werden.

Fauth et al. (2019) untersuchten in ihrer Studie a) die zeitliche Stabilität von Schülerurteilen über ein Jahr hinweg bei gleicher Lehrkraft, b) die Stabilität zweier unterschiedlicher Klassen, die jeweils von derselben Lehrkraft unterrichtet wurden (N = 190 Klassen, N = 95 Lehrkräfte) und c) die Stabilität von Urteilen von Schülerinnen und Schülern aus 30 verschiedenen Klassen, die alle von derselben Lehrkraft und zur selben Thematik unterrichtet wurden. Die Ergebnisse wiesen für Urteile von Schülerinnen und Schülern für die gleiche Lehrkraft über ein Jahr hinweg a) eine insgesamt moderate bis hohe Stabilität auf. Jedoch fanden sich b) für die Urteile von Schülerinnen und Schülern je zweier Klassen für dieselbe Lehrkraft relativ niedrige Übereinstimmungen. Beispielsweise korrelierten die Dimensionen mit r = .40 für Klassenführung, r = .36 für konstruktive Unterstützung und r = .27 für Klarheit der Instruktionen nur moderat. Schließlich zeigte sich c) auch hinsichtlich der Urteile unterschiedlicher Klassen, die alle von derselben Lehrkraft zum selben fachlichen Thema unterrichtet wurden, besonders für die Dimensionen konstruktive Unterstützung und Klarheit der Instruktionen eine hohe Variation, die darauf hindeutet, dass sich in Urteilen nicht allein die Qualität des Unterrichts widerspiegelt.

Die Ergebnisse dieser Studien zeigen einerseits, dass Merkmale der Unterrichtsqualität über unterschiedliche Kontexte hinweg von Schülerinnen und Schülern valide eingeschätzt werden können. Jedoch zeigten sich die Urteile stabiler für die Klassenführung als beispielsweise für konstruktive Unterstützung (siehe auch Evertson & Weade, 1989; Polikoff, 2015; Praetorius et al., 2014; Wagner et al., 2016). Dies weist drauf hin, dass Schülerurteile durch von der zu beurteilenden Unterrichtsqualität unabhängige Faktoren beeinflusst sein könnten. Im Folgenden werden einige dieser Faktoren und deren möglicher Einfluss auf Schülerurteile vorgestellt.

Verzerrende Faktoren

Eng verbunden mit der Validität von Schülerurteilen ist die Frage nach Faktoren, die *unabhängig* von der zu beurteilenden Unterrichtsqualität sind, die die Urteile von Schülerinnen und Schülern beeinflussen könnten. So sollten Schülerurteile zur Unterrichtsqualität nur Informationen enthalten, die auch durch die jeweiligen Skalen erfragt wurden. Beispielsweise zeigte die Studie von Fauth et al. (2019) auch, dass Urteile unterschiedlicher Klassen trotz konstanter Bedingungen stark variierten. In der Vergangenheit konnte für einige solcher Faktoren, die theoretisch abgrenzbar von der zu beurteilenden Unterrichtsqualität sind, gezeigt werden, dass sie Schülerurteile beeinflussen können. Dies wurde immer wieder von Kritikern vorgebracht, die die Validität von Schülerurteilen

anzweifelten (Abrami, 1989; Greenwald, 1997; Waldis et al., 2010). Im Folgenden werden einige zentrale Einflussfaktoren vorgestellt.

Der Halo-Effekt. Ein bekannter Einflussfaktor auf Urteile der persönlichen Umgebung, der häufig im Zuge der Frage nach der Dimensionalität von Schülerurteilen aufkommt, ist der sogenannte Halo-Effekt. Dieser kann verstanden werden als „die mangelnde Fähigkeit eines Urteilers, zwischen konzeptionell unterschiedlichen und potenziell unabhängigen Aspekten des Verhaltens eines zu Beurteilenden zu unterscheiden“ (Feeley, 2002, S. 226). In Bezug auf Schülerurteile bedeutet dies, dass die Urteile von Schülerinnen und Schülern durch ein Merkmal, das von der Unterrichtsqualität unabhängig ist, beeinflusst sein können. Beispielhaft wäre hier das Geschlecht einer Lehrkraft zu nennen. So zeigte sich, dass Schülerurteile etwas höher ausfallen, wenn Schülerinnen und Schüler dem gleichen Geschlecht wie die Lehrkraft angehören (Centra & Gaubatz, 2000; Feldman, 1998; Griffin, 2004). Aus psychometrischer Perspektive würde dies in erhöhten Interkorrelation der Subdimensionen resultieren, da der Halo-Effekt potenziell viele Schülerinnen und Schüler betrifft (Wagner, 2008). Es wird angenommen, dass der Halo-Effekt aus einem *true halo*, also einem tatsächlichen Zusammenhang zwischen Merkmalen, und einem *invalid halo*, einer Verzerrung, die auf einem Beurteilerfehler beruht, besteht (Fisicaro & Lance, 1990; Lance et al., 1994). Methodisch ist es jedoch nicht einfach, die jeweiligen Anteile des true halo und des invalid halo in Schülerurteilen nachzuweisen. Der Halo-Effekt wird daher häufig eher als theoretische Erklärungsgrundlage für beispielsweise hohe Interkorrelationen einzelner Qualitätssubdimensionen verwendet als statistisch nachgewiesen (De Jong & Westerhof, 2001; Fauth et al., 2014).

Grading Leniency. Der Begriff *grading leniency* besagt, dass Lehrkräfte höhere Bewertungen, z. B. hinsichtlich ihrer Unterrichtsqualität, erhalten, wenn sie aus der Perspektive ihrer Schülerinnen und Schüler bessere Noten vergeben. Umgekehrt erhalten sie niedrigere Bewertungen, wenn sie einer strengen Benotungspraxis folgen (Griffin, 2004). Dieser Zusammenhang zwischen der von Schülerinnen und Schülern wahrgenommenen Notengebung einer Lehrkraft und der Beurteilung der Unterrichtsqualität durch ihre Schülerinnen und Schüler wurde in mehreren Studien nachgewiesen (z. B. Greenwald & Gillmore, 1997; Griffin, 2004; Reyes et al., 2012). So zeigte Olivares (2001), dass eine wahrgenommene milde Notengebung der Lehrkraft höhere Schülerurteile vorhersagt. Konzeptuell ist den Studien jedoch gemein, dass die Notengebung der Lehrkraft nur indirekt

erfasst werden konnte, da die Notengebung ein Charakteristikum aufseiten der Lehrkraft ist, die Wahrnehmung der Milde oder Strenge sowie die Beurteilung der Unterrichtsqualität jedoch aufseiten der Schülerinnen und Schüler stattfindet. Daher konnte die wahrgenommene Notengebung nur indirekt durch Items wie „Wie würden Sie die Benotung dieser Lehrkraft im Vergleich zu allen anderen Lehrkräften, die Sie hatten, bewerten?“ (Olivares, 2001) erfasst werden. Jedoch wurde auch in Studien, die nicht explizit die Praxis der Notengebung im Fokus hatten, positive Zusammenhänge zwischen Noten und der Beurteilung der Unterrichtsqualität gefunden (z. B. Aldrup et al., 2018; Benton et al., 2013; Göllner et al., 2018; Marsh & Roche, 2000).

Merkmale der Klassenkomposition. Die Schülerschaft einer jeder Klasse variiert hinsichtlich verschiedener Merkmale, etwa der Geschlechterverteilung, der Verteilung des sozioökonomischen Status oder des mittleren Leistungsniveaus. In Studien, die die Einflüsse unterschiedlicher Kompositionsmerkmale untersuchten, konnten Zusammenhänge mit Schülerurteilen zur Unterrichtsqualität belegt werden (siehe dazu auch das „Vermittlungsmodell für Kontexteffekte“; Baumert et al., 2006). Beispielsweise zeigte sich, dass Klassen mit einem geringeren Jungenanteil und einem höheren mittleren Leistungsniveau über eine höhere Unterrichtsqualität berichteten (Fauth et al., 2019; Hochweber & Vieluf, 2018; Kunter et al., 2013; Rjosk et al., 2014).

Göllner et al. (2020) nahmen in ihrer Studie Unterschiede in den inhaltlichen Bezügen von Subdimensionen der Dimension Klassenführung in den Blick. Denn besonders Subdimensionen dieser Dimension unterscheiden sich häufig im Hinblick auf den Referenten, also darin, wessen Verhalten beurteilt werden soll. Während sich eine Subdimension wie Störungen häufig auf das Verhalten der Schülerschaft bezieht (z.B.: „Im Unterrichts ist es oft laut und alles geht durcheinander.“), bezieht sich beispielsweise Monitoring auf das Handeln und Verhalten der Lehrkraft (z.B. „Unsere Lehrkraft weiß immer genau, was im Unterricht vor sich geht.“). Die Autoren konnten zeigen, dass die Kompositionsmerkmale Geschlecht, Leistungsniveau und sozioökonomischer Status Subdimensionen beeinflussten, die sich auf das Verhalten der Schülerinnen und Schüler oder die Lehrkraft-Schüler-Interaktion bezogen (Abwesenheit von Störungen und effiziente Zeitnutzung).

Fauth et al. (2020) untersuchten in ihrer Studie Zusammenhänge unterschiedlicher Klassenkompositionen hinsichtlich des sozioökonomischen Status, der kognitiven Fähigkeiten und der Motivation von Drittklässlerinnen und Drittklässlern mit der eingeschätzten

Unterrichtsqualität. Die Ergebnisse zeigten, dass die kognitiven Fähigkeiten und die Motivation der Schülerinnen und Schüler in positivem Zusammenhang mit Schülerurteilen zur Unterrichtsqualität standen. Genauer zeigten sich positive Korrelationen zwischen der Klassenkomposition hinsichtlich kognitiver Fähigkeiten und der Motivation der Schülerinnen und Schüler und den Schülerurteilen zur Klassenführung, nicht jedoch für konstruktive Unterstützung und kognitive Aktivierung. Keine signifikanten Zusammenhänge mit der Unterrichtsqualität ließen sich für den sozioökonomischen Status finden.

Rjosk et al. (2014) untersuchten die Auswirkungen der Klassenzusammensetzung hinsichtlich des sprachlichen und sozioökonomischen Hintergrunds auf die Leseleistung unter Betrachtung dreier Subdimensionen der Unterrichtsqualität als Mediatoren. Sie fanden heraus, dass in Klassen mit einem durchschnittlich niedrigeren sozioökonomischen Status Lehrkräfte weniger Wert auf anspruchsvolle und korrekte Sprache legen. Dies wiederum stand in positivem Zusammenhang mit der Leseleistung, sodass eine mediierende Rolle der betrachteten Unterrichtsqualität im Sinne der verwendeten Sprache angenommen werden konnte.

Die Ergebnisse dieser Studien unterstreichen, dass die Beurteilung von Unterricht durch die Schülerinnen und Schüler nicht nur einseitig von der Lehrkraft, sondern auch wesentlich von der Schülerschaft abhängt. Der objektiv gleiche Unterricht kann von zwei verschiedenen Klassen ganz unterschiedlich beurteilt werden – auch abhängig von der jeweiligen Klassenkomposition. Diese Ergebnisse spiegeln auch die Annahmen des „Angebot-Nutzungs-Modells der Wirkungsweise des Unterrichts“ (Helmke, 2017) wider, in welchem Kontextfaktoren als Einflussfaktoren des Unterrichts berücksichtigt werden (siehe Kapitel 1.1.4). In den Studien wurde dies besonders bei der Beurteilung von Klassenführung sichtbar, denn häufig beziehen sich Subdimensionen auf das Verhalten der Schülerinnen und Schüler, etwa beim Ausmaß an Störungen (z. B. Fauth et al., 2019). Dies spricht dafür, bei der Nutzung von Schülerurteilen stets Merkmale der jeweiligen Klasse mitzuberücksichtigen.

Prädiktive Validität

Wie in Kapitel 1.1 beschrieben, wird erfolgreicher Unterricht häufig am Erreichen unterschiedlicher Zielkriterien wie dem fachbezogenen Interesse, der fachlichen Leistung oder der Lernmotivation der Schülerinnen und Schüler festgemacht. Folglich ist es sinnvoll, die Zusammenhänge zwischen Unterrichtsqualität und Zielkriterien des Unterrichts zu betrachten und eingehender der Frage nachzugehen, ob sich durch eine hohe wahrgenommene Unterrichtsqualität eine hohe Ausprägung von Zielkriterien vorhersagen lässt. In vielen

Studien, deren Konzeption von Unterrichtsqualität sich dem „Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität“ (Klieme et al., 2001) zuordnen lassen, ließ sich die prädiktive Kraft von Schülerurteilen empirisch belegen.

Für die Dimension Klassenführung konnte die Vorhersagekraft der Schülerurteile insbesondere für die Leistung und affektiv-motivationale Merkmale der Schülerinnen und Schüler in zahlreichen Studien gezeigt werden (Brophy, 2006; Dubberke et al., 2008; Emmer & Stough, 2001; Evertson et al., 1983; Fauth et al., 2014; Hattie, 2009; Klieme et al., 2010; Kunter & Baumert, 2006; Kunter et al., 2007; Lenske et al., 2016; Lipowsky, 2015; Marzano et al., 2000; Praetorius et al., 2018; Seidel & Shavelson, 2007; Wagner et al., 2016; Wallace et al., 2016). Dies bedeutet, dass ein gut strukturierter und störungsarmer Unterricht mit einem höheren Lernzuwachs und höher ausgeprägten motivationalen Merkmalen (z. B. dem Engagement) der Schülerinnen und Schüler einhergeht.

Die Vorhersagekraft der konstruktiven Unterstützung konnte in erster Linie für Zielkriterien wie die Lernmotivation, das akademische Selbstkonzept oder die Lernfreude gezeigt werden (Baumert et al., 2010; Cornelius-White, 2007; Davis, 2003; Fauth et al., 2014; Fendick, 1990; Hattie, 2009; Klieme et al., 2006; Kunter et al., 2013; Lipowsky, 2015; Scherer et al., 2016; Wagner et al., 2016). Demnach steht ein Unterricht, in welchem die Lehrkraft beispielsweise konstruktiv mit Schülerfehlern umgeht oder ihnen Rückmeldung gibt, wie sie sich verbessern können, in positivem Zusammenhang mit motivationalen Merkmalen der Schülerinnen und Schüler.

Auch für die kognitive Aktivierung konnte die prädiktive Kraft auf Zielkriterien des Unterrichts empirisch belegt werden. Besonders für den Lernzuwachs der Schülerinnen und Schüler hat sich ein kognitiv anregender Unterricht als relevant herausgestellt (Baumert et al., 2010; Kunter & Ewald, 2016; Kunter & Voss, 2011; Lipowsky et al., 2009; Praetorius et al., 2018). Regt also der Unterricht zu einer aktiven und auf dem jeweiligen Lernniveau herausfordernden Auseinandersetzung mit den Lerninhalten an, geht dies mit einem höheren Lernzuwachs der Schülerinnen und Schüler einher.

Jedoch lassen sich auch für andere zugrunde liegende Konzepte zur Erfassung von Unterrichtsqualität Zusammenhänge mit Zielkriterien finden. Beispielsweise ließ sich für den „Tripod 7Cs Student Perceptions Survey“ die fachliche Leistung in Mathematik und Englisch durch Schülerurteile vorhersagen (Ferguson, 2012; Kuhfeld, 2017; Wallace et al., 2016).

Zusammenfassend zeigen diese Befunde, dass Schülerinnen und Schüler Merkmale der Unterrichtsqualität sowohl auf Ebene individueller Schülerinnen und Schüler als auch auf

Klassenebene reliabel beurteilen können. Hinsichtlich der Validität von Schülerurteilen zeigte sich einerseits, dass wichtige Belege für die Dimensionalität und die Generalisierbarkeit von Schülerurteilen gefunden wurden. Schülerinnen und Schüler sind in vielerlei Hinsicht in der Lage, unterschiedliche Merkmale von Unterrichtsqualität valide einzuschätzen. Ihre Urteile erwiesen sich zudem als prädiktiv für affektive und leistungsbezogene Zielkriterien des Unterrichts. Jedoch zeigte sich auch, dass Schülerurteile durch Faktoren unabhängig der zu bewertenden Unterrichtsqualität wie der Klassenkomposition oder der Note beeinflusst sein können und Schülerurteile bei relativ konstanten Bedingungen eine hohe Variation zeigten. Da noch wichtige Fragen der Validität von Schülerurteilen unbeantwortet sind, lässt sich keine abschließende Aussage treffen. Einige dieser offenen Fragen werden im folgenden Kapitel adressiert und deren Relevanz für die Nutzung von Schülerurteilen dargelegt.

1.3 Offene Fragen zur Nutzung von Schülerurteilen

Auch wenn Schülerurteile häufig eingesetzt werden und gezeigt werden konnte, dass sie in weiten Teilen eine valide und reliable sowie informative Methode zur Erfassung von Unterrichtsqualität sind, bleiben aktuell noch wichtige Fragen hinsichtlich ihrer Nutzung offen. Diese betreffen beispielsweise die Erfassung von Unterrichtsqualität in gänzlich veränderten Kontexten. Im Frühjahr 2020 musste der traditionelle Unterricht seine gewohnten Bahnen verlassen, als die Schulen aufgrund der COVID-19-Pandemie geschlossen wurden. Doch bis zum jetzigen Zeitpunkt ist unklar, wie genau der Distanzunterricht stattfand und ob Schülerurteile auch geeignet sind, um die Qualität von Distanzunterricht zu erfassen. Auch im Hinblick auf die konkrete Formulierung von Items zur Erfassung von Unterrichtsqualität fanden wesentliche Aspekte kaum Berücksichtigung. So wurde der Frage, ob Schülerinnen und Schüler im Beurteilungsprozess zwischen unterschiedlichen Item-Adressaten differenzieren, in der Vergangenheit wenig Beachtung geschenkt. Schließlich wurde bereits dargelegt, dass Schülerinnen und Schüler bei der Beantwortung von Items auch Informationen unabhängig der eigentlichen Unterrichtspraxis heranziehen. Da dies bisher hauptsächlich innerhalb des jeweiligen Fachs untersucht wurde, stellt sich die Frage, ob auch Informationen anderer Unterrichtsfächer die Beurteilungen beeinflussen könnten. Diese drei offenen Fragen und deren Relevanz für die Nutzung von Schülerurteilen werden im Folgenden dargestellt.

1.3.1 Lässt sich auch die Qualität von Distanzunterricht durch Schülerurteile erfassen?

All die bisher aufgeführten Modelle und Befunde beziehen sich auf den schulischen Unterricht in Präsenz, wie er bereits seit vielen Jahrzehnten stattfindet. Jedoch wurden das deutsche Schulsystem sowie alle daran beteiligten Akteure im Frühjahr 2020 auf eine harte Probe gestellt, als aufgrund der COVID-19-Pandemie innerhalb weniger Tage der Präsenzunterricht, wie er schon eine lange Tradition hat, vollumfänglich auf Distanzunterricht umgestellt werden musste (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2020). In Deutschland und in vielen anderen Ländern wurde Distanzunterricht, in welchem Schülerinnen und Schüler von zu Hause aus lernen, vor den Schulschließungen nur in besonderen Ausnahmefällen angewandt, beispielsweise bei längerfristiger Krankheit (Deutscher Bundestag, 2016). Aus diesem Grund waren die wenigsten Schulen und Haushalte mit dem nötigen technischen Equipment ausgestattet. Auch gab es aufgrund der zuvor sehr seltenen Umsetzung kaum aus der Forschung abgeleitete und evaluierte Implikationen für die Unterrichtspraxis und Erkenntnisse darüber, welche Merkmale zu einem qualitätsvollen

Distanzunterricht führen. Schließlich war bislang ungeklärt, ob Schülerurteile für die Erfassung der Unterrichtsqualität für das Lehren und Lernen auf Distanz ebenfalls eine reliable und valide Methode sind und in Zusammenhang mit Zielkriterien des Unterrichts stehen.

Lehren und Lernen auf Distanz

Distanzunterricht kann bezeichnet werden als „institutionalisierter, formaler Unterricht, in welchem die Lerngruppen getrennt sind und in welchem interaktive Telekommunikationssysteme genutzt werden, um die Lernenden, Lernquellen und Lehrende zu verbinden“ (Simonson et al., 2019, S. 1). Im Vergleich zur Definition von schulischem Präsenzunterricht werden unmittelbar zwei wesentliche Unterschiede deutlich: zum einen die reguläre Trennung der Lerngruppe, zum anderen die Nutzung von digitalen Medien zur Interaktion. Lehren und Lernen auf Distanz hatte in den meisten Ländern vor den Schulschließungen im Frühjahr 2020 kaum systematisch stattgefunden. In Deutschland wird das reguläre Beschulen zu Hause vergleichsweise restriktiv gehandhabt und ist grundsätzlich verboten (Ladenthin, 2018). In anderen Ländern wie Großbritannien oder den Vereinigten Staaten von Amerika war das permanente Lernen auf Distanz eher verbreitet. Laut dem US-amerikanischen National Center of Education Statistics (NCES) lernten im Jahr 2016 knapp 1,7 Millionen Kinder und Jugendliche – dies entspricht 3,3 % aller Schulpflichtigen – von zu Hause aus (NCES, 2019). Als häufigsten Grund hierfür nannten Eltern zu 34 % Gründe bezüglich der Schulumgebung wie Sicherheit, Drogenkonsum oder problematische Peers (NCES, 2019).

Diese Zahlen änderten sich im Frühjahr 2020 rapide: Am 24. März 2020 titelte die UNESCO auf ihrer Website, dass nun 1,37 Milliarden Kinder aufgrund der COVID-19-bedingten Schulschließungen von zu Hause aus lernen (UNESCO, 2020). In Deutschland waren Schulen, Eltern sowie Schülerinnen und Schüler in technischer Hinsicht meist nicht entsprechend ausgestattet und geübt. Beispielsweise zeigte eine Umfrage der Gewerkschaft Erziehung und Wissenschaft (GEW), die im November 2019, also vier Monate vor den bundesweiten Schulschließungen veröffentlicht wurde, dass eine Mehrheit der Schulleitungen und Lehrkräfte mit einer durchschnittlich vergebenen Note von 3,8 unzufrieden mit der digitalen Ausstattung ihrer Schulen sind (Gewerkschaft Erziehung und Wissenschaft [GEW], 2019). Die Ergebnisse der „International Computer- and Information Literacy Study“ (ICILS) (Eickelmann et al., 2019) konkretisieren diese Wahrnehmung. Beispielsweise gaben nur 15,7 % der befragten Lehrkräfte an, dass an ihrer Schule jede oder einige Lehrkräfte mit

einem tragbaren, digitalen Endgerät ausgestattet sind (Dänemark: 93,7 %, USA: 86,1 %). Als Hindernisse beim Einsatz von digitalen Medien im Unterricht wurden besonders Schwierigkeiten mit dem Internet (36,8 %), zu wenige Computer (24,0 %) und Mangel an leistungsstarken Computern (18,6 %) angegeben. Während sich deutsche Lehrkräfte kompetent hinsichtlich der Suche von Unterrichtsmaterialien im Internet (Zustimmung 98,1 %) oder der Vorbereitung von Unterricht, der digitale Medien beinhaltet (Zustimmung 78,9 %) einschätzten, sind die Lehrkräfte skeptischer, was die Potenziale der Verbesserung schulischer Leistungen von Schülerinnen und Schülern durch den Einsatz digitaler Medien betrifft (Zustimmung Deutschland: 34,7 %; internationaler Mittelwert: 71,0 %) (Eickelmann et al., 2019).

Aufseiten der Schülerinnen und Schüler in Deutschland gaben nur 22,8 % der befragten Achtklässlerinnen und Achtklässler an, mindestens einmal pro Woche digitale Medien für schulbezogene Zwecke zu nutzen. In Dänemark bejahten dies 90,9 % der Befragten. Zudem zeigte sich, dass in Deutschland durchschnittlich ein Tablet auf 41 Schülerinnen und Schüler bzw. ein Laptop auf 68 Schülerinnen und Schüler kamen (in Dänemark auf 8 Schülerinnen und Schüler). Nur rund 26 % der befragten deutschen Schülerinnen und Schüler besuchten eine Schule, in welchem ein Internetzugang für Lehrkräfte *und* die Schülerschaft vorhanden ist, während dies dänische Schülerinnen und Schüler zu 100 % bejahten (Eickelmann et al., 2019). Bezüglich der computer- und informationsbezogenen Kompetenzen lagen deutsche Achtklässlerinnen und Achtklässler mit 518 Punkten im Mittelfeld (internationaler Mittelwert: 496 Punkte).

Diese Zahlen verdeutlichen, dass Deutschland, besonders was die technische Ausstattung anbelangt, in einer wenig vorteilhaften Ausgangslage war, als aufgrund der COVID-19-Pandemie innerhalb weniger Tage der Unterricht von Präsenz- auf Distanzunterricht umgestellt werden musste. Die technische Ausstattung ist nur eine Facette zur erfolgreichen Umsetzung von Distanzunterricht, wenn auch eine grundlegende. Denn damit einhergehend betraten auch hinsichtlich der Umsetzungsqualität von Distanzunterricht die meisten Lehrkräfte Neuland.

Unterrichtsqualität im Distanzunterricht

Aufgrund der weit selteneren Umsetzung von Distanzunterricht im Schulkontext und der – zumindest in Deutschland – unzureichenden Nutzung von digitalen Medien aufgrund der mangelhaften Ausstattung blickt die Unterrichtsforschung auf dem Gebiet des Distanzlernens auf eine weniger umfangreiche und systematische Tradition zurück. In Deutschland

fokussieren Veröffentlichungen beispielsweise eher auf spezifische Teilaspekte wie das Erstellen von Lernvideos (Schön & Ebner, 2013). Zudem befassen sie sich eher mit dem Hochschulbereich, da Distanzunterricht in diesem Bildungsbereich – beispielsweise im Rahmen von Fernstudiengängen – hierzulande eher verbreitet ist (Drossel et al., 2020; Gründler, 2018; Mayrberger, 2017). Auch in Ländern, in welchen Distanzunterricht erlaubt und vergleichsweise häufig umgesetzt wird, erschweren die sehr unterschiedlichen Teilnehmergruppen, Anbieter, Qualifikationen der Lehrenden oder Forschungsfokusse einen systematischen Vergleich von Qualitätsmerkmalen und Effektivität von Distanzunterricht (Barbour, 2019; Rice, 2006; Simonson et al., 2011).

Dennoch lassen sich insbesondere aus solchen Ländern Hinweise auf Qualitätsmerkmale finden. So formulierten beispielsweise Graham et al. bereits im Jahr 2001 sieben Prinzipien für gelingenden Distanzunterricht:

1. Lehrende sollten klare Richtlinien zur Interaktion mit Schülerinnen und Schülern bereitstellen.
2. Gut gestaltete Diskussionsaufgaben ermöglichen eine sinnvolle Zusammenarbeit zwischen den Schülerinnen und Schülern.
3. Schülerinnen und Schüler sollten eigene Projekte im Unterricht präsentieren.
4. Lehrende sollten zwei Arten von Feedback zur Verfügung stellen: Informatives Feedback (z. B. Note für einen Aufsatz) und bestätigendes Feedback (z. B. Aufsatz ist per E-Mail angekommen).
5. Onlinekurse brauchen Fristen.
6. Anspruchsvolle Aufgaben, Beispiele sowie Lob für gelungene Aufgaben vermitteln hohe Erwartungen an die Schülerinnen und Schüler.
7. Durch die Möglichkeit zur eigenen Themenwahl werden unterschiedliche Sichtweisen in den Distanzunterricht einbezogen.

Interessanterweise bauen diese Prinzipien auf den ursprünglich für den Präsenzunterricht formulierten „Seven Principles For Good Practice in Undergraduate Education“ (Chickering & Gamson, 1987) auf und zeigen erstaunlich viele Überschneidungen mit diesen Prinzipien. Dies deutet darauf hin, dass auch Merkmale, die für den Präsenzunterricht gelten, wichtig für qualitätsvollen Distanzunterricht sind.

Aktuell bieten die US-amerikanischen „National Standards for Quality Online Teaching“ (NSQ) umfangreiche Merkmale für gelingenden Distanzunterricht („Quality online

teaching“) im Schulkontext. Auf Basis eines Literatur-Reviews wurden acht Standards für qualitätsvolles Distanzlehren formuliert. Die vorgeschlagenen Standards beziehen sich auf Bereiche wie die Anwendung unterschiedlicher Lehrmethoden (Standard B), den aktiven Einbezug der Lernenden (Standard D) oder eine abwechslungsreiche, differenzierende Anleitung (Standard F) im Distanzunterricht (NSQ, 2019). Diese Standards befassen sich mit einer eher abstrakten Ebene von Unterrichtsmerkmalen, jedoch werden auch konkrete Beispiele für den Distanzunterricht genannt (NSQ, 2019):

> The online teacher actively participates and models both asynchronous and synchronous facilitation and interaction. This may include, but is not limited to, the following: instant messaging, text chat, audio and/or video conferencing, and other live exchange of information (synchronous); as well as email, discussion boards, blogs, and other non-live methods (asynchronous).

Betrachtet man die Inhalte der Prinzipien von Graham et al. (2001) und der NSQ (2019) genauer, lassen sich in vielen Bereichen Ähnlichkeiten mit Subdimensionen des „Rahmenmodells der drei Basisdimensionen der Unterrichtsqualität“ (Klieme et al., 2001) wie beispielsweise Feedback, Unterstützung im Lernprozess, Motivierung durch die Lehrkraft oder regelmäßige Leistungsrückmeldung ausmachen.

Zudem zeigen Studien, die sich mit einzelnen Merkmalen des Distanzunterrichts beschäftigten, wie ein qualitätsvoller Distanzunterricht gestaltet werden kann. So stellte sich heraus, dass Monitoring und unterstützende interaktive Strukturen durch die Lehrkraft die Selbstregulation von Schülerinnen und Schülern fördern (Cho & Kim, 2013; Cho & Shen, 2013). Weiterhin erwiesen sich klare Aufgabenstellungen und zeitnahes Feedback als bedeutende Merkmale für den Lernerfolg von Schülerinnen und Schülern (Hawkins et al., 2013; Liu & Cavanaugh, 2012). Schließlich fanden sich positive Zusammenhänge zwischen herausfordernden und kognitiv anspruchsvollen Aufgaben, die zur vertieften Auseinandersetzung mit dem fachlichen Inhalt anregen, und dem Lernerfolg der Schülerinnen und Schüler (Abrami et al., 2011). Eine besondere Bedeutung kommt der Möglichkeit zur Interaktion zwischen Schülerinnen und Schülern untereinander, aber auch mit der Lehrkraft zu. In zahlreichen Studien konnte gezeigt werden, dass die Möglichkeit zur Kommunikation in positivem Zusammenhang mit Zielkriterien wie dem Leistungszuwachs, dem

akademischen Selbstkonzept oder dem Abschluss des Kurses stehen (z. B. Carabajal et al., 2007; Hawkins et al., 2013; Kumi-Yeboah et al., 2017; Liu & Cavanaugh, 2012; Ouzts, 2006; Simonson et al., 2019). Auch beinhalten die „National Standards for Quality Online Teaching“ (2019) einen eigenen Themenbereich „Klassengemeinschaft“ (Standard C) mit Hinweisen und Anwendungsbeispielen zur Interaktion und Kommunikation.

Insgesamt lässt sich eine große Schnittmenge von Merkmalen für qualitätsvollen Distanz- und Präsenzunterricht erkennen. Die *Übersetzung* von Merkmalen qualitätsvollen Unterrichts in Präsenz in die digitale Lernumgebung (Graham et al., 2001) lässt darauf schließen, dass beide Lehr-Lernsettings grundlegende Gemeinsamkeiten besitzen. Bisher gibt es, anders als für den Präsenzunterricht, zur Evaluation der Qualität des Distanzunterrichts kein empirisch geprüftes Rahmenmodell, auf welches zurückgegriffen werden könnte. Aufgrund der Parallelen zwischen den Merkmalen von Präsenz- und Distanzunterricht (z. B. Monitoring und Struktur, Feedback, anspruchsvolle Aufgaben) kann jedoch davon ausgegangen werden, dass das „Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität“ für den Präsenzunterricht (Klieme et al., 2001) auch eine passende Grundlage zur Evaluation von Unterrichtsqualität im Distanzunterricht ist. Eine weitere und damit einhergehende Frage ist, ob sich auch für dieses neue und in vielerlei Hinsicht anders stattfindende Lehr-Lernsetting des Distanzunterrichts Schülerurteile zur Erfassung von Unterrichtsqualität eignen. Denn, anders als im Präsenzunterricht ist es im Distanzunterricht möglich, dass Schülerinnen und Schüler Unterricht ganz unterschiedlich erleben. Somit wäre die gemeinsame Grundlage, auf Basis derer Schülerinnen und Schüler die Qualität von Unterricht beurteilen, geringer ausgeprägt. Auch wäre es denkbar, dass Schülerinnen und Schüler über gewisse Qualitätsdimensionen keine genaue Aussage machen können. So könnten beispielsweise Merkmale der Klassenführung im Distanzunterricht für Schülerinnen und Schüler weniger offensichtlich sein als im Präsenzunterricht. Auch in bisherigen Studien wurden Schülerurteile zur Erfassung von Unterrichtsqualität genutzt. Jedoch wurden hierbei kaum psychometrische Gütekriterien berichtet, sodass unklar ist, wie gut Schülerurteile wirklich geeignet sind. Dies und die Uneinheitlichkeit der Messungen führte dazu, dass der Ruf nach reliablen und validen Messungen lauter wird (Mohammed, 2018). Die durch die Schulschließungen veränderten Bedingungen des Lehrens und Lernens bieten die Chance zu prüfen, ob Schülerurteile auch zur Erfassung der Unterrichtsqualität im Distanzunterricht geeignet und prädiktiv für Zielkriterien des Unterrichts sind.

1.3.2 Nutzen wir die richtigen Itemformulierungen zur Erfassung von Unterrichtsqualität?

Blickt man zurück auf die Anfänge der Nutzung von Schülerurteilen in der Unterrichtsforschung, so lassen sich ihre Ursprünge primär im Rahmen von sogenannten Large-Scale-Studien finden. In solchen meist repräsentativen Studien wie beispielsweise der PISA-Studie (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung [OECD], 2019) werden eine große Anzahl unterschiedlicher Merkmale von Schülerinnen und Schülern – wie deren schulische Kompetenzen sowie kulturelle, schulische oder häusliche Merkmale – als mögliche Einflussfaktoren auf die Leistung erfasst (Schulz-Heidorf & Gerick, 2017, S. 1).

Der Fokus solcher Large-Scale-Studien lag in den Anfängen meist auf der Erfassung fachlicher Kompetenzen von Schülerinnen und Schülern. Dies hatte zur Folge, dass die Aufmerksamkeit eher auf der Entwicklung und dem Einsatz von fachlichen Aufgaben lag (z. B. in Mathematik) und weniger auf Skalen und Items zur Erfassung von Unterrichtsqualität (Prenzel et al., 2013). Wirft man einen Blick auf die Instrumente aktueller Studien, wird ersichtlich, dass auch heute noch Items zur Erfassung von Unterrichtsqualität aus den ersten Studien in nahezu unveränderter Form eingesetzt werden. Beispielsweise werden die Items der Skala „Störungspräventive Überwachung der Schülertätigkeiten“ aus der Studie BIJU (Baumert et al., 1997) immer noch in aktuellen Studien wie PISA 2015 verwendet (Reiss et al., 2016). Eine langjährige Verwendung von Skalen und Items ist nicht zwangsläufig mit einer weniger guten Qualität gleichzusetzen – im Gegenteil, denn dies könnte bei entsprechenden Indizes eher als ein Indikator für passende Messungen gewertet werden. Jedoch wird bei der Betrachtung der unterschiedlichen Subdimensionen von Unterrichtsqualität ersichtlich, dass sich diese in mehrfacher Hinsicht unterscheiden. Ob und wie sehr diese Unterschiede relevant sein könnten, wurde in der Vergangenheit wiederum kaum systematisch untersucht.

Das folgende Beispiel zeigt erneut einen Ausschnitt eines Schülerfragebogens (siehe Abbildung 5, Kapitel 1.2):

		stimmt gar nicht	stimmt eher nicht	stimmt eher	stimmt genau
Feedback	Die Lehrkraft gibt mir regelmäßig Rückmeldung, was ich schon kann.	☐	☐	☐	☐
	Unsere Lehrkraft gibt uns regelmäßig Rückmeldung, was wir noch nicht so gut können.	☐	☐	☐	☐
	Die Lehrkraft zeigt immer wieder, wie ich mich verbessern kann.	☐	☐	☐	☐

In diesem Fragebogen wird ersichtlich, dass die Items sich darin unterscheiden, aus wessen Sicht das Verhalten der Lehrkraft beurteilt werden soll, also ob Items mit einem Ich-Adressaten oder einem Wir-Adressaten formuliert sind. Das Item „Die Lehrkraft zeigt immer wieder, wie *ich* mich verbessern kann" bezieht sich auf die individuelle Sicht einer jeden Schülerin und eines jeden Schülers. Das Item „Unsere Lehrkraft gibt *uns* regelmäßig Rückmeldung, was *wir* noch nicht so gut können" erfragt hingegen die Wahrnehmung aller Schülerinnen und Schüler.

Die Frage der Relevanz des Adressaten hängt eng mit dem Antwortprozess zusammen, den die Schülerinnen und Schüler bei der Bearbeitung eines Items durchlaufen. Grundsätzlich wird davon ausgegangen, dass das Beantworten von Items mit einem potenziell komplexen kognitiven Prozess einhergeht (Sudman et al., 1996; Krosnick, 1991; Tourangeau et al., 2000). Nach dem Modell von Tourangeau, Rips und Rasinski (2000) kann dieser Prozess in vier Hauptphasen unterteilt werden, die jeweils mehrere Unterphasen beinhalten.

1) *Verstehen des Items*, was die Identifizierung der gesuchten Information und die Verknüpfung von Schlüsselwörtern mit Konzepten beinhaltet. In dieser Phase lesen und interpretieren die Schülerinnen und Schüler das Item, wobei auch die Itemlänge und die Semantik[3] relevant sind. Dies setzt voraus, dass jedes Wort verstanden wird, um zu erkennen (z. B.: Was bedeutet „motiviert"?), um zu erkennen, welche Informationen gefordert sind, um das Item angemessen zu beantworten.

[3] Zur Rolle der Itemlänge und der Semantik siehe auch Baddeley et al., 1975; Göllner et al., 2016; Lenske, 2016; Shaftel et al., 2006.

2) Das *Abrufen der angeforderten Informationen* umfasst das Erinnern an relevante Ereignisse und das Füllen von Lücken durch Schlussfolgerungen. Bezieht sich ein Item beispielsweise nicht auf einen bestimmten Zeitraum (z. B. auf die letzte Unterrichtsstunde), sollen sich die Schülerinnen und Schüler an diejenigen Ereignisse erinnern, die ihnen Informationen über das entsprechende Verhalten geben.
3) Ein *Urteil wird abgegeben*, indem die Vollständigkeit und Genauigkeit der erinnerten Ereignisse bewertet und Inferenzen gezogen werden und man zu einem Urteil kommt. Basierend auf den verfügbaren Informationen in ihrem Langzeitgedächtnis kommen die Schülerinnen und Schüler zu einer Schlussfolgerung über die gesuchte Information.
4) Als letzter Schritt erfolgt die *Beantwortung des Item*s in Abhängigkeit von den Antwortkategorien und ggf. die Überarbeitung der Antwort. Die Schülerinnen und Schüler überprüfen die verfügbaren Antwortkategorien (z. B. 4-stufige Skala von „stimme gar nicht zu“ bis „stimme voll zu“) und kreuzen die entsprechende Kategorie an, je nachdem, zu welchem Urteil sie im vorherigen Schritt gekommen sind.

Somit würden bereits ab der ersten Stufe des Antwortprozesses ein Ich-Adressat oder ein Wir-Adressat aufgrund der Interpretation der unterschiedlichen Personalpronomen und der daraus resultierenden verschiedenen Informationsgrundlagen zu potenziell unterschiedlichen Urteilen führen. Beim Ich-Adressaten sind ausschließlich Ereignisse und Informationen, die die jeweilige Schülerin bzw. den jeweiligen Schüler betreffen, gefragt. Items mit Wir-Adressaten verlangen jedoch den Rückgriff auf Ereignisse, die alle Schülerinnen und Schüler einer Klasse betreffen. Dementsprechend verlangt ein solches Item theoretisch eine aufwendigere Suche nach Ereignissen, nach mehr Vergleichen und anspruchsvolleren Inferenzen (Wagner, 2008).

Ein Blick auf bestehende Instrumente zeigt, dass der „Ich-Adressat“ überwiegend für die Beurteilung der konstruktiven Unterstützung, der „Wir-Adressat“ häufiger für die Beurteilung von Klassenführung oder der kognitiven Aktivierung verwendet wird (z. B. Baumert et al., 1997; Kuhfeld, 2017; Schenke et al., 2018). Gleichzeitig zeigte sich, dass die Verwendung des Ich-Adressaten in Subdimensionen der konstruktiven Unterstützung in niedrigeren Übereinstimmungen (ICCs) innerhalb von Klassen sowie niedrigeren Zusammenhängen mit Zielkriterien auf Klassenebene resultierte als die Verwendung des Wir-Adressaten (Aldrup et al., 2018; Kunter et al., 2013). Unklar ist jedoch, ob dies der

Verwendung des Ich-Adressaten oder der zugrunde liegenden Erfassung von konstruktiver Unterstützung geschuldet ist, die besonders das individuell erlebte soziale Klima und die emotionale Unterstützung durch die Lehrkraft erfasst. So wurde die Relevanz der individuellen Wahrnehmung und der Lehrer-Schüler-Interaktion mehrfach belegt. Demnach ist die mangelnde Übereinstimmung innerhalb von Klassen auch auf die individuell unterschiedlichen Wahrnehmungen von Schülerinnen und Schülern zurückzuführen (Göllner et al., 2018; Kenny, 2004).

Die Rolle des Adressaten wurde in der Vergangenheit in nur wenigen Studien systematisch untersucht. Den Brok et al. (2006) fanden in ihrer experimentellen Studie höhere Mittelwerte für den Einfluss der Lehrkraft (z. B. durch ihre Strenge), deren Hilfsbereitschaft und ihr Verständnis für Items in einer *personal version* (Ich-Adressat) im Vergleich zur *class version* (Wir-Adressat). In einer weiteren experimentellen Studie fanden Fraser und Kollegen (1995), dass die Schülerinnen und Schüler den naturwissenschaftlichen Unterricht positiver bewerteten, wenn eine Klassenversion verwendet wurde (z. B.: „In unseren Laborsitzungen machen verschiedene Schüler verschiedene Experimente.“) als wenn eine persönliche Version verwendet wurde (z. B.: „In meinen Laborsitzungen mache ich andere Experimente als andere Schüler.“). Auch in der Studie von McRobbie und Kollegen (1991) fanden sich für die Beurteilung von Subdimensionen wie die Regelklarheit oder die Übertragbarkeit gelernter Konzepte im naturwissenschaftlichen Unterricht in den meisten Fällen positivere Beurteilungen für Items mit einem Wir-Adressaten.

Insgesamt wurde der Rolle des Item-Adressaten in der bisherigen Forschung wenig Aufmerksamkeit geschenkt, obwohl dieser theoretisch von großer Relevanz ist. Die Studien, welche sich der Thematik annahmen, zeigen keine einheitlichen Ergebnisse und haben sich größtenteils auf Mittelwertsunterschiede beschränkt. Jedoch könnten sich Unterschiede auch in einigen weiteren psychometrischen Eigenschaften der Schülerurteile niederschlagen. Daher ist es wichtig, die Rolle des Adressaten differenziert und systematisch zu untersuchen, um eine Aussage darüber treffen zu können, ob Schülerinnen und Schüler tatsächlich zwischen ihrer eigenen Wahrnehmung und der Wahrnehmung der gesamten Klasse unterscheiden und auf welche Weise man Schülerinnen und Schüler bestmöglich nach ihrer Wahrnehmung der Unterrichtsqualität befragen sollte.

1.3.3 Welche Informationen nutzen Schülerinnen und Schüler zur Beurteilung von Unterrichtsqualität?

Im Hinblick auf die Frage der Validität von Schülerurteilen wurde gezeigt, dass Schülerinnen und Schüler für ihre Urteile zur Unterrichtsqualität auch auf Informationen theoretisch abgrenzbarer Aspekte im jeweiligen Fach zurückgreifen (siehe Kapitel 1.2.3). Betrachtet man bisherige Studien, in welchen Merkmale der Unterrichtsqualität erfasst wurden, so fällt auf, dass sich diese zumeist auf ein Unterrichtsfach konzentrierten (z. B. IGLU, Bos et al., 2010; Pythagoras, Hugener et al., 2006; MARKUS, Jäger & Helmke, 2008). Aus diesem Grund beschränkte sich meist auch die Forschung zu möglichen Einflussfaktoren der Schülerurteile zur Unterrichtsqualität auf Zusammenhänge innerhalb des Unterrichtsfachs. Offen ist daher die Frage, welche Rolle Informationen eines anderen Faches, wie beispielsweise die Note, für die Beurteilung der Unterrichtsqualität spielen.

Für Schülerinnen und Schüler gibt es wohl wenig Wegweisenderes als ihre Noten. An ihnen werden Übergänge in die nächste Klassenstufe, der Erhalt von Schulabschlüssen oder Möglichkeiten zu Ausbildung und Studium festgemacht. Es wundert also nicht, dass Schülerinnen und Schüler ihren Noten eine besondere Bedeutung beimessen. Es ist hinlänglich und vielfach belegt, dass Noten als Indikatoren für die fachliche Leistung in positivem Zusammenhang mit unterschiedlichen Zielkriterien des Unterrichts stehen (Aldrup et al., 2018; Lüdtke et al., 2009; Marsh et al., 2018; Möller et al., 2009). Daher haben Noten nicht nur in Bezug auf Bildungsentscheidungen, sondern auch für das Lernen der Schülerinnen und Schüler eine zentrale Bedeutung. Für Schülerinnen und Schüler sind Noten stets präsent und ermöglichen soziale Vergleiche mit ihrer Bezugsgruppe wie beispielsweise den Mitschülerinnen und Mitschülern (Weidinger et al., 2015). Durch solche externen bzw. sozialen Vergleiche ordnen sich Schülerinnen und Schüler innerhalb ihrer Klasse hinsichtlich ihrer Leistung, widergespiegelt durch die erhaltene Note, ein. Resultat solcher Vergleiche ist die individuell unterschiedliche Entwicklung und Ausprägung des akademischen Selbstkonzepts. Dies bezeichnet die Wahrnehmung und die Kenntnis der eigenen fachlichen Fähigkeiten (Marsh & Martin, 2011; Marsh & Seaton, 2013).[4]

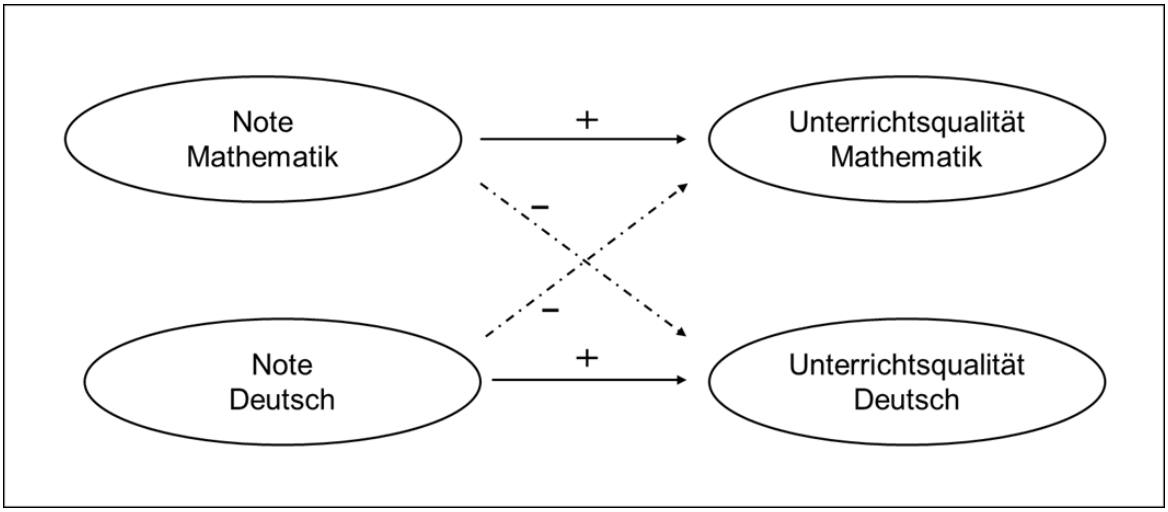
[4] Welche Rolle dabei das Leistungsniveau der Klasse spielt, wurde unter anderem in Bezug auf das akademische Selbstkonzept von Schülerinnen und Schülern erforscht und ist unter dem Begriff „Big-Fish-Little-Pond-Effect“ („Fischteicheffekt“) bekannt (Köller et al., 2006; Marsh, 1987). Demnach hängt das akademische Selbstkonzept von Schülerinnen und Schülern wesentlich vom Leistungsniveau der Referenzgruppe ab. So haben Schülerinnen und Schüler mit gleichen Fähigkeiten ein niedrigeres akademisches Selbstkonzept, wenn sie sich mit leistungsstärkeren Schülerinnen und Schülern vergleichen. Ein höheres akademisches Selbstkonzept liegt vor, wenn sie sich mit leistungsschwächeren Schülerinnen und Schülern vergleichen (Marsh et al., 2008).

Doch Schülerinnen und Schüler vergleichen sich nicht nur mit ihren Mitschülerinnen und Mitschülern: Es zeigte sich, dass sie auch interne Vergleiche hinsichtlich ihrer eigenen Leistungen in unterschiedlichen Fächern anstellen. Diese dimensionalen Vergleiche finden statt, wenn Schülerinnen und Schüler ihre Leistung in einem Fach im Verhältnis zu ihren eigenen Leistungen in einem anderen Fach evaluieren („Dimensional comparison theory"; Möller et al., 2009; Möller et al., 2015). Im Hinblick auf das akademische Selbstkonzept zeigte sich, dass soziale Vergleiche zu positiven Zusammenhängen innerhalb des Faches führen, dimensionale jedoch zu negativen Zusammenhängen zwischen zwei Fächern. Wenn beispielsweise Michelle besser in Deutsch ist als in Mathematik, dann ist ihr Selbstkonzept in Deutsch höher ausgeprägt als das für Mathematik. Wenn Jasmin in Deutsch gleich gut ist wie Michelle, jedoch besser in Mathematik, wird ihr akademisches Selbstkonzept in Deutsch niedriger ausgeprägt sein als das von Michelle, obwohl sie gleich gute Fähigkeiten haben. Diese dimensionalen (internen) und sozialen (externen) Vergleiche werden durch das „Internal/External-Frame-of-Reference-Modell" (I/E-Modell) abgebildet und wurden besonders für Zusammenhänge zwischen Mathematik und sprachlichen Fächern erforscht (Möller et al., 2006). Darüber hinaus zeigte sich, dass dieses Modell nicht nur in Bezug auf das akademische Selbstkonzept, sondern auch für andere Zielkriterien des Unterrichts, wie Motivation oder Lernfreude, Anwendung findet (Götz et al., 2008; Möller et al., 2015). Dies wurde unter dem Titel „generalisiertes I/E-Modell" zusammengefasst (Möller et al., 2015; Rösler et al., 2018).

Dass Noten für Zielkriterien wie etwa das akademische Selbstkonzept oder die Lernmotivation von großer Bedeutung sind, wurde vielfach belegt. Auch wurde gezeigt, dass Noten in positivem Zusammenhang mit Unterrichtsqualität stehen, d. h., je besser die Note, desto besser wird auch die Unterrichtsqualität des jeweiligen Fachs bewertet (siehe auch Kapitel 1.2.3). Weniger betrachtet wurde bisher jedoch der Zusammenhang von Note und Unterrichtsqualität in zwei Fächern, das heißt, ob die Annahmen der sozialen und dimensionalen Vergleiche auch für die Leistung von Schülerinnen und Schülern und deren Einschätzung von Unterrichtsqualität zutreffen. Abbildung 6 zeigt, welche Zusammenhänge dem Muster der sozialen Vergleiche (durchgezogene Pfeile) und der dimensionalen Vergleiche (gestrichelte Pfeile) zufolge angenommen werden können.

Abbildung 6

Zusammenhänge zwischen Noten und Unterrichtsqualität nach dem Internal/External-Frame-of-Reference-Modell (nach Marsh, 1986)



In diesem Modell werden positive Zusammenhänge zwischen einer Note und der Einschätzung der Unterrichtsqualität innerhalb eines Faches, jedoch negative Zusammenhänge zwischen Fächern angenommen. Diese Zusammenhänge nahmen Arens und Möller (2016) in den Blick. Sie untersuchten in ihrer Studie das Muster der dimensionalen Vergleiche auf Schülerebene zwischen Mathematik- und Deutschnoten und zwei Facetten der Unterrichtsqualität, der Schüler-Lehrer-Beziehung und Instruktionsqualität. Die Autorinnen und Autoren konnten die Annahmen der dimensionalen Vergleiche, nämlich positive Zusammenhänge innerhalb der Fächer und negative Zusammenhänge zwischen Fächern bestätigten. Dennoch bleiben wichtige Fragen unbeantwortet, wie beispielsweise, ob dieses Muster auch auf Merkmale der Unterrichtsqualität zutrifft, die weniger auf die persönliche Beziehung zwischen einem Schüler und seiner Lehrkraft abzielt (z. B. Strukturiertheit der Unterrichtsstunden) und ob dieses Muster nicht nur auf der Schülerebene, sondern auch auf Klassenebene zutrifft. Schließlich ist offen, ob diese Zusammenhänge aufgrund von tatsächlichen Leistungsunterschieden in den beiden Fächern auftreten oder aufgrund der dimensionalen Vergleiche.

Die Beantwortung dieser Fragen und damit, ob Schülerinnen und Schüler für die Beurteilung der Unterrichtsqualität eines Faches auch von der Lehrkraft unabhängige Informationen, nämlich die Note eines anderen Faches nutzen, leistet einen wichtigen Beitrag für das Verständnis der Validität von Schülerurteilen. Würden die Zusammenhänge des Modells bestätigt, so würde dies bedeuten, dass Schülerinnen und Schüler bei der Bewertung

von Unterrichtsqualität in zwei Fächern auch fächer- und lehrkraftunabhängige Informationen des jeweiligen Fachs heranziehen.

2 FRAGESTELLUNGEN

Schülerinnen und Schüler erleben tagtäglich Unterricht in vielen unterschiedlichen Facetten. Herrn Schmidt in Geschichte kann man so richtig auf der Nase rumtanzen und alles geht drunter und drüber, aber Frau Müller hat in Mathe die Klasse voll im Griff und keiner kann am Ende der Stunde von sich behaupten, er hätte nichts zu Papier gebracht. Zur Erfassung der Qualität von Unterricht bieten Schülerurteile einige wesentliche Vorteile. Beispielsweise, dass sie niederschwellig einsetzbar und kostengünstig sind und damit eine große Anzahl von Beurteilern erreicht werden kann (Kunter et al., 2008; Wagner, 2008). Unterricht wird heute als interaktionistisches Geschehen zwischen Lehrkräften und ihrer Schülerschaft verstanden (Klieme, 2019; Reusser, 2009). Daher vermittelt das Einholen von Schülerurteilen zur Unterrichtsqualität auch Wertschätzung gegenüber der Schülerschaft, die den Unterricht täglich erlebt und mitgestaltet.

Das „Angebot-Nutzungs-Modell der Wirkungsweise des Unterrichts“ (Helmke, 2017) bietet eine hilfreiche Grundlage, um Unterricht und dessen Qualität sowie mögliche Einflussfaktoren zu verorten. Studien mit Fokus auf der Konzeptualisierung von Unterrichtsqualität konnten empirisch zeigen, dass das „Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität“ durch die generischen Basisdimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung das Unterrichtsgeschehen abgebildet werden kann und diese prädiktiv für unterschiedliche Zielkriterien des Unterrichts sind (Baumert et al., 2010; Göllner et al., 2018; Klieme et al., 2001; Lipowsky et al., 2018; Praetorius et al., 2018; Wagner et al., 2013). Schülerurteile als *Werkzeug* zur Erfassung von Unterrichtsqualität im Allgemeinen und der drei Basisdimensionen im Besonderen wurden vielfach als reliabel befunden (De Jong & Westerhof, 2001; Kane et al., 2013; Praetorius et al., 2018). Auch hinsichtlich unterschiedlicher Perspektiven der Validität konnte gezeigt werden, dass Schülerinnen und Schüler Unterrichtsqualität verlässlich einschätzen können (Fauth et al., 2014; Wagner et al., 2013). Dennoch sind besonders in Bezug auf die Validität von Schülerurteilen, die aus unterschiedlichen Blickwinkeln betrachtet werden muss, noch wichtige Fragen unbeantwortet. Die vorliegende Arbeit hat zum Ziel, drei dieser Fragen zur Nutzung von Schülerurteilen zu adressieren und so einen wichtigen Beitrag zum Verständnis der Validität von Schülerurteilen zu leisten.

Studie 1 nimmt zunächst eine nie da gewesene Unterrichtssituation in den Blick: Die Schulschließungen zur Eindämmung der Corona-Pandemie im Frühjahr 2020 führten zu einer massiven Verschiebung der Unterrichtssituation – nämlich vom Klassenzimmer ins heimische Wohn- oder Kinderzimmer. Jedoch haben sich nicht nur die örtlichen Gegebenheiten verändert. Vielmehr ist davon auszugehen, dass auch der Unterricht selbst in seiner Struktur und Umsetzung anders als gewohnt stattgefunden hat. Wie also wurde Unterricht auf Distanz konkret umgesetzt? Welche digitalen Tools, Methoden und Formen der Interaktion hat es gegeben? Und: Worauf kommt es im Distanzunterricht wirklich an? Um diese Fragen zu beantworten, wurde die Onlinestudie „Digitaler Unterricht im Homeschooling“ (CUNITAS) durchgeführt, an der im Frühsommer 2020, also unmittelbar nach der ersten landesweiten Schulschließung 3.159 Schülerinnen und Schüler der Sekundarstufe, 1.688 Eltern und 227 Lehrkräfte teilnahmen. Diese Daten ermöglichen es, herauszufinden, wie der Distanzunterricht aus organisatorischer Sicht umgesetzt und wie die Qualität dieser unterschiedlichen Umsetzungsformen in mehreren Fächern eingeschätzt wurde. Zudem konnten diese Umsetzungsformen mit Zielkriterien des Unterrichts in Zusammenhang gebracht werden, um Aussagen über die Vorhersage unterschiedlicher Merkmale des Distanzunterrichts auf das schulische Lernen treffen zu können. Schließlich lässt sich anhand dieser Daten erstmals untersuchen, ob Schülerurteile auch in veränderten Kontexten, nämlich dem Distanzunterricht, eine geeignete Methode sind, um Unterrichtsqualität auf Basis des „Rahmenmodells der drei Basisdimensionen der Unterrichtsqualität“ (Klieme et al., 2001) zu erfassen und deren Zusammenhänge mit Zielkriterien des Unterrichts zu betrachten.

Studie 2 widmet sich der Rolle der Itemformulierungen. Welchen Unterschied macht es, wenn Schülerinnen und Schüler mit Items mit einem Ich-Adressaten befragt werden, die also auf die *individuelle* Wahrnehmung der Schülerinnen und Schüler abzielen, oder mit Items mit einem Wir-Adressaten, die die *geteilte* Wahrnehmung aller Schülerinnen und Schüler erfassen sollen? In Anbetracht des Antwortprozesses bei der Bearbeitung von Items (Tourangeau et al., 2000) lässt sich aus theoretischer Perspektive begründen, dass sich beide Arten der Befragung auf unterschiedliche Erfahrungen und Informationsgrundlagen beziehen. Auch wurde gezeigt, dass Schülerurteile auf der Schülerebene individuelle und auf der Klassenebene geteilte Wahrnehmungen der Schülerinnen und Schüler abbilden. Somit ließe sich ein Ich-Adressat aufgrund der individuellen Informationen eher auf Schülerebene und ein Wir-Adressat aufgrund der geteilten Informationen eher auf Klassenebene als relevant erachten. Ob diese theoretischen Annahmen zutreffen und sich Unterschiede beider Adressaten in Schülerurteilen widerspiegeln, wird anhand einer experimentellen Variation des

Adressaten untersucht. Vier unterschiedliche Versionen des Item-Adressaten (Ich-Adressat, Wir-Adressat, zwei Mischformen mit wechselndem Adressaten) in der UNITAS-Studie ermöglichten es, die Rolle des Item-Adressaten grundlegend zu untersuchen. In der UNITAS-Studie wurden rund 6.500 Schülerinnen und Schüler der Klassenstufen 5 bis 10 zu einer Vielzahl an Subdimensionen der Unterrichtsqualität sowie zu Zielkriterien des Unterrichts in den Fächern Deutsch und Mathematik befragt. Potenzielle Unterschiede in einer Vielzahl von psychometrischen Eigenschaften von Schülerurteilen geben Aufschluss darüber, wie Items zur Erfassung von Unterrichtsqualität formuliert sein sollten, um valide Informationen zu erhalten.

Studie 3 geht der Frage nach, ob Schülerurteile zur Unterrichtsqualität in einem Fach durch Informationen eines anderen Faches, nämlich der Note, beeinflusst sein können. Noten spielen für Schülerinnen und Schüler eine bedeutende Rolle (Weidinger et al., 2015). Durch dimensionale Vergleiche ihrer Noten in zwei Fächern ordnen Schülerinnen und Schüler ihre schulische Leistung ein. Die Zusammenhänge innerhalb und zwischen Fächern wurden primär in Bezug auf das schulische Selbstkonzept von Schülerinnen und Schülern untersucht (Möller et al., 2009; Möller et al., 2015). Zudem ließen sich vielfach positive Zusammenhänge zwischen der Note und der Einschätzung der Unterrichtsqualität im selben Fach finden (Aldrup et al., 2018; Lüdtke et al., 2009; Marsh et al., 2018). In dieser Studie werden beide Ansätze kombiniert und der Frage nachgegangen, ob für die Beurteilung der Unterrichtsqualität in zwei Fächern nicht nur die Note im gleichen Fach, sondern auch die Note des zweiten zu beurteilenden Fachs von Bedeutung ist. Hierfür wurden ebenfalls Daten der UNITAS-Studie, in welcher Unterrichtsqualität sowie eine Reihe weiterer Merkmale für die Fächer Deutsch und Mathematik erhoben wurden, genutzt. Um sicherzustellen, dass mögliche Vergleiche nicht auf tatsächliche Leistungsunterschiede zurückzuführen sind, sondern aufgrund der dimensionalen Vergleiche angestellt werden, wird zusätzlich für Testwerte standardisierter Leistungstests kontrolliert.

Zusammengefasst haben diese drei Studien zum Ziel, wichtige Erkenntnisse für die Nutzung von Schülerurteilen von Unterrichtsqualität zu gewinnen. Ließe sich auch für den veränderten Kontext des Distanzunterrichts zeigen, dass Schülerinnen und Schüler die Unterrichtsqualität einschätzen können und ihre Urteile in Zusammenhang mit Zielkriterien des Unterrichts stehen, könnten Schülerurteile wichtige Aussagen über diesen neu stattfindenden Lernkontext und die Bedeutung unterschiedlicher Umsetzungsformen machen. Die Beantwortung der Frage, ob Schülerinnen und Schüler tatsächlich zwischen Items mit

Ich-Adressaten und Wir-Adressaten differenzieren, liefert wichtige Erkenntnisse zur grundsätzlichen Frage, welche Rolle der Adressat in Items zur Erfassung von Unterrichtsqualität spielt. Schließlich gibt die Betrachtung von Einflussfaktoren eines anderen Faches auf Schülerurteile Aufschluss darüber, ob Schülerinnen und Schüler bei der Beurteilung von Unterrichtsqualität in einem Fach auch auf Informationen eines anderen Fachs zurückgreifen, die theoretisch unabhängig von der zu bewertenden Unterrichtsqualität sind.

3

STUDIE 1: DIGITAL TEACHING DURING THE COVID-19 CRISIS: SOCIAL CONNECTEDNESS MATTERS MOST FOR TEACHING QUALITY AND STUDENTS' LEARNING

Jaekel, A., Wagner, W., Trautwein, U., & Göllner, R. (2021). Digital Teaching During the COVID-19 Crisis: Social Connectedness Matters Most for Teaching Quality and Students' Learning. *Manuskript eingereicht zur Publikation.*

Abstract

In spring 2020, school closures were a core action to slow the spread of the COVID-19 pandemic. Students and teachers faced the challenge of organizing digital teaching and learning without sufficient preparation time. In this study, we investigated how teachers implemented teaching at a distance and how these different implementations were associated with students' and their parents' perceptions of teaching quality as well as students' social involvement, enjoyment of learning, academic effort, and perceived competence. To this end, we examined data from 277 teachers, 3,159 students and 1,688 parents who rated classes in mathematics, German language arts, and English language during the school closures. The results showed that teachers exhibited great variety in their implementations of distance teaching. Teaching methods enabling social connectedness (e.g., video meetings, learning videos created by the teacher) revealed the most consistent positive associations with students' and parents' teaching quality ratings as well as student outcomes.

Introduction

"The only positive thing was the silence while working. Besides that, I missed the social contact very much, and of course being able to ask the teacher spontaneously when I had questions."

Male student, 7th grade

The worldwide school closures in spring 2020 were a huge challenge for everyone involved in school life. Students and teachers had to get used to a very different teaching and learning format, and parents were more involved than ever before in their children's learning processes (Garbe et al., 2020). Due to the suddenness of the closures, teachers in most cases had no concepts for effective distance learning and teaching upon which to rely. Consequently, it can be assumed that distance learning was implemented very differently not only across schools, but also by different individual teachers, with unknown consequences for teaching quality and students' learning.

The quality of teaching in the school context has been extensively investigated in the past - but mainly in face-to-face teaching situations (e.g., Doyle, 2013; Hamre & Pianta, 2010; Hattie 2009). In such work, three basic domains of teaching quality, namely classroom management, supportive climate, and cognitive activation, each operationalized in terms of several quality dimensions (e.g., how structured classes are or how regularly student receive feedback), were found to be relevant for a variety of student learning outcomes (Hamre & Pianta, 2010; Hattie, 2009). There is no corresponding, systematically evaluated framework that could be used to identify components of quality distance education. However, there are some guidelines and recommendations for the effective implementation of distance education (Graham et al., 2001; NSQ, 2019). Some of these indicators (e.g., monitoring students' learning progress, regular feedback) have also been confirmed empirically, as they have been found to be positively associated with student outcomes (e.g., Hawkins et al., 2013). These factors exhibit great overlap with indicators of face-to-face teaching quality in the school context. Therefore, we apply the three basic domains of teaching quality framework to systematically evaluate the quality of different distance education formats implemented during the COVID-19 pandemic as well as their associations with student outcomes.

In the present study, we aimed to investigate the implementation of distance teaching and learning in three subjects during school closures in Germany. First, we investigated which methods teachers applied. Second, we studied the associations between the applied methods and students' and their parents' perceptions of teaching quality. Third, we investigated the

associations between the applied methods and student outcomes, that is, students' social involvement, enjoyment of learning, academic effort, and student-perceived competence. Finally, we examined the associations between teaching quality and student outcomes. Answering these questions gives us more insight into the extent to which teachers' teaching practices are associated with students' learning and therefore have important implications for similar situations in the future.

The Quality of Teaching

Teaching quality can be understood as a teacher's actual behavior, but also the teacher's interactions with students (Doyle, 2013; Fauth et al., 2019). In the face-to-face teaching context, teaching quality has been found to be vital for a wide range of student outcomes like achievement, learning enjoyment, and academic effort (e.g., Lam et al., 2015; Wagner et al., 2016). The concept of teaching quality can be considered from two perspectives: First, one can investigate how teachers implement their lessons from an organizational perspective, with reference to the learning materials, devices, or social arrangements teachers use and apply in their teaching. For instance, does the teacher use a traditional workbook, worksheets they create themselves, or the latest technical devices? Are the lessons taught in a lecture format or do students work in small groups? Generally, wide variability has been found in the use of different devices and methods in face-to-face teaching (Pauli et al., 2003; Seidel & Shavelson, 2007); most importantly, these observable aspects of how lessons are organized have been found to not predict student outcomes (Hattie, 2009; Kunter et al., 2011).

Second, teaching quality can be considered with respect to different characteristics related to the students' learning process. For instance, how well does the teacher monitor the students' work? To what extent does he/she support the students emotionally in their learning process? From this perspective, the three domains of teaching quality offer a framework to describe and evaluate teaching. In this framework, classroom management, supportive climate, and cognitive activation are applied as indicators describing the quality of teaching (Hamre & Pianta, 2010; Praetorius, 2018). Classroom management refers to an efficient way of teaching and using instructional time; it can result from rule clarity, a well-structured lesson, or the absence of disturbances, for example (Kunter et al., 2007). Supportive climate builds on a positive student-teacher relationship and a learning environment in which, for example, students receive constructive feedback on how to improve their performance and experience the relevance of the subject matter (Brophy, 2000). Finally, cognitive activation

aims to have students actively engage with the subject matter. This can be facilitated, for instance, by providing challenging tasks that clarify the connection between different concepts or linking new learning content with prior knowledge (Lipowsky et al., 2009). The theoretical framework of the three domains of teaching quality has received empirical support from several studies (e.g., Authors, 2018; Authors, 2020; Fauth et al., 2014; Kunter & Voss, 2013; Wagner et al., 2013). A large number of studies have revealed the three dimensions' predictive power for student outcomes such as motivation, achievement, and academic self-concept (Hattie, 2009; Praetorius et al., 2018; Seidel & Shavelson, 2007; Wagner et al., 2016). Accordingly, the organizational component of teaching quality, including which devices, materials, or methods teachers use, is less important for student learning; rather, as described in this framework, it is characteristics related to the students' learning process that matter. For instance, Hattie (2009) found in a meta-analysis that methods like individual work (d = .04) or adapted learning methods (d = .19) exhibited rather small effects on students' learning, whereas teaching characteristics like feedback (d = .73) or effective practicing (d = .71) were much more important and yielded larger effect sizes on students' learning outcomes.

These two perspectives on teaching quality should be relevant and applicable to distance education during the pandemic-induced school closures. Teachers might have selected different tools and methods to implement their lessons, for instance, how work assignments were transmitted to students or whether virtual meetings took place. However, the influence of organizational aspects such as teaching methods is unclear, because their deployment in distance learning situations might strongly depend how the teacher structures and organizes lessons. For example, students are only able to ask questions or work together with their classmates if the teacher provides them with structures to do so, such as video meetings or collaborative tasks. Furthermore, the framework of the three basic domains highlights important features that can also be helpful for evaluating the quality of distance education. As is the case for face-to-face teaching, structured lessons with clear rules, regular feedback by the teacher, and challenging tasks can be assumed to be relevant for students' successful learning in distance learning as well.

Distance Teaching and Learning

Modern distance education is defined as "institution-based, formal education where the learning group is separated, and where interactive telecommunication systems are used to connect learners, resources, and instructors" (Simonson & Schlosser, 2009, p. 1). Typically, distance education at the K-12 level is implemented through synchronous or asynchronous

text, audio, or video courses, often supplemented with print or digital learning materials (Barbour & Reeves, 2009; Watson et al., 2015). The way these formats are organized varies greatly regarding aspects such as the overall course design (e.g., blended learning, full-time online learning), the digital tools used (e.g., email, telephone, the cloud), and the teaching methods applied (e.g., learning videos created by the teacher, groupwork) (Burch et al., 2016; Kumi-Yeboah et al., 2018; Watson et al., 2015). For this reason, other terms like e-learning, virtual learning, remote learning, or web-based learning are also commonly used, either interchangeably or to refer to specific facets of distance education (Simonson, 2019).

Although distance education had been implemented in many countries prior to the COVID-19 pandemic only under special circumstances (for instance, for students living in rural areas; Picciano et al., 2010), several guidelines exist describing aspects of effective distance education and providing recommendations for implementation. For instance, Graham and colleagues (2001) proposed seven principles of effective teaching in online undergraduate courses. These principles promote a clear structure with deadlines, opportunities for interaction and cooperation, meaningful feedback, and active learning, which includes, for instance, challenging tasks or project presentations. The National Standards for Quality Online Teaching (2019) defined eight standards for effective online teaching, such as "digital pedagogy", "community building", and "learner engagement", based on an extensive review of research. Each of these standards is subdivided into different indicators and underpinned with explanations and examples. For instance, one indicator of "digital pedagogy" reads: "The online teacher uses different types of tools to interact in online courses in order to nurture learner relationships, encourage learner interaction, and monitor and motivate learner engagement." (p. 12). Thus, this indicator focuses on the teacher's interaction with his/her students, which enables the teacher to monitor and motivate them. Teacher-student interactions in distance education have also been found to be important for students' learning in several studies. For instance, teachers' monitoring and providing supportive structures can help students regulate their learning (Cho & Shen, 2013) and enable teachers to keep abreast of their students' learning progress (Moore & Kearsley, 2011). Moreover, in distance education, the quantity and quality of interactions not only between the teacher and students but also among the students themselves has exhibited positive associations with student outcomes such as achievement, academic self-concept, and course satisfaction (Baturay & Yükseltürk, 2015; Borup et al., 2014; Cavanaugh et al., 2009; Hawkins et al., 2013; Kuo et al., 2014; Kumi-Yeboah et al., 2018; Liu & Cavanaugh, 2012). Meaningful and timely feedback to students was found to be another important characteristic for students' learning

(Cavanaugh et al., 2009; Hawkins et al., 2013; Kumi-Yeboah et al., 2018; Liu & Cavanaugh, 2012). For instance, Hawkins and colleagues (2013) examined the associations between the quantity and quality of interactions (e.g., feedback) and course completion rate in a virtual high school and found that high-quality and more frequent interactions increased the proportion of course completers.

Overall, previous research has identified several important aspects of effective distance education. Specifically, promoting interactions and students' feeling of social involvement and inclusion has been found to be a core component of students' distance learning. There is still little research systematically examining the association between different forms of implementing distance education, perceived teaching quality, and student outcomes. Moreover, there is no theoretical framework which can be used to evaluate the effectiveness of distance education formats along validated quality dimensions. Frameworks based on face-to-face teaching offer helpful suggestions but need to be adapted to this different teaching and learning situation with its unique requirements. There is great overlap among the indicators in guidelines for effective online teaching, existing research of important factors for student outcomes in distance education (e.g., monitoring, feedback, challenging tasks), and the three domains of teaching quality framework. For this reason, we draw upon the framework in our study, which allows to systematically evaluate teaching quality in distance education and its associations between different implementation formats as well as student outcomes.

The Present Investigation

In the present study, we sought to shed light on the question of how teaching and learning took place during the period of school closures as well as the relevance of these different implementations for students' learning. Studies have examined indicators of effective teaching and learning in distance education and found aspects like monitoring, motivation, and challenging tasks to be relevant for students' learning outcomes. Several studies particularly emphasized the role of interactions with the teacher or with other students. However, there is a lack of research systematically investigating the associations between how teaching is implementation, perceived teaching quality, and the impact on students' learning. In this study, we sought to contribute to this research field by investigating which methods teachers used during the school closures in spring 2020 and how these implementation formats were associated with students' perceptions of teaching quality. Because parents were more involved in their children's learning process than ever before during the period of school closures, we

also took their perspective on teaching quality into account. Furthermore, we examined the associations between the applied teaching methods as well as students' perceptions of teaching quality with several outcome variables. Our work addressed the following research questions:

1) Which teaching methods (e.g., groupwork, video meetings, teacher-created learning videos) did teachers use in mathematics, German language arts, and English as a foreign language during the school closures?
2) How is the use of different teaching methods associated with students' and parents' perceptions of teaching quality along the teaching quality dimensions of monitoring, structuredness, rule clarity, differentiation, learning support, feedback, challenging tasks, and practicing?
3) How are the teaching methods associated with student outcomes in terms of social involvement, enjoyment of learning, academic effort, and perceived competence?
4) How are the students' perceptions of teaching quality associated with the student outcomes of social involvement, enjoyment of learning, academic effort, and perceived competence?

Method

The present study was part of a larger research project about the validity of students' teaching quality for predicting learning (Teaching Quality From the Students' Perspective, UNITAS; Jaekel et al. 2021) which was approved by the Ministry of Culture, Youth, and Sport of Baden-Württemberg. In addition, the ethics committee of Economics and Social Sciences at the University of Tübingen confirmed that the procedures were in line with the ethical standards for research with human subjects (File number A2.5.4-074_aa).

Sample

The data for this study stem from the "Distance Education in Homeschooling (CUNITAS)" study, which was conducted in June/July 2020 in the federal state of Baden-Württemberg, Germany. The CUNITAS study examined the implementation and quality of distance education during the spring 2020 school closures in mathematics, German language arts, and English as a foreign language. A total of 3,159 students in 241 classes from Grades 5 to 12 participated in the study. 52.2% of the students were female, 47.2% male, and 0.6% other. Students came from five academic-track schools (n = 1,719) and seven non-academic-track schools (n = 1,440) and provided ratings on their mathematics, German language arts, and English language classes. Additionally, a total of 227 teachers provided ratings on 327 classes (mathematics: n = 125; German language arts: n = 112; English language: n = 90). 69.1% of the teachers were female, 30.4% male, 0.5% other. Finally, 1,688 parents and legal

guardians (mother: *n* = 1,348; father: *n* = 293; other: *n* = 19) rated their children's learning situation during the school closures as well as their learning background.

Instruments

Teacher Ratings on Digital Tools and Teaching Methods

Teachers were asked about the digital tools they used and teaching methods they applied during the school closure period starting in March 2020. To create a broad list of options, we referred to the current literature (Hillmayr et al., 2017; Robinson et al., 2019) and conducted pretests with teachers. Based on their feedback, we adapted the different options available. Concerning their use of digital tools, teachers could choose options such as email, telephone, messaging services, or the cloud. With respect to methods used, they could choose from among video meetings, groupwork, or links to third-party learning videos, for example. Teachers could select multiple responses to both questions.

Teaching Quality

Teaching quality from the students' and their parents' perspective was assessed along the three domains of teaching quality with six quality dimensions: monitoring (e.g., "My math teacher always knew exactly what I was working on") and structuredness (e.g., "Assignments and learning materials were always provided on time for math") for *classroom management*; learning support (e.g., "My math teacher encouraged me to ask questions") and feedback (e.g., "My math teacher gave me regular feedback on my tasks") for *supportive climate*; challenging tasks (e.g., "My math teacher assigned tasks that I had to think about very carefully") and practicing (e.g., "The practice exercises allowed me to see if I had mastered the material") for *cognitive activation*. Each dimension was assessed with 3 to 4 items, for a total number of 21 items. The items from the student and parent perspectives were worded the same, e.g., "My math teacher always knew exactly what I was working on" and "The math teacher always knew exactly what my child was working on". Most of the items had previously been used in large-scale studies such as PISA (Authors, 2020). Additionally, we adapted the specific subject named in the items to address each of the three subjects. All items were rated on a 4-point Likert scale from 1 (*strongly disagree*) to 4 (*strongly agree*). To keep effort for the students low but obtain as much data as possible, students were randomly assigned to rate the subject-specific items for two out of the three subjects (mathematics: *n* = 1,319; German language arts: *n* = 1,317; English language: *n* = 1,228). Intraclass correlations (ICCs) for teaching quality ratings from the students' perspective ranged from .05 (challenging tasks in English language) to .37 (monitoring in mathematics), and from the

parents' perspective from .08 (challenging tasks in mathematics) to .37 (monitoring in mathematics) (Table 1). Reliability of the dimensions ranged from α = .67 to α = .88 for the students' perspective and α = .78 to α = .93 for the parents' perspective (Table 2).

Student Outcomes

We assessed students' learning outcomes in the three subjects with four scales, which were each rated on a 4-point Likert scale from 1 (*strongly disagree*) to 4 (*strongly agree*). S*tudents' perceived competence* (Ramm et al., 2006) was assessed with four items, e.g. "In math class, I was also given difficult tasks.". Scale reliability was α = .76 (mathematics), α = .80 (German language arts), and α = .82 (English language). *Academic effort* (Jonkmann et al., 2013) was assessed with four items, e.g., "I tried hard to learn a lot" (mathematics: α = .87; German language arts: α = .88; English language α = .90). *Learning enjoyment* (adapted from Ramm et al., 2006) was also assessed with four items, e.g., "I enjoyed math classes", with good reliability (mathematics: α = .87; German language arts: α = .88; English language α = .88). Finally, we assessed students' perceived *social involvement* (self-development) with four items, e.g. "In math classes, I experienced our class as a class community" (mathematics: α = .82; German language arts: α = .85; English language α = .86). In terms of intraclass correlations, students' reports on their academic effort exhibited the lowest values (mathematics: *ICC* = .08; German language Arts; *ICC* = .08; English language *ICC* = .06), whereas social involvement exhibited the highest values (mathematics: *ICC* = .19; German language Arts; *ICC* = .17; English language: *ICC* = .13). Descriptive statistics are shown in Table 3.

Analyses

All data analysis was conducted using SPSS 24.0 for Windows. Descriptive statistics were calculated for the variables used in the analyses and to present teachers' reports on the digital tools and teaching methods used. In line with our research questions, we then inspected bivariate relations for all three subjects combined (a) between teaching methods and dimensions of teaching quality from the students' and parents' perspective, (b) between teaching methods and student learning outcomes, as well as (c) between the dimensions of teaching quality from the students' and parents' perspective and students' learning outcomes. We conducted partial correlations controlling for the potential impact of the covariates school type (academic-track vs. non-academic-track), grade level (5 to 12), and class size. We also took differences between the three subjects into account, with mathematics as the reference category (Table S1 to S3 in the appendix). In analyzing the associations between teaching

Table 1

Descriptive Statistics for Teaching Quality Dimensions from Students' and Parents' Perspectives for Mathematics, German Language Arts, and English Language

	Mathematics						German Language Arts						English Language					
	Students			Parents			Students			Parents			Students			Parents		
	M	*SD*	*ICC*	*M*	*SD*	*ICC*	*M*	*SD*	*ICC*	*M*	*SD*	*ICC*	*M*	*SD*	*ICC*	*M*	*SD*	*ICC*
Monitoring	2.52	0.60	.37	2.53	0.68	.37	2.72	0.60	.30	2.78	0.62	.35	2.83	0.58	.30	2.95	0.60	.30
Structuredness	3.44	0.41	.18	3.41	0.42	.17	3.41	0.38	.27	3.46	0.38	.27	3.44	0.42	.24	3.46	0.44	.22
Learning Support	2.92	0.53	.25	2.71	0.57	.25	2.86	0.60	.25	2.73	0.61	.28	2.89	0.57	.18	2.79	0.61	.22
Feedback	2.36	0.63	.32	2.25	0.62	.27	2.53	0.69	.30	2.42	0.68	.29	2.67	0.70	.24	2.61	0.63	.24
Challenging Tasks	3.13	0.33	.07	3.10	0.32	.08	2.89	0.35	.12	2.99	0.35	.12	3.02	0.33	.05	3.04	0.42	.10
Practicing	3.10	0.44	.14	3.09	0.39	.09	2.85	0.47	.18	2.94	0.40	.15	3.11	0.48	.12	3.11	0.42	.12

Table 2

Cronbach's Alpha for Teaching Quality Dimensions from Students' and Parents' Perspective for Mathematics, German Language Arts, and English Language

	Mathematics		German Language Arts		English Language	
	Students	Parents	Students	Parents	Students	Parents
Monitoring	.82	.91	.86	.92	.86	.92
Structuredness	.72	.84	.79	.87	.82	.91
Learning Support	.88	.92	.88	.93	.88	.93
Feedback	.82	.89	.86	.91	.88	.91
Challenging Tasks	.69	.78	.67	.81	.68	.83
Practicing	.75	.80	.79	.82	.81	.82

Table 3

Descriptive Statistics for Student Outcomes in Mathematics, German Language Arts, and English Language

	Mathematics				German Language Arts				English Language			
	M	*SD*	*ICC*	*α*	*M*	*SD*	*ICC*	*α*	*M*	*SD*	*ICC*	*α*
Perceived Competence	2.91	0.41	.09	.76	2.82	0.42	.14	.80	2.90	0.47	.11	.82
Academic Effort	3.38	0.37	.08	.87	3.33	0.36	.08	.88	3.34	0.41	.06	.90
Enjoyment of Learning	2.63	0.46	.11	.87	2.65	0.48	.14	.88	2.77	0.49	.12	.88
Social Involvement	1.92	0.48	.19	.82	1.89	0.52	.17	.85	1.97	0.55	.13	.86

methods, teaching quality dimensions and student outcomes, we used the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) to control the false discovery rate for multiple comparisons.

Results

Teaching Methods Applied in Distance Learning

To gain a broader understanding of how distance teaching was implemented, we first examined the digital tools teachers reported having used. The frequencies are shown in Table 4. In all subjects, teachers most frequently used email (mathematics: 86.5%; German language arts: 93.1%; English language: 99.0%) and telephone (mathematics: 39.9%; German language arts: 44.8%; English language: 43.0%). For some tools, teachers' reports varied between subjects: For instance, teachers in mathematics used YouTube (37.6%) and the video tool BigBlueButton (15.2%) more often than teachers in German language arts (23.4% and 8.3%, respectively). This shows that distance teaching was implemented differently in different subjects.

For our first research question, we then investigated which teaching methods teachers applied during the school closures. Teachers most frequently reported having conducted video meetings (mathematics: 66.3%; German language arts: 59.3%; English language: 66.9%) and meeting with single students or in small groups (mathematics: 48.9%; German language arts: 46.9.2%; English language: 43.7%) (Table 5). We again found differences between subjects: For instance, 52.8% of mathematics teachers used learning videos they themselves had created, compared to 10.3% of German language arts teachers and 26.8% of English teachers. Groupwork was applied by 5.6% of mathematics teachers, 15.2% of German language arts teachers 15.2% and 23.9% of English teachers, again reflecting the different implementations of distance teaching in different subjects.

Associations Between Teaching Methods and Students' and Parents' Perceptions of Teaching Quality

For our second research question, we investigated the associations between the applied teaching methods and students' and parents' perceptions of teaching quality along six quality dimensions. For this, we conducted partial correlations controlling for the context variables school type, class size, grade level, and school subject. Overall, the results revealed low to moderate associations between the applied methods and perceived teaching quality, which,

Table 4

Frequency of Digital Tool Use in Mathematics, German Language Arts, and English Language

	Mathematics		German Language Arts		English Language	
	N	%	*N*	%	*N*	%
Email	154	86.5	135	93.1	125	99.0
Telephone	71	39.9	65	44.8	61	43.0
YouTube	67	37.6	34	23.4	47	33.1
MS Teams	59	33.1	49	33.8	50	35.2
Moodle	57	32.0	48	33.1	47	33.1
Messengers	51	28.7	36	24.8	26	18.3
School-internal platform	47	26.4	41	28.3	45	31.7
BigBlueButton	27	15.2	12	8.3	11	7.7
Cloud-based platform	20	11.2	14	9.7	12	8.5
Skype	6	3.4	3	2.1	5	3.5
WebEx	4	2.2	1	0.7	3	2.1
Jitsi	4	2.2	5	3.4	5	3.5
Zoom	3	1.7	1	0.7	1	0.7
Wikis	2	1.1	2	1.4	0	0
Blogs	1	0.6	2	1.4	3	2.1
Instagram	0	0	0	0	0	0
Twitter	0	0	0	0	0	0
ILIAS	0	0	0	0	0	0
Facebook	0	0	0	0	0	0

Table 5

Frequency of Teaching Method Use in Mathematics, German Language Arts, and English Language

	Mathematics		German Language Arts		English Language	
	N	%	*N*	%	*N*	%
Video meetings	118	66.3	86	59.3	95	66.9
Teacher-generated learning videos	94	52.8	15	10.3	38	26.8
Groupwork	10	5.6	22	15.2	34	23.9
Meetings with students	87	48.9	68	46.9	62	43.7
Online student presentations	5	2.8	6	4.1	12	8.5
Third-party learning videos	12	6.7	6	4.1	3	2.1

however, were consistent across students' and parents' ratings (Table 6 and Table 7). Furthermore, the results showed that video meetings und virtual meetings with single students or in small groups, which seek to foster social connectedness with and between the students, are important for how students and parents perceive support by the teachers (learning support: $.16 \leq r \leq .18$; feedback: $.15 \leq r \leq .18$). There were no or few statistically significant associations with the teaching methods of groupwork, online student presentations, and third-party learning videos for both students' and parents' perceptions of teaching quality. The strongest and most consistent findings were revealed for teacher-created learning videos: Whereas the use of third-party learning videos was not associated with the dimensions of teaching quality, the use of videos created by the teacher him-/herself was linked to higher ratings on all examined teaching quality dimensions ($.11 \leq r \leq .24$).

Associations Between Teaching Methods and Student Outcomes

Our third research question addressed the extent to which the applied teaching methods were associated with student outcomes in terms of students' perceived competence, academic effort, enjoyment of learning, and social involvement. We found that teachers' reported use of video meetings and meetings with single students or in small groups were significantly linked to higher student outcomes ratings ($.10 \leq r \leq .39$) (Table 8). We did not find any statistically significant associations for the teaching methods of groupwork, online presentations by students, and third-party learning videos. Again, teacher-created learning videos were associated most consistently with student outcomes (students' perceived competence, $r = .15$; academic effort, $r = .14$; enjoyment of learning, $r = .13$).

Table 6

Correlations of Teaching Methods and Teaching Quality Dimensions in Mathematics, German Language Arts, and English Language from Students' Perspective

	Monitoring	Structuredness	Learning Support	Feedback	Challenging Tasks	Practicing
	r	*r*	*r*	*r*	*r*	*r*
Video meetings	.12	-.01	**.18**	**.18**	-.01	**.15**
Teacher-generated learning videos	**.11**	**.17**	**.18**	**.12**	**.12**	**.15**
Groupwork	.09	**.09**	.07	.06	.04	.04
Meetings with students	.13	.09	**.17**	.12	.07	.07
Online student presentations	-.06	.01	-.01	-.02	.01	-.01
Third-party learning videos	-.05	.01	-.05	**-.11**	-.03	-.03

Table 7

Correlations of Teaching Methods and Teaching Quality Dimensions from Parents' Perspective

	Monitoring	Structuredness	Learning Support	Feedback	Challenging Tasks	Practicing
	r	*r*	*r*	*r*	*r*	*r*
Video meetings	.10	-.03	**.18**	**.15**	-.01	-.00
Teacher-generated learning videos	**.16**	**.21**	**.24**	**.19**	**.18**	**.20**
Groupwork	.06	.08	.09	.07	.05	.04
Meetings with students	**.11**	.09	**.16**	.10	.01	.03
Online student presentations	-.04	-.08	-.05	-.02	-.07	-.01
Third-party learning videos	-.05	.03	-.05	-.05	.01	-.02

Table 8

Correlations of Teaching Methods and Student Outcomes from Students' Perspective

	Perceived Competence	Academic Effort	Enjoyment of Learning	Social Involvement
	r	*r*	*r*	*r*
Video meetings	**.13**	.00	.08	**.39**
Teacher-generated learning videos	**.15**	**.14**	**.13**	.02
Groupwork	.07	.06	.08	.16
Meetings with students	**.12**	.03	**.10**	**.12**
Online student presentations	-.06	-.09	-.10	-.05
Third-party learning videos	-.05	-.04	-.03	-.04

Associations Between Teaching Quality Dimensions and Student Outcomes

In our fourth research question, we examined the associations between students' and their parents' perceptions of teaching quality and student outcomes (Table 9). We found that all teaching quality dimensions from both the students' and parents' perspective were statistically significant associated with higher student outcomes ratings ($p \leq .05$). The statistically significant associations with teaching quality ranged from $r = .17$ to $r = .74$ from the students' perspective and from $r = .13$ to $r = .56$ from the parents' perspective. Overall, teaching quality dimensions from the students' perspective exhibited higher associations with the student outcomes than teaching quality dimensions from the parents' perspective (e.g., for the association students' perceived competence and practicing, $r = .74$ and $r = .47$, respectively).

Table 9

Correlations of Teaching Quality Dimensions from Students' Perspective and Student Outcomes

	Perceived Competence		Academic Effort		Enjoyment of Learning		Social Involvement	
	r		*r*		*r*		*r*	
	Students	Parents	Students	Parents	Students	Parents	Students	Parents
Monitoring	**.51**	**.41**	**.42**	**.35**	**.44**	**.27**	**.41**	**.33**
Structuredness	**.56**	**.42**	**.40**	**.32**	**.42**	**.34**	**.17**	**.13**
Learning Support	**.68**	**.56**	**.43**	**.38**	**.57**	**.48**	**.41**	**.43**
Feedback	**.59**	**.41**	**.41**	**.27**	**.48**	**.34**	**.46**	**.43**
Challenging Tasks	**.58**	**.35**	**.41**	**.34**	**.34**	**.31**	**.28**	**.22**
Practicing	**.74**	**.47**	**.45**	**.41**	**.55**	**.35**	**.38**	**.25**

Discussion

The school closures due to the COVID-19 pandemic were a unique action taken worldwide to slow down the spread of the virus and ease the burden on health systems. Never before had schools been required required to completely shift their teaching concepts and to deal with the organizational and legal consequences that went along with these changes. Before the pandemic-induced school closures, distance teaching and learning was implemented only in very exceptional cases in most countries (e.g., a long student illness). Therefore, administrations, schools, and households were not prepared or equipped for the switch to a distance education setting. In the present study, we were interested in how teaching and learning took place during this unique period in three different subjects (mathematics, German language arts, and English language) and how these different implementations of teaching at a distance were linked to students' learning. Our analyses built on a large dataset of ratings encompassing the perspectives of teachers, students, and their parents on teaching during the period of school closures. We found that the tools and methods used in digital teaching tremendously differed across teachers as well as across subjects. Furthermore, we found that teaching methods enabling social connectedness between teachers and their students as well as among students themselves were consistently associated with more favorable teaching quality reports by students and their parents as well as with better student outcomes. The role of social connectedness was particularly manifest in the use of teacher-generated learning videos, for which we found positive associations with all dimensions of teaching quality from students' and their parents' perspective and with nearly all student outcomes. In contrast, we found no or even significant negative associations for third-party learning videos (e.g., from YouTube). Finally, we found that the three basic domains of teaching quality framework is suitable for evaluating students' learning in distance education.

Implementation of Distance Education During the School Closures

Whereas distance education in the higher education setting is more common and can draw upon a broad, systematic research base, distance education at the secondary level had rarely been conducted before in many countries (e.g., German Federal Paliament, 2009; Watson, 2015). Consequently, the school closures in March 2020 provided an opportunity to learn more about distance education at the secondary school level. Our findings showed that teachers in all subjects largely used tools with which they were more familiar and that were widely available, such as email and telephone. In terms of teaching methods, we found that

video meetings or meetings with individually students were used equally often in all subjects, whereas we found larger differences between subjects for teaching methods such as teacher-generated learning videos or groupwork. One explanation for this could be that, as in face-to-face teaching, certain methods might be perceived by teachers as more or less appropriate for different subjects and topics in distance education (Ufer et al., 2015). For instance, if the English language curriculum asks students to read and analyze a certain novel, the use of a learning video might be less suitable for helping students comprehend written English texts. Instead, it would be better for students read the novel and complete related assignments over a longer time span. Conversely, learning videos seem to be better suited for mathematics, as certain rules and procedures need to be explained more concretely and explicitly. Finally, groupwork was perceived as more appropriate for verbal subjects, such as having students discuss a chapter of the novel they had read, and thus might be seen as a way to additionally support students' communication skills through listening and speaking activities within a group discussion. That is, tools and teaching methods differ in the opportunities they provide for distance teaching and learning and how they help accomplish a particular teaching objective (Richards & Rogers, 2014; Ufer et al., 2015).

Teaching Quality and Social Connectedness in Distance Education

The present study also investigated whether and to what extent different teaching methods in distance education were associated with well-known quality dimensions (Pauli et al., 2003; Seidel & Shavelson, 2007) and student learning outcomes in terms of perceived competence, academic effort, enjoyment of learning, and social involvement. Contrary to previous findings for face-to-face teaching (Hattie, 2009; Kunter et al., 2011), a large number of methods were found to be relevant for for the teaching quality dimensions and students' experienced learning. Most importantly, our results showed that methods fostering social connectedness between the teacher and students as well as among students themselves were most consistently linked to student learning outcomes. This might be because that social connectedness in face-to-face teaching arises from the classroom-based teaching and learning setting (Hirschy & Wilson, 2002), whereas in distance education it needs to be actively structured and provided by the teacher (Hawkins et al., 2011; Thurmond & Wambach, 2004). For this reason, organizational elements that are typically not associated with students' learning in face-to-face teaching become more important in distance learning. At the same time, these results suggest that teaching methods in distance education can serve other functions than the same methods in face-to-face teaching. For instance, groupwork in distance

education not only means that students work with each other; rather, students also have the possibility to meet and chat with each other, thereby creating a social group context which is otherwise rather challenging in distance learning. Therefore, teaching methods in distance education depend on the available and used tools and can also serve other functions, such as social interaction (Thurmond & Wambach, 2004).

In this vein, the results for teacher-generated learning videos need to be particularly emphasized, as they serve as an impressive example ofg the relevance of social connectedness. In fact, teacher-generated learning videos showed the most consistent associations with the teaching quality dimensions and student outcomes, whereas third-party learning videos showed no or even negative associations. This is surprising, as one could argue that a large number of very professional and high-quality learning videos are available (e.g., Ranga, 2017; Richtberg & Girwidz, 2017). However, the fact that the teacher put a lot of effort into creating such videos and the resulting personal connection students might feel to their teacher carries more weight than a high-quality video by an unfamiliar person. These findings are in line with previous research on learning videos in higher education: For instance, Diwanji and colleagues (2014) investigated components of effective learning videos in massive open online courses (MOOCs) in a meta-analysis. They found that videos in which the professor was visible were more engaging, perceived as less monotonous, and provided a more personal touch.

Finally, we found consistently strong associations between the dimensions of teaching quality in both students' and parents' reports and student learning outcomes. Therefore, the present findings provide very strong empirical support that the well-established dimensions of teaching quality can also serve as a framework describing good teaching in distance education (Cho & Shen, 2013; Kumi-Yeboah et al., 2018; NSQ, 2019). Furthermore, these pronounced associations with students' learning outcomes further suggest that teaching quality in terms of monitoring, feedback to students, and varying, challenging tasks may be even more important for the quality of distance teaching than for face-to-face teaching, as the classroom context is no longer available to provide students with orientation and structure (Hirschy & Wilson, 2002).

Limitations and Further Research

The present study revealed important aspects of successful distance education at the secondary level and highlighted the importance of connection and interaction in distance learning during pandemic-induced school closures. To our knowledge, this is the first study to

investigate distance education during the unique period of COVID-19-related school closures from students', parents', and teachers' perspectives. Nevertheless, there are some limitations that need to be considered.

Despite many advantages of online surveys (e.g., low costs and time efficiency), this method of data collection also has some disadvantages. Our sample might be biased towards schools with greater interest in evaluating their teaching practices, in which teachers may practice at a higher quality level and be more motivated than in other schools. Furthermore, this study might have drawn people who are more tech-savvy and use technology more frequently, as such persons would find it easier to participate in such an online study. In particular, parents and students who do not possess digital devices or are not willing or able to complete the required technical steps may have been excluded from participating in the survey (Huebener et al., 2020). Such parents and students could have provided valuable insights into their learning situation and would also benefit from successfully implemented distance learning. Future research should consider how to reach these less accessible schools and families.

We assessed data in mathematics, German language arts, and English as a foreign language. Instruction in these three subjects was mandatory in the German federal state of Baden-Württemberg during the school closures, and the present study's data has provided valuable insights into the different ways distance education was implemented. As we found large differences between subjects for some variables, such as the use of learning videos, it is possible that examining other school subjects would have revealed even more subject specificities. Therefore, it would be interesting to investigate how teaching in other subjects took place during the pandemic-induced school closures as well.

Finally, we assessed multiple learning outcomes which have been shown to be associated with students' achievement (Kunter et al., 2013; Lam et al., 2015; Wagner et al., 2016). Nevertheless, students' achievement on standardized achievement tests would have been a valuable component to explain further differences in students' perceptions of distance learning. In future research, it would be interesting to link students' perceived quality of distance learning with their achievement.

References

Authors, 2018

Authors, 2020

Barbour, M. K., & Reeves, T. C. (2009). The reality of virtual schools: A review of the literature. *Computers and Education, 52*(2), 402-416. https://doi.org/10.1016/j.compedu.2008.09.009

Baturay, M.H. & Yukselturk, E. (2015). The Role of Online Education Preferences on Student's Achievement. *Turkish Online Journal of Distance Education, 16*(3), 3-12. https://doi.org/10.17718/tojde.47810

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B. 57*, 289–300.

Borup, J., West, R. E., Graham, C. R., & Davies, R. S. (2014). The adolescent community of engagement framework: A lens for research on K–12 online learning. *Journal of Technology and Teacher Education, 22*(1), 107–129.

Brophy, J. (2000). Teaching. *Educational Practices Series, 1.* International Academy of Education (IAE).

Burch, P., Good, A., & Heinrich, C., (2016). Improving access to, quality, and the effectiveness of digital tutoring in K-12 education. *Educational Evaluation and Policy Analysis 38*(1), 65-87. https://doi.org/10.3102/0162373715592706

Cavanaugh, C., Barbour, M. K., & Clark, T. (2009). Research and practice in K–12 online learning: A review of literature. *International Review of Research and Open and Distance Learning, 10*(1). https://doi.org/10.19173/irrodl.v10i1.607

Cho, M.-H., & Shen, D. (2013). Self-regulation in online learning. *Distance Education, 34*(3), 290–301. https://doi.org/10.1080/01587919.2013.835770.

Diwanji, P., Simon B.P., Märki, M., Korkut, S., Dornberger, R. (2014). Success factors of online learning videos. In Interactive Mobile Communication Technologies and Learning (IMCL), *2014 International Conference on IEE*, 125-132.

Doyle, W. (2013). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management* (pp. 107–136). Routledge.

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. http://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff, J., Lüdtke, O.,… Trautwein, U. (2019). Don't blame the teacher? The need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology*. Advanced online publication. http://doi.org/10.1037/edu0000416

Garbe, A., Ogurlu, U., Logan, N., & Cook, P. (2020). Parents' Experiences with Remote Education during COVID-19 School Closures. *American Journal of Qualitative Research, 4*(3), 45-65. https://doi.org/10.29333/ajqr/8471

German Federal Parliament (2009). *Homeschooling in westlichen Industrieländern* [Homeschooling in Western industrialized nations]. https://www.bundestag.de/resource/blob/415424/dbc64afb565391f883ebe737ba44475f/wd-8-047-09-pdf-data.pdf

Graham, C., Cagiltay, K., Lim, B., Craner, J., & Duffy, T. M. (2001). Seven principles of effective teaching: A practical lens for evaluating online courses. *The Technology Source, 30*(5), 50.

Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. Meece & J. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 25– 41). New York: Routledge.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hawkins, A., Graham, C. R., Sudweeks, R. R. & Barbour, M. K. (2013). Academic performance, course completion rates, and student perceptions of the quality and frequency of interaction in a virtual high school. *Distance Education, 34*, 1, 64–83. http://doi.org/10.1080/01587919.2013.770430

Hillmayr, D., Reinhold, F., Ziernwald, L., & Reiss, K. (2017). *Digitale Medien im mathematisch-naturwissenschaftlichen Unterricht der Sekundarstufe: Einsatzmöglichkeiten, Umsetzung und Wirksamkeit [Teaching mathematics and science with digital media in secondary schools: usability, implementation, and effectiveness]*. Waxmann.

Hirschy, A. S. & Wilson, M. E. (2002). The Sociology of the Classroom and its influence on student learning, *Peabody Journal of Education, 77*(3), 85-100. http://doi.org/10.1207/S15327930PJE7703_5

Huebener, M., Spieß, K., & Zinn, S. (2020). *SchülerInnen in Corona-Zeiten: Teils deutliche Unterschiede im Zugang zu Lernmaterial nach Schultypen und -trägern [Students in Corona times: Partly significant differences in access to learning materials according to school types and authorities]*. German Institute for Economic Research. https://www.diw.de/sixcms/detail.php?id=diw_01.c.804559.de

Jonkmann, K., Rose, N., & Trautwein, U. (Eds.). (2013). *Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen - Abschlussbericht für die Länder Baden-Württemberg und Sachsen [Tradition and Innovation: Developmental Trajectories at Hauptschulen and Realschulen in Baden-Württemberg and Mittelschulen in Saxony - Final Report for the States of Baden-Württemberg and Saxony*]. Project report to the ministries of education and cultural affairs of the federal states.

Kumi-Yeboah, A., Dogbey, J., & Yuan, G. (2017). Exploring factors that promote online learning experiences and academic self-concept of minority high school students. *Journal of Research on Technology in Education, 50*(1), 1-17. https://doi.org/10.1080/15391523.2017.1365669

Kunter, M., Frenzel, A., Nagy, G., Baumert, J., & Pekrun, R. (2011). Teacher enthusiasm: Dimensionality and context specificity. *Contemporary Educational Psychology, 36*, 289–301. https://doi.org/10.1016/j.cedpsych.2011.07.001

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*, 494–509. http://doi.org/10.1016/j.learninstruc.2007.09.002

Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV: A multicriteria analysis. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers – results from the COACTIV project* (pp. 97–124). Springer. http://doi.org/10.1007/978-1-4614-5149-5_6

Kuo, Y., Belland, B. R., Schroder, K. E. E., & Walker, A. E. (2014). K-12 teachers' perceptions of and their satisfaction with interaction type in blended learning

environments. *Distance Education, 35*(3), 360-381. http://doi.org/10.1080/01587919.2015.955265

Lam, A. C., Ruzek, E. A., Schenke, K., Conley, A. M., & Karabenick, S. A. (2015). Student perceptions of classroom achievement goal structure: Is it appropriate to aggregate? *Journal of Educational Psychology, 107*(4), 1102–1115. https://doi.org/10.1037/edu0000028

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. http://doi.org/10.1016/j.learninstruc.2008.11.001

Liu, F., & Cavanaugh, C. (2012). Factors influencing student academic performance in online high school algebra. *Open Learning, 27*(2), 149–167. http://doi.org/10.1080/02680513.2012.678613

Moore, M. G., & Kearsley, G. (2011). *Distance education: A systems view of online learning. (3rd ed.)*. Wadsworth Cengage Learning.

National Standards for Quality (2019). *Quality online teaching*. https://www.nsqol.org/the-standards/quality-online-teaching/

Pauli, C. & Reusser, K. (2003). Unterrichtsskripts im schweizerischen und im deutschen Mathematikunterricht. *Unterrichtswissenschaft, 31*(3), 238-272.

Picciano, A. G., Seaman, J., Allen, I. E. (2010). Educational transformation through online learning: To be or not to be. *Journal of Asynchronous Learning Networks, 14*(4), 17-35. http://doi.org/10.24059/olj.v14i4.147

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM, 50*, 407–426. http://doi.org/10.1007/s11858-018-0918-4

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D.,… Schiefele, U. (Eds.). (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente* [PISA 2003: Documentation of the instruments]. Waxmann.

Ranga, J. S. (2017). Customized videos on a YouTube Channel: A beyond the classroom teaching and learning platform for general chemistry courses. *Journal of Chemistry Education, 94*, 867-872. https://doi.org/10.1021/acs.jchemed.6b00774

Richards, J & Rogers, Th. S. (2014). *Approaches and Methods in Language Teaching* (Third edition). Cambridge University Press.

Richtberg, S., & Girwidz, R. (2019). Learning physics with interactive videos – possibilities, perception, and challenges. *Journal of Physics: Conference Series, 1287*(1). IOP Publishing. https://doi.org/10.1088/1742-6596/1287/1/012057

Robinson, J., Dusenberry, L., Hutter, L, Lawrence, L., Frazee, A. and Burnett, R. (2019). State of the Field: Teaching with Digital Tools in Writing and Communication. *Computers and Composition (54),* 1-19. https://doi.org/10.1016/j.compcom.2019.102511

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77,* 454-499. http://doi.org/10.3102/003465430731031

Simonson, M., & Schlosser, L., (2009*). Distance education. Definitions and Glossary of terms* (3rd eds.). Information Age Publishing.

Simonson, M. (2019). Research in distance education: A summary. *Quarterly Review of Distance Education, 20*(3), 31-43.

Thurmond, V., & Wambach, K. (2004, January). Understanding interactions in distance education: A review of literature. *International Journal of Instructional Technology & Distance Learning*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.9189&rep=rep1&type=pdf#page=17

Ufer, S., Heinze, A., & Lipowsky, F. (2015). *Unterrichtsmethoden und Instruktionsstrategien* [Teaching methods and stretegies of instruction]. Springer Link.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*, 1–11. http://doi.org/10.1016/j.learninstruc.2013.03.003

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, 108*, 705–721. http://doi.org/10.1037/edu0000075

Watson, J., Pape, L., Murin, A., Gemin, B., & Vashaw, L. (2015). *Keeping pace with K–12 digital learning: An annual review of policy and practice*. Evergreen Education Group.

Supplemental Material

Table S1

Correlations of Teaching Methods and Background Variables

	Video meetings	Teacher-generated learning videos	Group work	Meetings with single students	Online student presentations	Third-party learning videos
	r	*r*	*r*	*r*	*r*	
School type	-.11	.05	**.20**	**-.21**	.02	-.00
Class size	-.03	.05	.04	-.02	.00	-.03
Grade level	.03	**-.20**	.08	-.02	-.01	-.05
German language arts	-.07	**-.43**	**.13**	-.02	.03	-.06
English language	.00	**-.25**	**.25**	-.05	**.12**	**-.10**

Table S2

Correlations of Teaching-Quality Dimensions and Background Variables

	School Type	Class Size	Grade Level	German Language Arts	English Language
	r	*r*	*r*	*r*	*r*
Monitoring	-.02	.09	.04	**.16**	**.23**
Structuredness	.06	.03	**-.18**	-.05	-.01
Learning Support	**-.18**	.01	-.04	-.05	-.03
Feedback	**-.22**	-.02	.02	**.12**	**.21**
Challenging Tasks	.06	-.01	**-.22**	**-.33**	**-.15**
Practicing	-.01	.08	**-.15**	**-.25**	.01

Table S3

Correlations of Student Outcomes and Background Variables

	School Type	Class Size	Grade Level	German Language Arts	English Language
	r	*r*	*r*	*r*	*r*
Perceived Competence	.01	.06	**-.17**	**-.12**	-.02
Academic Effort	-.03	**.14**	**-.27**	-.08	-.05
Enjoyment of Learning	.01	.02	**-.32**	.01	**.13**
Social Involvement	-.03	-.09	.06	-.02	.04

Table S4

Correlations of Digital Tools and Teaching Quality Dimensions in Mathematics, German Language Arts, and English Language from Students' Perspective

	Monitoring	Structuredness	Learning Support	Feedback	Challenging Tasks	Practicing
	r	*r*	*r*	*r*	*r*	*r*
Email	-.01	-.01	.02	.01	.02	-.01
Telephone	.06	.04	**.14**	**.13**	.01	.03
Messenger	.07	.07	**.20**	**.14**	**.11**	**.12**
MS Teams	**.14**	-.04	**.27**	**.24**	-.03	.07
Cloud-based platform	**-.13**	.04	.02	.00	.04	-.02
Moodle	**-.17**	-.07	**-.21**	**-.13**	-.04	-.07
School-internal platform	**.10**	**.13**	-.02	-.08	.07	.03
YouTube	-.08	**.10**	.02	-.04	**.12**	**.11**
BigBlueButton	**-.13**	.01	-.08	-.09	.05	.04
Video meeting tool	.09	.00	**.24**	**.18**	.02	**.12**

Table S5

Correlations of Digital Tools and Teaching Quality Dimensions in Mathematics, German Language Arts, and English Language from Parents' Perspective

	Monitoring	Structuredness	Learning Support	Feedback	Challenging Tasks	Practicing
	r	*r*	*r*	*r*	*r*	*r*
Email	-.02	.02	.02	.03	.04	.02
Telephone	.06	.03	**.17**	**.11**	.03	.04
Messenger	.07	.03	**.18**	**.11**	-.03	.03
MS Teams	**.13**	-.03	**.19**	**.19**	**-.11**	-.04
Cloud-based platform	**-.15**	.08	.02	-.07	.08	.06
Moodle	**-.11**	-.04	**-.10**	-.09	**.11**	.03
School-internal platform	.08	**.11**	.00	-.06	.03	.04
YouTube	-.04	**.10**	.04	.01	.08	**.11**
BigBlueButton	-.06	.04	-.01	-.02	**.14**	.08
Video meeting tool	.08	-.02	**.18**	**.16**	-.03	.00

Table S6

Correlations of Digital Tools and Student Outcomes in Mathematics, German Language Arts, and English Language from Students' Perspective

	Perceived Competence	Academic Effort	Enjoyment of Learning	Class Community
	r	*r*	*r*	*r*
Email	-.04	-.09	-.06	-.02
Telephone	.07	**.10**	**.12**	.06
Messenger	**.14**	.02	.07	**.21**
MS Teams	.07	.09	.09	**.16**
Cloud-based platform	-.02	.05	-.02	**-.16**
Moodle	-.08	**-.19**	-.08	**.15**
School-internal platform	.02	.07	.00	**-.24**
YouTube	.08	.00	.02	.01
BigBlueButton	.00	-.04	-.03	**.11**
Video meeting tool	**.10**	.05	.08	**.23**

4

STUDIE 2: "THE TEACHER MOTIVATES US – OR ME?" – THE ROLE OF THE ADDRESSEE IN STUDENT RATINGS OF TEACHER SUPPORT

Jaekel, A., Wagner, W., Trautwein, U., & Richard, G. (2021). How students' perceptions of teaching quality in one subject are impacted by the grades they receive in another subject: dimensional comparisons in student evaluations of teaching quality. *Manuskript eingereicht zur Publikation.*

Abstract

Student ratings have become a standard way to assess teaching quality. However, little is known about whether using a we/us-addressee (“The teacher motivates us”) or an I/me-addressee (“The teacher motivates me”) makes a difference for the information that is obtained. In this study, we experimentally varied the addressee in teaching-quality items capturing six dimensions of teacher support in two school subjects. We investigated differences between the two addressee versions in mean levels, level of agreement, associations between dimensions using the same versus different addressees, and correlations with a variety of student outcome variables. We found that the item addressee was relevant for most psychometric properties in question, and differences were more pronounced for mathematics than for German language arts.

Keywords: student ratings of teacher support, item addressee, level of agreement, associations within subjects

Introduction

Student ratings have become a standard way to assess teaching quality in research on students' learning environments and in school practice (Authors, 2016; Fraser, 2012; Greenwald, 1997). They are easy to implement in daily school life, they enable the collection of data from many different raters within one classroom, and students experience many different teaching situations that they can compare. Another interesting characteristic of student ratings is that they yield information at both the student and classroom levels, allowing researchers to differentiate between class average ratings of teachers and individual deviations from this average level (Lam et al., 2015; Lüdtke et al., 2006).

In general, student ratings have been shown to reliably and validly assess teaching quality in the school setting (De Jong & Westerhof, 2001; Spooren et al., 2013). However, recent research has also highlighted the need to better understand the potential limitations of student ratings, including modest agreement among students (Mainhard et al., 2018), high intercorrelations among theoretically distinct teaching quality dimensions at the class level (Authors, 2013), and student characteristics that impact teaching quality ratings (Authors, 2018; Benton & Cashin, 2012; Griffin, 2004). One aspect that might be relevant to each of these issues in teaching quality assessment is the *addressee* of the measures used. More specifically, does it make a difference whether an item reads "The teacher gives me regular feedback on what *I* can already do" versus "Our teacher gives *us* regular feedback on what *we* can already do"? And are the items "The teacher gives *me* additional support when *I* need help" and "Our teacher gives *us* additional support when *we* need help" really equivalent?

A quick look at prominent teaching quality measures shows that both the "I/me" and "we/us" formulations are frequently used (e.g., Den Brok et al., 2006; Fauth et al., 2014; Kuhlfeld, 2017; OECD, 2017). In fact, both formulations can be found in survey instruments that assess teaching quality from a student perspective; however, their use is typically confounded with the respective teaching quality dimension a researcher intends to measure. For instance, dimensions of teacher support are frequently assessed with items that have an I-addressee, whereas dimensions of classroom management tend to be worded with a we-addressee (e.g., Downer et al., 2015; Kuhfeld, 2017). Therefore, there is no clear answer to whether and to what extent differences in addressees make a difference for student ratings of teaching quality. In addition, the addressee might also play a role in the extent to which information can be obtained at different levels. For example, whereas the I-addressee might theoretically be more important for providing information about differences between students

within a classroom, the we-addressee might be more appropriate for gaining insights into differences between classrooms. In this context, the question of whether teachers' behavior addresses the individual student or all students in the classroom can be assumed to matter for the assessment of teaching quality from the students' perspective.

In the present study, we examined whether the choice of addressee matters in frequently used teaching quality measures in the school setting for the information that is obtained. To this end, we focused on dimensions of teacher support, a crucial domain within teaching quality research that refers to the support of students' learning development by the teacher and includes aspects such as motivating students to behave in certain ways and providing adequate feedback. Consequently, differences between the I-addressee and the we-addressee might be specifically relevant here. We experimentally varied the addressee in assessments of multiple dimensions of teacher support in a large-scale study and investigated differences in the means of the teaching quality dimensions using either a we- or an I-addressee, students' level of agreement between the two addressee versions, as well as differences in the associations among teacher support dimensions when assessed using the same addressee version and different addressee versions in two subjects: mathematics and German language arts. Finally, we investigated the associations with crucial outcome variables, such as academic self-concept and grades.

Student Ratings of Teaching Quality

Teaching quality is widely understood to encompass empirically observable aspects of teachers' behavior which are associated with students' learning and development (Doyle, 2013; Pianta & Hamre, 2009). One prominent stream of research describes teaching quality as composed of three overarching quality domains, namely, classroom management, teacher support, and cognitive activation (Hamre & Pianta, 2010, 2018; Klieme et al., 2009; Praetorius et al., 2018). Whereas classroom management and cognitive activation focus on the classroom environment (e.g., disturbances in the classroom) and cognitive load in class, respectively, the domain of teacher support aims to capture students' emotional and social experience and the extent to which they feel fostered regarding their achievement-related development (Hamre & Pianta, 2018). The teacher support domain has consistently been found to be relevant for student outcomes such as students' motivation and interest (e.g., Hattie, 2009; Kuhfeld, 2017; Rakoczy, 2008).

Alongside ratings given by external observers and teachers' self-reports, student ratings are frequently used to assess teaching quality. Student ratings have a number of

advantages (Authors, 2016; Authors, 2018; Hattie, 2009; Kunter et al., 2013; Mainhard et al., 2018): Students can base their ratings on a wide variety of teaching and learning situations because they are taught by many different teachers in different subjects during their time in school. Student ratings allow scholars and practitioners to assess a wide variety of aspects of teaching quality and have been shown to be predictive of several student outcomes (e.g., Authors, 2018; Aldrup et al., 2018; Hamre & Pianta, 2010). Teacher support is particularly suitable for assessment via student ratings, as the individually perceived support is difficult to observe. One specific benefit of student ratings is that they provide information about teaching quality at the classroom level on the basis of students' shared perceptions, which are assumed to be experienced similarly by all students at the class level (e.g., Michelle and Yasmine agree that the mathematics teacher is able to motivate students very well) as well as students' nonshared perceptions of teaching quality at the student level (e.g., Michelle feels much more motivated by the mathematics teacher than Yasmine does; Lüdtke et al., 2006).

Challenges in Assessing Teaching Quality With Student Ratings

Despite the benefits of student ratings and their importance for students' learning outcomes, scholars have also expressed doubts about the validity of student ratings, as important questions have not yet been fully answered (e.g., Abrami et al., 2007; Kunter, 2007; Marsh et al., 2012). These limiting factors are assumed to be reflected in the psychometric properties of student ratings. Three persistent issues that require further research concern agreement among students within classrooms, the associations between theoretically distinct teaching quality factors, and the appropriate consideration of the information obtained at different levels of analysis.

First, with respect to the agreement in students' perceptions of the same construct, student ratings are often aggregated at the classroom level (Lüdtke et al., 2006). However, many studies have found only low to moderate ICCs for teaching quality dimensions, meaning that only a small proportion of the variance that is shared between students can be explained by the fact that students belong to the same classroom (Authors, 2013; Schenke et al., 2017; Schweig, 2014). This is even true for relatively objective quality aspects, such as the extent of disruptions in the classroom (e.g., Authors, 2013; Lüdtke et al., 2006).

Second, studies have shown that student ratings of theoretically different quality aspects are much more strongly associated than one might expect from theory or other data sources (e.g., teacher self-reports or observational data; e.g., Authors, 2013). Such strong associations might be explained by a general impression that affects students' ratings on many

different items. For example, the halo effect is a form of systematic rater bias in which students form a general impression on the basis of a single teacher characteristic (Benton & Cashin, 2012). In this case, a student's ratings on all teaching quality dimensions might be influenced by his or her rather positive general impression of the teacher due to the teacher's gender or popularity (Bennett, 1982; Fauth et al., 2018). Consequently, the student's ratings on different teaching quality dimensions would exhibit higher intercorrelations due to the student's overall more positive responding.

Third, typically, students' individual ratings are aggregated at the classroom level to obtain information about the shared perceptions of students within a given classroom and to explain differences between classrooms (Aldrup et al., 2018; Marsh et al., 2012). By contrast, much less attention has been paid to differences in students' perceptions of teaching quality analyzed within one classroom. This is surprising given that effects are often more pronounced at the student level than at the class level and that rating differences between students within the same classroom have been found to provide meaningful information about students' individual learning. For instance, Aldrup et al. (2018) found that perceived learning support at the student level predicted student outcomes, such as satisfaction, self-esteem, truancy, and achievement. Mainhard et al. (2018) found that the teacher-student relationship strongly predicted students' emotions of enjoyment and anxiety at the student level. Authors (2018) found that students' perceived support by the teacher predicted students' achievement in and enjoyment of mathematics at the student level.

However, in all of these studies, the item wording, or more precisely, the item addressee, as one essential property of teaching quality measures, was not considered systematically. Items were sometimes worded with a we-addressee and sometimes with an I-addressee, or even mixed within a single instrument. At the same time, the relevance of the item addressee has not been investigated systematically for the dimensions of teacher support, making it unclear whether the identified associations were due to the addressee used in the items or due to the individual nature of students' perceived support and its associations with students' learning outcomes.

The Addressee in Items Assessing Teaching Quality

The addressee of items within teaching quality measures often varies between measurement instruments (Authors, 2016; Schenke et al., 2017), between dimensions (Den Brok, 2006; Praetorius, 2018), and in some cases, even within a single dimension (Wallace et al., 2016). Theoretically, items worded with a we-addressee ("We feel motivated by the

teacher") versus an I-addressee ("I feel motivated by the teacher") encourage students to refer to different experiences. This assumption is closely linked to the response process students go through when responding to an item (Authors, 2008; Sudman et al., 1996; Tourangeau et al., 2000). For instance, the item "We feel motivated by the teacher" refers to the perceptions of all students in the classroom, namely, whether all students think the teacher has the ability to motivate them. In this case, all students refer to similar sources of information when making their judgments. That is, students try to think of relevant situations that affect all the students in the classroom, add them up, and then—without talking to their classmates about the topic to obtain first-hand information—each student draws conclusions about the extent to which all students in the classroom feel motivated by the teacher. By contrast, the source of information for the item "I feel motivated by the teacher" is more unique to each student, as it reflects the individually experienced ability of the teacher to motivate the respective student. Thus, when responding to the item, students theoretically draw on only their individual experiences with the teacher.

These potential differences in the relevant cognitive processes that take place when students respond to these two items might be reflected in the psychometric properties of students' teaching quality ratings, such as the mean values of the two versions with different addressees. Given that items with a we-addressee refer to experiences that affect all students in the classroom, even very rare, specific occurrences between a teacher and an individual student might be perceived by all the students in the classroom. These occurrences might accumulate in students' perceptions and thus lead to higher mean values for items with a we-addressee. On the other hand, one might expect that relevant situations that affect all students in a classroom are less observable for each student than situations affecting only one individual student, such as one-on-one talks between the teacher and a student. Therefore, students' judgments can be assumed to be subject to some degree of uncertainty, as students have to draw different inferences about teachers' behavior. This could potentially lead to biases in student ratings. For instance, the concept of acquiescence describes a respondent's general tendency to agree with items regardless of their content (Krosnick, 1991; Weijters et al., 2013). This bias is more pronounced when more information must be retrieved from memory and information need to be compared, which makes responding to the items more cognitively challenging (Krosnick, 1991). Furthermore, this bias also applies when items are rather imprecise and vague and has been found to lead to more positive ratings (Cabooter et al., 2010; Moors, 2008). These higher ratings might lead to higher mean values for items using a we-addressee because the relevant information requires many comparisons and is not

as precise as items using an I-addressee. Conversely, responding to an item with an I-addressee can be assumed to require a much higher degree of self-evaluation by respondents (Dunning et al., 1989), and such higher level of self-evaluation might be associated with more positive self-perceptions. In personality research, this phenomenon is called self-enhancement, which means that individuals see themselves more favorably than others do, hold more optimistic views about their own future, and have a more positive view of their own achievement than of the achievements of others (e.g., Kim & Di Domenico, 2019; Paulhus, 2002). Given such views, items worded with a I-addressee can be expected to be more likely to trigger this self-enhancement effect, leading to higher mean values for items with an I-addressee than for items with a we-addressee.

In addition, the two addressee versions might also differ in terms of the level of agreement among students within classes. If all students in a classroom refer to the same source of information when responding to items using a we-addressee, the result should be higher agreement in student ratings and therefore higher intraclass correlations (ICCs). ICCs are a commonly used indicator of agreement among students within a classroom and also reflect the reliability of student ratings aggregated at the class level (Lüdtke et al., 2006). By contrast, one might expect differences between students' individual ratings to be more pronounced for items with an I-addressee, resulting in lower agreement among students and therefore lower ICCs.

In fact, from previous studies, there are some indications that the choice of addressee does in fact contribute to the information obtained from student ratings of teaching quality. Three studies have suggested that class versions of questionnaires lead to more positive mean teaching quality ratings compared with personal versions: In an experimental study, Fraser et al. (1995) found that student ratings of science classes were more positive when a class version was used (e.g., "In our laboratory sessions, different students do different experiments") than when a personal version was used (e.g., "In my laboratory sessions, I do different experiments than some of the other students"). McRobbie et al. (1998) likewise found more positive values in most cases for a class version compared with a personal version. In another experimental study, Den Brok et al. (2006) found higher mean values for teachers' influence (e.g., teachers' leadership and strictness) in a personal version compared with a class version. Likewise, for the dimension of teachers' proximity, which includes aspects that are likely related to teacher support (e.g., helpfulness and understanding), mean values were higher for the personal version than the class version, but the differences were not statistically significant. However, both of the dimensions that were considered, teachers'

influence and proximity, also included scales that were not explicitly part of teacher support (e.g., teachers' uncertainty or strictness).

Likewise, findings have suggested the importance of the item addressee for agreement among student ratings within classrooms. In previous nonexperimental studies, dimensions of teacher support, which tended to be worded with an I-addressee (e.g., "My teacher helps me solve problems myself"), exhibited lower agreement within classes and lower associations with student outcomes at the classroom level than teacher support items that used a we-addressee (e.g., Aldrup et al., 2018; Fauth et al., 2014; Kunter et al., 2013). Furthermore, studies have found lower associations between dimensions of teacher support and student outcomes compared with items or scales for classroom management or cognitive activation (Downer et al., 2015; Schweig, 2014). However, it is not clear whether this was the result of unique characteristics of teacher support or because an I-addressee is typically used in items from the support dimension, whereas a we-addressee tends to be used for other dimensions. Moreover, these studies were not designed to systematically address the impact of different item addressees. Therefore, such results rely on correlational data and run the risk of confounding the effects of teacher support dimensions and item formulations due to variation in the addressee. Den Brok et al. (2006) investigated differences between a class version and a personal version of a questionnaire and found that the ICCs from the class version were slightly higher than the ICCs from the personal version.

In sum, the few existing studies that have used varying addressees have provided important insights into potential differences in teaching quality measures, indicating that the addressee is a relevant item property. Although these studies were conducted with large samples and systematically distinguished between teachers' behavior toward individual students versus the entire class, further research is still needed on some points. For instance, teacher support was assessed with a very limited number of scales or referred only to science class. Furthermore, because these studies used two versions of the questionnaire containing items with either a we-addressee or an I-addressee, it was not possible to investigate the associations between teaching quality dimensions as measured with both addressee versions (e.g., feedback with a we-addressee and an I-addressee), which would provide information about how the two versions affected students' ratings. Moreover, it was not possible to investigate differences between the two versions at the student level because students received only one of the two versions. Furthermore, these studies provided no information about whether and to what extent the use of different addressee versions is linked to differences in the associations between teacher support and students' learning at each level of analysis. Most

importantly for the present study, in most existing studies, different addressee versions were confounded with the quality dimensions in question. Therefore, an investigation of a larger number of quality dimensions while systematically varying the addressee is still needed.

The Present Investigation

In this study, we sought to extend existing research on the relevance of the item addressee in teaching quality measures. To this end, we examined differences in the mean levels of two addressee versions (i.e., I and we) as well as the degree of agreement between students' ratings of the teacher support dimensions and investigated the associations between student ratings using different addressees for one and the same support dimension in two subjects: mathematics and German language arts. Additionally, we investigated potential differences in the associations between teacher support dimensions assessed using both versions with prominent student outcome variables in both subjects. For this, we experimentally varied the addressee in items that were used to assess students' perceived support by the teacher. We derived several hypotheses that we present and explain next.

1) Only a few experimental studies have examined whether mean differences arise from the use of different addressee versions. The findings from these studies have not been fully consistent, but they have revealed a tendency in the direction of more positive mean values for dimensions referring to the whole classroom (we-addressee) compared with individual students (I-addressee). We aimed to systematically investigate possible differences in a large number of dimensions of teacher support in two different subjects: mathematics and German language arts. We hypothesized that using a we-addressee would result in higher mean values for all dimensions than using an I-addressee (Hypothesis 1).
2) On the basis of theoretical assumptions about students' response processes, one can assume that students refer to shared experiences when responding to items using a we-addressee but to their own individual experiences when responding to items using an I-addressee. Among students in a single classroom, this should lead to higher agreement when a we-addressee is used and lower agreement when an I-addressee is used. Therefore, we expected to find higher intraclass correlations for dimensions of teacher support for items using a we-addressee than for items using an I-addressee in both subjects (Hypothesis 2).
3) Assuming that responding to dimensions with a we-addressee is more cognitively challenging because students have to consider their classmates' experiences and

integrate these considerations into their ratings, students' ratings of teaching quality may be more similar across dimensions when a we-addressee is used than when an I-addressee is used. For this reason, we hypothesized that using a we-addressee would result in higher intercorrelations across dimensions of teacher support than using an I-addressee would (Hypothesis 3).

4) Given that items with different addressees are assumed to capture different perspectives on teaching quality, we expected to find positive but not perfect correlations between measures of the same dimension using the two addressee versions (e.g., motivation using a we-addressee and motivation using an I-addressee; Hypothesis 4).
5) Finally, we were interested in the associations between teacher support dimensions and subject-specific student outcome variables in mathematics and German language arts (academic self-concept, engagement, grades, and achievement test scores). In line with our overall assumption that a we-addressee better captures students' shared perceptions at the classroom level, whereas an I-addressee better captures individual students' perceptions at the within level, we expected to find differences in the associations between teacher support dimensions and students' outcome variables depending on the addressee. Specifically, we hypothesized that we would find more pronounced associations at the student level when using an I-addressee and more pronounced associations at the classroom level when using a we-addressee (Hypothesis 5).

Method

The Ministry of Education and Cultural Affairs of the German federal state of Baden-Württemberg approved the study and the data collection (date of approval: February 12, 2018, file number: 31-6600.0/279). The ethics committee of [Institution; blinded for review] confirmed that the procedures were in line with ethical standards for research with human subjects (date of approval: May 4, 2018, file number: A2.5.4-074_aa).

Sample

The data for this study stemmed from the large-scale study "Teaching Quality from the Students' Perspective (UNITAS)," which was conducted in spring/summer 2018 in the federal state of Baden-Württemberg, Germany. The UNITAS study examines teaching quality on the secondary level in two different subjects, mathematics and German language arts, with

a special focus on student ratings. A total of 6,479 students in Grades 5 to 10 from 401 classes in 27 schools participated in the study. On average, 16 students per class participated in the study. A total of 50.8% of the students were female, and 87% of the students reported that the language they usually spoke at home was German. In order to examine student ratings of teaching quality in mathematics and German language arts as distinctly as possible, 16 classes were excluded from the analysis because both subjects were taught by the same teacher. The final sample for the present study comprised 6,317 students from 385 classes. The lower secondary education system in Baden-Württemberg consists of four main school types: academic track (Gymnasium), intermediate track (Realschule), lower track (Hauptschule), and multitrack (Gemeinschaftsschule) schools. Students from 11 academic-track schools (n = 3,847), eight intermediate-track schools (n = 1,795), seven multitrack schools (n = 652), and one lower track secondary school (n = 23) provided ratings on their mathematics and German language arts classes.

Design

One essential limitation of many previous studies (for exceptions, see Den Brok et al., 2006; Fraser et al., 1995; McRobbie et al., 1998) is that student ratings can be confounded by a correlation between the addressee that is used and the dimension of teacher support that is assessed. In our study, we used experimental variation to avoid these effects. Prior to data assessment, students were randomly assigned to one of four conditions on the individual level in a between-participants design. Specifically, the addressee was experimentally varied within classrooms by using four different versions of a survey, allowing us to estimate the associations between different addressee versions in the same dimension. The first version exclusively comprised items using a we-addressee (n = 1,584). The second version exclusively comprised items using an I-addressee (n = 1,568). The third (n = 1,603) and fourth versions (n = 1,562) contained items with both addressees but alternated which specific items used which addressee. For instance, in Version 3, the first and third items used a we-addressee, and the second and fourth items used an I-addressee. In Version 4, the first and third items used an I-addressee, and the second and fourth items used a we-addressee. On average, each version was completed by roughly four students per class. Because the students responded to the individual items in only one of the addressee versions, Versions 3 and 4 with alternating addressees were used to deal efficiently with missing data. Furthermore, these two versions enabled us to investigate differences not only at the classroom level but also at the

individual student level. In other words, the experimental variation allows us to model each quality dimension as assessed with different item addressees.

Instruments

Teaching Quality Dimensions

Students were asked to rate teaching quality with regard to six dimensions of teacher support, namely, handling mistakes, motivation, learning support, clarity, autonomy support, and feedback. Each dimension was assessed with three or four items (see Tables 1 and S1 for detailed descriptions). Example items are "The teacher can sometimes really motivate me"/ "Our teacher can sometimes really enthuse us" for the motivation dimension and "The teacher encourages me to work autonomously"/ "Our teacher encourages us to work autonomously" for the autonomy support dimension. Students responded to a total of 20 parallel items for mathematics and German language arts. All items were rated on a 4-point rating scale ranging from 1 (*completely disagree*) to 4 (*completely agree*). Internal consistencies showed good values in both subjects. For mathematics, Cronbach's alpha ranged from .70 to .83 (handling mistakes: $.70 \leq \alpha \leq .75$; motivation: $.79 \leq \alpha \leq .82$; learning support: $.72 \leq \alpha \leq .76$; clarity: $.80 \leq \alpha \leq .83$; autonomy support: $.70 \leq \alpha \leq .73$; feedback: $.76 \leq \alpha \leq .81$). For German language arts, Cronbach's alpha ranged from .70 to .79 (handling mistakes: $.71 \leq \alpha \leq .73$; motivation: $.79 \leq \alpha \leq .79$; learning support: $.71 \leq \alpha \leq .75$; clarity: $.70 \leq \alpha \leq .73$; autonomy support: $.71 \leq \alpha \leq .73$; feedback: $.75 \leq \alpha \leq .79$).

Student Outcome Variables. ***Student Grades.*** Student achievement in mathematics and German language arts was assessed with student-reported school grades from their latest report card, which they had received after the first half of the academic year. As school grades in Germany range from 1 (*best grade*) to 6 (*lowest grade*), we reverse-scored the grades for easier interpretation so that higher grades represented higher achievement.

Standardized Achievement. Additionally, we administered achievement tests to assess students' achievement in German language arts (LGVT 5-12+; Schneider et al., 2017) and mathematics (MBK 5-12+; Ennemoser et al.,, 2011; Krajewski & Ennemoser, 2013). The LGVT 5-12+ is used to assess the speed, accuracy, and comprehension with which students are able to process a written text. The MBK 5-12+ consists of several timed tasks, such as increasingly more difficult calculation tasks or dictation of large numbers. In the present study, both achievement tests revealed satisfactory internal consistencies (LGVT 5-12+: KR-20 = .98; MBK 5-12+: KR-20 = .83).

Academic Self-Concept. Students' academic self-concept in each subject was assessed with four items each (e.g., "I just have no talent for mathematics"; Gaspard et al., 2016). All items were rated on a 4-point rating scale. Cronbach's alpha was .93 for mathematics and .88 for German language arts. The means were 2.82 for mathematics and 2.86 for German language Arts. The ICCs were .05 for mathematics and .07 for German language arts.

Engagement. Students' affective engagement was measured with five items on a 4-point rating scale (e.g., "How often does it occur in German lessons that you enjoy the tasks in class?"; adapted from Fredricks et al., 2005). Internal consistencies were $\alpha = .77$ for mathematics and $\alpha = .75$ for German language arts. The means were 3.22 for mathematics and 3.18 for German language Arts. The ICCs were .12 for mathematics and .13 for German language arts.

Analysis

Multilevel Factor Analyses

In the present study, we used a factor analysis procedure to examine whether the item addressee makes a difference for the assessment of teaching quality. As we were interested in differences in student ratings of teacher support within classrooms and differences in students' shared perceptions between classrooms, we conducted a multilevel confirmatory factor analysis, modeling latent factors for each support dimension and each addressee version on two levels. The first level was that of individual students within classrooms, also known as the student level or the within level; the second level was that of classes or the between level. Multilevel confirmatory factor analysis applies a conventional factor analysis structure to each of the levels simultaneously rather than a single "overall" model as in a single-level analysis. The items are assumed to be indicators of factors that vary across students within classes as well as between classes, that is, item (co-)variances are assumed to result from factors and residual variances at both levels (Muthén, 1994).

Measurement Invariance

Before addressing our research questions, we tested a series of measurement invariance models. Establishing measurement invariance was a crucial prerequisite for investigating mean-level differences between the addressee versions (Hypothesis 1), students' level of agreement in terms of intraclass correlations (Hypothesis 2), factor correlations at each level of analysis (Hypotheses 3 and 4), and the associations between the teaching quality factors and students' outcome variables at each level of analysis (Hypothesis 5). To test for measurement invariance, we imposed equality constraints, meaning that factor loadings had to

be equal across levels and addressees, and compared the model fit results with those of freely estimated models. In order to examine mean-level differences between the two addressee versions, we additionally constrained the item intercepts, allowing us to compare latent factor means. To ensure that our tests were sufficiently sensitive to detecting potential violations of measurement invariance, all measurement models were created separately for each of the quality dimensions. Within these models, cross-version constraints were consecutively imposed for the we-addressee and the I-addressee, and cross-level constraints were consecutively imposed at the student and classroom levels. Finally, we constructed combined models with model constraints across addressee versions at the two analytical levels and model constraints across analytical levels for the two addressee versions. Model fit was evaluated using the χ^2-difference test, comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR) following the guidelines developed for conventional single-level structural equation models in which changes in the CFI $\leq .01$, RMSEA $\geq .02$, and SRMR $\geq .01$ indicate no substantial change in model fit (Chen, 2007). The results of the invariance tests are presented in detail in Section S2 of the online supplemental material. We assumed measurement invariance across both analytical levels and addressee versions. Both the models assuming cross-version invariance at the two analytical levels and the models assuming cross-level invariance for the two addressee versions exhibited only relatively small differences from the baseline models, allowing us to address the questions of mean-level differences, students' agreement, the intercorrelations of teaching quality factors, and the associations between the teaching quality factors and students' outcome variables at each level of analysis.

Missing Data

The issue of missing values required careful consideration in the present study. As the four addressee versions were experimentally varied within classrooms and students were not given the same items for the I- and we-addressee versions, missing data appeared at the student level, and correlations for the same item with different addressees could not be computed at the student level. Model identification at the student level was achieved by setting the limit for the coverage of a single item pair differing only in the addressee to 0%, which allowed us to estimate the amount of missing data in the total sample. The correlations between factors assessed with different addressees were computed using data from the two experimental groups with varying addressees (Versions 3 and 4), assuming that the findings would be equivalent for groups receiving only the we-addressee or only the I-addressee. In order to make full use of the available data, we additionally modeled correlations between

factors with different addressees and the two analytical levels for each subject. All additional correlations between item indicators within subjects were constrained to be zero. The full information maximum likelihood (FIML) procedure was used to handle both planned and unplanned missing data (Arbuckle, 1996).

For all of the analysis, we used the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) to control the false discovery rate for multiple comparisons.

Results

Mean-Level Differences in Teacher Support Dimensions Using an I-Addressee Versus a We-Addressee

To test our first hypothesis, we compared the means of the teacher support dimensions assessed with different item addressees separately for the two subjects. Overall, students had positive perceptions of teacher support in both subjects ($2.58 \leq M \leq 3.26$; Table 1). Across subjects and addressee versions, the means were descriptively highest for motivation and lowest for feedback (mathematics: we-addressee: $2.62 \leq M \leq 3.26$; I-addressee: $2.65 \leq M \leq 3.24$; German language arts: we-addressee: $2.58 \leq M \leq 3.23$; I-addressee: $2.59 \leq M \leq 3.21$), with no substantial differences found between the means in mathematics and German language arts. In line with our hypothesis, we found that the means of most dimensions were descriptively higher when a we-addressee was used than when an I-addressee was used (mathematics: $0.02 \leq \Delta M \leq 0.12$; German language arts $0.02 \leq \Delta M \leq 0.11$). In each subject, we computed an overall mean for the six dimensions of teacher support using a we-addressee, and we computed an overall mean for the for the six dimensions of teacher support using an I-addressee. Then we computed the difference between the two overall means in each subject. When testing the overall mean difference for significance, we found that, in both subjects, the mean for dimensions using a we-addressee was significantly higher (mathematics, we-addressee: $M = 2.91$, I-addressee: $M = 2.87$; $p < .01$; German language arts, we-addressee: $M = 2.91$, I-addressee $M = 2.87$; $p < .01$). Therefore, our first hypothesis was confirmed.

Agreement on Teacher Support Dimensions Using an I-Addressee Versus a We-Addressee

In order to investigate whether the use of a we-addressee would result in higher agreement among students within classes, we investigated differences in within-class agreement on the teaching support dimensions in both addressee versions (Table 2; for the single items see Table S3). The ICCs in mathematics ranged from .24 to .39 (we-addressee)

Table 1

Descriptive Statistics for the Teacher Support Dimensions

Dimension	Sample item with we- and I-addressee	Mathematics		German language arts	
		M	*SD*	*M*	*SD*
Handling mistakes	Our teacher is patient when someone makes a mistake in class./	3.26	0.33	3.23	0.38
	The teacher is patient when I make a mistake in class.	3.24	0.37	3.21	0.37
Motivation	Our teacher can sometimes really motivate us./	2.77	0.46	2.78	0.44
	The teacher can sometimes really motivate me.	2.71	0.43	2.70	0.44
Learning support	Our teacher gives us additional support when we need help./	2.97	0.35	2.92	0.34
	The teacher gives me additional support when I need help.	2.88	0.36	2.84	0.36
Clarity	Our teacher teaches so comprehensibly that we understand even difficult things./	2.83	0.48	2.95	0.34
	The teacher teaches so comprehensibly that I understand even difficult things.	2.81	0.45	2.97	0.30
Autonomy support	Our teacher encourages us to work autonomously./	3.02	0.27	3.00	0.25
	The teacher encourages me to work autonomously.	2.90	0.29	2.89	0.25
Feedback	Our teacher gives us regular feedback on what we can already do./	2.62	0.28	2.58	0.22
	The teacher gives me regular feedback on what I can already do.	2.65	0.27	2.60	0.21

Note. *M* = mean; *SD* = standard deviation.

Table 2

Intraclass Correlations for the Teacher Support Dimensions in Mathematics and German Language Arts

	Student level				Classroom level				ICC			
	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ_{MW}	σ_{MI}	σ_{GW}	σ_{GI}
Handling mistakes	0.36	0.32	0.31	0.29	0.14	0.11	0.14	0.13	0.28	0.26	0.31	0.31
Motivation	0.35	0.40	0.33	0.40	0.22	0.18	0.20	0.20	0.39	0.31	0.38	0.33
Learning support	0.29	0.31	0.26	0.27	0.12	0.13	0.12	0.13	0.29	0.30	0.32	0.33
Clarity	0.43	0.48	0.28	0.24	0.23	0.21	0.12	0.09	0.35	0.30	0.30	0.27
Autonomy support	0.23	0.28	0.21	0.21	0.08	0.09	0.06	0.07	0.26	0.24	0.22	0.25
Feedback	0.26	0.28	0.25	0.26	0.08	0.07	0.05	0.05	0.24	0.20	0.17	0.16

Note. MW = Mathematics, we-addressee; MI = Mathematics, I-addressee; GW = German language arts, we-addressee; GI = German language arts, I-addressee; ICC = Intraclass correlation.

and from .20 to .31 (I-addressee). In German language arts, the ICCs ranged from .17 to .38 (we-addressee) and from .16 to .33 (I-addressee). We found that across all dimensions, the ICCs in mathematics were descriptively higher for a we-addressee (we-addressee: M_{ICC} = .30; I-addressee: M_{ICC} = .27). In German language arts, we found no difference in the ICCs between the two addressee versions (we-addressee: M_{ICC} = .28; I-addressee: M_{ICC} = .28). Testing these overall mean differences for significance, we found that for mathematics, the teacher support dimensions using a we-addressee resulted in significantly higher ICCs (p < .05). Therefore, our hypothesis that we would find higher ICCs for dimensions using a we-addressee could be partly confirmed.

Associations Between Teacher Support Dimensions Using the Same Addressee Version

To test Hypothesis 3, we compared the intercorrelations among the teacher support dimensions using the same addressee version (e.g., between motivation and feedback using an I-addressee in mathematics and between motivation and feedback using a we-addressee in mathematics, respectively) at the student level and at the classroom level, separately for the two subjects. For mathematics, the intercorrelation of the dimensions when using an I-addressee ranged from .40 to .84 at the student level and from .51 to .95 at the classroom level (Table 3). When using a we-addressee, the intercorrelations of the dimensions ranged from .42 to .85 (student level) and from .61 to .97 (classroom level).

Table 3

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in Mathematics

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.73/.68	.83/.84	.75/.73	.75/.74	.51/.61
2. Motivation	.64/.58		.88/.89	.92/.91	.93/.95	.81/.85
3. Learning support	.81/.80	.79/.79		.88/.92	.95/.97	.86/.91
4. Clarity	.64/.64	.76/.81	.72/.81		.89/.92	.73/.77
5. Autonomy support	.60/.63	.84/.85	.79/.84	.76/.78		.85/.89
6. Feedback	.40/.42	.61/.67	.62/.67	.50/.60	.71/.79	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant (p < .001).

For German language arts, the intercorrelations of the dimensions when an I-addressee was used ranged from .36 to .86 at the student level and from .51 to .96 at the classroom level (Table 4). When a we-addressee was used, the intercorrelations ranged from .39 to .86 at the student level and from .51 to .95 at the classroom level (Table 4). Overall, in mathematics, the intercorrelations combined for all dimensions were descriptively higher when a we-addressee was used than when an I-addressee was used at the student level (I-addressee: $M_r = .67$; we-addressee: $M_r = .71$) and at the classroom level (I-addressee: $M_r = .82$; we-addressee: $M_r = .84$). We also found this pattern for German language arts but only at the student level (I-addressee: $M_r = .67$; we-addressee: $M_r = .71$). Testing these overall differences for statistical significance, we found that for both subjects, using an I-addressee resulted in higher correlations between teacher support dimensions at the student level ($p < .01$), whereas the difference at the classroom level in mathematics did not reach statistical significance. For this reason, our hypothesis could only partly be confirmed.

Table 4

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in German Language Arts

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.71/.76	.86/.85	.86/.85	.80/.76	.51/.51
2. Motivation	.56/.59		.90/.89	.92/.95	.95/.94	.83/.75
3. Learning support	.77/.76	.75/.80		.89/.93	.96/.92	.85/.80
4. Clarity	.68/.69	.77/.80	.83/.84		.92/.91	.72/.70
5. Autonomy support	.58/.55	.86/.86	.79/.85	.83/.80		.87/.84
6. Feedback	.36/.39	.53/.66	.59/.66	.50/.58	.67/.77	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$).

Associations Between the Same Teacher Support Dimension Using Different Addressee Versions

To test Hypothesis 4, we examined the correlations between the same teacher support dimensions measured with the different addressee versions (e.g., association between clarity using a we-addressee and clarity using an I-addressee). We found very high associations for all dimensions, with no differences between the two subjects (Table 5). At the student level, the correlations ranged from .97 to 1.00 for mathematics and reached 1.00 for all dimensions

in German language arts. At the classroom level, the correlations ranged from .93 to 1.00 for mathematics and from .94 to .99 for German language arts, without any statically significant difference between the revealed correlation coefficients. Therefore, our hypothesis was not confirmed.

Table 5

Associations of Teacher Support Dimensions Between Addressee Versions

	Mathematics		German language arts	
	r (SL/CL)	SE (SL/CL)	*r* (SL/CL)	SE (SL/CL)
Handling mistakes	**1.00/.99**	0.04/0.02	**1.00/.98**	0.04/0.03
Motivation	**1.00/.99**	0.02/0.01	**1.00/.99**	0.02/0.01
Learning Support	**1.00/.99**	0.03/0.02	**1.00/.98**	0.03/0.02
Clarity	**1.00/1.00**	0.03/0.01	**1.00/.97**	0.03/0.02
Autonomy support	**.97/.93**	0.04/0.03	**1.00/.94**	0.04/0.04
Feedback	**1.00/.98**	0.02/0.02	**1.00/.96**	0.02/0.04

Note. SL = student level; CL = classroom level. Statistically significant results ($p < .001$) after the correction are presented in bold.

Associations With Student Outcome Variables

Finally, we examined possible differences in the associations between the teacher support dimensions and subject-specific student outcome variables, separately for the two addressee versions and the two subjects (Hypothesis 5). The results are shown in Table 6 (mathematics) and Table 7 (German language arts). Overall, we found statistically significant associations between the teacher support dimensions and the subject-specific student outcomes self-concept, engagement, grades, and test scores at the student level (mathematics: $.05 \leq r \leq .69$; German language arts: $.06 \leq r \leq .69$) and the classroom level (mathematics: $.13 \leq r \leq .94$; German language arts: $.21 \leq r \leq .94$).

Regarding the differences in the associations for the different addressee versions, in most cases, we found descriptively stronger associations in mathematics at the student level when an I-addressee was used than when a we-addressee was used (self-concept: I-addressee: $M_r = .30$, we-addressee: $M_r = .25$; engagement: I-addressee: $M_r = .52$, we-addressee: $M_r = .48$; grade: I-addressee: $M_r = .22$, we-addressee: $M_r = .18$; test score: I-addressee: $M_r = .11$, we-addressee: $M_r = .09$). At the classroom level, the opposite pattern was revealed: In most cases, we found higher associations when an we-addressee was used than when a I-addressee was

Table 6

Correlations Between Students' Outcome Variables and Teacher Support Variables in Mathematics

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.19/.19**	**.56/.53**	**.34/.29**	**.68/.73**	**.20/.21**	**.63/.67**	**.52/.38**	**.72/.73**	**.39/.22**	**.67/.65**	**.18/.19**	**.59/.61**
Engagement	**.39/.33**	**.71/.70**	**.69/.63**	**.93/.94**	**.47/.45**	**.86/.89**	**.61/.56**	**.88/.88**	**.59/.49**	**.90/.92**	**.38/.40**	**.77/.82**
Grades	**.19/.15**	**.42/.41**	**.22/.21**	**.45/.53**	**.14/.15**	**.38/.45**	**.42/.28**	**.44/.49**	**.27/.17**	**.43/.44**	**.07/.09**	**.33/.39**
Test Scores	**.09/.07**	.09/.10	**.10/.06**	.04/.03	.03/**.07**	**.13**/.09	**.25/.16**	.05/.09	**.16/.09**	.03/.08	**.05/.06**	.01/-.06

Note. SL = student level; CL = classroom level; SES = Socioeconomic status. Statistically significant results ($p < .05$) after the correction are presented in bold.

Table 7

Correlations Between Students' Outcome Variables and Teacher Support Dimensions in German Language Arts

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.22/.22**	**.44/.47**	**.35/.32**	**.50/.48**	**.26/.25**	**.43/.44**	**.45/.37**	**.55/.51**	**.38/.29**	**.43/.38**	**.14/.19**	**.33/.28**
Engagement	**.33/.35**	**.65/.69**	**.69/.63**	**.94/.92**	**.48/.45**	**.75/.74**	**.55/.52**	**.87/.88**	**.63/.52**	**.85/.83**	**.35/.37**	**.72/.74**
Grades	**.18/.19**	**.30/.24**	**.18/.14**	**.26**/**.26**	**.14/.14**	**.21**/.17	**.30/.21**	**.29/.24**	**.20/.16**	**.21**/**.26**	.03/**.08**	.11/.10
Test Scores	**.08/.11**	.09/.06	.00/.02	.01/.07	.02/.04	.00/.03	**.15/.09**	.08/.06	**.11/.06**	-.04/.12	.01/.01	-.07/.01

Note. SL = student level; CL = classroom level; SES = Socioeconomic status. Statistically significant results ($p < .05$) after the correction are presented in bold.

used (self-concept: I-addressee: $M_r = .64$; we-addressee: $M_r = .65$; engagement: I-addressee: $M_r = .84$, we-addressee: $M_r = .86$; grade: I-addressee: $M_r = .41$, we-addressee: $M_r = .45$). In German language arts, we found higher associations when using the I-addressee at the student level for academic self-concept (I-addressee: $M_r = .30$; we-addressee: $M_r = .27$), engagement (I-addressee: $M_r = .51$; we-addressee: $M_r = .47$), and grade (I-addressee: $M_r = .17$; we-addressee: $M_r = .15$). At the classroom level, we found a more varying result pattern: academic self-concept (I-addressee: $M_r = .45$; we-addressee: $M_r = .43$), engagement (I-addressee: $M_r = .80$; we-addressee: $M_r = .80$), grade (I-addressee: $M_r = .23$; we-addressee: $M_r = .21$), test score (I-addressee: $M_r = .01$; we-addressee: $M_r = .06$).

When we tested the differences for statistical significance in mathematics, we found that the associations between each of the student outcome variables (i.e., academic self-concept, engagement, grade, and test scores) and the teacher support dimensions revealed statistically significantly higher values for I-addressee items than for we-addressee items at the student level (*p*s' < .01). At the classroom level, we found no statistically significant differences between the two addressee versions. For German language arts at the student level, we found that the associations between academic self-concept and engagement and the teacher support dimensions revealed statistically significantly higher values when an I-addressee was used than when a we-addressee was used ($p < .05$). We found no statistically significant differences between the two addressee versions at the classroom level. On the basis of these results, our hypothesis could be partly confirmed.

Robustness Analyses for Students' Grades and School Track

Finally, we conducted additional analyses to investigate possible differences with respect to students' grade level and school tracks for both subjects. For the grade levels, we combined two grade levels for each group (Grades 5 and 6; Grades 7 and 8; Grades 9 and 10), which resulted in three different groups for the analysis. With regard to the school tracks, we examined differences between the two groups of academic-track students and non-academic-track students. We then conducted all of the analyses separately. The results are shown in Tables S4 to S38. In general, we could not identify notable deviations from the overall findings. The resulting differences were quite small, and the difference tests were not statistically significant. Neither school type nor grade level impacted the findings.

Discussion

Student ratings are a valuable source of information about students' perceptions of their learning environment. Although student ratings are frequently used in research and educational practice, the role of the item addressee, which often differs between and even within measurement instruments, has rarely been investigated so far. The aim of the present study was to systematically examine differences in the information obtained from students when using two different addressee versions (*we*-addressee vs. *I*-addressee) of frequently used dimensions of teacher support. Data from a large-scale study employing an experimental variation enabled us to investigate the impact of using a we-addressee versus an I-addressee in items from six dimensions of teacher support in two subjects: mathematics and German language arts.

The results supported the assumption that the item addressee shapes the information we get from students about teachers' support. Overall, our findings showed that, in most cases, teacher support dimensions using a we-addressee revealed higher mean values, ICCs, and intercorrelations at the classroom level. Dimensions using an I-addressee usually resulted in higher associations with students' outcome variables at the student level. This pattern was more pronounced for mathematics than for German language arts.

Student Ratings as a Primary Source for Assessing Teacher Support

The domain of teacher support aims to assess the emotional and social facet of students' learning environment (Kunter & Voss, 2013). In contrast to the domains of classroom management and cognitive activation, students' individually experienced support is much more difficult to rate for external observers compared with, for example, how well the rules are followed or which exercises are used in class. Therefore, student ratings offer a valuable method for obtaining information about students' individual experiences of support from their teacher. In this study, we assessed teacher support on a large number of different dimensions. Overall, our results confirmed that student ratings are a reliable method for assessing teaching quality. In line with previous research, we found that students were able to distinguish the different dimensions of teacher support in both subjects (e.g., Authors, 2018; Authors, 2020; Downer et al., 2015; Schweig, 2014). Additionally, our findings support existing research as we found substantial differences between classrooms and associations of dimensions of teacher support with important student outcomes at the student level and at the classroom level (e.g., Authors, 2013; Kuhfeld, 2017; Schenke et al., 2017).

The Role of the Item Addressee

Even at first glance, measures of teaching quality substantially differ not only with regard to the assessed dimensions, but also with respect to the item addressee. Only a few studies have investigated the role of the item addressee; however, these studies spoke in favor of its relevance (Den Brok et al., 2006; Fraser, 1995; McRobbie et al., 1998). Our study extended this research by experimentally varying the addressee in a large number of teacher support dimensions in two subjects, thus putting the role of the item addressee to an extensive test.

With regard to differences in the overall mean levels (i.e., the mean of all the teacher support dimensions) between the two kinds of addressees, we found statistically significantly higher mean values in both subjects when a we-addressee was used. These results are in line with existing research in which higher mean values were found in most cases for the class version (we-addressee) compared with the personal version (I-addressee). These previous results were likewise rather small (Fraser, 1995; McRobbie et al., 1998). Based on our theoretical assumptions, higher mean values for using a we-addressee could be explained by students' tendency to give items with a rather vague frame of reference higher ratings than items concerning only themselves (Cabooter et al., 2010; Krosnick, 1991; Moors, 2008). Another explanation could be that a student's experience consists of not only teacher behavior that affects all students in the class in the same way (e.g., "chalk-and-talk" or lecture-style teaching) but also more rare teacher behaviors that are directed toward individual students. For instance, Yasmine might feel that the teacher praises Michelle a lot but that she herself does not receive such praise. This would mean that rare dyadic occurrences between the teacher and individual other students (i.e., not the rater) have a higher likelihood of being reflected in dimensions measured using a we-addressee. This might lead to higher values for the we-addressee version, which would then include even rare occurrences that affect only individual students in the classroom.

With regard to student agreement, we found higher associations when using a we-addressee. This finding is in line with the study by Den Brok et al. (2006), who also found that the ICCs were slightly higher for the class version than for the personal version. We also found more pronounced intercorrelations between dimensions using a we-addressee. These results raise questions about the underlying reasons and whether these findings speak in favor of using a we-addressee than a I-addressee. For instance, existing research has shown that students' perceptions of teaching quality can be impacted by an overall general impression, such as the halo effect or teacher popularity, which could lead to a higher level of agreement

across the students (Benton & Cashin, 2012; Bennett, 1982; Fauth et al., 2018). If such general impressions are more "triggered" by using the we-addressee, the findings might speak for a lower specificity in ratings using a we-addressee. In line with this, an experimental study conducted by Roch et al. (2009) found that different than expected, students tend to agree more about rather imprecise and vague aspects of teaching than about behavioral and observable aspects, which could also apply to dimensions of teacher support. That is, ratings which need more inferences, like items using a we-addressee, might result into a higher agreement, which however might also be at the cost of the specificity of ratings.

Investigating the associations with students' learning outcomes, in many cases, we found statistically significantly higher associations at the student level when using an I-addressee. We sincerely believe that this is an important finding that shows that dimensions using an I-addressee seem to be more appropriate when teachers or researchers want to take students' individual achievement-related variables into account. In previous studies, teaching quality was oftentimes assessed with items that had a we-addressee (e.g., Aldrup et al., 2018; Fauth et al., 2014; Kunter et al., 2013). For instance, Authors (2018) investigated students' idiosyncratic perceptions of teaching quality. The results emphasized the importance of the relationship between a single student and the student's teacher for student ratings of teaching quality, which, however, were assessed with items using a we-addressee. Our findings point to the relevance of the item addressee, and thus, it would be interesting to differentiate between the two addressee versions to investigate the associations between teaching quality and students' learning at the student level and at the classroom level.

Experimentally varying the item addressee in this study provided a unique opportunity to analyze the associations between the same dimension assessed with different addressees because the item addressee is usually not varied within studies. We found almost perfect correlations between the two addressee versions for each dimension (e.g., feedback using a we-addressee and feedback using an I-addressee), which suggests that students rate the two addressee versions of teacher support dimensions almost equivalently. We had expected to find lower associations because teacher support is a teaching quality domain that refers to students' very individual aspects of being supported in their learning by the teacher (Hamre & Pianta, 2018).

Limitations and Further Research

For several reasons, we believe that our study makes a strong contribution to the fundamental question of the role of the addressee in assessments of teacher support

dimensions, which has been given insufficient attention in research on the assessment of teaching quality. However, our study also has some limitations that should be considered in future research.

First, we formulated the two addressee versions in parallel (e.g., “The teacher can sometimes really motivate me”/ “Our teacher can sometimes really motivate us”), and students were instructed regarding both versions before the survey. However, the students responded to a large number of items, and the extent to which a single word (“I” vs. “we”) catches a readers’ eye has yet to be determined. Indeed, given that students typically take only a few seconds to respond to each item, they might not have read as carefully as necessary. Nevertheless, future research could use more extreme variations (e.g., with personal pronouns printed in bold) or investigate students’ eye movements, which would provide additional information about how accurately students read the different versions of the items.

Second, our findings provide evidence that the addressee matters to some extent for student ratings of teacher support. To gain further insights into the higher mean values and stronger associations for dimensions using a we-addressee, interviewing students would be a very promising approach that could be applied to uncover what information students actually refer to when responding to such items. This approach has also been called for in previous research, as interview ratings for teaching quality items have rarely been applied in research (Leighton, 2019; Lenske & Praetorius, 2020).

Finally, in the present study, we investigated the idea that different addressee versions are accompanied by differences in the psychometric properties of students’ teaching quality ratings, including mean levels, level of agreement among students within classes, intercorrelations among teacher support dimensions, and associations with students’ learning outcomes. However, in order to answer the questions of whether and to what extent using different addressee versions is beneficial when assessing teaching quality, there is a need to further evaluate its consequences for other purposes. For instance, more recent analytical procedures are aimed at modeling classroom heterogeneity in student ratings as an additional indicator of good teaching (e.g., Schenke et al, 2018). Applying these modeling procedures to surveys with an I-addressee might be a better way to assess student-teacher fit and teacher adaptivity than surveys with a we-addressee.

Conclusion

We investigated the relevance of the item addressee in teaching quality measures. Our results showed that the addressee that was used in items measuring teacher support made a

difference for the information that could be obtained from student ratings of teachers' support at the level of students' individual perceptions and students' shared perceptions at the classroom level. Even though the differences we found were relatively small, these findings further underline the need to consider the role of item framing when one is interested in assessing teaching quality from the perspective of students in research and practice.

References

Authors, 2008

Authors, 2013

Authors, 2016

Authors, 2018

Authors, 2018

Authors, 2020

Abrami P. C., d'Apollonia S., Rosenfield S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In: Perry R. P., Smart J. C. (Eds), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385–445). Springer. https://doi.org/10.1007/1-4020-5742-3_10

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: issues and techniques* (pp. 243–277). Lawrence Erlbaum Associates, Inc.

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology, 74*(2), 170–179. https://doi.org/10.1037/0022-0663.74.2.170

Benton, S. L., & Cashin, W. E. (2011). Student ratings of teaching: A summary of research and literature. *IDEA paper no. 50. Center for faculty education and development*. IDEA Center, Kansas State University. Retrieved from https://www.semanticscholar.org/paper/Student-Ratings-of-Teaching-%3A-A-Summary-of-Research-Benton-Cashin/f15397d5900e85579bede0dc160888f86b4d5f5b

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cabooter, E. F. K. (2010). *The impact of situational and dispositional variables on response styles with respect to attitude measures*. Doctoral dissertation, Ghent University.

De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*, 51–85. http//doi.org/10.1023/A:1011402608575

Den Brok, P., Brekelmans, M., & Wubbels, T. (2006). Multilevel issues in research using students' perceptions of learning environments: The case of the Questionnaire on Teacher Interaction. *Learning Environments Research, 9*(3), 199–213. https://doi.org/10.1007/s10984-006-9013-9

Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E. (2015). Measuring effective teacher-student interactions from a student perspective. *The Journal of Early Adolescence, 35*(5-6), 722–758. https://doi.org/10.1177/0272431614564059

Doyle, W. (2013). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management* (pp. 107–136). Routledge.

Dunning, D., Meyerowitz, J., Holzberg, A. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*(6), 1082–1090. https://doi.org/10.1037/0022-3514.57.6.1082

Ennemoser, M., Krajewski, K., & Schmidt, S. (2011). Entwicklung und Bedeutung von Mengen-Zahlen-Kompetenzen und eines basalen Konventions- und Regelwissens in den Klassen 5 bis 9 [Development and importance of sets and numbers competencies and of a basic knowledge of conventions and rules in Grades 5 through 9]. *Zeitschrift Für Entwicklungspsychologie und Pädagogische Psychologie, 43*(4), 228–242. https://doi.org/10.1026/0049-8637/a000055

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. http://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2018). Exploring teacher popularity: Associations with teacher characteristics and student outcomes in primary school. *Social Psychology of Education, 21*(5), 1225–1249. https://doi.org/10.1007/s11218-018-9462-x

Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff, J., Lüdtke, O., Polikoff, M. S., Klusmann, U., & Trautwein, U. (2019). Don't blame the teacher? The need to account

for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology*. Advance online publication. https://doi.org/10.1037/edu0000416

Fraser, B. J., Giddings, G. J., & McRobbie, C. J. (1995). Evolution and validation of a personal form of an instrument for assessing science laboratory classroom environments. *Journal of Research in Science Teaching, 32*(4), 399–422. https://doi.org/10.1002/tea.3660320408

Fredericks, K. A., & Durland, M. M. (2005). The historical evolution and basic concepts of social network analysis. *New Directions for Evaluation, 2005*(107), 15–23. https://doi.org/10.1002/ev.158

Gaspard, H., Dicke, A.-L., Flunger, B., Häfner, I., Brisson, B. M., Trautwein, U., & Nagengast, B. (2016). Side effects of motivational interventions? Effects of an intervention in math classrooms on motivation in verbal domains. *AERA Open, 2*(2), 1–14. doi:10.1177/2332858416649168

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182–1186. http://dx.doi.org/10.1037/ 0003-066X.52.11.1182

Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology, 29*(4), 410–425. https://doi.org/10.1016/j.cedpsych.2003.11.001

Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes. In J. L. Meece & J. S. Eccles (Eds.), *Handbook of research on schools, schooling, and human development*. Routledge. https://doi.org/10.4324/9780203874844.ch3

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Kim, H., Di Domenico, S. I., & Connelly, B. S. (2019). Self-other agreement in personality reports: A meta-analytic comparison of self- and informant-report means. *Psychological Science, 30*(1), 129–138. https://doi.org/10.1177/0956797618810000

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.

Krajewski, K., & Ennemoser, M. (2013). Entwicklung und Diagnostik der Zahl-Größen-Verknüpfung zwischen 3 und 8 Jahren [Development and diagnostics of the link between numbers and quantities between ages 3 and 8]. In M. Hasselhorn, A. Heinze, W. Schneider, & U. Trautwein (Eds.), *Diagnostik mathematischer Kompetenzen. Jahrbuch der pädagogisch-psychologischen Diagnostik. Tests und Trends* (pp. 41–65). Hogrefe.

Krosnick J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology 5*(3), 213–236

Kuhfeld, M. (2017). When students rate their teachers: A validity analysis of the Tripod student survey. *Educational Assessment, 22*(4), 253-274. https://doi.org/10.1080/10627197.2017.1381555

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*, 494–509. http://doi.org/10.1016/j.learninstruc.2007.09.002

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820. https://doi.org/10.1037/a0032583

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M.,… Stanat, P. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente [PISA 2000: Documentation of Scales].* Heenemann GmbH & Co.

Lam, A. C., Ruzek, E. A., Schenke, K., Conley, A. M., & Karabenick, S. A. (2015). Student perceptions of classroom achievement goal structure: Is it appropriate to aggregate? *Journal of Educational Psychology, 107*(4), 1102–1115. https://doi.org/10.1037/edu0000028

Leighton, J. (2019). Students' Interpretation of formative assessment feedback: Three claims for why we know so little about something so important. *Journal of Educational Measurement, 56*(4), 793–814. https://doi.org/10.1111/jedm.12237

Lenske, G. & Praetorius, A.-K. (2020). Schülerfeedback- was steckt hinter dem Kreuz auf dem Fragebogen [Student feedback - what is behind the cross on the questionnaire?] *Empirische Pädagogik, 34*(1), 11–29.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9*(3), 215–230. https://doi.org/10.1007/s10984-006-9014-8

Mainhard, T., Oudman, S., Hornstra, L., Bosker, R. J., & Goetz, T. (2018). Student emotions in class: The relative importance of teachers and their interpersonal relations with students. *Learning and Instruction, 53*, 109–119. https://doi.org/10.1016/j.learninstruc.2017.07.011

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106–124. https://doi.org/10.1080/00461520.2012.670488

McRobbie, C. J., Fisher, D. L., & Wong, A. F. L. (1998). Personal and class forms of classroom environment instruments. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 581–594). Kluwer.

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality and Quantity, 42*(6), 779–794. https://doi.org/10.1007/s11135-006-9067-x

Muthén, B. O. (1994). Multilevel Covariance Structure Analysis. *Sociological Methods & Research, 22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

OECD (Organization for Economic Cooperation and Development) (2017). *Student Questionnaire for PISA 2018 - Main Survey Version*. Retrieved from https://www.oecd.org/pisa/data/2018database/CY7_201710_QST_MS_STQ_NoNotes_final.pdf

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic 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. doi:10.3102/0013189X09332374

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM, 50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht: Unterricht aus der Perspektive von Lernenden und Beobachtern [Motivation support in mathematics teaching: Teaching from the perspectives of learners and observers]*. Waxmann.

Schenke, K., Ruzek, E., Lam, A. C., Karabenick, S. A., & Eccles, J. S. (2017). Heterogeneity of student perceptions of the classroom climate: a latent profile approach. *Learning Environments Research, 20*(3), 289–306. https://doi.org/10.1007/s10984-017-9235-z

Schneider, W., Schlagmüller, M., & Ennemoser, M. (2017). *LGVT 5-12+: Lesegeschwindigkeits- und Verständnistest für die Klassen 5–12+ Manual [Reading speed and reading comprehension test for Grades 5-12+ manual]*. Hogrefe.

Schweig, J. (2014). Cross-Level Measurement Invariance in School and Classroom Environment Surveys. *Educational Evaluation and Policy Analysis, 36*(3), 259–280. https://doi.org/10.3102/0162373713509880

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching. *Review of Educational Research, 83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology* (1. ed.). Jossey-Bass. Retrieved from http://www.loc.gov/catdir/bios/wiley043/95016504.html

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal, 53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*(3), 320–334. https://doi.org/10.1037/a0032121

Supplemental Material

Table S1

Items for Measuring the Teacher Support Dimensions

Dimension	I-addressee		We-addressee	
Handling mistakes	I1	The teacher is patient when I make a mistake in class.	W1	Our teacher is patient when someone makes a mistake in class.
	I2	For the teacher, it's not a bad thing if I make a mistake.	W2	For our teacher, it's not a bad thing if we make a mistake.
	I3	The teacher makes sure that I am not laughed at for making a mistake.	W3	Our teacher makes sure that nobody in our class is laughed at for making a mistake.
Motivation	I1	The teacher can sometimes really motivates me.	W1	Our teacher can sometimes really motivates us.
	I2	The teacher can make even dry subject material really interesting to me.	W2	Our teacher can make even dry material really interesting to us.
	I3	The teacher points out to me that the material I am learning is useful in my everyday life.	W3	Our teacher points out to us that the material we are learning is useful in our everyday lives.
	I4	For me, this teacher's lessons are quite varied.	W4	Our teacher makes the lessons varied.
Learning support	I1	The teacher is interested in my learning progress.	W1	Our teacher is interested in individual students' learning progress.
	I2	The teacher gives me additional support when I need help.	W2	Our teacher gives us additional support when we need help.
	I3	The teacher takes time for me when I want to discuss something.	W3	Our teacher takes time when we want to discuss something.

Table S1 (continued)

Clarity	I1	The teacher teaches so comprehensibly that I understand even difficult things.	W1	Our teacher teaches so comprehensibly that we understand even difficult things.
	I2	I understand the teacher's explanations.	W2	Our teacher expresses him/herself clearly.
	I3	The teacher explains difficult subject matter to me very slowly and carefully.	W3	Our teacher explains very slowly and carefully, especially difficult topics.
Autonomy support	I1	The teacher encourages me to work autonomously.	W1	Our teacher encourages us to work autonomously.
	I2	The teacher gives me the opportunity to explore interesting topics more intensively.	W2	Our teacher gives us the opportunity to explore interesting topics more intensively.
	I3	I feel encouraged to think for myself about how to proceed.	W3	Our teacher encourages us to think for ourselves about how to proceed.
Feedback	I1	The teacher regularly informs me about where I stand with my performance.	W1	The teacher regularly informs us about where we stand with our performance.
	I2	The teacher gives me regular feedback on what I already can do.	W2	Our teacher gives us regular feedback on what we already can do.
	I3	The teacher gives me regular feedback on what I am not yet good at.	W3	Our teacher gives us regular feedback on what we are not yet good at.
	I4	The teacher consistently shows me how I can improve.	W4	Our teacher consistently shows us how we can improve.

Note. I1 = Item 1, I-addressee; I2 = Item 2, I-addressee; I3 = Item 3, I-addressee; I4 = Item 4, I-addressee; W1 = Item 1, we-addressee; W2 = Item 2, we-addressee; W3 = Item 3, we-addressee; W4 = Item 4, we-addressee.

Table S2

Model Fit Results for Teacher Support Dimensions

	Model	χ^2	*df*	SCF	RMSEA	CFI	TLI	$SRMR_w$	$SRMR_b$
Handling mistakes	1 Baseline model	159.94	88	0.63	0.01	1.00	0.99	0.05	0.07
	2 Cross-version loading constraints (student level)	178.28	92	0.66	0.01	0.99	0.99	0.05	0.07
	3 Cross-version loading constraints (classroom level)	170.49	92	0.65	0.01	0.99	0.99	0.05	0.07
	4 Cross-version loading constraints (student and classroom level)	188.88	96	0.68	0.01	0.99	0.99	0.05	0.07
	5 Cross-version intercept constraints (classroom level)	207.18	100	0.69	0.01	0.99	0.99	0.05	0.07
	6 Cross-level invariance (We)	184.83	92	0.66	0.01	0.99	0.99	0.05	0.07
	7 Cross-level invariance (I)	166.50	92	0.67	0.01	1.00	0.99	0.05	0.07
	8 Cross-level invariance (We and I)	189.52	96	0.69	0.01	0.99	0.99	0.06	0.07
Motivation	1 Baseline model	404.17	183	0.71	0.01	0.99	0.99	0.06	0.08
	2 Cross-version loading constraints (student level)	410.71	189	0.72	0.01	0.99	0.99	0.06	0.08
	3 Cross-version loading constraints (classroom level)	409.18	189	0.72	0.01	0.99	0.99	0.06	0.08
	4 Cross-version loading constraints (student and classroom level)	416.51	195	0.73	0.01	0.99	0.99	0.06	0.08
	5 Cross-version intercept constraints (classroom level)	431.04	201	0.74	0.01	0.99	0.99	0.06	0.08
	6 Cross-level invariance (We)	423.10	189	0.72	0.01	0.99	0.99	0.06	0.08
	7 Cross-level invariance (I)	415.82	189	0.72	0.01	0.99	0.99	0.06	0.08
	8 Cross-level invariance (We and I)	434.54	195	0.73	0.01	0.99	0.99	0.05	0.08
Learning support	1 Baseline model	137.92	87	0.62	0.01	1.00	0.99	0.05	0.05
	2 Cross-version loading constraints (student level)	142.75	912	0.65	0.01	1.00	1.00	0.05	0.05
	3 Cross-version loading constraints (classroom level)	135.51	91	0.64	0.01	1.00	1.00	0.05	0.05

Table S2 (continued)

	4 Cross-version loading constraints (student and classroom level)	140.98	95	0.66	0.01	1.00	1.00	0.05	0.05
	5 Cross-version intercept constraints (classroom level)	192.12	99	0.67	0.01	0.99	0.99	0.05	0.05
	6 Cross-level invariance (We)	145.87	91	0.65	0.01	1.00	0.99	0.06	0.05
	7 Cross-level invariance (I)	135.95	91	0.64	0.01	1.00	1.00	0.05	0.05
	8 Cross-level invariance (We and I)	143.95	95	0.67	0.01	1.00	1.00	0.06	0.05
Clarity	1 Baseline model	240.52	88	0.62	0.02	0.99	0.99	0.06	0.06
	2 Cross-version loading constraints (student level)	288.54	92	0.65	0.02	0.99	0.98	0.06	0.06
	3 Cross-version loading constraints (classroom level)	235.55	92	0.65	0.02	0.99	0.99	0.06	0.06
	4 Cross-version loading constraints (student and classroom level)	284.78	96	0.67	0.02	0.99	0.98	0.06	0.06
	5 Cross-version intercept constraints (classroom level)	359.39	100	0.69	0.02	0.98	0.98	0.06	0.06
	6 Cross-level invariance (We)	257.88	92	0.64	0.02	0.99	0.99	0.06	0.06
	7 Cross-level invariance (I)	271.36	92	0.65	0.02	0.99	0.98	0.06	0.06
	8 Cross-level invariance (We and I)	280.96	96	0.67	0.02	0.99	0.98	0.06	0.06
Autonomy support	1 Baseline model	136.73	90	0.69	0.01	1.00	0.99	0.06	0.07
	2 Cross-version loading constraints (student level)	140.51	94	0.71	0.01	1.00	0.99	0.06	0.07
	3 Cross-version loading constraints (classroom level)	140.09	94	0.71	0.01	1.00	0.99	0.06	0.07
	4 Cross-version loading constraints (student and classroom level)	143.51	98	0.73	0.01	1.00	0.99	0.06	0.07
	5 Cross-version intercept constraints (classroom level)	221.57	102	0.74	0.01	0.99	0.99	0.06	0.07
	6 Cross-level invariance (We)	136.73	90	0.69	0.01	1.00	0.99	0.06	0.07
	7 Cross-level invariance (I)	137.62	94	0.71	0.01	1.00	0.99	0.06	0.07
	8 Cross-level invariance (We and I)	144.10	98	0.73	0.01	1.00	0.99	0.06	0.07
Feedback	1 Baseline model	600.21	184	0.67	0.02	0.98	0.98	0.05	0.11

Table S2 (continued)

2 Cross-version loading constraints (student level)	592.74	190	0.68	0.02	0.98	0.98	0.05	0.11
3 Cross-version loading constraints (classroom level)	583.15	190	0.70	0.02	0.98	0.98	0.04	0.11
4 Cross-version loading constraints (student and classroom level)	579.12	196	0.71	0.02	0.99	0.98	0.05	0.11
5 Cross-version intercept constraints (classroom level)	883.81	200	0.70	0.02	0.97	0.97	0.05	0.12
6 Cross-level invariance (We)	668.78	190	0.70	0.02	0.98	0.98	0.05	0.10
7 Cross-level invariance (I)	691.89	190	0.70	0.02	0.98	0.97	0.05	0.10
8 Cross-level invariance (We and I)	720.03	196	0.72	0.02	0.98	0.97	0.05	0.10

Table S3

Factor Loadings, Intraclass Correlations, and Loading Ratios for Mathematics and German Language Arts

		Student level		Classroom level		ICC		Loading ratio	
		M	GL	M	GL	M	GL	M	GL
Handling	I1	0.57	0.53	0.32	0.39	.24	.35	1.76	1.36
mistakes	I2	0.57	0.57	0.31	0.34	.23	.26	1.84	1.67
	I3	0.40	0.39	0.23	0.24	.25	.28	1.75	1.62
	W1	0.58	0.58	0.39	0.37	.31	.30	1.49	1.54
	W2	0.59	0.56	0.31	0.34	.21	.26	1.93	1.68
	W3	0.37	0.32	0.23	0.22	.28	.31	1.62	1.49
Motivation	I1	0.63	0.62	0.42	0.46	.30	.36	1.52	1.34
	I2	0.70	0.64	0.44	0.42	.28	.29	1.60	1.55
	I3	0.42	0.43	0.27	0.31	.28	.34	1.59	1.39
	I4	0.55	0.48	0.41	0.33	.36	.31	1.34	1.48
	W1	0.58	0.57	0.48	0.46	.41	.40	1.20	1.24
	W2	0.69	0.64	0.49	0.47	.33	.35	1.41	1.37
	W3	0.43	0.43	0.29	0.31	.31	.34	1.49	1.39
	W4	0.54	0.48	0.44	0.34	.40	.33	1.22	1.42
Learning support	I1	0.51	0.49	0.33	0.33	.29	.32	1.55	1.47
	I2	0.57	0.57	0.34	0.37	.26	.29	1.68	1.55
	I3	0.52	0.48	0.35	0.32	.31	.31	1.50	1.49
	W1	0.50	0.46	0.33	0.33	.30	.34	1.52	1.38
	W2	0.59	0.59	0.33	0.36	.23	.27	1.80	1.64
	W3	0.49	0.44	0.35	0.30	.33	.32	1.42	1.46
Clarity	I1	0.69	0.54	0.45	0.31	.30	.24	1.53	1.78
	I2	0.65	0.51	0.39	0.29	.26	.24	1.67	1.79
	I3	0.49	0.42	0.38	0.33	.37	.37	1.29	1.30
	W1	0.68	0.58	0.47	0.32	.33	.23	1.44	1.84
	W2	0.58	0.46	0.41	0.31	.33	.32	1.43	1.47
	W3	0.57	0.44	0.43	0.32	.36	.35	1.32	1.37
Autonomy	I1	0.54	0.48	0.29	0.24	.23	.20	1.84	1.98
support	I2	0.52	0.48	0.28	0.26	.22	.23	1.89	1.83
	I3	0.48	0.48	0.26	0.28	.23	.25	1.83	1.72
	W1	0.48	0.44	0.29	0.24	.26	.23	1.67	1.83
	W2	0.49	0.51	0.31	0.25	.29	.20	1.58	2.02

Table S3 (continued)

	W3	0.48	0.49	0.23	0.26	.18	.23	2.13	1.85
Feedback	I1	0.51	0.50	0.29	0.25	.24	.21	1.79	1.96
	I2	0.69	0.69	0.28	0.22	.14	.09	2.48	3.09
	I3	0.65	0.66	0.26	0.19	.14	.08	2.49	3.48
	I4	0.57	0.55	0.36	0.33	.29	.27	1.58	1.66
	W1	0.49	0.47	0.28	0.22	.25	.18	1.74	2.15
	W2	0.65	0.65	0.31	0.25	.19	.12	2.09	2.66
	W3	0.62	0.64	0.25	0.19	.14	.08	2.46	3.32
	W4	0.52	0.48	0.36	0.31	.32	.30	1.47	1.54

Note. M = mathematics; GL = German language arts; ICC = Intraclass correlation; I1 = Item 1, I-addressee; I2 = Item 2, I-addressee; I3 = Item 3, I-addressee; I4 = Item 4, I-addressee; W1 = Item 1, we-addressee; W2 = Item 2, we-addressee; W3 = Item 3, we-addressee; W4 = Item 4, we-addressee.

Table S4

Descriptive Statistics of the Teacher Support Dimensions of Non-Academic Tracks

Dimension	Sample item with we- and I-addressee	Mathematics		German language arts	
		M	*SD*	*M*	*SD*
Handling mistakes	Our teacher is patient when someone makes a mistake in class./	3.26	0.36	3.24	0.28
	The teacher is patient when I make a mistake in class.	3.24	0.31	3.23	0.28
Motivation	Our teacher can sometimes really motivate us./	2.85	0.42	2.89	0.38
	The teacher can sometimes really motivate me.	2.75	0.40	2.75	0.37
Learning support	Our teacher gives us additional support when we need help./	3.09	0.33	3.03	0.31
	The teacher gives me additional support when I need help.	3.03	0.34	2.95	0.31
Clarity	Our teacher teaches so comprehensibly that we understand even difficult things./	2.82	0.44	2.93	0.31
	The teacher teaches so comprehensibly that I understand even difficult things.	2.80	0.45	2.94	0.27
Autonomy support	Our teacher encourages us to work autonomously./	3.07	0.27	3.08	0.23
	The teacher encourages me to work autonomously.	2.95	0.28	2.96	0.24
Feedback	Our teacher gives us regular feedback on what we can already do./	2.75	0.27	2.66	0.22
	The teacher gives me regular feedback on what I can already do.	2.73	0.27	2.66	0.22

Table S5

Descriptive Statistics of the Teacher Support Dimensions of the Academic Track

Dimension	Sample item with we- and I-addressee	Mathematics		German language arts	
		M	*SD*	*M*	*SD*
Handling mistakes	Our teacher is patient when someone makes a mistake in class./	3.27	0.38	3.22	0.43
	The teacher is patient if I make a mistake in class.	3.24	0.35	3.20	0.43
Motivation	Our teacher can sometimes really motivate us./	2.69	0.44	2.70	0.47
	The teacher can sometimes really motivate me.	2.75	0.48	2.65	0.49
Learning support	Our teacher gives us additional support when we need help./	2.96	0.35	2.83	0.35
	The teacher additionally supports me when I need help.	2.87	0.37	2.75	0.38
Clarity	Our teacher teaches so comprehensibly that even difficult things are understood./	2.83	0.50	2.95	0.37
	The teacher teaches so comprehensibly that I understand even difficult things.	2.81	0.47	2.98	0.33
Autonomy support	Our teacher encourages us to work autonomously./	3.02	0.31	2.92	0.24
	The teacher encourages me to work autonomously.	2.90	0.27	2.82	0.25
Feedback	Our teacher gives us regular feedback on what we can already do./	2.61	0.26	2.51	0.19
	The teacher gives me regular feedback on what I can already do.	2.62	0.24	2.54	0.18

Table S6

Intraclass Correlations for the Teacher Support Dimensions in Mathematics and German Language Arts of Non-Academic Tracks

	Student level				Classroom level				ICC			
	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ_{MW}	σ_{MI}	σ_{GW}	σ_{GI}
Handling mistakes	0.32	0.27	0.32	0.27	0.13	0.10	0.08	0.08	0.29	0.27	0.20	0.23
Motivation	0.30	0.35	0.31	0.36	0.18	0.16	0.15	0.14	0.38	0.31	0.33	0.28
Learning support	0.27	0.32	0.32	0.30	0.11	0.11	0.09	0.10	0.29	0.26	0.22	0.25
Clarity	0.43	0.46	0.30	0.26	0.20	0.20	0.09	0.07	0.32	0.30	0.23	0.21
Autonomy support	0.24	0.29	0.21	0.22	0.07	0.08	0.05	0.06	0.23	0.22	0.19	0.21
Feedback	0.29	0.33	0.29	0.33	0.07	0.07	0.05	0.05	0.19	0.18	0.15	0.13

Note. MW = Mathematics, we-addressee; MI= Mathematics, I-addressee; GW = German language arts, we-addressee; GI = German language arts, I-addressee; ICC = Intraclass correlation.

Table S7

Intraclass Correlations for the Teacher Support Dimensions in Mathematics and German Language Arts of the Academic Track

	Student level				Classroom level				ICC			
	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ_{MW}	σ_{MI}	σ_{GW}	σ_{GI}
Handling mistakes	0.38	0.35	0.31	0.30	0.15	0.13	0.19	0.18	0.28	0.27	0.38	0.38
Motivation	0.38	0.44	0.35	0.43	0.24	0.20	0.22	0.24	0.39	0.31	0.39	0.36
Learning support	0.30	0.30	0.23	0.24	0.12	0.13	0.12	0.14	0.29	0.30	0.34	0.37
Clarity	0.43	0.49	0.26	0.24	0.25	0.22	0.14	0.11	0.37	0.31	0.35	0.31
Autonomy support	0.22	0.28	0.21	0.20	0.08	0.09	0.06	0.07	0.27	0.24	0.22	0.26
Feedback	0.23	0.26	0.24	0.23	0.06	0.05	0.04	0.03	0.21	0.16	0.14	0.12

Note. MW = Mathematics, we-addressee; MI= Mathematics, I-addressee; GW = German language arts, we-addressee; GI = German language arts, I-addressee; ICC = Intraclass correlation.

Table S8

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in Mathematics of Non-Academic Tracks

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.72/.71	.86/.91	.78/.79	.75/.82	.50/.72
2. Motivation	.68/.60		.88/.89	.94/.92	.89/.93	.85/.84
3. Learning support	.83/.80	.80/.82		.88/.93	.89/.97	.84/.90
4. Clarity	.67/.65	.76/.82	.74/.83		.92/.91	.78/.82
5. Autonomy support	.65/.61	.89/.84	.77/.79	.83/.72		.74/.85
6. Feedback	.43/.48	.70/.71	.61/.69	.53/.65	.79/.81	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$).

Table S9

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in Mathematics of the Academic Track

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.73/.68	.81/.83	.72/.70	.75/.71	.58/.61
2. Motivation	.62/.58		.88/.88	.92/.90	.95/.94	.81/.85
3. Learning support	.81/.81	.79/.77		.90/.92	.96/.94	.88/.89
4. Clarity	.62/.64	.76/.81	.71/.80		.88/.93	.79/.78
5. Autonomy support	.57/.64	.82/.86	.82/.89	.72/.83		.91/.91
6. Feedback	.39/.38	.56/.65	.64/.66	.48/.56	.67/.77	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$).

Table S10

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in German Language Arts of Non-Academic Tracks

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.68/.78	.82/.89	.86/.91	.77/.79	.56/.62
2. Motivation	.64/.63		.90/.88	.88/.94	.95/.90	.89/.77
3. Learning support	.83/.83	.78/.84		.82/.93	.87/.91	.85/.82
4. Clarity	.74/.71	.76/.86	.85/.84		.90/.86	.74/.75
5. Autonomy support	.64/.59	.84/.91	.78/.84	.85/.85		.87/.75
6. Feedback	.40/.47	.60/.73	.59/.67	.55/.67	.67/.83	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$). All correlations are statistically significant ($p < .001$).

Table S11

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in German Language Arts of the Academic Track

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.74/.78	.91/.89	.84/.83	.84/.82	.59/.56
2. Motivation	.52/.56		.88/.87	.93/.96	.94/.95	.82/.74
3. Learning support	.73/.73	.73/.77		.92/.94	.97/.91	.82/.75
4. Clarity	.63/.69	.77/.77	.82/.83		.94/.97	.77/.74
5. Autonomy support	.56/.53	.88/.83	.81/.86	.82/.76		.87/.86
6. Feedback	.33/.35	.49/.61	.59/.65	.46/.52	.66/.73	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$).

Table S12

Associations for the Teacher Support Dimensions Between Addressee Versions of Non-Academic Tracks

	Mathematics		German language arts	
	r (SL/CL)	SE (SL/CL)	*r* (SL/CL)	SE (SL/CL)
Handling mistakes	**1.00**/.97	0.11/0.63	**1.00/.**96	0.10/0.50
Motivation	**1.00/.97**	0.07/0.26	**1.00/.98**	0.04/0.03
Learning Support	**.99/.96**	0.05/0.25	**1.00/.95**	0.04/0.09
Clarity	**1.00/0.99**	0.05/0.16	**1.00/.95**	0.06/0.19
Autonomy support	**.98/.91**	0.07/0.08	**.96/.88**	0.07/0.23
Feedback	**1.00/.94**	0.04/0.06	**1.00/.**86	0.04/0.80

Note. SL = student level; CL = classroom level. Statistically significant results ($p < .05$) are presented in bold.

Table S13

Associations for the Teacher Support Dimensions Between Addressee Versions of the Academic Track

	Mathematics		German language arts	
	r (SL/CL)	SE (SL/CL)	*r* (SL/CL)	SE (SL/CL)
Handling mistakes	**1.00/.99**	0.04/0.04	**1.00/.99**	0.05/0.02
Motivation	**1.00/.99**	0.02/0.01	**1.00/.98**	0.03/0.01
Learning Support	**1.00/.99**	0.04/0.02	**1.00/.98**	0.05/0.02
Clarity	**1.00/1.00**	0.03/0.01	**1.00/.98**	0.04/0.02
Autonomy support	**.97/.91**	0.05/0.04	**1.00/.95**	0.06/0.05
Feedback	**1.00/.99**	0.03/0.05	**1.00/1.00**	0.03/0.08

Note. SL = student level; CL = classroom level. Statistically significant results ($p < .05$) are presented in bold.

Table S14

Correlations Between Students' Background Variables and Teacher Support Variables in Mathematics of *Non-Academic Tracks*

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.15/.17**	**.**43/**.40**	**.33/.30**	**.**57/**.67**	**.18/.19**	.53/**.55**	**.52/.35**	**.62/.**63	**.36/.19**	**.66/.54**	**.18/.20**	**.52/.50**
Engagement	**.38/.29**	.67/**.64**	**.62/.58**	**.90/.91**	**.44/.41**	.84/**.84**	**.57/.50**	**.89/.87**	**.55/.38**	**.94/.91**	**.38/.42**	**.78/.79**
Grades	**.14/.14**	**.31/.30**	**.20/.18**	**.39/.55**	**.11/.14**	.23/**.39**	**.39/.24**	**.39/.48**	**.26/.15**	.39/.38	**.08/.08**	.18/.26
Test Scores	.05/.04	.05/-.01	**.**06/.03	**-.28/-.22**	.00/**.07**	**-.**04/**-.**12	**.24/.14**	-.18/-.09	**.13/.08**	**-.24/-.**08	**.**05/**.06**	**-.27/-.24**

Note. SL = student level; CL = classroom level; SES = Socioeconomic status. Statistically significant results ($p < .05$) are presented in bold.

Table S15

Correlations Between Students' Background Variables and Teacher Support Variables in Mathematics of the Academic Track

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.22/.20**	**.56/.54**	**.35/.29**	**.77/.79**	**.22/.23**	**.58/.65**	**.51/.40**	**.76/.75**	**.40/.26**	**.66/.64**	**.19/.19**	**.56/.56**
Engagement	**.39/.35**	**.63/.65**	**.74/.67**	**.94/.94**	**.50/.48**	**.76/.81**	**.63/.60**	**.83/.83**	**.61/.57**	**.86/.82**	**.37/.38**	**.74/.75**
Grades	**.22/.16**	**.39/.38**	**.24/.21**	**.48/.52**	**.16/.16**	**.30/.35**	**.43/.30**	**.44/.45**	**.27/.20**	**.40/.37**	**.07/.10**	**.37/.29**
Test Scores	**.11/.10**	.06/.07	**.13/.07**	**-.17/-.22**	**.06/.08**	**.06/.06**	**.25/.17**	-.02/-.02	**.17/.10**	-.09/.04	**.05/.06**	-.01/.04

Note. SL = student level; CL = classroom level; SES = Socioeconomic status. Statistically significant results ($p < .05$) are presented in bold.

Table S16

Correlations Between Students' Background Variables and Teacher Support Dimensions in German Language Arts of the Non-Academic Tracks

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.27/.23**	**.**44/.40	**.33/.31**	.**55/.52**	**.26/.27**	**.51/.48**	**.40/.37**	**.**56/**.56**	.**38/.26**	**.56/.**41	**.17/.18**	**.**50/.49
Engagement	**.34/.34**	**.64/.65**	**.65/.60**	**.90/.88**	**.45/.44**	**.74/.74**	**.50/.48**	**.84/.88**	**.54/.44**	**.90/.79**	**.34/.40**	**.75/.70**
Grades	**.17/.24**	.28/.06	**.14/.11**	.04/.18	**.14/.14**	.13/.08	**.22/.15**	.27/.17	**.20/.12**	.07/.14	.03/**.**02	.01/-.06
Test Scores	**.09/.12**	-.09/-.14	-.04/-.05	**-.28/-.23**	-.03/.01	**-.28/-.**31	**.08/.07**	-.10/**-.**16	.07/.00	**-.**41/**-.29**	-.03/-.02	**-.**40/**-.**27

Note. SL = student level; Cl = classroom level; SES = Socioeconomic status. Statistically significant results ($p < .05$) are presented in bold.

Table S17

Correlations Between Students' Background Variables and Teacher Support Dimensions in German Language Arts of the Academic Track

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.19/.22**	**.49/.52**	**.36/.32**	**.58/.54**	**.26/.24**	**.49/.52**	**.48/.37**	**.56/.53**	**.39/.32**	**.45/.47**	**.13/.18**	**.**32/.
Engagement	**.33/.37**	**.68/.73**	**.72/.66**	**.97/.95**	**.51/.46**	**.79/.78**	**.59/.55**	**.87/.87**	**.68/.57**	**.86/.86**	**.36/.36**	**.73/.73**
Grades	**.18/.16**	**.39/.41**	**.20**/**.16**	.05/.18	**.13/.14**	.13/.08	**.34/.24**	.27/.17	**.21/.19**	.07/.14	.03/**.10**	.02/-.05
Test Scores	**.08/.10**	-.08/-.11	.02/**.05**	-.28/-.23	.02/**.06**	**-.29**/-.31	**.19/.10**	-.11/**-.**16	**.14/.09**	**-.**42/**-.30**	.03/.03	**-.**40/-.27

Note. SL = student level; Cl = classroom level; SES = Socioeconomic status. Statistically significant results ($p < .05$) are presented in bold.

Table S18

Descriptive Statistics for the Teacher Support Dimensions of Grades 5 and 6

Dimension	Sample item with we- and I-addressee	Mathematics		German language arts	
		M	*SD*	*M*	*SD*
Handling mistakes	Our teacher is patient when someone makes a mistake in class./	3.24	0.30	3.30	0.29
	The teacher is patient when I make a mistake in class.	3.19	0.30	3.26	0.28
Motivation	Our teacher can sometimes really motivate us./	2.84	0.40	2.99	0.35
	The teacher can sometimes really motivate me.	2.93	0.34	2.88	0.31
Learning support	Our teacher gives us additional support when we need help./	2.92	0.33	3.05	0.28
	The teacher gives me additional support when I need help.	2.99	0.32	3.03	0.27
Clarity	Our teacher teaches so comprehensibly that we understand even difficult things./	2.94	0.42	3.03	0.27
	The teacher teaches so comprehensibly that I understand even difficult things.	2.91	0.38	3.03	0.29
Autonomy support	Our teacher encourages us to work autonomously./	3.01	0.26	3.04	0.21
	The teacher encourages me to work autonomously.	2.91	0.26	2.94	0.19
Feedback	Our teacher gives us regular feedback on what we can already do./	2.58	0.26	2.56	0.21
	The teacher gives me regular feedback on what I can already do.	2.65	0.25	2.65	0.19

Table S19

Descriptive Statistics for the Teacher Support Dimensions of Grades 7 and 8

Dimension	Sample item with we- and I-addressee	Mathematics		German language arts	
		M	*SD*	*M*	*SD*
Handling mistakes	Our teacher is patient when someone makes a mistake in class./	3.29	0.34	3.22	0.35
	The teacher is patient if I make a mistake in class.	3.28	0.33	3.22	0.36
Motivation	Our teacher can sometimes really enthuse us./	2.70	0.47	2.74	0.45
	The teacher can sometimes really enthuse me.	2.66	0.43	2.65	0.45
Learning support	Our teacher gives us additional support when we need help./	2.96	0.34	2.93	0.35
	The teacher additionally supports me when I need help.	2.89	0.36	2.93	0.38
Clarity	Our teacher teaches so comprehensibly that even difficult things are understood./	2.75	0.46	2.94	0.36
	The teacher teaches so comprehensibly that I understand even difficult things.	2.77	0.44	2.91	0.30
Autonomy support	Our teacher encourages us to work autonomously./	2.99	0.30	2.96	0.29
	The teacher encourages me to work autonomously.	2.88	0.32	2.85	0.28
Feedback	Our teacher gives us regular feedback on what we can already do./	2.67	0.26	2.61	0.23
	The teacher gives me regular feedback on what I can already do.	2.67	0.28	2.62	0.22

Table S20

Descriptive Statistics for the Teacher Support Dimensions of Grades 9 and 10

Dimension	Sample item with we- and I-addressee	Mathematics		German language arts	
		M	*SD*	*M*	*SD*
Handling mistakes	Our teacher is patient when someone makes a mistake in class./	3.25	0.50	3.13	0.49
	The teacher is patient if I make a mistake in class.	3.25	0.39	3.13	0.46
Motivation	Our teacher can sometimes really enthuse us./	2.62	0.47	2.59	0.45
	The teacher can sometimes really enthuse me.	2.59	0.49	2.53	0.49
Learning support	Our teacher gives us additional support when we need help./	2.95	0.39	2.75	0.35
	The teacher additionally supports me when I need help.	2.81	0.42	2.67	0.39
Clarity	Our teacher teaches so comprehensibly that even difficult things are understood./	2.80	0.55	2.88	0.37
	The teacher teaches so comprehensibly that I understand even difficult things.	2.74	0.56	2.92	0.32
Autonomy support	Our teacher encourages us to work autonomously./	3.01	0.27	2.98	0.21
	The teacher encourages me to work autonomously.	2.92	0.29	2.87	0.26
Feedback	Our teacher gives us regular feedback on what we can already do./	2.62	0.29	2.58	0.22
	The teacher gives me regular feedback on what I can already do.	2.64	0.32	2.51	0.22

Table S21

Intraclass Correlations for the Teacher Support Dimensions in Mathematics and German Language Arts of Grades 5 and 6

	Student level				Classroom level				ICC			
	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ_{MW}	σ_{MI}	σ_{GW}	σ_{GI}
Handling mistakes	0.34	0.38	0.31	0.26	0.09	0.09	0.08	0.08	0.28	0.26	0.31	0.32
Motivation	0.36	0.34	0.34	0.40	0.16	0.11	0.12	0.09	0.38	0.31	0.37	0.33
Learning support	0.26	0.24	0.26	0.24	0.11	0.10	0.08	0.08	0.29	0.29	0.31	0.33
Clarity	0.44	0.42	0.29	0.25	0.15	0.18	0.08	0.08	0.35	0.32	0.30	0.29
Autonomy support	0.28	0.26	0.21	0.20	0.07	0.07	0.04	0.04	0.24	0.23	0.23	0.24
Feedback	0.29	0.29	0.28	0.28	0.07	0.06	0.04	0.04	0.23	0.20	0.16	0.14

Note. MW = Mathematics, we-addressee; MI = Mathematics, I-addressee; GW = German language arts, we-addressee; GI = German language arts, I-addressee; ICC = Intraclass correlation.

Table S22

Intraclass Correlations for the Teacher Support Dimensions in Mathematics and German Language Arts of Grades 7and 8

	Student level				Classroom level				ICC			
	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ_{MW}	σ_{MI}	σ_{GW}	σ_{GI}
Handling mistakes	0.32	0.30	0.30	0.30	0.12	0.11	0.12	0.13	0.27	0.27	0.29	0.30
Motivation	0.37	0.40	0.33	0.37	0.22	0.18	0.20	0.20	0.37	0.31	0.38	0.35
Learning support	0.32	0.32	0.27	0.25	0.13	0.11	0.15	0.12	0.29	0.26	0.36	0.32
Clarity	0.48	0.42	0.25	0.23	0.21	0.19	0.09	0.13	0.30	0.31	0.26	0.36
Autonomy support	0.26	0.23	0.23	0.22	0.10	0.09	0.08	0.08	0.28	0.28	0.26	0.27
Feedback	0.26	0.24	0.24	0.23	0.07	0.08	0.05	0.05	0.21	0.25	0.17	0.18

Note. MW = Mathematics, we-addressee; MI= Mathematics, I-addressee; GW = German language arts, we-addressee; GI = German language arts, I-addressee; ICC = Intraclass correlation.

Table S23

Intraclass Correlations for the Teacher Support Dimensions in Mathematics and German Language Arts of Grades 9 and 10

	Student level				Classroom level				ICC			
	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ^2_{MW}	σ^2_{MI}	σ^2_{GW}	σ^2_{GI}	σ_{MW}	σ_{MI}	σ_{GW}	σ_{GI}
Handling mistakes	0.37	0.34	0.35	0.31	0.25	0.16	0.24	0.21	0.27	0.27	0.29	0.30
Motivation	0.34	0.45	0.46	0.31	0.22	0.24	0.24	0.20	0.37	0.31	0.38	0.35
Learning support	0.32	0.37	0.27	0.28	0.15	0.18	0.15	0.13	0.29	0.26	0.36	0.32
Clarity	0.43	0.54	0.26	0.32	0.30	0.32	0.10	0.14	0.30	0.31	0.26	0.36
Autonomy support	0.20	0.34	0.20	0.21	0.08	0.09	0.05	0.07	0.28	0.28	0.26	0.27
Feedback	0.23	0.29	0.28	0.26	0.10	0.09	0.05	0.05	0.21	0.25	0.17	0.18

Note. MW = Mathematics, we-addressee; MI= Mathematics, I-addressee; GW = German language arts, we-addressee; GI = German language arts, I-addressee; ICC = Intraclass correlation.

Table S24

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in Mathematics of Grades 5 and 6

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.73/.83	.78/.85	.80.80	.76/.85	.42/.61
2. Motivation	.67/.62		.85/.86	.92/.93	.93/.95	.77/.76
3. Learning support	.81/.83	.81/.84		.90/88	.96/.97	.80/.92
4. Clarity	.63/.63	.74/.82	.70/.84		.89/.95	.67/.76
5. Autonomy support	.60/.64	.85/.82	.86/.88	.78/.76		.84/.87
6. Feedback	.42/.38	.62/.65	.67/.68	.49/.56	.76/.78	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$).

Table S25

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in Mathematics of Grades 7 and 8

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.75/.67	.84/.85	.71/.68	.76/.76	.58/.60
2. Motivation	.65/.58		.93/.92	.92/.89	.94/.97	.83/.88
3. Learning support	.82/.78	.85/.73		.87/.88	.96/.97	.90/.87
4. Clarity	.68/.68	.80/.81	.76/.82		.91/.88	.74/.71
5. Autonomy support	.65/.67	.92/.88	.83/.87	.80/.80		.87/.92
6. Feedback	.41/.48	.64/.69	.62/.69	.48/.62	.74/.82	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$).

Table S26

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in Mathematics of Grades 9 and 10

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.78/.71	.85/.83	.77/.76	.78/.69	.56/.66
2. Motivation	.58/.54		.89/.90	.92/.94	.89/.92	.83/.92
3. Learning support	.80/.80	.70/.83		.91/.94	.91/.87	.86/.88
4. Clarity	.59/.59	.72/.81	.68/.78		.86/.92	.78/.84
5. Autonomy support	.51/.56	.74/.83	.69/.79	.67/.79		.81/.88
6. Feedback	.37/.39	.57/.67	.62/.67	.55/.61	.60/.73	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$).

Table S27

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in German Language Arts of Grades 5 and 6

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.59/.68	.76/.79	.81/.80	.74/.65	.34/.49
2. Motivation	.59/.63		.81/.79	.89/.96	.84/.89	.74/.79
3. Learning support	.77/.77	.73/.84		.86/.84	.98/.87	.79/.78
4. Clarity	.64/.71	.75/.84	.84/.89		.85/.81	.62/.70
5. Autonomy support	.59/.57	.86/.80	.79/.86	.79/.83		.81/.95
6. Feedback	.33/.40	.55/.62	.56/.64	.45/.59	.69/.79	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$).

Table S28

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in German Language Arts of Grades 7 and 8

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.82/.84	.89/.90	.90/.94	.81/.80	.59/.58
2. Motivation	.55/.57		.90/.91	.90/.94	.95/.93	.84/.75
3. Learning support	.78/.79	.77/.78		.84/.95	.91/.91	.88/.83
4. Clarity	.73/.74	.79/.84	.87/.86		.87/.92	.67/.72
5. Autonomy support	.60/.56	.91/.92	.84/.86	.92/.78		.88/.82
6. Feedback	.35/.37	.54/.72	.59/.69	.54/.59	.68/.78	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$).

Table S29

Intercorrelations for the Teacher Support Dimensions for Both Addressee Versions at the Student and Classroom Levels in German Language Arts of Grades 9 and 10

	1. (I/we)	2. (I/we)	3. (I/we)	4. (I/we)	5. (I/we)	6. (I/we)
1. Handling mistakes		.62/.70	.84/.83	.73/.79	.75/.81	.53/.50
2. Motivation	.53/.55		.86/.88	.92/.93	.90/.95	.84/.76
3. Learning support	.77/.72	.75/.78		.91/.93	.90/.97	.79/.81
4. Clarity	.65/.62	.77/.71	.76/.76		.95/.97	.81/.68
5. Autonomy support	.57/.52	.84/.87	.76/.84	.77/.79		.86/.73
6. Feedback	.40/.44	.51/.62	.64/.67	.50/.57	.63/.74	

Note. Student-level correlations are presented below the diagonal. Classroom-level correlations are presented above the diagonal. All correlations are statistically significant ($p < .001$).

Table S30

Associations for the Teacher Support Dimensions Between Addressee Versions of Grades 5 and 6

	Mathematics		German language arts	
	r (SL/CL)	SE (SL/CL)	*r* (SL/CL)	SE (SL/CL)
Handling mistakes	**1.00**/.97	0.08/0.85	**1.00/.**93	0.09/0.48
Motivation	**1.00/.98**	0.04/0.05	**1.00/.95**	0.04/0.48
Learning Support	**1.00/.98**	0.08/0.24	**1.00/.**95	0.08/0.52
Clarity	**1.00/.98**	0.05/0.03	**1.00/.93**	0.06/0.12
Autonomy support	**.98/.93**	0.07/0.07	**1.00/.**90	0.15/1.85
Feedback	**1.00/.96**	0.04/0.11	**1.00/.**94	0.05/2.11

Note. SL = student level; CL = classroom level. Statistically significant results ($p < .05$) are presented in bold.

Table S31

Associations for the Teacher Support Dimensions Between Addressee Versions of Grades 7 and 8

	Mathematics		German language arts	
	r (SL/CL)	SE (SL/CL)	*r* (SL/CL)	SE (SL/CL)
Handling mistakes	**1.00**/0.98	0.07/1.06	**1.00/.98**	0.07/0.31
Motivation	**1.00**/0.98	0.04/0.93	**1.00/.98**	0.04/0.12
Learning Support	**1.00/.97**	0.05/0.04	**1.00/.97**	0.05/0.03
Clarity	**1.00/0.98**	0.04/0.17	**1.00/.98**	0.06/0.22
Autonomy support	**1.00/.90**	0.18/0.09	**1.00/.92**	0.07/0.13
Feedback	**1.00/.97**	0.05/0.30	**1.00/.95**	0.04/0.15

Note. SL = student level; CL = classroom level. Statistically significant results ($p < .05$) are presented in bold.

Table S32

Associations for the Teacher Support Dimensions Between Addressee Versions of Grades 9 and 10

	Mathematics		German language arts	
	r (SL/CL)	SE (SL/CL)	*r* (SL/CL)	SE (SL/CL)
Handling mistakes	**1.00/.98**	0.06/0.10	**1.00/.98**	0.04/0.07
Motivation	**1.00/.99**	0.04/0.03	**0.96/.99**	0.08/0.04
Learning Support	**1.00/1.00**	0.06/0.05	**1.00/.99**	0.06/0.06
Clarity	**1.00/1.00**	0.04/0.07	**1.00/.98**	0.06/0.21
Autonomy support	**.81/.95**	0.07/0.38	**1.00/.93**	0.03/0.17
Feedback	**1.00/.97**	0.02/0.30	**1.00/.92**	0.07/0.28

Note. SL = student level; CL = classroom level. Statistically significant results ($p < .05$) are presented in bold.

Table S33

Correlations Between Students' Background Variables and Teacher Support Variables in Mathematics Version of Grades 5 and 6

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.19/.23**	.43/.42	**.23/.29**	**.48/.53**	**.22/.27**	.17/.20	**.49/.34**	**.47/.53**	**.33/.23**	.28/.33	**.16/.17**	-.02/.09
Engagement	**.37/.29**	**.73/.77**	**.64/.63**	**.92/.94**	**.48/.48**	**.73/.73**	**.58/.57**	**.87/.90**	**.57/.49**	**.80/.84**	**.39/.37**	**.56/.66**
Grades	**.12/.12**	.20/.15	**.10/.13**	.09/.19	.05/**.09**	-.13/-.11	**.36/.21**	.13/.18	**.16/.11**	-.04/-.01	.04/.06	-.**32**/-.24
Test Scores	**.10/.11**	-.05/-.05	.05/.06	**-.26/-.**18	.01/**.07**	**-.31/-.36**	**.24/.15**	-.13/-.13	**.12/.11**	**-.27**/-.23	.05/.05	**-.60/-.49**

Note. SL = student level; CL = classroom level; Statistically significant results ($p < .05$) are presented in bold.

Table S34

Correlations Between Students' Background Variables and Teacher Support Variables in Mathematics Version of Grades 7 and 8

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.20/.18**	.33/.40	**.40/.29**	.40/.45	**.23/.20**	.36/.39	**.51/.38**	.57/.52	**.41/.22**	.44/.43	**.17/.22**	.18/.18
Engagement	**.38/.35**	**.64/.66**	**.71/.63**	**.92/.91**	**.49/.43**	**.86/.88**	**.59/.58**	**.88/.84**	**.59/.49**	**.91**/.89	**.36/.41**	**.73/.**74
Grades	**.20/.14**	.18/.29	**.24/.20**	.05/.12	**.17/.15**	-.03/.07	**.39/.26**	.19/.19	**.29/.17**	.02/.15	.08/.09	-.28/-.20
Test Scores	**.09/.07**	.08/.19	**.13/.**04	-.15/-.15	.06/**.08**	-.16/-.20	**.24/.14**	-.03/-.02	**.21/.**06	**-.**18/-.09	.04/.07	-.43/-.35

Note. SL = student level; CL = classroom level; Statistically significant results ($p < .05$) are presented in bold.

Table S35

Correlations Between Students' Background Variables and Teacher Support Variables in Mathematics Version of Grades 9 and 10

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.19/.16**	.52/**.47**	**.38/.30**	**.43/.45**	**.15/.16**	.50/.51	**.56/.41**	**.60/.60**	**.43/.23**	.45/.40	**.23/.21**	.30/.31
Engagement	**.42/.34**	**.79/.75**	**.73/.64**	**.91/.90**	**.45/.46**	**.88/.91**	**.66/.57**	**.89/.89**	**.61/.47**	**.89/.89**	**.39/.43**	**.74/.82**
Grades	**.24/.19**	**.46/.41**	**.32/.28**	**.37/.38**	**.20/.21**	.35/.34	**.52/.36**	**.50/**.51	**.37/.27**	.37/.35	**.12/.15**	.19/.23
Test Scores	.07/.03	.06/.03	**.13/.09**	-.01/-.06	.02/.06	-.01/-.07	**.27/.19**	.15/.12	**.16/.11**	-.00/.04	.05/**.10**	-.19/-.23

Note. SL = student level; CL = classroom level; Statistically significant results ($p < .05$) are presented in bold.

Table S36

Correlations Between Students' Background Variables and Teacher Support Dimensions in German Language Arts Version of Grades 5 and 6

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.22/.19**	.38/.37	**.33/.31**	**.36/.**30	**.24/.26**	.20/.08	**.43/.36**	.37/**.40**	**.38/.27**	**.16/.05**	**.14/.18**	-.04/.15
Engagement	**.31/.35**	.61/**.63**	**.71/.60**	**.93/.89**	**.45/.46**	**.**69/**.56**	**.54/.52**	**.85/.87**	**.62/.49**	.73/.70	**.35/.37**	**.56/.**66
Grades	**.14/.11**	.21/.12	**.11/.**05	.00/-.06	.03/.04	-.16/**-.29**	**.22/.12**	.10/.04	**.10/.**10	-.16/-.30	-.05/**.**00	-.38/-.32
Test Scores	**.07/.09**	.11/.10	-.02/-.00	-.18/-.21	-.02/-.01	-.28/-.28	**.12/.07**	-.02/**-.**09	**.**05/.06	-.35/-.43	-.04/-.03	**-.54**/-.45

Note. SL = student level; CL = classroom level; Statistically significant results ($p < .05$) are presented in bold.

Table S37

Correlations Between Students' Background Variables and Teacher Support Dimensions in German Language Arts Version of Grades 7 and 8

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.25/.30**	**.56/.49**	**.35/.32**	**.56/.50**	**.28/.28**	**.52/.55**	**.46/.38**	**.69/.58**	**.38/.31**	**.52/.52**	**.13/.20**	**.40/.40**
Engagement	**.33/.35**	**.73/.71**	**.65/.63**	**.95/.90**	**.50/.45**	**.79/.82**	**.54/.52**	**.87/.89**	**.62/.52**	**.89/.84**	**.35/.39**	**.75/.77**
Grades	**.21/.25**	**.32/.**23	**.18/.15**	**.**10/.10	**.17/.18**	.06/.06	**.32/.28**	**.30/.**17	**.19/.16**	-.05/.05	.03/**.10**	-.20/-.15
Test Scores	**.10/.13**	-.02/.00	.03/.02	**-.27/-.23**	.04/**.08**	**-.32/-.28**	**.15/.12**	-.03/**-.**09	**.15/.**06	**-.35/-.28**	.01/.01	**-.56/-.4**

Note. SL = student level; CL = classroom level; Statistically significant results ($p < .05$) are presented in bold.

Table S38

Correlations Between Students' Background Variables and Teacher Support Dimensions in German Language Arts Version of Grades 9 and 10

	Handling mistakes		Motivation		Learning support		Clarity		Autonomy support		Feedback	
	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL	I/we SL	I/we CL
Self-concept	**.15/.16**	**.**33/**.**41	**.37/.32**	**.43/.44**	**.25/.23**	**.36/.**44	**.47/.37**	**.**45/.42	**.41/.30**	**.**31/.30	**.17/.18**	**.28/.18**
Engagement	**.36/.36**	**.60/.70**	**.73/.69**	**.93/.96**	**.50/.47**	**.75/.80**	**.59/.55**	**.87/.86**	**.66/.56**	**.79/.83**	**.73/.65**	**.70/.70**
Grades	**.17/.18**	**.**15/.18	**.25/.24**	**.**10/.10	**.21/.18**	.02/.05	**.36/.24**	.14/.16	**.36/.25**	.10/.16	**.12/.15**	-.07/-.22
Test Scores	**.**05/**.18**	.00/-.02	.00/.04	-.03/-.04	.02/.03	-.05/-.09	**.20/.09**	.00/**-.**03	**.14/.**06	-.07/-.12	.06/.05	-.08/-.19

Note. SL = student level; CL = classroom level; Statistically significant results ($p < .05$) are presented in bold.

5

STUDIE 3: HOW STUDENTS' PERCEPTIONS OF TEACHING QUALITY IN ONE SUBJECT ARE IMPACTED BY THE GRADES THEY RECEIVE IN ANOTHER SUBJECT: DIMENSIONAL COMPARISONS IN STUDENT EVALUATIONS OF TEACHING QUALITY

Jaekel, A., Göllner, R., Trautwein, U. (2020). How students' perceptions of teaching quality in one subject are impacted by the grades they receive in another subject: dimensional comparisons in student evaluations of teaching quality. *Journal of Educational Psychology*. Advance online publication. https://doi.org/10.1037/edu0000488

Abstract

According to dimensional comparison theory (DCT), students evaluate their ability in one domain (e.g., math) by comparing their achievement in that domain with their achievement in other domains (e.g., English). Primarily in research on students' academic self-concept, these comparison processes have been found to lead to positive associations within subjects (e.g., the better the student's achievement in math, the higher that student's math self-concept) but negative associations between subjects (e.g., better skills in math than in English lead to a relatively lower self-concept in English than in math; Möller & Marsh, 2013). However, less is known about dimensional comparison effects in evaluations of others, e.g., students' ratings of their teachers. In the present study, we used data from a large-scale assessment of teaching quality in Germany (N = 6,479 students from 401 classes in Grades 5 to 10) and examined the associations between students' grades and their teaching-quality ratings in mathematics and German language classes. In line with DCT, the results revealed positive within-subject associations and negative between-subject associations: Most importantly, dimensional comparison effects were also found at the classroom level and remained stable even after achievement test scores were controlled for.

Keywords: student ratings, teaching quality, dimensional comparison effects

Educational Impact And Implications Statement

Student ratings are frequently used in assessments of teaching quality and evaluations of teachers and institutions in schools and higher education. At the same time, the use of student ratings is also approached with some caution as student ratings might function differently from ratings made by adult observers. In the present study, we examined students' ratings of teaching quality in two subjects (mathematic and the German language) in a sample of 6,749 students in Grades 5 to 10 and investigated the extent to which student grades are associated with their teaching-quality ratings in the same subject but also associated with students' ratings of another subject. Our findings showed that the better a student's grade in one subject-area course, the better he or she rated the quality of the teaching in that same course. Additionally, the better the student's grade in one subject-area course, the relatively lower he or she rated the quality of the teaching in the other subject-area course. That is, student ratings of one subject are impacted by grades they receive in another subject. This phenomenon should especially be taken into account when using student ratings for teacher evaluations.

Introduction

Student ratings provide a valuable and inexpensive source of information for research on teaching quality (De Jong & Westerhof, 2001; Spooren, Brockx, & Mortelmans, 2013). They are time- and cost-effective, reliable, and predictive of students' achievement. Thus, they are frequently used in educational research and practice. However, several scholars (e.g., Abrami, d'Apollonia, & Cohen, 1990) have expressed doubts about the validity of student ratings. On the side of the students, there might be biasing factors (e.g., gender or achievement) that affect students' perceptions of their learning environment (Griffin, 2004; Marsh & Roche, 2000). In addition, student ratings may be impacted by factors related to the teacher. Most importantly, previous research has provided empirical evidence of effects of grading leniency (Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Greenwald & Gilmore, 1997a, 1997b; Marsh & Roche, 2000) such that teachers who give higher grades tend to receive more favorable teaching-quality ratings. However, much less is known about whether and to what extent student ratings are impacted not only by students' grades in the same subject but also by their grades in another subject. According to dimensional comparison theory (DCT), students evaluate their achievement in one domain on the basis of their relative achievement between that domain and other domains. In research on students' academic self-concept, these internal comparison processes have been found to lead to positive associations within subjects but negative associations between subjects (Möller, Helm, Müller-Kalthoff, Nagy & Marsh, 2015; Möller & Marsh, 2013).

Previous research has investigated dimensional comparison effects primarily in relation to students' academic self-concept (Ehm, Lindberg, & Hasselhorn, 2014), whereas few studies have indicated that DCT also applies to student characteristics other than self-concept, for instance, regarding students' perception of teaching quality (Arens & Möller, 2016). However, it is still unclear whether and how dimensional comparison effects differ on an individual level (students' individual perceptions of teaching quality) and on the classroom level (the perception shared by students in one classroom). Additionally, studies have yet to investigate whether the DCT pattern is due to differences in students' achievement or due to comparison processes between subjects. In line with the assumptions of DCT, in the present study, we specifically investigated the existence of dimensional comparison effects in student ratings of teaching

quality in mathematics and the German language for a variety of teaching-quality aspects on both the student and classroom levels.

Student Ratings of Teaching Quality

Teaching quality is widely understood as based on a teacher's actual behavior, but it is also based on the teacher's interactions with the students (Doyle, 2013; Fauth et al., 2019; Göllner, Fauth, Lenske, Praetorius, & Wagner, in press; Hamre & Pianta, 2010; Kunter et al., 2013). This also means that the context and the conditions in which teaching takes place always need to be considered. Regarding the content of teaching quality, there are a number of different frameworks from which to describe and assess teaching quality, many of which show a great deal of overlap (Hamre & Pianta, 2010). This research employs a widely used conception of teaching quality (see Praetorius, Klieme, Herbert, & Pinger, 2018) that categorizes the aspects of teaching quality into three superordinate quality domains including classroom management, supportive climate, and cognitive activation. Classroom management refers to an efficient way of teaching and use of time, which results, for example, from rule clarity, a well-structured lesson, or the absence of disturbances (Kunter, Baumert, & Köller, 2007). Supportive climate builds on a positive student-teacher relationship and a learning environment in which, for example, students get constructive feedback on how to improve or experience the relevance of the subject matter (Brophy, 2000). Finally, cognitive activation includes, for example, the provision of challenging tasks that clarify the connection between different concepts or link new learning content with prior knowledge (Lipowsky et al., 2009). The theoretical framework of the three domains of teaching quality has received empirical support from several studies (e.g., Fauth et al., 2014; Kunter & Voss, 2013). However, in many of these studies, teaching quality was assessed in mathematics or science classes (e.g., Pinger, Rakoczy, Besser, & Klieme, 2017) or by focusing on one of the three teaching domains (Kunter, Baumert, Blum, Klusmann, Krauss, & Neubrand, 2013; Schweig, 2014).

One widely used approach for assessing teaching quality is student questionnaires in which students are asked to rate several theoretically driven aspects of teaching quality. In the meantime, a variety of student questionnaires have been used in both educational research and practice (Den Brok, Brekelmans, & Wubbels, 2007; Fauth et al., 2014; Turner & Meyer, 2000). Student ratings are easy to implement and result in extensive sets of data that cover a fairly long

evaluation period (e.g., the last school year) and can be collected in a short period of time. Furthermore, in general, the reliability and validity of student ratings of teaching quality have been found to be high (Kunter & Baumert, 2006; Wagner et al., 2013). For instance, students appear to be able to differentiate between several aspects of their learning environment and rate teaching quality subject-specifically (Wagner, Göllner, Helmke, Trautwein, & Lüdtke, 2013; Wallace, Kelcey, & Ruzek, 2016). Furthermore, student ratings provide information not only about students' shared perceptions but also about students' individual perceptions of teaching quality, which have been shown to predict students' academic achievement and learning (e.g., Aldrup, Klusmann, Lüdtke, Göllner, & Trautwien, 2018; Hattie, 2009; Kunter et al., 2013; Mainhard, Oudman, Hornstra, Bosker, & Goetz, 2018; Wagner, Göllner, Werth, Voss, Schmitz, & Trautwein, 2016).

However, the use of student ratings of teaching quality has also been criticized (e.g., Abrami, 1989). This is particularly true in the context of high-stakes evaluations (Nichols, Glass, & Berliner, 2006), where rating results are used in the career-development process. Critics argue that students have no didactical training and are therefore unable to adequately assess teaching methods. Research has identified teacher-independent factors such as students' harshness or leniency, the halo error, or classroom composition features (Benton & Cashin, 2012; Göllner et al., in press; Kunter & Baumert, 2006), revealing an effect on student ratings of teaching quality. It has also been argued that teachers' grading practices have a tremendous impact on student ratings of teaching quality (Griffin, 2004; Spooren et al., 2013). Specifically, teachers are rewarded with higher ratings for engaging in lenient grading practices or conversely penalized with lower ratings when viewed as harsh graders. A number of previous studies have shown that direct measures of grading leniency were related to undergraduate students' teaching-quality ratings. For instance, Griffin (2004) found that undergraduate students' perceptions of grading leniency were positively associated with their ratings of different statements about teaching quality such as useful feedback or fairness. However, previous research on the association between students' grades and their teaching-quality ratings have been restricted to one and the same subject. In other words, prior studies have analyzed the association between grades in a specific subject (e.g., math) and teaching-quality ratings in the same subject (e.g., math) but not between grades in one subject and teaching-quality ratings in a different subject.

Thus, the question arises as to whether students' teaching-quality ratings might also be impacted by the grading of a teacher in another subject. In fact, there are some good reasons to expect exactly such an effect. More specifically, arguments for claiming such a between-subject association can be found in the theory of dimensional comparisons, which argues that students evaluate their ability in one domain by comparing their achievement in that domain with their achievement in another domain, and this may affect student characteristics such as academic self-concept (Möller & Marsh, 2013) but might also be used as one evaluative cue for rating teaching quality.

Dimensional Comparison Effects on Student Ratings of Teaching Quality: Dimensional Comparison Theory

Humans' judgments of their own abilities are often based on comparisons (Marsh & Craven, 2006; Mussweiler, Rüter, & Epstude, 2006). These comparison processes are ubiquitous in daily life and occur every time one evaluates other people or oneself. Students draw comparisons to assess their own abilities, and to do so, they use different frames of reference. The internal/external frame of reference model (I/E model; Marsh, 1986) proposes that students compare their achievement not only with other students but also with their own achievement in another subject. With regard to students' academic self-concept, these comparison processes lead to positive within-subject associations between students' achievement and their academic self-concept but at the same time to negative between-subject associations. For example, if Hannah is better in English than she is in math, her English self-concept will be higher than her math self-concept. If Maria is as good as Hannah in English, but Maria is even better in math, her English self-concept will be lower than Hannah's even though they have the same ability level in this subject. In general, studies have found stronger effects for school grades than for standardized test scores (Köller, Trautwein, Lütdke, & Baumert, 2006; Marsh, 1987; Marsh et al., 2018; Möller, Pohlmann, Köller, & Marsh, 2009) as grades provide a more salient, local source of feedback to students about their achievement and therefore tend to be more relevant for dimensional comparison processes.

The generalized I/E model (gI/E model; Möller, Müller-Kalthoff, Helm, Nagy, & Marsh, 2015) further develops and extends the I/E model. More specifically, research on the gI/E model is not restricted to self-concept as the outcome but also examines other academic outcomes such

as students' intrinsic motivation and effort (Skaalvik & Rankin, 1995) or emotions that occur in academic settings such as anxiety or enjoyment (Goetz, Frenzel, Hall, & Pekrun, 2008; Marsh, 1988). In general, these studies supported the idea that dimensional comparison processes take place in many instances. Also in line with the assumptions of the gI/E model, additional research has shown that dimensional comparisons are relevant not only for students' self-evaluations but also for their evaluations of others, for instance, students' ratings of teaching quality. A study by Dietrich, Dicke, Kracke, and Noack (2015) found that teacher support was positively related to intrinsic value and effort within subjects and negatively related to these constructs between subjects, thus supporting the theory that students' perceptions of their learning environment in one subject might affect their perceptions in another subject. In addition, Arens and Möller (2016) examined dimensional comparison effects between students' math and language grades and student ratings of student-teacher relationships and instructional quality in both subjects. Consistent with the model of dimensional comparisons, they found that students with higher grades in math and language provided more positive ratings of teaching quality in the corresponding subject but more negative ratings in the other subject.

However, there are important issues that were not considered in this study. First, the study addressed only the students' individual perspectives of teaching quality. Technically, student ratings of teaching quality contain several different components that should be considered separately (Lüdtke, Robitzsch, Trautwein, & Kunter, 2009). That is, students' teaching-quality ratings reflect both common opinions shared by all students in a classroom and an individual component reflecting students' nonshared perceptions. Combining these two components yields individual perceptions, that is, what students actually say about their teacher's teaching quality. On the other hand, however, the question of whether dimensional comparison effects exist on both of these two levels has not yet been investigated. Second, Arens and Möller (2016) examined the association only for aspects of teaching quality that belong to teachers' learning support (e.g., structuredness). Because these aspects address the perspective of the student-teacher relationship, the question of whether dimensional comparison effects will also emerge for other aspects of teaching quality have remained unanswered (Arens & Möller, 2016). Third, the authors used only school grades to operationalize students' achievement. School grades are an important source of feedback on students' achievement, and as studies have shown, dimensional comparison effects were stronger for grades than for achievement test scores (e.g.,

Köller, Trautwein, Lüdtke, & Baumert, 2006; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005). However, it is not clear whether these effects are due to actual differences in students' achievement in these subjects or to dimensional comparison processes caused by the grades the students received in different subjects. In addition, teachers typically grade on a curve, such that the best and worst students in each class tend to get the highest and lowest grades, respectively, independent of the average ability levels of the students in each class. This tendency undermines the comparability of grades across classrooms and schools and might thereby limit the use of grades for dimensional comparison effects at the classroom level. An objective form of measurement such as standardized achievement tests offers the opportunity to investigate whether the effects appear on the basis of differences in students' achievement or on the basis of dimensional comparison processes caused by grading differences.

The Present Investigation

In the present study, we sought to address these major open questions about possible dimensional comparison processes and student ratings of teaching quality. On the basis of DCT, in this study we investigated that students' dimensional comparisons of their achievements in different subjects would have an impact on their ratings of teaching quality. Figure 1 depicts the model that we tested in the present study. In accordance with the pattern from DCT, the paths from grades to teaching-quality ratings within subjects represent positive associations at the student and classroom levels. The dotted paths reflect possible dimensional comparison effects between subjects.

For this, we used a recently collected large data set containing students' ratings of teachers' performances in a variety of teaching-quality domains and dimensions in mathematics and the German language (see Table 1). We used students' grades and their achievement measured with standardized test scores to predict students' teaching-quality ratings in the two subjects. In addition, we examined dimensional comparison effects at both the student and classroom levels in order to find out whether dimensional comparison effects are primarily the result of individual perceptions or also affect students' shared perceptions at the classroom level. We derived the following three hypotheses as explained below:

1. Previous studies have consistently found positive associations between students' grades and teaching-quality ratings in the same subject. However, teaching quality

Table 1

Descriptions of the Teaching-Quality Domains and Dimensions

Dimension	Description	Source
Classroom management		
Monitoring	Refers to how well the teacher has an eye for what is going on in the classroom and anticipates possible disturbances (4 items).	Adapted from Baumert, Gruehn, Heyn, Köller, and Schnabel (1997)
Disturbances	The degree to which the beginning of the class is delayed and the class is disrupted (6 items).	Adapted from Kunter et al. (2002) Adapted from Baumert et al. (1997)
Structuredness	Reflects the planning and structure of a lesson and the clarity of what the lesson will be about (4 items).	Bos, Gröhlich, Dudas, Guill, and Scharenberg (2010) Kunter et al. (2002) Adapted from Rakoczy, Buff, and Lipowsky (2005)
Rule clarity	Students know exactly what rules apply and what happens if they do not follow them (3 items).	Baumert et al. (1997)
Supportive climate		
Handling mistakes	Refers to how the teacher deals with students' mistakes and is able to ensure that no one is laughed at for making a mistake (3 items).	Ramm et al. (2006)
Differentiation	The teacher adapts questions and tasks to the students' achievement levels (4 items).	Baumert et al. (1997)

Table 1 (continued)

Motivation	Encompasses teachers' ability to motivate students, to point out the usefulness of the lesson, and to keep the lessons varied (4 items).	Baumert et al. (1997) Adapted from Wagner, Helmke, and Rösner (2009)
General support	The degree to which the teacher supports students in their learning processes (3 items).	Kunter et al. (2002) Ramm et al. (2006)
Clarity	Refers to teachers' ability to explain comprehensibly and at an appropriate speed (4 items).	Adapted from Baumert et al. (2009) Ramm et al. (2006) Baumert et al. (1997)
Autonomy support	Measures how much the teacher supports and encourages students in learning autonomously (4 items).	Based on Rakoczy et al. (2005)
Feedback	Encompasses how students receive feedback on their current achievement level and how they can improve (4 items).	Adapted from Klimczak et al. (2012)
Expectations	Reflects teachers' expectations of students and whether they can meet them (3 items).	Ditton (2000)
Cognitive activation		
Challenging tasks	The provision of challenging tasks that fosters students to think about solution processes and to apply what they have learned (4 items).	Based on Ramm et al. (2006) Based on Baumert et al. (1997) Based on Ramm et al. (2006); Mang et al. (2018)

Table 1 (continued)

Regularity	Teachers' requests for students to find regularities and distinctions in what they have learned and to transfer the rules they have learned to other concepts (3 items).	Self-development
Practicing	How important practicing is to the teacher and the variety with which he or she organizes the exercises (4 items).	Based on Baumert et al. (1997); Ramm et al. (2006); Rakoczy, Buff, & Lipowsky (2005)
Socratic dialogue	The degrees to which the teacher lets students think for themselves, lets them explore false assumptions until they recognize their own mistakes, and asks the students to justify their answers (4 items).	Baumert et al. (1997) Based on Baumert et al. (2009)

was limited to dimensions of teacher support and to the student level. We aimed to replicate these within-subject associations and extend them by investigating student ratings of classroom management and cognitive activation at the student and classroom levels. We hypothesized that students' grade in one subject is positively associated with their teaching-quality ratings in the same subject on both levels and hold for both the three broad teaching-quality domains and the teaching-quality dimensions (Hypothesis 1).

2. To date, the effects of student grades in one subject on student ratings of teaching quality in another subject have been investigated only for a limited number of teaching-quality dimensions and only on the student level. We sought to address this gap in research and extended the investigation of dimensional comparison effects between subjects at the student and classroom levels. We hypothesized that students' grade in one subject is negatively related to their ratings of teaching-quality domains and dimensions in the other subject on both levels (Hypothesis 2).
3. Because student grades may contain information other than students' achievement, it is not clear whether the dimensional comparison effects are due to differences in students' achievement or to dimensional comparison processes. To examine this, we controlled for standardized achievement test scores in both subjects. We hypothesized that the dimensional comparison effects are more pronounced for student grades than for the achievement test scores and remain stable after we controlled for the achievement test scores (Hypothesis 3).

Method

The ethical review of the study "Teaching quality from the students' perspective (UNITAS)," which took place in 2018, was composed of a two-stage procedure. All sampling procedures and materials were reviewed and approved by the ministry of education and cultural affairs in the state of Baden-Württemberg (file number 31-6600.0/279). Afterwards, the ethics committee from the faculty of Economics and Social Sciences at the University of Tübingen (file number A2.5.4-074_aa) gave its approval.

Sample

The analyses drew on data from the large "Teaching quality from the students' perspective" (UNITAS) study carried out in the German federal state of Baden-Württemberg in spring/summer 2018. UNITAS is about the validity of students' teaching-quality ratings and about the prospective relations between teaching quality as rated by teachers and students for predicting learning over time. A total of 27 schools volunteered to participate, and 401 classes and 6,479 students from Grades 5 to 10 were assessed. On average, 16 students per class provided ratings of their mathematics and German-language teachers' teaching quality. In Germany, a common school system in the federal states consists of four tracks of secondary schools: academic track (Gymnasium), intermediate track (Realschule), lower academic track (Hauptschule), and multitrack (Gemeinschaftsschule) schools. German students typically remain with the same group of classmates throughout secondary school, but they have different teachers in different subjects who may change from year to year. Students came from 11 academic track schools (n = 3,847), eight intermediate track schools (n = 1,837), seven multitrack schools (n = 715), and one lower track secondary school (n = 80). The assessments were conducted at the end of the school year. Descriptive statistics are shown in Table 2.

Instruments

Student ratings

Student ratings of teaching quality were assessed along the three domains of teaching quality (classroom management, supportive climate, cognitive activation), and each domain contained four to eight dimensions (yielding a total of 16 quality dimensions) assessed with up to six items. In total, the questionnaire consisted of 61 items for each of the two subjects (math and German language) under investigation. The median Cronbach's alpha value was .70 for mathematics and .69 for the German language. For both subjects, we used the same item wording (e.g., "Our [math/German language] teacher always knows exactly what goes on in the classroom"). Most of the items had already been applied in large studies such as the Programme for International Student Assessment (PISA) study (Organization for Economic Cooperation and Development, 2004). One subscale (Regularity) and some single items in three out of 16 quality dimensions were self-developed, and the wording of several items was adapted. Students rated all items on a 4-point scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). Descriptive results are shown in Table 3.

Table 2

Descriptive Statistics and Sample Sizes by School Type

Grade	Total	Classes	Nonacademic	Multitrack	Intermediate	Academic
5	1,111	66	3 (50% female)	117 (55% female)	323 (51% female)	668 (51% female)
6	1,206	76	7 (60% female)	121 (41% female)	383 (50% female)	695 (48% female)
7	1,360	80	13 (33% female)	199 (51% female)	385 (52% female)	763 (50% female)
8	1,110	70	20 (71% female)	159 (48% female)	277 (43% female)	654 (50% female)
9	1,094	73	18 (41% female)	87 (53% female)	415 (54% female)	574 (53% female)
10	598	36	19 (31% female)	32 (59% female)	54 (52% female)	493 (55% female)
Sum	6,479	401	80 (39% female)	715 (49% female)	1837 (50% female)	3847 (50% female)

Table 3

Descriptive Statistics for the Teaching-Quality Domains and Dimensions for Mathematics and German Language

	Mathematics			German language		
	M	*SD*	*ICC*	*M*	*SD*	*ICC*
Classroom management						
Monitoring	2.94	0.63	.43	2.97	0.64	.40
Disturbances	2.51	0.69	.44	2.50	0.67	.40
Structuredness	2.63	0.66	.25	2.63	0.66	.25
Rule clarity	3.30	0.63	.14	3.30	0.63	.13
Supportive climate						
Handling mistakes	3.33	0.70	.25	3.33	0.68	.29
Differentiation	2.37	0.63	.28	2.14	0.64	.28
Motivation	2.66	0.79	.33	2.67	0.75	.33
General support	3.13	0.73	.27	3.09	0.71	.30
Clarity	2.86	0.76	.32	3.02	0.61	.27
Autonomy support	2.77	0.67	.23	2.72	0.65	.18
Feedback	2.51	0.75	.20	2.49	0.71	.22
Expectations	2.61	0.70	.22	2.59	0.69	.15
Cognitive activation						
Challenging tasks	3.13	0.55	.29	2.85	0.58	.18
Regularity	2.90	0.62	.13	2.68	0.64	.17
Practicing	3.00	0.62	.22	2.88	0.62	.26
Socratic dialolgue	2.91	0.67	.13	2.69	0.65	.13

Student achievement

Student grades. To assess student achievement in the two subjects, we used student-reported school grades from their last report card, which they had received after their first semester in school. As school grades in Germany range from 1 (*best grade*) to 6 (*lowest grade*), we reverse-scored the grades for easier interpretation, so a higher grade represented higher achievement.

Standardized achievement. Additionally, we administered achievement tests to assess students' ability in the German language (LGVT 5-12+) by Schneider, Schlagmüller, and Ennemoser (2017) and mathematics (MBK 5-12+) (Ennemoser, Krajewski, & Schmidt, 2011; Krajewski & Ennemoser, 2013). The LGVT 5-12+ is a test that determines the speed, accuracy, and comprehension with which students are able to process a text within 6 min. The MBK 5-12+ consists of several tasks each with limited time, such as marking numbers on a number line or number dictation. In the present study, both achievement tests revealed satisfactory internal consistencies (LGVT 5-12+: KR-20 = .98; MBK 5-12+: KR-20 = .83).

Analysis

Multilevel analyses

Because we were interested in investigating dimensional comparison effects at both the student and class levels, we applied multilevel modeling (Raudenbush & Bryk, 2002). For this, we specified multilevel latent models with items used as measurement indicators of latent factors. We postulated the same factors measured by the same indicators at both levels of analysis, whereby the modeling of the latent factors was done in two different ways: First, items from the respective teaching-quality domains were used to model the three domains (classroom management: 17 items in total; supportive climate: 29 items in total; cognitive activation: 15 items in total, see Table 1). Second, measurement indicators were used to model 16 latent teaching-quality factors with three to seven items per dimension (see Table 1).

Given that one prerequisite for multilevel regression models in the present study is the comparability of constructs across levels and subjects, we then tested for the measurement invariance of the teaching-quality domains and dimensions across the student and classroom levels as well as the two subjects. Specifically, we empirically compared a factor model with freely estimated parameter loadings in both subjects with a model that included both subjects with constrained cross-level and cross-subject factor loadings (i.e., strong measurement invariance; Stapleton, Yang, & Hancock, 2016). Due to the complexity of the model, all multilevel models were specified separately for each of the teaching-quality domains and dimensions.

Dimensional comparison model

Within the multilevel framework shown in Figure 1, we then predicted student ratings of teaching quality from students' achievement measured by their grades and their achievement test scores in the two corresponding subjects. That is, students' ratings of teaching quality in mathematics were regressed on students' achievement in the corresponding subject and students' achievement in the German language. At the same time, students' ratings of teaching quality in the German language were simultaneously regressed on students' achievement in the German language and students' achievement in mathematics. In order to fully explore potential differences between students' grades and achievement test scores, we first conducted two separate analyses for students' grades and achievement test scores before combining the grades and achievement test scores. Additionally, we included students' grade level (i.e., Grades 5 to 10) and their academic track (i.e., academic and nonacademic) as covariates in the analytical model. These analyses were conducted for both the teaching-quality domains and teaching-quality dimensions.

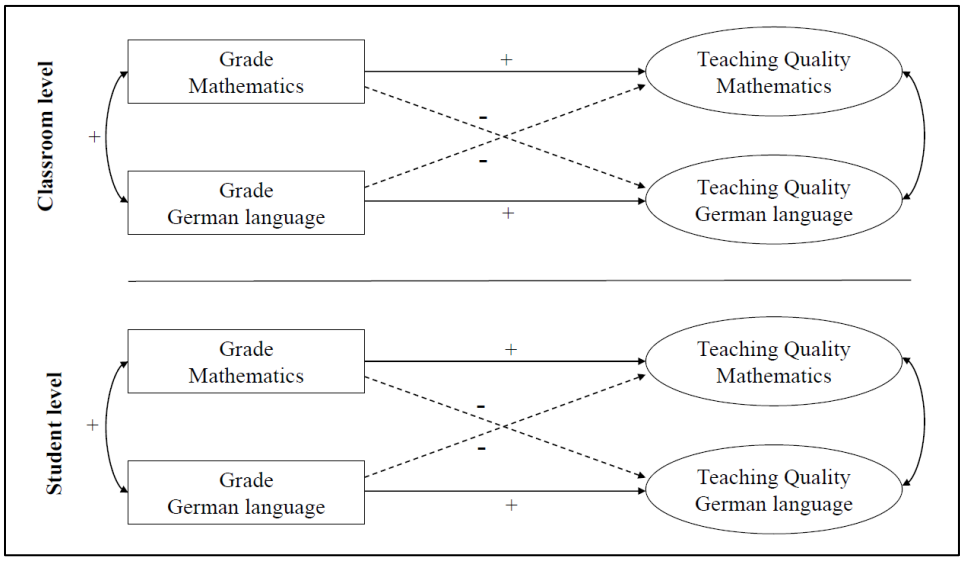


Figure 1. Expected associations within and between subjects on the student and classroom levels based on DCT (Möller & Marsh, 2013).

Goodness-of-fit indices

Goodness of fit was evaluated using the following criteria: the Comparative Fit Index (CFI), the Tucker Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA),

and the Standardized Root Mean Square Residual (SRMR) as level-specific fit indices for both the within- and between-models. For single-level CFA models, Yu (2002) proposed cutoff values close to .95, .95, .05, and .07 for the CFI, TLI, RMSEA, and SRMR, respectively. Prior to analysis, all items were checked for non-normality. The skewness and kurtosis of the data were negligible (for all indicators, skewness ranged from -0.58 to 0.30; kurtosis from -0.86 to 0.06), and therefore, (largely) unbiased model fit estimates could be expected (West, Finch, & Curran, 1995). Moreover, as a precaution, the regular Chi-Square value was adjusted by applying a scaling correction factor to avoid inflated overall goodness-of-fit test statistics (Bryant & Satorra, 2012) due to violations of the multivariate normality assumption. We conducted all analyses with the Mplus 8 software (Muthén & Muthén, 1998-2017) using the full information maximum likelihood algorithm to deal with missing values. All significance testing was performed at the .05 level. For the analyses of dimensional comparison effects, we used the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) to control the false discovery rate for multiple comparisons across the teaching-quality domains and dimensions as well as at the student and classroom levels.

Results

Factor Analysis and Measurement Invariance Models

We began by testing for the measurement invariance of teaching-quality domains. For this, we specified separate models for each of the teaching-quality domains and tested for strong measurement invariance across both levels and school subjects. The results for the unconstrained models showed an acceptable model fit for cognitive activation, $\chi^2(429) = 1730.39$, CFI = .97, TLI = .96, RMSEA = .02, $SRMR_w = .03$, $SRMR_b = .11$, whereas model fits for classroom management ($\chi^2(687) = 5980.75$, CFI = .92, TLI = .90, RMSEA = .04, $SRMR_w = .06$, $SRMR_b = .14$) and supportive climate ($\chi^2(2570) = 14096.89$, CFI = .92, TLI = .91, RMSEA = .03, $SRMR_w = .05$, $SRMR_b = .15$) were substantially lower. This was due to relatively low factor loadings for the indicators of disturbances (classroom management), and differentiation and expectations (supportive climate). For these reasons, we then specified models that excluded these indicators, resulting in improved model fits for the two teaching-quality domains (classroom management, $\chi^2(280) = 1194.40$, CFI = .98, TLI = .97, RMSEA = .02, $SRMR_w = .04$,

$SRMR_b$ = .10; supportive climate, $\chi^2(1337) = 6241.81$, CFI = .96, TLI = .95, RMSEA = .02, $SRMR_w$ = .04, $SRMR_b$ = .10).

Constraining the factor loading to be invariant across level and subjects showed acceptable model fits for classroom management, $\chi^2(298) = 1477.56$, CFI = .97, TLI = .96, RMSEA = .03, $SRMR_w$ = .04, $SRMR_b$ = .11; supportive climate, $\chi^2(1375) = 6436.50$, CFI = .96, TLI = .95, RMSEA = .02, $SRMR_w$ = .04, $SRMR_b$ = .10; and cognitive activation, $\chi^2(453) = 1897.23$, CFI = .96, TLI = .95, RMSEA = .02, $SRMR_w$ = .03, $SRMR_b$ = .11 (Table S1).

Based on the constrained model, the results revealed positive correlations for the teaching-quality domains between subjects at the student level ($.36 \leq r \leq .54$) and negative correlations at the classroom level ($-.26 \leq r \leq -.07$; Table 4). Furthermore, the student ratings of all teaching-quality dimensions varied substantially between classrooms (mathematics: $.20 \leq$ ICC $\leq .42$; German language: $.22 \leq$ ICC $\leq .39$).

Table 4

Correlations between the Teaching-Quality Domains in Mathematics and German Language on the Student and Classroom Levels

	r (SL)	*r* (CL)
Classroom management	.54***	-.07
Supportive climate	.48***	-.26**
Cognitive activation	.36***	-.07

SL = student level. CL = classroom level.

* $p < .05$. ** $p < .01$. *** $p < .001$.

After analyzing the teaching-quality domains, we tested for the measurement invariance of the teaching-quality dimensions. The fully unconstrained models for each dimension revealed good to excellent model fit (see Table S2). Only with respect to teacher differentiation did the factor model reveal a relatively lower fit for the baseline model without any model constraints, $\chi^2(34) = 370.04$, CFI = .97, TLI = .94, RMSEA = .04, $SRMR_w$ = .03, $SRMR_b$ = .11. Constraining the factor loadings to be invariant across levels and subjects did not result in a significantly lower model fit for 15 of the 16 dimensions (see Table S2), and the changes were in the range proposed by Cheung and Rensvold (2002). Again, teacher differentiation revealed a relatively lower model

fit, $\chi^2(44) = 469.85$, CFI = .96, TLI = .94, RMSEA = .04, $SRMR_w$ = .03, $SRMR_b$ = .16, and therefore, this dimension was excluded from all further analyses.

For all of the remaining teaching-quality dimension, the results provided empirical support for the comparability of measures at the student and classroom levels as well as across the two subjects. On the basis of the final invariance models, Table 3 presents the descriptive results for the teaching-quality dimensions. In addition, the results showed positive correlations for teaching quality dimensions between subjects at the student level ($.23 \leq r \leq .71$) and negative as well as positive correlations at the classroom level ($-.37 \leq r \leq .55$; Table 5). Furthermore, the

Table 5

Correlations between the Teaching-Quality Dimensions in Mathematics and German Language on the Student and Classroom Levels

	r (SL)	*r* (CL)
Monitoring	.39***	-.17*
Disturbances	.59***	.24**
Structuredness	.64***	.30***
Rule clarity	.71***	.15
Handling mistakes	.41***	-.05
Motivation	.26***	-.01
General support	.37***	-.03
Clarity	.23***	-.37***
Autonomy support	.44***	.06
Feedback	.63***	.29***
Expectations	.69***	.55***
Challenging tasks	.40***	-.23**
Regularity	.58***	-.13
Practicing	.42***	-.13
Socratic dialogue	.49***	-.14

SL = student level. CL = classroom level.

* $p < .05$. ** $p < .01$. *** $p < .001$.

student ratings of all teaching-quality dimensions varied substantially between classrooms (mathematics: $.13 \leq ICC \leq .44$; German language: $.13 \leq ICC \leq .40$).

Within-Subject Associations between Grades and Students' Teaching-Quality Ratings

To address our three hypotheses, we applied multilevel regression models predicting students' teaching-quality ratings in the two subjects from their grades. With regard to our first hypothesis, we expected a positive association between students' grades in the two subjects and the corresponding teaching-quality ratings (Hypothesis 1). As predicted, for all teaching-quality domains, we found positive within-subject associations at the student level showing that higher achieving students within classrooms reported higher teaching quality[5] (mathematics: $0.10 \leq \beta \leq 0.23$, *p*s < .05; German language: $0.14 \leq \beta \leq 0.20$, *p*s < .05; Table 6).

At the classroom level, five out of six domains showed statistically significant associations (mathematics: $0.19 \leq \beta \leq 0.37$; *p*s < .05; German language: $\beta = 0.15$ and $\beta = 0.22$; *p*s < .05; Table 6). Classrooms with higher grades in mathematics and the German language provided more positive teaching-quality ratings within the same subject. Only for students' grades in German language regressed on classroom management in German language showed statistically nonsignificant associations on the classroom level ($\beta = 0.02$, *ns*).

The results for the teaching-quality dimensions at the student level showed a similar pattern (mathematics: $-0.12 \leq \beta \leq 0.37$; *p*s < .05; German language: $-0.09 \leq \beta \leq 0.27$; *p*s < .05; Table 7). The association was not statistically significant only for structuredness in mathematics ($\beta = 0.01$, *ns*), and the results for expectations revealed negative associations in both subjects (mathematics: $\beta = -0.09$, $p < .05$; German language: $\beta = -0.10$, $p < .05$).

Similar results for the associations between students' grades and the teaching-quality dimensions were found at the classroom level. In mathematics, only four of the 15 dimensions showed a statistically non-significant association (disturbances, structuredness, rule clarity, expectations) between students' grades and their teaching-quality ratings (Table 7). For all other dimensions, the associations between students' grades and teaching-quality ratings in mathematics were

[5] As a robustness check, we ran additional models in which we also included the items from the Disturbances, Differentiation, and Expectations dimensions when modelling the teaching-quality domains. Whereas model fit decreased, we could not identify notable deviations from the substantive findings from models that were based on the reduced number of indicators. The absolute differences between the regression coefficients were smaller than 0.02 at the student and classroom levels. All statistically significant results for the teaching-quality domains remained unchanged.

statistically significant and ranged from 0.19 to 0.41, *p*s < .05. For the German language, eight of the 15 dimensions showed statistically significant associations with teaching-quality ratings ($0.17 \leq \beta \leq 0.29$; *p*s < .05). Again, for expectations, we found negative associations with student grades ($\beta = -0.37$, $p < .05$). For monitoring, disturbances, structuredness, rule clarity, feedback, regularity, and Socratic dialogue, the results revealed no statistically significant associations (-0.05 to 0.13, *ns*).

Between-Subject Associations between Grades and Students' Teaching-Quality Ratings

To address our second hypothesis, we next regressed students' teaching-quality ratings in one subject on their grades in the other subject. In line with Hypothesis 2, the analyses at the student level revealed negative associations between students' grades in one subject and their teaching-quality ratings in the other subject. For teaching-quality domains the results showed negative associations between students' grades in one subject and their teaching-quality ratings in another subject ($-0.07 \leq \beta \leq -0.04$; *p*s < .05; Table 6). Only for students' grades in the German language and their teaching-quality ratings in mathematics for classroom management and cognitive activation were no statistically significant negative associations.

On the classroom level, for all teaching-quality domains, we found statistically significant negative associations in both subjects (mathematics: $-0.20 \leq \beta \leq -0.18$; *p*s < .05; German language: $-0.33 \leq \beta \leq -0.21$; *p*s < .05; Table 6) showing that classes with higher grades in one subject gave relatively lower teaching-quality ratings in the other subject.

For teaching-quality dimensions at the student level, eight of the 15 quality dimensions showed statistically significant negative associations between students' grades in mathematics and their teaching-quality ratings in the German language ($-0.13 \leq \beta \leq -0.04$; *p*s < .05; see Table 7). Furthermore, we found negative associations between students' grades in their German language classes and their ratings of teaching quality in mathematics for three dimensions (motivation, clarity, feedback; -0.07 to -0.05, *p*s < .05). All other dimensions revealed statistically nonsignificant results.

Table 6

Student Grades in Mathematics and German Language Predicting the Teaching-Quality Domains within and between Subjects on the Student and Classroom Levels

	Within subjects		*Between subjects*	
	M, Grade → M, TQ	G, Grade → G, TQ	M, Grade → G, TQ	G, Grade → M, TQ
	β (SL/CL)	β (SL/CL)	β (SL/CL)	β (SL/CL)
Classroom management	**0.10** / **0.19**	**0.14** / 0.02	**-0.07** / **-0.18**	0.00 / **-0.21**
Supportive climate	**0.23** / **0.33**	**0.20** / **0.22**	**-0.05** / **-0.20**	**-0.04** / **-0.26**
Cognitive activation	**0.17** / **0.37**	**0.17** / **0.15**	**-0.07** / **-0.19**	0.00 / **-0.33**

Note. M = mathematics; G = German language. TQ = teaching quality. β = standardized regression coefficients. SL = student level. CL = classroom level. The Benjamini-Hochberg correction was used to control the false-positive rate due to multiple tests (Benjamini & Hochberg, 1995). Statistically significant results ($p < .05$) after correction are presented in bold.

Table 7

Student Grades in Mathematics and German Language Predicting the Teaching-Quality Dimensions within and between Subjects on the Student and Classroom Levels

	Within subjects		*Between subjects*	
	M, Grade → M, TQ	G, Grade → G, TQ	M, Grade → G, TQ	G, Grade → M, TQ
	β (SL/CL)	β (SL/CL)	β (SL/CL)	β (SL/CL)
Monitoring	**0.12** / **0.19**	**0.10** / 0.05	-0.04 / **-0.17**	-0.01 / **-0.20**
Disturbances	**-0.12** / -0.11	**-0.09** / -0.05	-0.01 / 0.08	0.00 / 0.05
Structuredness	0.01 / 0.17	**0.13** / 0.13	**-0.13** / **-0.23**	0.01 / **-0.22**
Rule clarity	**0.08** / 0.15	**0.08** / 0.04	0.01 / -0.11	0.02 / -0.09
Handling mistakes	**0.15** / **0.30**	**0.17** / **0.29**	-0.01 / **-0.16**	0.00 / -0.09
Motivation	**0.23** / **0.36**	**0.20** / **0.22**	**-0.10** / **-0.17**	**-0.07** / **-0.24**
General support	**0.14** / **0.31**	**0.13** / **0.17**	-0.03 / **-0.19**	-0.02 / **-0.23**
Clarity	**0.37** / **0.41**	**0.27** / **0.28**	**-0.05** / **-0.27**	**-0.07** / **-0.30**
Autonomy support	**0.21** / **0.35**	**0.21** /**0.22**	**-0.08** / **-0.19**	-0.02 / **-0.25**
Feedback	**0.09** / **0.23**	**0.05** / 0.04	-0.01 / **-0.19**	**-0.05** / **-0.31**
Expectations	**-0.09** / -0.10	**-0.10** / **-0.37**	0.00 / 0.09	-0.03 / **-0.27**
Challenging tasks	**0.19** / **0.28**	**0.19** / **0.19**	**-0.06** / **-0.17**	**0.07** / **-0.29**
Regularity	**0.13** / **0.32**	**0.14** / 0.12	**-0.07** / **-0.19**	0.02 / **-0.30**
Practicing	**0.18** / **0.35**	**0.14** / **0.17**	**-0.08 / -0.18**	-0.03 / **-0.31**
Socratic dialogue	**0.06** / **0.24**	**0.07** / -0.02	**-0.04** / -0.13	-0.01 / **-0.32**

Note. M = mathematics; G = German language. TQ = teaching quality. β = standardized regression coefficients. SL = student level. CL = classroom level. The Benjamini-Hochberg correction was used to control the false-positive rate due to multiple tests (Benjamini & Hochberg, 1995). Statistically significant results ($p < .05$) after correction are presented in bold.

At the classroom level, we found negative associations between students' grades in mathematics and teaching quality in German language classes for 11 dimensions ($-0.27 \leq \beta \leq -0.16$; ps < .05; Table 7), indicating that classes with better grades in mathematics gave lower ratings to their German-language teacher. In the same vein, classrooms' average grade in the German language revealed negative associations with students' teaching-quality ratings in mathematics (-0.32 to -0.20, ps < .05). Only three of the 15 dimensions showed statistically nonsignificant associations (i.e., disturbances, handling mistakes, and expectations; see Table 7.)

Controlling for Students' Achievement Test Scores

To test whether the associations would remain statistically significant after we controlled for students' achievement measured as achievement test scores (Hypothesis 3), we included students' achievement test scores in the two subjects as additional predictors of students' teaching-quality ratings. For the three teaching-quality domains, the results remained unchanged within (student level: $0.11 \leq \beta \leq 0.23$, ps < .05; classroom level: $0.16 \leq \beta \leq 0.32$, ps < .05) and between subjects (student level: $-0.07 \leq \beta \leq -0.04$, ps < .05; classroom level: $-0.16 \leq \beta \leq -0.34$, ps < .05) after including achievement test scores (Table 8).

Similarly, the findings for the teaching-quality dimensions on the student level remained largely unchanged (Table 9). As the only exception, the association between mathematics grades and student ratings of Socratic dialogue in mathematics was no longer statistically significant. On the classroom level, within-subject associations remained unchanged, whereas between-subject associations between mathematics grades and motivation, autonomy support, feedback, challenging tasks, and regularity in the German language were no longer statistically significant. All other negative associations remained unchanged after including achievement test scores.

Table 8

Student Grades and Achievement in Mathematics and German Language Predicting the Teaching-Quality Domains within and between Subjects on the Student and Classroom Levels

	Within subjects		*Between subjects*	
	M, Grade → M, TQ	G, Grade → G, TQ	M, Grade → G, TQ	G, Grade → M, TQ
	β (SL/CL)	β (SL/CL)	β (SL/CL)	β (SL/CL)
Classroom management	**0.11** / **0.18**	**0.14** / 0.06	**-0.07** / **-0.16**	0.01 / **-0.20**
Supportive climate	**0.23** / **0.32**	**0.20** / **0.23**	**-0.04** / **-0.17**	**-0.04** / **-0.25**
Cognitive activation	**0.15** / **0.31**	**0.16** / **0.16**	**-0.06** / **-0.16**	0.00 / **-0.34**

Note. M = mathematics; G = German language. TQ = teaching quality. β = standardized regression coefficients. SL = student level. CL = classroom level. The Benjamini-Hochberg correction was used to control the false-positive rate due to multiple tests (Benjamini & Hochberg, 1995). Statistically significant results ($p < .05$) after correction are presented in bold.

Table 9

Student Grades and Achievement in Mathematics Predicting the Teaching-Quality Dimensions in Mathematics and German Language on the Student and Classroom Levels

	M, Grade → M, TQ	G, Grade → G, TQ	M, Grade → G, TQ	G, Grade → M, TQ
	β (SL/CL)	β (SL/CL)	β (SL/CL)	β (SL/CL)
Monitoring	**0.12** / **0.18**	**0.11** / 0.07	-0.04 / **-0.16**	-0.00 / **-0.20**
Disturbances	**-0.11** / -0.08	**-0.09** / -0.04	-0.02 / 0.09	0.00 / 0.08
Structuredness	0.01 / **0.19**	**0.13** / **0.16**	**-0.12** / **-0.20**	0.02 / **-0.20**
Rule clarity	**0.09** / 0.14	**0.07** / 0.02	0.01 / -0.09	0.01 / -0.11
Handling mistakes	**0.15** / **0.29**	**0.15** / **0.27**	-0.01 / **-0.14**	-0.00 / -0.11
Motivation	**0.23** / **0.36**	**0.21** / **0.24**	**-0.08** / -0.14	**-0.06** / **-0.23**
General support	**0.14** / **0.29**	**0.13** / **0.19**	-0.02 / **-0.17**	-0.01 / **-0.23**
Clarity	**0.34** / **0.40**	**0.25** / **0.28**	**-0.05** / **-0.24**	**-0.08** / **-0.30**
Autonomy support	**0.20** / **0.34**	**0.20** / **0.24**	**-0.07** / -0.15	-0.02 / **-0.24**
Feedback	**0.08** / **0.23**	**0.05** / 0.07	-0.01 / -0.16	**-0.05** / **-0.29**
Expectations	**-0.10** / -0.12	**-0.08** /**-0.33**	-0.02 / 0.08	-0.01 / **-0.23**
Challenging tasks	**0.17** / **0.25**	**0.17** / **0.21**	**-0.06** / -0.15	0.05 / **-0.30**
Regularity	**0.11** / **0.28**	**0.13** / 0.12	**-0.05** / -0.16	0.00 / **-0.30**
Practicing	**0.17** / **0.32**	**0.15** /**0.19**	**-0.07** / **-0.16**	-0.04 / **-0.31**
Socratic dialogue	0.04 / 0. 21	**0.06** / -0.00	**-0.05** / -0.12	-0.02 / **-0.33**

Note. M = mathematics; G = German language. TQ = teaching quality. β = standardized regression coefficients. SL = student level. CL = classroom level. The Benjamini-Hochberg correction was used to control the false-positive rate due to multiple tests (Benjamini & Hochberg, 1995). Statistically significant results ($p < .05$) after correction are presented in bold.

Discussion

Whereas the existence of dimensional comparison processes has been found in numerous studies on students' academic self-concept, a few studies have also provided empirical support for dimensional comparison processes that have applied to student characteristics other than academic self-concept. The aim of the present study was to investigate possible dimensional comparison effects of student grades on their ratings of teaching quality. Specifically, we looked at the associations between grades in both mathematics and the German language and student ratings of three teaching-quality domains and 15 teaching-quality dimensions at both the student and classroom levels. To investigate whether possible dimensional comparison effects occur because of students' achievement, we controlled for students' achievement by using standardized achievement test scores. In general, the results supported the pattern of the dimensional comparison model for most of the teaching-quality dimensions on both levels and could also be found for the three basic dimensions of teaching quality.

First, in line with Hypothesis 1, for all teaching-quality domains and a large number of teaching-quality dimensions, analyses on the student level revealed positive associations between student grades and teaching-quality ratings in both subjects. For most of the domains and dimensions, we also found negative associations between subjects, showing that a relatively higher grade in one subject resulted in a lower rating of the other subject. Second, we found the same pattern and even stronger associations on the classroom level within and between subjects for most of the dimensions and domains (Hypothesis 2). Between subjects, the results were stronger for student grades in the German language and student ratings in mathematics for both the teaching-quality domains and dimensions. Third, most of the results remained statistically significant after we controlled for standardized achievement test scores in both subjects and on both levels (Hypothesis 3).

Empirical Support for Dimensional Comparison Theory

Our findings support the predictions of DCT (Möller et al., 2015) and show that student ratings of teaching quality in one subject are impacted by the grading practices of a teacher in another subject. This is true for individual students as well as for whole classrooms. Student ratings are a valuable source for assessing students' perceptions of teaching quality. They have been shown to be reliable and valid and are economical because many students can be reached in

a brief amount of time (De Jong & Westerhof, 2001). However, studies have also shown that student ratings might be biased, for instance, due to a teacher's leniency in grading: Teachers who give higher grades in general have been found to receive more positive feedback on their teaching quality (Greenwald & Gillmore, 1997a; Griffin, 2004).

Biases in student ratings may also be the result of comparison processes within and between subjects. In this study, we aimed to link research on DCT with research on teaching quality. The results of the within- and between-subject analyses that extended the work from previous studies to include the teaching-quality domains and 15 teaching-quality dimensions showed good fits and confirmed the assumptions of DCT and findings from previous studies (see Arens & Möller, 2016, for the student level) because we found positive within-subject and negative between-subject associations on both levels for most domains and dimensions. Within subjects, the results showed that individuals and whole classes with higher grades also gave higher ratings of teaching quality. These positive associations seem plausible because positive within-subject associations have been found on the student level in a number of studies on students' achievement using an academic outcome (e.g., Marsh et al., 2018) and because teachers apply their grading and teaching practices to the whole class and not only to single students.

Between subjects, the negative associations showed that students and classes with higher grades in one school subject than in the other school subject gave relatively lower ratings of teaching quality in the school subject in which they had the lower grades. In the context of biasing factors such as grading leniency, the results provide evidence for the impact of students' classroom experiences in one subject for their teaching-quality ratings in another subject and that ratings might have a social component (Benton & Cashin, 2012). Teachers receive better teaching-quality ratings by grading their students higher than their colleagues do. Furthermore, teachers can negatively affect their colleagues' ratings by giving out higher grades. Therefore, teaching-quality ratings also seem to depend on the grading practices of colleagues and not only on a teacher's own teaching behavior.

Finally, by controlling for standardized achievement test scores, we tested whether dimensional comparison effects emerged from differences in students' achievement or due to teachers' different grading practices while controlling for achievement test scores. For the larger parts of the domains and dimensions, the associations remained positive within subjects and negative between subjects. This shows that the dimensional comparison effects emerged because of students' comparisons of the grades they received in both subjects and not due to comparisons

of their actual competences in the two subjects. These findings are in line with previous studies (e.g., Möller, Pohlmann, Köller, & Marsh, 2009) because, for students, grades are the criterion that is relevant for their school careers, and grades are ubiquitous in daily school life. The results were stronger on the classroom level than on the student level and for the effect of student grades in the German language on their teaching-quality ratings in mathematics. Subject differences in the perceived importance and significance might provide a potential explanation. Students may feel that mathematics is a more crucial and more suppositional subject than the German language, and therefore, they might be more willing to depreciate the quality of teaching in mathematics when they get higher grades in their German language classes than the other way around. In addition, the differences we found might be explained by different modes of instruction in the two subjects. Studies have found that teaching in mathematics is more sequential and structured than in social science lessons (Grossmann & Stodolsky, 1995; Stodolsky, 1988). This could also apply to language lessons, which could be made more varied through, for example, role-plays, discussions, or creative writing. Students might feel more actively involved and more responsible for their achievement in rather student-centered German language lessons and might subsequently be less likely to attribute their achievement to the teacher. Hence, in the present study, higher or lower grades in mathematics might have shown less pronounced dimensional comparison effects on students' teaching quality perceptions in German language classes.

Limitations and Future Research

Our study addresses important open questions concerning possible dimensional comparison effects between student grades and teaching-quality ratings on the student and classroom levels, including the impact of achievement in terms of test scores. We assessed a large number of students in different grades and school types and showed dimensional comparison effects for teaching-quality domains as well as a variety of teaching-quality dimensions. Nevertheless, this study has some limitations that need to be considered in future research. First, for reasons of effectiveness, we had to limit ourselves to two subjects: mathematics and the German language. These subjects are very different, and further research is necessary to investigate whether and to what extent the DCT also applies to more similar subjects. The GI/E model (Möller et al., 2015) proposes that dimensional comparison processes between more similar subjects such as mathematics/physics or English/French might not

necessarily lead to the negative associations but to positive assimilation effects which might also be the case for student ratings of teaching quality.

Second, by including standardized achievement tests as supplements of students' reported grades, we took special care to ensure that the findings were dimensional comparison effects and did not result from differences between students' actual achievements in the two subjects. However, both tests have some restrictions because they were applicable to all grade levels and school types; in other words, they were not specifically tailored to the curriculum in one specific school year. In future research, more specific tests should be used to provide more precise achievement scores.

Third, to further shed light on the "grading" effect, future research might also include additional measures of teaching quality (e.g., observer ratings). Such additional measures would allow researchers to study associations between teaching quality as rated by observers and grading policies. Furthermore, such observer reports could also be included in the regression models, thus allowing for the identification of the net effect of grading, after any potential differences in observer-rated teaching quality have been removed.

Conclusion

The present study investigated whether and to what extent students' grades in two subjects (mathematics and the German language) are related to student ratings of teaching quality. In line with DCT, we found that a higher grade in one subject than in the other subject was associated with a more positive rating in the same subject but a less favorable rating in the other subject. Most importantly, the same results were found at the student and classroom levels and even after we controlled for students' achievement in the form of test scores. That is, student ratings of teaching quality are impacted not only by the grading of the teacher they are evaluating but also by the grading of a teacher of another subject. These effects need to be considered in evaluations of teachers and their professional development (e.g., when student ratings are used to provide feedback to teachers and to improve their instructional practices). Future research should further examine whether subject differences (e.g., in terms of the prevalent modes of learning) are associated with the extent to which dimensional comparison effects in students' perceptions of teaching quality exist.

References

Abrami, P. C. (1989). How should we use student ratings to evaluate teaching? *Research in Higher Education, 30*, 221–227. http://doi.org/10.1007/BF00992718

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*, 219–231. http://doi.org/10.1037/0022-0663.82.2.219

Aldrup, K., Klusmann, U., Lüdtke, O., Göllner, R., & Trautwein, U. (2018). Social support and classroom management are related to secondary students' general school adjustment: A multilevel structural equation model using student and teacher ratings. *Journal of Educational Psychology, 110*, 1066–1083. http://doi.org/10.1037/edu0000256

Arens, A. K., & Möller, J. (2016). Dimensional comparisons in students' perceptions of the learning environment. *Learning and Instruction, 42,* 22–30. http://doi.org/10.1016/j.learninstruc.2015.11.001

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., . . . Tsai, Y. (2009). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente* [Professional competence of teachers, cognitively activating instruction, and development of students' mathematical literacy (COACTIV): Documentation of the instruments]. Berlin, Germany: Max-Planck-Institut für Bildungsforschung.

Baumert, J., Gruehn, S., Heyn, S., Köller, O., & Schnabel, K.-U. (1997). *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU). Dokumentation – Band 1. Skalen Längsschnitt I.* [Educational and psychosocial development in adolescence. Documentation – Volume I, scales, longitudinal data collection I]. Berlin: Max-Planck-Institut für Bildungsforschung.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B. 57,* 289–300.

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature*. IDEA Paper No. 50. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Bos, W., Gröhlich, C., Dudas, D., Guill, K., & Scharenberg, K. (2010). *KESS 8 - Skalenhandbuch zur Dokumentation der Erhebungsinstrumente* [KESS 8 - Documentation of the instruments]. Münster: Waxmann.

Brophy, J. (2000). Teaching. *Educational Practices Series, 1.* Brüssel: International Academy of Education (IAE).

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi- square testing. *Structural Equation Modeling: A Multidisciplinary Journal, 19*, 372–398. http://doi.org/10.1080/10705511.2012.687671

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5

De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4,* 51–85. http//doi.org/10.1023/A:1011402608575

Den Brok, P., Brekelmans, M., & Wubbels, T. (2004). Interpersonal teacher behavior and student outcomes. *School Effectiveness and School Improvement*, *15*, 407–442. http://doi.org/10.1080/09243450512331383262

Dietrich, J., Dicke, A.-L., Kracke, B., & Noack, P. (2015). Teacher support and its influence on students' intrinsic value and effort: Dimensional comparison effects across subjects. *Learning and Instruction, 39*, 45–54. http://doi.org/10.1016/j.learninstruc.2015.05.007

Ditton, H. (2000). *„Qualität von Schule und Unterricht"- QuaSSU Skalenbildung Hauptuntersuchung* [Quality of school and teaching – QuaSSU scales development main assessment]. Retrieved from http://daqs.fachportal-paedagogik.de/search/show/instrument/3555_43 on July 18, 2019.

Doyle, W. (2013). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management* (pp. 107–136). London: Routledge.

Ehm, J.-H., Lindberg, S., & Hasselhorn, M. (2014). Reading, writing, and math self-concept in elementary school children: influence of dimensional comparison processes. *European Journal of Psychology of Education, 29,* 277–294. http://doi.org/10.1007/s10212-013-0198-x

Ennemoser, M., Krajewski, K., & Schmidt, S. (2011). Entwicklung und Bedeutung von Mengen-Zahlen-Kompetenzen und eines basalen Konventions-und Regelwissens in den Klassen 5 bis 9 [Development and importance of sets and numbers competencies and of a basic knowledge of conventions and rules in classes 5 through 9]. *Zeitschrift für*

Entwicklungspsychologie und Pädagogische Psychologie, 43, 228–242. http://doi.org/10.1026/0049-8637/a000055

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. http://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff, J., Lüdtke, O.,… Trautwein, U. (2019). Don't blame the teacher? The need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology*. Advanced online publication. http://doi.org/10.1037/edu0000416

Goetz, T., Frenzel, A. C., Hall, N. C., & Pekrun, R. (2008). Antecedents of academic emotions: Testing the internal/external frame of reference model for academic enjoyment. *Contemporary Educational Psychology, 33*, 9–33. http://doi.org/10.1016/j.cedpsych.2006.12.002

Göllner, R., Fauth, B., Lenske, G., Praetorius, A.K., & Wagner, W. (in press). Do student ratings of classroom management tell us more about teachers or about classroom composition? *Zeitschrift für Pädagogik*.

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

Greenwald, A. G., & Gillmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209–1217. http://dx.doi.org/10.1037/0003-066X.52.11.1209

Greenwald, A. G., & Gillmore, G. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*, 743–751.

Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology, 29,* 410–425. http://doi.org/10.1016/j.cedpsych.2003.11.001

Grossman, P., & Stodolsky, S. (1995). Content as context: the role of school subjects in secondary teaching. *Educational Researcher, 24*, 5–11. https://doi.org/10.3102/0013189X024008005

Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. Meece & J. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 25– 41). New York: Routledge.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Klimczak, M., Kampa, M., Bürgermeister, A., Harks, B., Rakoczy, K., Besser, M.,…, Leiss, D. (2012). *Dokumentation der Befragungsinstrumente der Interventionsstudie im Projekt "Conditions and Consequences of Classroom Assessment"*(Co²CA) [Documentation of the intervention study questionnaires in the project "Conditions an Consequences of Classroom Assessment"]. Frankfurt am Main: DIPF.

Köller, O., Trautwein, U., Lüdtke, O., & Baumert, J. (2006). Zum Zusammenspiel von schulischer Leistung, Selbstkonzept und Interesse in der gymnasialen Oberstufe [On the interplay of academic achievement, self-concept, and interest in upper secondary schools]. *Zeitschrift für Pädagogische Psychologie, 20*, 27−39. http://doi.org/10.1024/1010-0652.20.12.27

Krajewski, K., & Ennemoser, M. (2013). Entwicklung und Diagnostik der Zahl-Grössen-Verknüpfung zwischen 3 und 8 Jahren [Development and diagnostics of the linkage of numbers and quantities between ages 3 and 8]. In M. Hasselhorn, A. Heinze, W. Schneider, & U. Trautwein (Eds.), *Diagnostik mathematischer Kompetenzen. Jahrbuch der pädagogisch-psychologischen Diagnostik. Tests und Trends* (pp. 41–65). Göttingen, Germany: Hogrefe.

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. http://doi.org/10.1007/s10984-006-9015-7

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Eds.). (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project*. Dordrecht: Springer.

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*, 494–509. http://doi.org/10.1016/j.learninstruc.2007.09.002

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*, 805–820. http://doi.org/10.1037/a0032583

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M.,… Stanat, P. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente* [PISA 2000: Documentation of Scales]. Berlin: Heenemann GmbH & Co.

Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV: A multicriteria analysis. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers – results from the COACTIV project* (pp. 97–124). New York: Springer. http://doi.org/10.1007/978-1-4614-5149-5_6

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. http://doi.org/10.1016/j.learninstruc.2008.11.001

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. http://doi.org/10.1016/j.cedpsych.2008.11.002

Mainhard, T., Oudman, S., Hornstra, L., Bosker, R. J., & Goetz, T. (2018). Student emotions in class: The relative importance of teachers and their interpersonal relations with students. *Learning and Instruction*, *53*, 109–119. http://doi.org/10.1016/j.learninstruc.2017.07.011

Mang, J., Ustjanzew, N., Schiepe-Tiska, A., Prenzel, M., Sälzer, C., Müller, K., & Rodríguez, E. G. (2018). *PISA 2012 Skalenhandbuch. Dokumentation der Erhebungsinstrumente* [PISA 2012: Documentation of the instruments]. Münster: Waxmann.

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23,* 129–149. http://doi.org/10.3102/00028312023001129

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79*, 280–295. http://doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W. (1988). The content specificity of math and English anxieties: the high school and beyond study. *Anxiety Research, 1*, 137–149.

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1*, 133–163. http://doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Kuyper, H., Seaton, M., Parker, P. D., Morin, A. J. S., Möller, J., & Abduljabbar, A. S. (2014). Dimensional comparison theory: An extension of the internal/external frame of reference effect on academic self-concept formation. *Contemporary Educational Psychology, 39,* 326–341. http://doi.org/10.1016/j.cedpsych.2014.08.003

Marsh, H. W., Pekrun, R., Murayama, K., Arens, K. A., Parker, P. D., Guo, J., & Dicke, T. (2018). An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over six years. *Developmental Psychology, 54*, 263–280. http://doi.org/10.1037/dev0000393

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluation of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology, 92*, 202–228. http://doi.org/10.1037/0022- 0663.92.1.202

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76*, 397–416. http://dx.doi.org/10.1111/j.1467-8624.2005.00853

Möller, J., Helm, F., Müller-Kalthoff, H., Nagy, N. & Marsh, H. W. (2015). Dimensional comparisons and their consequences for self-concept, motivation, and emotion. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (2nd ed., pp. 430–436). Amsterdam: Elsevier. http://doi.org/10.1016/B978-0-08-097086-8.26092-3

Möller, J., & Köller, O. (2001). Dimensional comparisons: An experimental approach to the internal/external frame of reference model. *Journal of Educational Psychology, 93*, 826–835. http://doi.org/10.1037//0022-0663.93.4.826

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review, 120*, 544–560. http://doi.org/10.1037/a0032459

Möller, J., Müller-Kalthoff, H., Helm, F., Nagy, N., & Marsh, H. W. (2015). The generalized internal/external frame of reference model: An extension to dimensional comparison theory. *Frontline Learning Research, 4,* 1–11. http://doi.org/10.14786/flr.v4i2.169

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79*, 1129–1167. http://doi.org/10.3102/0034654309337522

Mussweiler, T., Rüter, K., & Epstude, K. (2006). The why, who, and how of social comparison: A social-cognition perspective. In S. Guimonde (Ed.), *Social comparison and social psychology: Understanding cognition, intergroup relations, and culture* (pp. 33–54).

Cambridge, England: Cambridge University Press. http//doi.org/ 10.1017/CBO9780511584329.004

Muthén, L. K. & Muthén, B. O. (1998-2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Nichols, S. L. & Glass, G. V., & Berliner, D.C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*. Retrieved from http://epaa.asu.edu/ojs/article/view/72/198

OECD Organization for Economic Cooperation and Development (2004). *Learning for tomorrow's world: First results from PISA 2003.* Paris: OECD.

Pinger, P., Rakoczy, K., Besser, M., & Klieme, E. (2018a). Interplay of formative assessment and instructional quality—Interactive effects on students' mathematics achievement. *Learning Environments Research, 21*, 61–79. http://doi.org/10.1007/s10984-017-9240-2

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM, 50*, 407–426. http://doi.org/10.1007/s11858-018-0918-4

Rakoczy, K., Buff, A., & Lipowsky, F. (2005). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie. Befragungsinstrumente* [Technical report of the German – Swiss video study]. Frankfurt, Germany: GFPF.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D.,… Schiefele, U. (Eds.). (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente* [PISA 2003: Documentation of the instruments]. Münster: Waxmann.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: CA: Sage.

Schneider, W., Schlagmüller, M., & Ennemoser, M. (2017). *LGVT 5-12+ Lesegeschwindigkeits- und verständnistest für die Klassen 5-12+ Manual* [Reading speed and reading comprehension test for grades 5–12+ manual]. Göttingen: Hogrefe.

Schurtz, I. M., Pfost, M., Nagengast, B., & Artelt, C. (2014). Impact of social and dimensional comparisons on student's mathematical and English subject-interest at the beginning of secondary school. *Learning and Instruction, 34*, 32–41. http://doi.org/10.1016/j.learninstruc.2014.08.001

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 36*, 259–280. http://doi.org/10.3102/0162373713509880

Skaalvik, E. M., & Rankin, R. J. (1995). A test of the internal/external frame of reference model at different levels of math and verbal self-perception. *American Educational Research Journal, 32*, 161–184. http://doi.org/10.3102/00028312032001161

Spooren, P., Bockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Review of Educational Research, 83*, 598–642. http://doi.org/10.3102/0034654313496870

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*, 481–520. http://doi.org/10.3102/1076998616646200

Stodolsky, S. (1988*). The subject matters: Classroom activity in math and social studies*. Chicago, IL: University of Chicago Press.

Turner, J. C., & Meyer, D. K. (2000). Studying and understanding the instructional contexts of classrooms: Using our past to forge our future. *Educational Psychologist, 35*, 69–85. http://doi.org/10.1207/S15326985EP3502_2

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*, 1–11. http://doi.org/10.1016/j.learninstruc.2013.03.003

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, 108*, 705–721. http://doi.org/10.1037/edu0000075

Wagner, W., Helmke, A., & Rösner, E. (2009). *Deutsch Englisch Schülerleistungen International. Dokumentation der Erhebungsinstrumente für Schülerinnen und Schüler, Eltern und Lehrkräfte* [German English Student Achievement International. Documentation of the instruments]. Frankfurt am Main: DIPF.

Wallace, L. W., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Education Research Journal, 53*, 1834–1868. http://doi.org/10.3102/0002831216671864

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA, US: Sage Publications, Inc.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Unpublished doctoral dissertation. LA: University of California.

Supplemental Material

Table S1

Unconstrained and Constrained Model Solutions for Teaching Quality Domains in Mathematics and German Language

	Unconstrained model solution						Constrained model solution					
	χ^2 (*df*)	CFI	TLI	RMSEA	$SRMR_w$	$SRMR_b$	χ^2 (*df*)	CFI	TLI	RMSEA	$SRMR_w$	$SRMR_b$
Classroom Management	1194.40 (280)	.98	.97	.02	.04	.10	1477.56 (298)	.97	.96	.03	.04	.11
Supportive Climate	6241.81 (1337)	.96	.95	.02	.04	.10	6436.50 (1375)	.96	.95	.02	.04	.10
Cognitive Activation	1730.39 (429)	.97	.96	.02	.03	.11	1897.23 (453)	.96	.95	.02	.03	.11

Table S2

Unconstrained and Constrained Model Solutions for Teaching Quality Dimensions in Mathematics and German Language

	Unconstrained model solution						Constrained model solution					
	χ^2 (*df*)	CFI	TLI	RMSEA	$SRMR_w$	$SRMR_b$	χ^2 (*df*)	CFI	TLI	RMSEA	$SRMR_w$	$SRMR_b$
Monitoring	68.16 (34)	1.00	1.00	.01	.01	.03	107.70 (43)	.99	.99	.02	.02	.03
Disturbances	573.40 (95)	.98	.98	.03	.03	.08	613.04 (110)	.98	.98	.03	.03	.08
Structuredness	93.18 (13)	.99	.98	.03	.01	.10	109.04 (19)	.99	.98	.03	.01	.10
Rule clarity	156.59 (15)	.98	.97	.04	.01	.13	174.73 (21)	.98	.98	.03	.02	.11
Handling mistakes	59.66 (15)	.99	.99	.02	.02	.07	69.74 (21)	.99	.99	.02	.02	.07
Differentiation	370.04 (34)	.97	.94	.04	.03	.11	469.85 (44)	.96	.94	.04	.03	.16
Motivation	140.48 (36)	.99	.99	.02	.02	.08	191.68 (45)	.99	.99	.02	.02	.08
General support	34.48 (15)	1.00	1.00	.01	.01	.04	53.30 (21)	1.00	.99	.02	.01	.04
Clarity	51.05 (14)	1.00	.99	.02	.01	.05	80.42 (20)	.99	.99	.02	.01	.05
Autonomy support	31.99 (13)	1.00	.99	.02	.01	.05	60.06 (19)	.99	.99	.02	.01	.06
Feedback	175.07 (36)	.99	.99	.03	.02	.10	272.07 (42)	.99	.98	.03	.02	.09
Expectations	18.35 (15)	1.00	1.00	.01	.00	.08	49.65 (21)	1.00	1.00	.02	.01	.06
Challenging tasks	30.76 (15)	1.00	1.00	.01	.01	.08	119.38 (20)	.98	.98	.03	.01	.08
Regularity	29.50 (14)	1.00	.99	.01	.01	.11	34.32 (19)	1.00	1.00	.01	.01	.11
Practicing	60.71 (15)	.99	.98	.02	.01	.08	77.52 (21)	.99	.98	.02	.01	.08
Socratic dialogue	43.36 (15)	1.00	.99	.02	.01	.04	72.34 (21)	.99	.99	.02	.01	.05

Table S3

Students' Achievement in Mathematics and German predicting Teaching Quality in Mathematics and German Language on the Student Level

	M, Ach → M, TQ	M, Ach → G, TQ	G, Ach → M, TQ	G, Ach → G, TQ
	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Monitoring	0.06** (0.02)	0.00 (0.02)	-0.02 (0.02)	-0.03 (0.02)
Disturbances	-0.06*** (0.02)	-0.02 (0.02)	0.01 (0.02)	0.00 (0.02)
Structuredness	0.01 (0.02)	-0.05** (0.02)	-0.06** (0.02)	0.00 (0.02)
Rule clarity	0.03 (0.02)	0.01 (0.02)	0.04* (0.02)	0.05* (0.02)
Handling mistakes	0.07*** (0.02)	0.02 (0.02)	0.03* (0.02)	0.08*** (0.02)
Motivation	0.09*** (0.02)	-0.05** (0.02)	-0.04** (0.01)	0.03 (0.02)
General support	0.06** (0.02)	-0.02 (0.02)	-0.01 (0.02)	0.03 (0.02)
Clarity	0.20*** (0.02)	0.01 (0.02)	0.01 (0.01)	0.12*** (0.02)
Autonomy support	0.12*** (0.02)	-0.04* (0.02)	0.01 (0.02)	0.09*** (0.02)
Feedback	0.06*** (0.02)	0.00 (0.01)	-0.03 (0.02)	0.01 (0.02)
Expectations	-0.01 (0.02)	0.04* (0.02)	-0.09*** (0.02)	-0.10*** (0.02)
Challenging tasks	0.13*** (0.02)	-0.03 (0.02)	0.07*** (0.02)	0.09*** (0.02)
Regularity	0.08*** (0.02)	-0.03 (0.02)	0.05** (0.02)	0.04* (0.02)
Practicing	0.09*** (0.02)	-0.04 (0.02)	0.02 (0.02)	0.03 (0.02)
Socratic dialogue	0.05** (0.02)	-0.01 (0.02)	-0.02 (0.02)	-0.01 (0.02)

Note. M = mathematics. G = German language. TQ = teaching quality. Ach = Achievement test scores. SL = student level. CL = classroom level.

*$p < .05$. **$p < .01$. ***$p < .001$.

Table S4

Students' Achievement in Mathematics and German predicting Teaching Quality in Mathematics and German Language on the Classroom Level

	M, Ach → M, TQ	M, Ach → G, TQ	G, Ach → M, TQ	G, Ach → G, TQ
	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Monitoring	0.08 (0.15)	-0.11 (0.15)	-0.03 (0.12)	-0.11 (0.13)
Disturbances	-0.44** (0.16)	-0.27 (0.16)	-0.05 (0.11)	-0.11 (0.14)
Structuredness	-0.37* (0.16)	-0.43** (0.15)	0.05 (0.13)	-0.01 (0.14)
Rule clarity	0.15 (0.18)	-0.26 (0.18)	0.20 (0.14)	0.29 (0.16)
Handling mistakes	0.09 (0.19)	-0.06 (0.17)	0.24 (0.15)	0.20 (0.14)
Motivation	-0.03 (0.16)	-0.17 (0.14)	-0.01 (0.13)	0.05 (0.13)
General support	0.16 (0.17)	-0.20 (0.16)	-0.07 (0.13)	-0.02 (0.14)
Clarity	0.08 (0.16)	-0.22 (0.16)	0.04 (0.13)	0.18 (0.15)
Autonomy support	0.05 (0.19)	-0.42** (0.16)	-0.01 (0.14)	0.19 (0.15)
Feedback	-0.11 (0.16)	-0.35 (0.19)	-0.10 (0.12)	0.02 (0.15)
Expectations	-0.07 (0.19)	-0.08 (0.18)	-0.41* (0.16)	-0.34* (0.14)
Challenging tasks	0.48* (0.19)	-0.11 (0.17)	0.08 (0.15)	0.15 (0.16)
Regularity	0.56** (0.19)	-0.33 (0.18)	-0.11 (0.19)	0.23 (0.18)
Practicing	0.30 (0.17)	-0.22 (0.15)	0.00 (0.13)	-0.01 (0.14)
Socratic dialogue	0.16 (0.19)	-0.22 (0.17)	0.03 (0.18)	-0.13 (0.15)

Note. M = mathematics. G = German language. TQ = teaching quality. Ach = Achievement test scores. SL = student level. CL = classroom level.

*$p < .05$. **$p < .01$. ***$p < .001$

Table S5

Student Achievement in Mathematics and German Language controlled for Student Grades predicting Teaching Quality in Mathematics and German Language on the Student Level

	M, Ach → M, TQ	G, Ach → G, TQ	M, Ach → G, TQ	G, Ach → G, TQ
	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Monitoring	0.01 (0.02)	-0.05** (0.02)	0.01 (0.02)	-0.05** (0.02)
Disturbances	-0.01 (0.02)	0.03 (0.02)	-0.00 (0.02)	0.03 (0.02)
Structuredness	-0.00 (0.02)	-0.03 (0.02)	-0.02 (0.02)	-0.03 (0.02)
Rule clarity	-0.01 (0.02)	0.02 (0.02)	0.00 (0.02)	0.02 (0.02)
Handling mistakes	0.01 (0.02)	0.04* (0.02)	0.01 (0.02)	0.04* (0.02)
Motivation	0.00 (0.02)	-0.03 (0.02)	-0.04* (0.02)	-0.03 (0.02)
General support	-0.00 (0.02)	-0.01 (0.02)	-0.02 (0.02)	-0.01 (0.02)
Clarity	0.07*** (0.02)	0.05** (0.02)	0.00 (0.02)	0.05** (0.02)
Autonomy support	0.04 (0.02)	0.04* (0.02)	-0.03 (0.02)	0.04* (0.02)
Feedback	0.04* (0.02)	-0.01 (0.02)	0.00 (0.02)	-0.01 (0.02)
Expectations	0.03 (0.02)	-0.08*** (0.02)	0.06** (0.02)	-0.08*** (0.02)
Challenging tasks	0.05* (0.02)	0.05** (0.02)	-0.02 (0.02)	0.05** (0.02)
Regularity	0.04 (0.02)	0.01 (0.02)	-0.03 (0.02)	0.01 (0.02)
Practicing	0.03 (0.02)	-0.01 (0.02)	-0.02 (0.02)	-0.01 (0.02)
Socratic dialogue	0.04* (0.02)	-0.00 (0.02)	0.00 (0.02)	-0.00 (0.02)

Note. M = mathematics. G = German language. TQ = teaching quality. Ach = Achievement test scores. SL = student level. CL = classroom level.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table S6

Student Achievement in Mathematics and German Language controlled for Student Grades predicting Teaching Quality in Mathematics and German Language on the Student Level

	M, Ach → M, TQ	G, Ach → G, TQ	M, Ach → G, TQ	G, Ach → M, TQ
	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Monitoring	0.04 (0.15)	-0.14 (0.14)	-0.05 (0.15)	0.04 (0.12)
Disturbances	-0.42** (0.16)	-0.10 (0.14)	-0.30 (0.17)	-0.07 (0.12)
Structuredness	-0.41** (0.16)	-0.06 (0.14)	-0.39* (0.15)	0.11 (0.13)
Rule clarity	0.11 (0.18)	0.28 (0.17)	-0.23 (0.18)	0.23 (0.14)
Handling mistakes	-0.01 (0.18)	0.12 (0.14)	-0.01 (0.16)	0.29* (0.14)
Motivation	-0.12 (0.15)	-0.01 (0.13)	-0.15 (0.14)	0.08 (0.12)
General support	0.09 (0.16)	-0.08 (0.14)	-0.17 (0.16)	-0.01 (0.12)
Clarity	-0.03 (0.15)	0.09 (0.14)	-0.17 (0.16)	0.14 (0.12)
Autonomy support	-0.04 (0.18)	0.12 (0.15)	-0.41* (0.16)	0.08 (0.14)
Feedback	-0.15* (0.15)	-0.01 (0.15)	-0.30 (0.18)	-0.01 (0.12)
Expectations	0.01 (0.18)	-0.25 (0.14)	-0.06 (0.17)	-0.36* (0.16)
Challenging tasks	0.44* (0.18)	0.06 (0.16)	-0.10 (0.17)	0.17 (0.15)
Regularity	0.50 (0.19)	0.20 (0.18)	-0.28 (0.18)	-0.01 (0.19)
Practicing	0.22 (0.17)	-0.07 (0.14)	-0.18 (0.15)	0.10 (0.13)
Socratic dialogue	0.14 (0.19)	-0.13 (0.16)	-0.18 (0.18)	0.12 (0.17)

Note. M = mathematics. G = German language. TQ = teaching quality. Ach = Achievement test scores. SL = student level. CL = classroom level.
*$p < .05$. **$p < .01$. ***$p < .001$.

6 GESAMTDISKUSSION

Informationen von Schülerinnen und Schülern, wie sie ihren täglichen Unterricht wahrnehmen, bieten Lehrkräften, Schulleitungen, aber auch der Bildungsadministration und der Wissenschaft wichtige Erkenntnisse über das schulische Lernen. John Hattie, der eine prominente Synthese von Forschungsarbeiten zu Wirkmechanismen für den Schulerfolg von Schülerinnen und Schülern veröffentlicht hat, begründet die Berücksichtigung der Perspektive der Schülerinnen und Schüler folgendermaßen (Hattie, 2009, S. 116):

> The students sit in the classes, they know whether the teacher sees learning through their eyes, and they know the quality of the relationship. The visibility of learning from the students' perspective needs to be known by the teachers so that they can have a better understanding of what learning looks and feels like for the students.

Schülerinnen und Schüler sind nicht nur Rezipienten des Unterrichts. Vielmehr interagieren sie untereinander und mit der Lehrkraft und gestalten somit Unterricht aktiv mit (Klieme, 2019; Kunter & Voss, 2011; Reusser, 2009). Schülerurteile sind daher eine wertvolle Ressource, um Einblicke in die Wahrnehmung des Unterrichts aus Sicht einzelner Schülerinnen und Schüler, aber auch aus Sicht der gesamten Klasse zu erhalten (Ferguson, 2012; Mullis et al., 2009; Reiss et al., 2016). Aber inwiefern sind Schülerurteile geeignet, um Unterrichtsqualität reliabel und valide zu erfassen? Besonders im Hinblick auf die Validität von Schülerurteilen wurden in der bisherigen Forschung wichtige Fragen nicht oder nur teilweise untersucht. Die vorliegende Dissertation adressierte drei dieser offenen Fragen und leistet so einen Beitrag zur verlässlichen Nutzung von Schülerurteilen.

Studie 1 adressierte erstmals die Frage, ob Schülerinnen und Schüler auch in veränderten Kontexten, nämlich dem Distanzunterricht, eine geeignete Methode zur Erfassung von Unterrichtsqualität sind. Die Ergebnisse zeigten, dass Schülerinnen und Schüler differenziert Auskunft zu unterschiedlichen Subdimensionen der Unterrichtsqualität des Distanzunterrichts geben können und ihre Urteile prädiktiv für Zielkriterien des Unterrichts sind. Studie 2 widmete sich der Frage, welche Rolle der Adressat in Items zur Erfassung von Unterrichtsqualität spielt. Eine systematische Betrachtung von

Subdimensionen der Unterrichtsqualität, in welchen Items mit Ich-Adressat und Items mit Wir-Adressat verwendet wurden, zeigte, dass der Item-Adressat mit kleinen, jedoch sehr systematischen Unterschieden in verschiedenen psychometrischen Eigenschaften wie dem Mittelwert oder den ICCs der Subdimensionen einherging. Studie 3 ging schließlich der Frage nach, ob Schülerurteile zur Beurteilung der Unterrichtsqualität eines Faches durch die Note eines anderen Faches beeinflusst sind. Dem Muster dimensionaler Vergleiche folgend zeigten sich für die Fächer Mathematik und Deutsch positive Zusammenhänge zwischen der Note und der Beurteilung von Unterrichtsqualität innerhalb eines Faches, jedoch negative Zusammenhänge beider Merkmale zwischen den Fächern. Dies zeigt, dass Schülerurteile zur Unterrichtsqualität in einem Fach durch die Note eines zweiten zu bewertenden Fachs beeinflusst sein können.

Im Folgenden werden zunächst die Ergebnisse der drei Studien diskutiert und in einen breiten Forschungskontext eingeordnet (5.1). Schlussfolgerungen, die aus den einzelnen Studien gezogen werden können, werden am Ende jedes Unterkapitels formuliert. Anschließend werden Stärken und Grenzen der einzelnen Studien und der vorliegenden Arbeit diskutiert (5.2). Diesen Ausführungen folgen Implikationen für die zukünftige Forschung (5.3.1) und die praktische Nutzung von Schülerurteilen (5.3.2). Ein Fazit über die vorliegende Dissertation (5.4) schließt das Kapitel ab.

6.1 ZUSAMMENFASSUNG DER ERGEBNISSE

Unterricht wird im „Angebot-Nutzungs-Modell der Wirkungsweise des Unterrichts“ (Helmke, 2017) als ein Angebot verstanden, dessen Ertrag unter anderem davon abhängt, ob und wie dieses von den Schülerinnen und Schülern wahrgenommen und interpretiert wird und in welchem Ausmaß dieses Angebot in Lernaktivitäten aufseiten der Schülerschaft resultiert, also genutzt wird. Damit bildet das Modell die Annahme ab, dass Unterricht ein konstruktivistisches und interaktives Geschehen ist und gemeinsam durch die Lehrkraft und die Schülerinnen und Schüler gestaltet wird (Klieme, 2019). Unterricht wird in der empirischen Bildungsforschung vorrangig im Hinblick auf Prozessmerkmale wie das Lehrerverhalten oder die Lehrer-Schüler-Interaktion betrachtet, die in Zusammenhang mit unterschiedlichen Lernzielen der Schülerinnen und Schüler stehen (Brunner et al., 2006). Ein zentrales Prozessmerkmal stellt dabei die Unterrichtsqualität dar.

Zur Operationalisierung von Unterrichtsqualität wird insbesondere in der deutschsprachigen Forschungslandschaft auf das „Rahmenmodell der drei Basisdimensionen der Unterrichtsqualität“ (Klieme et al., 2001) zurückgegriffen, welches Unterricht durch die drei Dimensionen Klassenführung, konstruktive Unterstützung und kognitive Aktivierung abbildet. Diese Dimensionen, die häufig anhand verschiedener Subdimensionen erfasst werden, haben sich als bedeutsam für unterschiedliche Zielkriterien des Unterrichts wie Leistungszuwachs, Motivation oder Lernfreude erwiesen (Fauth et al., 2014; Göllner et al., 2018; Hattie 2009; Kunter & Trautwein, 2013; Praetorius et al., 2018; Seidel & Shavelson, 2007). Eine zentrale Herausforderung der Unterrichtsforschung jedoch ist es, Unterrichtsqualität reliabel und valide zu erfassen. Hierbei erwiesen sich die Urteile von Schülerinnen und Schülern in vielerlei Hinsicht als geeignete Methode (Fauth et al., 2014; Kane et al., 2013; Kuhfeld, 2017; Wagner et al., 2013). In Bezug auf deren Validität sind jedoch noch wichtige Fragen unbeantwortet, zu welchen die drei Studien dieser Arbeit einen Beitrag leisteten.

6.1.1 Die Nutzung von Schülerurteilen zur Erfassung von Unterrichtsqualität im Distanzunterricht

Regulärer Distanzunterricht im schulischen Kontext wurde in Deutschland und in vielen anderen Ländern in der Vergangenheit nur in sehr wenigen Ausnahmefällen durchgeführt (Ladenthin, 2018), weshalb sich die bisherige Unterrichtsforschung primär auf Qualitätsmerkmale des Präsenzunterrichts konzentrierte. Zwar ließen sich in Ländern wie den USA oder Australien, in welchen Distanzunterricht eher verbreitet war, Hinweise auf

Merkmale eines qualitätsvollen Distanzunterrichts finden, wie beispielsweise regelmäßiges Feedback oder die Möglichkeit zur Interaktion (Hawkins et al., 2013; Kumi-Yeboah et al., 2017; Liu & Cavanaugh, 2012; NSQ, 2019). Jedoch gibt es kein Rahmenmodell, das, ähnlich wie für den Präsenzunterricht, Qualitätsmerkmale von Distanzunterricht systematisiert und operationalisiert. Zudem wurden in bisherigen Studien die Gütekriterien der verwendeten Schülerurteile kaum berichtet, sodass unklar ist, ob und inwiefern diese tatsächlich für die Erfassung von Unterrichtsqualität im Distanzunterricht geeignet sind. Studie 1 nahm deshalb in den Blick, ob Schülerurteile auch für den Distanzunterricht geeignet sind, um Unterrichtsqualität auf Grundlage der drei Basisdimensionen zu erfassen. Angaben einer großen Anzahl von Schülerinnen und Schülern gaben in dieser Studie nicht nur wertvolle Einblicke in das Unterrichtsgeschehen der Fächer Mathematik, Deutsch und Englisch während der Schulschließungen im Frühjahr 2020. Die Ergebnisse dieser Studie zeigten auch, dass Schülerinnen und Schüler die Qualität von Distanzunterricht in unterschiedlichen Fächern reliabel beurteilen können und dass diese Urteile mit den Zielkriterien Kompetenzerleben, Anstrengungsbereitschaft, Lernfreude und Klassengemeinschaft assoziiert sind.

Bereits in früheren Studien wurden Angaben von Schülerinnen und Schülern zur Erfassung von Unterrichtsmerkmalen im Distanzunterricht genutzt (Hawkins et al., 2013; Kumi-Yeboah et al., 2017; Liu & Cavanaugh, 2012). Während die Angabe von Indizes zu Reliabilität und Validität in Studien zum Präsenzunterricht gängig ist (Lüdtke et al., 2006; Göllner et al., 2016), wurden diese in Studien zum Distanzunterricht überraschenderweise nur selten berichtet (z. B. Hawkins et al., 2013; Kumi-Yeboah et al., 2017; Liu & Cavanaugh, 2012). Aus diesem Grund gibt es kaum Kenntnis darüber, wie gut Schülerurteile für den Distanzunterricht tatsächlich geeignet sind. Das Fehlen solcher Informationen ist umso überraschender, da auch innerhalb des Forschungsfelds solche Informationen gefordert werden. Im *Handbook of Research on K-12 Online and Blended Learning* (Kennedy & Ferdig, 2018) merkt die Wissenschaftlerin Sarojani S. Mohammed an (Mohammed, 2018, S. 111):

> Along with emerging learning environments have come a need to measure academic and non-academic variables that have not traditionally been central to understanding effectiveness (Tough, 2013). This brings another challenge to measurement in emerging learning environments. The rigor and validity of any study hinges on the measures used, findings are

> only as good as the data on which they are based. There is an urgent need for valid and reliable measures of constructs of interest […].

Gerade weil bislang kaum Daten zur Umsetzung und Qualität des Distanzunterrichts vorliegen, ist es wichtig, verlässliche Informationen über das Lernen von Schülerinnen und Schülern zu erhalten. Die Gütekriterien für alle Subdimensionen der drei Fächer in Studie 1 zeigten, dass Schülerurteile eine geeignete Methode zur Erfassung von Unterricht im Distanzunterricht sind und wichtige Erkenntnisse über das Unterrichtserleben der Schülerinnen und Schüler bieten. Die Studie leistete somit einen wichtigen Beitrag im Hinblick auf die Nutzung von Schülerurteilen im Distanzunterricht.

Ein überraschendes Ergebnis von Studie 1 war, dass sich Sichtstrukturen im Distanzunterricht in Form von Unterrichtsmethoden als bedeutsam für die Zielkriterien des Unterrichts herausstellten. Für den Präsenzunterricht haben sich diese als wenig bedeutsam für das Lernen von Schülerinnen und Schülern erwiesen (Hattie 2009; Seidel & Shavelson, 2007). Als besonders relevant zeigten sich dabei Unterrichtsmethoden wie Treffen mit Schülerinnen und Schülern oder die Nutzung von selbst generierten Lernvideos, denen die Möglichkeit der sozialen Interaktion zugrunde liegt. Soziale Interaktion wird im Distanzunterricht erst möglich, wenn die Lehrkraft entsprechende Unterrichtsmethoden initiiert. Die Ergebnisse zur Relevanz der Interaktion bestätigen Befunde früherer Studien zum Distanzunterricht (Hawkins et al., 2013; Kumi-Yeboah et al., 2017; Liu & Cavanaugh, 2012; Ouzts, 2006; Simonson et al., 2019). Somit zeigten diese Ergebnisse, dass im Distanzunterricht Unterrichtsmethoden durchaus mit Zielkriterien des Unterrichts assoziiert sind – jedoch primär, wenn hierbei soziale Interaktion ermöglicht wird.

Eine wichtige Erkenntnis von Studie 1 im Hinblick auf die Erfassung von Unterrichtsqualität ist, dass sich diese im Distanzunterricht durch Subdimensionen der drei Basisdimensionen der Unterrichtsqualität (Klieme et al., 2001) erfassen lässt. Da eine gemeinsame Wahrnehmung des Unterrichts, wie dies im Präsenzunterricht täglich der Fall ist, im Distanzunterricht nicht oder nur eingeschränkt gegeben ist, stellt dies für die Erfassung von Unterrichtsqualität im Distanzunterricht eine zentrale Erkenntnis dar. Weiterhin zeigte sich, dass die Subdimensionen der Unterrichtsqualität positiv mit Zielkriterien wie der Lernfreude oder der Anstrengungsbereitschaft assoziiert sind. Wenngleich in dieser Studie keine Leistungstests eingesetzt werden konnten, zeigen diese Ergebnisse, dass auch im Distanzunterricht die Unterrichtsqualität relevant für das schulische Lernen der Schülerinnen und Schüler zu sein scheint.

Verlässliche Aussagen über Schülerurteile und deren Zusammenhänge mit Beurteilungen der Unterrichtsqualität sowie Zielkriterien des Unterrichts setzen belastbare Daten voraus. Die Ergebnisse von Studie 1 zeigten, dass Schülerurteile im Sinne der Reliabilität und Validität geeignet sind, um ein Verständnis im Hinblick auf den Distanzunterricht während der Schulschließungen im Frühjahr 2020, aber auch in zukünftigen Lehr-Lernsettings auf Distanz zu erhalten.

6.1.2 Die Rolle des Item-Adressaten zur Erfassung von Unterrichtsqualität aus Schülersicht

Bereits auf den ersten Blick wird ersichtlich, dass sich Items zur Erfassung von Unterrichtsqualität hinsichtlich ihrer Formulierungen unterscheiden (Baumert et al., 1997; Kuhfeld, 2017). Jedoch wurde der möglichen Relevanz dieser Unterschiede und insbesondere der Rolle des Item-Adressaten in der bisherigen Forschung erstaunlich wenig Beachtung geschenkt. Aus theoretischer Perspektive beziehen sich der Ich-Adressat und der Wir-Adressat in Items zur Erfassung von Unterrichtsqualität auf unterschiedliche Informationen: Während der Ich-Adressat die individuelle Sicht der Schülerinnen und Schüler adressiert, bezieht sich der Wir-Adressat auf die Wahrnehmung aller Schülerinnen und Schüler. Damit basieren beide Versionen des Adressaten auf verschiedenen Informationsgrundlagen und führen nach Modellen des Antwortprozesses zu potenziell unterschiedlichen Resultaten (Tourangeau et al., 2000). Studie 2 widmete sich deshalb der Frage, ob sich die Verwendung eines Ich-Adressaten und eines Wir-Adressaten in psychometrischen Eigenschaften von Schülerurteilen widerspiegelt. Die Ergebnisse zeigten folgendes Muster: Während Items mit Wir-Adressaten zu höheren Mittelwerten, ICCs und Interkorrelationen unterschiedlicher Subdimensionen bei gleichem Adressaten auf Klassenebene führten, zeigten sich höhere Zusammenhänge für Zielkriterien des Unterrichts auf Schülerebene, wenn Items mit Ich-Adressaten verwendet wurden.

Nur wenige Studien nahmen bisher die Rolle des Item-Adressaten in den Blick. Diese kommen zu uneinheitlichen Befunden: Während die Ergebnisse von Studie 2 die Befunde der Studien von McRobbie und Kollegen (1991, 1998) und Fraser und Kollegen (1995), die ebenfalls höhere Mittelwerte für den Wir-Adressaten fanden, bestätigen, fanden Den Brok et al. (2006) höhere Mittelwerte für den Ich-Adressaten. Da sich diese Studien jedoch primär auf Mittelwertdifferenzen und sehr unterschiedliche Facetten der Unterrichtsqualität (z. B. Hilfsbereitschaft der Lehrkraft in Den Brok et al. (2006) und Regelklarheit bei McRobbie et al., 1991) konzentrierten, lassen sich die Ergebnisse nur eingeschränkt vergleichen. Unterschiede in den Urteilen für beide Versionen des Adressaten lassen sich auch in

theoretische Annahmen des Antwortprozesses einordnen. Da beiden Versionen des Adressaten unterschiedliche Informationen zugrunde liegen, ist es schlüssig, dass sich diese in unterschiedlichen Werten widerspiegeln (Tourangeau et al., 2000). Im Hinblick auf die Einordnung höherer Werte für Items mit Wir-Adressat zeigten bisherige Studien, dass Items, die weniger eindeutige und eher unpräzise Inhalte erfragen, zu höheren Mittelwerten führen (Cabooter, 2010; Moors, 2008). Dies trifft auf Items mit Wir-Adressat zu, da Schülerinnen und Schüler zur Beantwortung der Items theoretisch viele Informationen abwägen und Inferenzen ziehen müssen. Auch zeigten Studien, dass Schülerinnen und Schüler eher hinsichtlich wenig beobachtbarer Merkmale übereinstimmen (Roch et al., 2009). Die Autoren begründen dies mit dem Rückgriff auf eine „general impression" über die Lehrkraft, was zu höheren Übereinstimmungen in den Urteilen führe (Roch et al., 2009, S. 403). Dies trifft auf die Verwendung eines Wir-Adressaten zu, da Schülerinnen und Schüler hierbei mehr Informationen abwägen und Inferenzen ziehen müssen, als wenn sie ausschließlich Angaben über ihr individuelles Erleben machen. Somit können diese Urteile durch einen Gesamteindruck über die Lehrkraft beeinflusst und daher weniger valide sein. Schließlich wäre es plausibel, dass ein Verhalten der Lehrkraft, das alle Schülerinnen und Schüler der Klasse betrifft, in Summe häufiger vorkommt als ein Verhalten, das nur einzelne Schülerinnen und Schüler betrifft. Dieses könnte von den Schülerinnen und Schülern aufgrund des häufigeren Vorkommens mehr wahrgenommen werden und somit zu höheren Werten für Items mit einem Wir-Adressaten führen.

Weiterhin bestätigen die Ergebnisse zu den Zusammenhängen mit Zielkriterien des Unterrichts in Studie 2 die Befunde zahlreicher Studien, die positive Zusammenhänge zwischen der wahrgenommenen Unterstützung und Zielkriterien des Unterrichts fanden (Fauth et al., 2014; Göllner et al., 2018; Kunter et al., 2008; Lipowsky, 2015; Praetorius, Klieme et al., 2020; Wagner et al., 2013). Die höheren Zusammenhänge für Items mit Ich-Adressat auf Schülerebene unterstreichen die Bedeutung der Dimension der konstruktiven Unterstützung für das individuelle Lernen der Schülerinnen und Schüler. Beispielsweise konnte auch die Bedeutung der individuellen Schüler-Lehrer-Beziehung in der Vergangenheit gezeigt werden (Göllner et al., 2018), wobei in dieser Studie Items mit Wir-Adressat verwendet wurden. Es wäre daher weiter aufschlussreich, zu prüfen, ob sich die Zusammenhänge mit der Verwendung des Ich-Adressaten weiter verstärken.

Der Frage, welche der beiden Versionen zu verlässlicheren Antworten führt, kann sich nur unter Rückgriff auf theoretische Modelle genähert werden. So gehen Modelle der Persönlichkeitspsychologie wie das „Self-Other Knowledge Asymmetry"-Modell (Vazire,

2010) oder das „Realistic Accuracy“-Modell (Funder, 1995) davon aus, dass je mehr Informationen verfügbar sind, desto präziser die Urteile ausfallen. Items mit Ich-Formulierung zielen auf das Verhalten der Lehrkraft ab, wie es die Schülerinnen und Schüler je individuell wahrnehmen. Deshalb sind die Informationen, die zur Beantwortung des Items benötigt werden, für die individuellen Schülerinnen und Schüler vollumfänglich verfügbar, während für Items mit Wir-Adressat viele Informationen herangezogen und Inferenzen gezogen werden müssen. In Anbetracht dieser Annahmen und den zusätzlich eher kleinen Unterschieden in den psychometrischen Eigenschaften beider Versionen des Adressaten lässt sich darauf schließen, dass aus Items mit Ich-Adressat verlässlichere Urteile resultieren.

6.1.3 Der Einfluss der Note eines anderen Faches auf Schülerurteile zur Unterrichtsqualität

In der Vergangenheit wurden auch Zweifel an der Eignung von Schülerurteilen zur Erfassung von Unterrichtsqualität geäußert (Aleamoni, 1999; Greenwald & Gillmore, 1997). Diese bezogen sich meist darauf, dass Schülerurteile von Merkmalen, die unabhängig von der zu erfassenden Unterrichtsqualität sind, beeinflusst sein können (Griffin, 2004; Reyes et al., 2012). Gleichzeitig zeigte sich, dass die Note nicht nur von großer Bedeutung für die Einordnung der eigenen Leistung von Schülerinnen und Schülern ist (Weidinger et al., 2015), sondern dass sie auch in Zusammenhang mit unterschiedlichen Zielkriterien des Unterrichts steht (Arens & Möller, 2016; Lüdtke et al., 2009; Marsh et al., 2018). Der Einfluss der Note auf Zielkriterien nicht nur innerhalb eines Faches, sondern auch zwischen zwei Fächern wurde bisher insbesondere in Bezug auf das akademische Selbstkonzept erforscht (Marsh & Martin, 2011; Möller et al., 2009; Möller et al., 2015). Bislang war jedoch offen, welchen Einfluss die Note eines Faches auf die Beurteilung der Unterrichtsqualität eines anderen Faches hat. In Studie 3 wurde dieser Frage nachgegangen, indem die Zusammenhänge zwischen der Note, kontrolliert durch die Testleistung, und Schülerurteilen zur Unterrichtsqualität innerhalb und zwischen den beiden Fächern Deutsch und Mathematik untersucht wurden. Die Ergebnisse bestätigten das Muster der dimensionalen Vergleiche: Es zeigten sich sowohl auf Schüler- als auch auf Klassenebene positive Zusammenhänge zwischen der Note und den Urteilen zur Unterrichtsqualität innerhalb eines Faches sowie negative Zusammenhänge zwischen den Fächern. Somit können die Schülerurteile in einem Fach durch die Note in einem anderen Fach beeinflusst sein.

Die Ergebnisse von Studie 3 bestätigen bisherige Befunde zu dimensionalen Vergleichen, in welchen positive Zusammenhänge zweier Merkmale (z. B. Note und Selbstkonzept) innerhalb von Schulfächern, jedoch negative Zusammenhänge zwischen den

Fächern gefunden wurden (Marsh, 1987; Möller et al., 2016; Dietrich et al., 2015). Auch Befunde dieses Musters zwischen der Note und der eingeschätzten Unterrichtsqualität innerhalb und zwischen Fächern konnte Studie 3 bestätigen und erweitern (Arens & Möller, 2016).

Die Ergebnisse dieser Studie könnten Zweifel an der Nutzung von Schülerurteilen aufkommen lassen. Jedoch muss zunächst beachtet werden, wie sich die Ergebnisse in die gesamte Befundlage zur Nutzung von Schülerurteilen einordnen lassen. Schülerurteile haben sich in zahlreichen Studien als reliabel gezeigt, sowohl hinsichtlich der verwendeten Skalen als auch hinsichtlich der Übereinstimmung von Schülerinnen und Schülern (De Jong & Westerhof, 2001; Fauth et al., 2014, 2019; Göllner et al., 2018). Bezüglich der Validität konnte die Generalisierbarkeit über Kontexte hinweg, wie in unterschiedliche Klassen oder Fächer, sowie die Dimensionalität von Schülerurteilen belegt werden (Polikoff, 2015; Wagner et al., 2013). Schließlich zeigte sich, dass sie prädiktiv für eine Vielzahl von Zielkriterien des Unterrichts sind (Hattie, 2009; Scherer et al, 2016; Wagner et al., 2016). Dies sind starke Befunde, die für eine Nutzung von Schülerurteilen sprechen. Mit Zweifeln an der Nutzung von Schülerurteilen geht auch die Frage nach alternativen Erfassungsmethoden einher, die nicht von Vergleichsprozessen betroffen sein könnten. Zwei weitere gängige Methoden zur Erfassung von Unterrichtsqualität sind Urteile externer Beobachter und Lehrerselbstberichte (Kunter & Baumert, 2006; Praetorius et al., 2014; Wagner et al., 2016). Da Vergleichsprozesse jedoch sehr häufig und in vielen Bereichen stattfinden und zudem zu Kontrasteffekten führen können (Festinger, 1954; Mussweiler, 2003; Pohlmann et al., 2004), ist davon auszugehen, dass sich diese auch in jenen beiden Erfassungsmethoden zeigen könnten. Beurteilt beispielsweise ein externer Beobachter zuerst einen aus seiner Sicht sehr qualitätsvollen Unterricht und anschließend einen Unterricht, in welchem es eher unruhig zugeht, könnte letzterer relativ gesehen negativer beurteilt werden, als wenn der Beobachter zuvor einen noch schlechteren Unterricht beurteilt hätte. Oder: Beurteilen Lehrkräfte zuerst eine Klasse, in der die Schülerinnen und Schüler oft stören und insgesamt ein unruhiges Klima herrscht, und anschließend eine sehr disziplinierte Klasse, würde letztere wohl relativ betrachtet höhere Beurteilungen erhalten.

Der Frage, welche Rolle die Note eines anderen Faches für die Beurteilung der Unterrichtsqualität spielt, wurde in der bisherigen Forschung nicht ausreichend nachgegangen. Studie 3 leistet somit einen wichtigen Beitrag zum Verständnis, ob Schülerurteile durch Merkmale, die von der Unterrichtsqualität unabhängig sind, beeinflusst sein können. Auch wenn die Ergebnisse zeigen, dass Schülerurteile durch die Note eines

anderen Faches beeinflusst sein können, spricht dies nicht vollumfänglich gegen die Nutzung von Schülerurteilen. Vielmehr zeigen die Ergebnisse dieser Studie Grenzen in der Nutzung von Schülerurteilen auf, wenn Schülerinnen und Schüler zur Unterrichtsqualität in zwei Fächern befragt werden.

6.2 STÄRKEN UND GRENZEN DER STUDIEN

Die drei Studien dieser Dissertation haben unterschiedliche Stärken, aber auch Grenzen, die bei der Interpretation der Ergebnisse berücksichtigt werden müssen. Alle drei Studien basieren auf großen Stichproben sowie auf Angaben zu einer großen Anzahl unterrichtsbezogener Variablen. Für die Beantwortung der Fragestellungen der vorliegenden Arbeit, aber auch für weiterführende Forschung bieten diese eine breite Grundlage. Die CUNITAS-Studie erfasste das Unterrichtsgeschehen während der Schulschließungen im Frühjahr 2020 hinsichtlich der Umsetzung des Distanzunterrichts, jedoch auch im Hinblick auf die Qualität und Zusammenhänge mit Zielkriterien des Unterrichts. Dies ermöglichte die umfassende Betrachtung von Distanzunterricht in unterschiedlichen Fächern sowie die Eignung von Schülerurteilen zur Erfassung von Unterrichtsqualität (Studie 1). Die UNITAS-Studie ist eine der größten deutschen Unterrichtsstudien der vergangenen Jahre. Eine breite Erfassung von Unterrichtsqualität und Zielkriterien in zwei Fächern ermöglicht die Betrachtung wichtiger offener Fragen zur Nutzung von Schülerurteilen wie die der Itemformulierung (Studie 2) und die des Einflusses der Note eines anderen Faches (Studie 3). Nichtsdestotrotz weist die vorliegende Arbeit auch Grenzen auf.

Erstens erfassten sowohl die CUNITAS-Studie als auch die UNITAS-Studie Unterrichtsqualität mit den gleichen Items in mehreren Fächern. CUNITAS erfasste den Unterricht in den Fächern Deutsch, Mathematik und Englisch, UNITAS in den beiden Fächern Deutsch und Mathematik. Es stellt sich jedoch bei der Verwendung fächerübergreifender Items die Frage, ob die Qualitätsmerkmale tatsächlich über die Fächer hinweg uneingeschränkt vergleichbar sind. Insbesondere für die Erfassung der kognitiven Aktivierung, die einen fachlichen und inhaltlichen Fokus auf Unterricht hat, ist es möglich, dass fachspezifische Merkmale des jeweiligen Schulfaches mit den verwendeten Items und Subdimensionen nicht ausreichend abgedeckt wurden. Diese Frage ist auch Gegenstand aktueller Diskussionen, nach welchen fachspezifische Erweiterungen und Ergänzungen der generischen Qualitätsdimensionen gefordert werden (Lipowsky & Bleck, 2019; Praetorius, Rogh et al., 2020). So ist es vorstellbar, dass im Englischunterricht mehr die aktive Anwendung der Sprache zur vertieften Auseinandersetzung mit sprachlichen Inhalten führt, während im Mathematikunterricht eher im Vordergrund stehen könnte, Bezüge zwischen Inhalten herzustellen oder Phänomene auf Alltagssituationen zu übertragen (siehe dazu auch Messner, 2019).

Zweitens wurde in der vorliegenden Dissertation die Komplexität von Items zur Erfassung von Unterrichtsqualität nur bedingt berücksichtigt. Zwar nahm Studie 2 explizit die Rolle des Item-Adressaten in den Blick. Jedoch gibt es darüber hinaus weitere Merkmale von Items, die zu möglichen Unterschieden in Schülerurteilen führen könnten (Wagner, 2008). So können Items im Hinblick darauf unterschieden werden, wer ein Item beurteilt (Lehrkräfte, Schülerinnen und Schüler, Beobachter) und auf wessen Verhalten sich das Item bezieht (Lehrkraft, Schülerinnen und Schüler oder beide). Fauth, Göllner et al. (2020) erstellten hierfür ein Raster, das Unterschiede in Items hinsichtlich des inhaltlichen Bezugs und der angewandten Erfassungsmethoden berücksichtigt. Beispielsweise ließen sich die in den Studien 2 und 3 dieser Arbeit verwendeten Items der Subdimensionen von Klassenführung in unterschiedliche Kategorien einordnen: Einerseits beziehen sich diese auf das Verhalten der Lehrkraft (Subdimension Monitoring), andererseits auch auf das Verhalten der Schülerinnen und Schüler (Subdimension Störungen im Unterricht). Diesem Raster zufolge hätten die unterschiedlichen Bezüge der Items und Subdimensionen zur Konsequenz, dass sich diese hinsichtlich ihres evaluativen Charakters unterscheiden.

Schließlich wurde in den drei Studien dieser Arbeit angenommen, dass Schülerurteile zur Erfassung von Unterrichtsqualität zur Erklärung der schulischen Leistung sowie weiterer Zielkriterien des Unterrichts herangezogen werden können. Jedoch konnte dies in allen Studien nur teilweise durch Korrelationen bzw. Regressionen geprüft werden. In der bisherigen Forschung gibt es insgesamt nur wenige Studien, die Aussagen über die Vorhersage von Zielkriterien des Unterrichts erlauben. Ob Schülerurteile auch Unterschiede in der Entwicklung von schulischer Leistung, dem Interesse oder der Motivation zulassen, wäre deshalb durch längsschnittliche Analysen mit großen Stichproben weiterhin zu untersuchen.

6.3 IMPLIKATIONEN FÜR FORSCHUNG UND PRAXIS

Aus den Ergebnissen der Studien dieser Dissertation können unterschiedliche Implikationen für die zukünftige Forschung und die praktische Nutzung von Schülerurteilen abgeleitet werden. Nachfolgend werden Vorschläge für die weitere Forschung in Bezug auf die Validität von Schülerurteilen dargelegt, die an aktuelle Fragestellungen anschließen. Im Anschluss folgen Implikationen für die Nutzung von Schülerurteilen in der Schulpraxis, die aus den drei Studien dieser Dissertation abgeleitet werden können.

6.3.1 Implikationen für zukünftige Forschung

Die vorliegende Arbeit hat in drei empirischen Studien wichtige Fragen zur Validität von Schülerurteilen adressiert und so zum Verständnis der Nutzung von Schülerurteilen zur Erfassung von Unterrichtsqualität beigetragen. Studie 1 ermöglichte wichtige Einblicke in das Lernen von Schülerinnen und Schülern während der Schulschließungen im Frühjahr 2020 und zeigte, dass Schülerurteile zur Erfassung von Unterrichtsqualität auf Grundlage der drei Basisdimensionen geeignet sind. Diese Erkenntnisse können als Grundlage zur Identifikation von Gelingensbedingungen im digitalen Unterricht beitragen, um die Weiterentwicklung von Unterricht mit digitalen Medien im Sinne einer qualitätsvollen Umsetzung zu begleiten. Der regelmäßige Einsatz digitaler Anwendungen und Medien im schulischen Unterricht war zuvor in Deutschland nur wenig verbreitet (Eickelmann et al., 2019). Gerade weil die Umstellung auf Distanzunterricht aufgrund der Schulschließungen immense Anstrengungen mit sich brachte, sollten auch über die Schulschließungen hinaus diejenigen Errungenschaften von Distanzunterricht, die sich als praktikabel und förderlich erwiesen haben, beibehalten werden. So könnten auch weiterhin Lernvideos zu wichtigen Inhalten gestaltet werden, die sich Schülerinnen und Schüler mehrfach anschauen können. Auch die Nutzung von digitalen Medien wie Tablets oder Laptops ist zu einer Selbstverständlichkeit geworden. Diese könnten beispielsweise im Hinblick auf adaptives Lehren und Lernen auch im regulären Unterricht Anwendung finden. Es bedarf jedoch einer kontinuierlichen und systematischen Evaluation deren Einsatzes, um sowohl über die Umsetzungsqualität als auch über Zusammenhänge mit Zielkriterien des Unterrichts weitere Erkenntnisse zu erhalten.

Die vorliegende Arbeit hat neben wichtigen Erkenntnissen auch Grenzen der Nutzung von Schülerurteilen aufgezeigt. Um ein tiefgreifenderes Verständnis davon zu erhalten, welche Merkmale der Unterrichtsqualität Schülerinnen und Schüler besonders wahrnehmen und wie sie diese bewerten, können experimentelle Studien mit Unterrichtsvideos weitere

Erkenntnisse liefern. Unterrichtsvideos bieten den Vorteil, dass sie, im Gegensatz zu direkten Beobachtungen im Unterricht, von mehreren unterschiedlichen Personen hinsichtlich verschiedener Merkmale beurteilt werden können (Krammer et al., 2012; Lotz et al., 2013). Zudem wäre es möglich, solche Unterrichtsvideos selbst herzustellen und durch geskriptete Unterrichtssequenzen hinsichtlich eines bestimmten Merkmals auszurichten (z. B. Ausmaß des Lobes oder des Umgangs mit Fehlern). Solche standardisierten Abläufe könnten nicht nur für Schulungen von Lehrkräften eingesetzt werden, sondern würden auch weitere Kenntnisse darüber liefern, welche Merkmale von Unterricht tatsächlich aus Sicht der Schülerinnen und Schüler wahrgenommen werden und was sie im Sinne eines qualitätsvollen Unterrichts als besonders relevant erachten. Da die vollständige Herstellung und die Auswertung von Unterrichtsvideos sehr zeit- und kostenaufwendig sind, ist das Thin-Slices-Verfahren eine hilfreiche Alternative (Ambady et al., 2000; Begrich et al., 2017). In diesem Verfahren werden meist ungeschulten Urteilern kurze Ausschnitte (z. B. 30 Sekunden) des Unterrichtsgeschehens gezeigt, die sie hinsichtlich festgelegter Merkmale beurteilen sollen (Begrich et al., 2017). Analysen von Unterrichtsvideos würden es beispielsweise ermöglichen, die Dimensionalität von Unterrichtsqualität zu überprüfen: Würden Schülerinnen und Schüler die Unterrichtssequenzen getrennt entweder hinsichtlich der Klassenführung oder der konstruktiven Unterstützung oder der kognitiven Aktivierung beurteilen, ließe sich nicht nur zeigen, welche Merkmale der jeweiligen Dimension sie im Unterrichtsgeschehen als relevant erachten, sondern auch, ob sich auch in diesem Verfahren eine abgrenzbare Struktur der einzelnen Qualitätsdimensionen zeigt. Dies würde weitere wichtige Erkenntnisse über die Validität von Schülerurteilen ermöglichen.

Die vorliegende Dissertation hat gezeigt, dass sich die Nutzung unterschiedlicher Item-Adressaten in psychometrischen Eigenschaften von Schülerurteilen widerspiegelt. Daher sollten diese Ergebnisse auch in zukünftigen Studien zu Unterrichtsqualität im Sinne konsistenter Formulierungen berücksichtigt werden. Während beispielsweise in der Diagnostik im Bereich der pädagogischen Psychologie einheitliche Instrumente eingesetzt werden (z. B. zur Feststellung von Verhaltensauffälligkeiten; Wilhelm & Kunina-Habenicht, 2015), gibt es in der empirischen Unterrichtsforschung zur Erfassung von Unterrichtsqualität eine große Variationsbreite. Klassenführung besteht beispielsweise in der Studie von Kunter et al. (2013) aus den zwei Subdimensionen „Prävention von Störungen“ und „Effektive Zeitnutzung“ (je drei Items). In der Studie von Kleickmann et al. (2020) wurden zur Erfassung von Klassenführung fünf Items verwendet, die im Kern auf Unterrichtsstörungen

abzielen. Ist also Klassenführung wirklich gleich Klassenführung? Selbst auf Basis der Items genügt ein kurzer Blick in die verwendeten Instrumente, um zu erkennen, dass für gleiche Subdimensionen unterschiedliche Items verwendet werden. Beispielsweise wird die Disziplin im Unterricht in der PISA-Studie 2015 durch fünf Items, in der DESI-Studie durch sieben Items erfasst (DESI-Konsortium, 2008; Mang et al, 2019). Zudem sind die Items unterschiedlich formuliert (z. B. in PISA: „Die Schülerinnen und Schüler hören der Lehrerin/dem Lehrer nicht zu.“; in DESI: „Wir hören nicht auf das, was unser Englischlehrer/unsere Englischlehrerin uns sagt.“).

Die Frage der Vergleichbarkeit stellt sich darüber hinaus nicht nur für unterschiedliche Operationalisierungen zwischen den Studien, sondern auch für einzelne Dimensionen in unterschiedlichen Kontexten. Dies trifft insbesondere auf die Erfassung von kognitiver Aktivierung zu, da sich für diese Dimension eine hohe Variabilität zwischen Unterrichtsstunden zeigte (Lipowsky & Bleck, 2019; Praetorius et al., 2014). Zudem ist aufgrund der stark fachlichen und inhaltlichen Ausrichtung der Dimension fraglich, ob diese generisch anwendbar ist. Falls nicht, wären Schülerurteile unterschiedlicher Fächer nicht ohne Weiteres vergleichbar.

6.3.2 Praktische Implikationen

Trotz der Notwendigkeit weiterführender Forschung lassen sich aus den Studien dieser Dissertation Implikationen für die Schulpraxis und Bildungsadministration ableiten. Alle drei Studien haben gezeigt, dass Schülerinnen und Schüler sowohl im Präsenzunterricht als auch im Distanzunterricht in der Lage sind, Unterrichtsqualität differenziert zu beurteilen. Aus diesem Grund können die in den Studien verwendeten Items und Skalen in beiden Unterrichtsformen eingesetzt werden, um von den Schülerinnen und Schülern Rückmeldung zum Unterricht zu erhalten. Besonders die in der UNITAS-Studie verwendete Bandbreite verschiedener Subdimensionen und Variablen kann einen umfassenden Überblick über die Wahrnehmung des Unterrichts aus Schülersicht geben.

In Studie 1 zeigte sich, dass insbesondere Unterrichtsmethoden, die eine soziale Interaktion ermöglichen, in positivem Zusammenhang mit den Schülerurteilen zur Unterrichtsqualität und zu Zielkriterien des Unterrichts stehen. Aus diesem Grund sollten Lehrkräfte im Distanzunterricht den Schülerinnen und Schülern regelmäßig die Möglichkeit geben, sich untereinander und mit der Lehrkraft in Videomeetings oder Treffen in kleineren Gruppen auszutauschen. Besonders die konsistenten Zusammenhänge mit Lernvideos, die die

Lehrkraft selbst erstellt hat, zeigen, dass sowohl Schülerinnen und Schüler als auch deren Eltern die Mühe und persönliche Präsenz der Lehrkraft wertschätzen. Daher sind selbst gestaltete Lernvideos den Videos externer Quellen vorzuziehen. Doch nicht nur für Extremfälle wie der Schließung gesamter Schulen lassen sich Erkenntnisse aus dieser Studie ziehen. Beispielsweise könnten digitale Tools zum kooperativen Arbeiten auch im Präsenzunterricht weiterhin genutzt werden. Auch könnte die Lehrkraft Erklärungen wichtiger Inhalte auf Video aufnehmen, sodass Schülerinnen und Schüler diese erneut anschauen können, wenn sie im Unterricht etwas nicht verstanden haben. Zudem könnten Kinder, die längerfristig den Unterricht nicht besuchen können, durch Videomeetings online am Unterricht teilnehmen. Insgesamt sollten diejenigen Errungenschaften, die sich im Distanzunterricht als praktikabel und lernförderlich herausgestellt haben, auch nach der Rückkehr zum Präsenzunterricht beibehalten werden.

Im Hinblick auf die Formulierung von Items zur Erfassung von Unterrichtsqualität zeigte Studie 2, dass in vielen Fällen Items mit Wir-Adressat zu leicht höheren Werten der eingeschätzten konstruktiven Unterstützung führen. Aus theoretischer Perspektive geht dieser Beantwortungsprozess jedoch mit einer größeren Unsicherheit einher. Zudem zeigte sich, dass Items mit Ich-Adressat Zusammenhänge mit Zielkriterien des Unterrichts individueller Schülerinnen und Schüler besser abzubilden scheinen. Beides lässt darauf schließen, dass es bei der Erfassung von konstruktiver Unterstützung zu verlässlicheren Ergebnissen führt, wenn Items mit Ich-Adressat verwendet werden. Möchte sich also eine Lehrkraft von ihrer Klasse Rückmeldung zur Unterrichtsqualität einholen und ist sie zudem interessiert an der Bedeutung der Unterrichtsqualität für das individuelle Lernen der Schülerinnen und Schüler, sollte sie in ihrer Befragung für die Dimension der konstruktiven Unterstützung Items mit einem Ich-Adressaten nutzen.

Schließlich sollte bei der Erfassung von Unterrichtsqualität darauf verzichtet werden, zwei Fächer gleichzeitig in einem Fragebogen abzufragen. Dies wäre beispielsweise im Zuge einer Schulevaluation denkbar, in welcher die Unterrichtsqualität in mehreren Fächern erfasst werden soll. Eine solche Abfrage könnte zu Vergleichsanstellungen der Schülerinnen und Schüler führen, sodass die Unterrichtsqualität desjenigen Faches, in dem sie die bessere (schlechtere) Note erhalten haben, relativ betrachtet aufgewertet (abgewertet) wird. Dies zeigte sich in bisherigen Studien insbesondere dann, wenn naturwissenschaftliche und sprachliche Fächer gegenübergestellt wurden. Daher sollten bei der Erfassung von

Unterrichtsqualität entweder nur einzelne Fächer erfragt werden oder die Abfrage mehrerer Fächer mit zeitlichem Abstand erfolgen.

6.4 FAZIT

Die vorliegende Dissertation leistet einen wichtigen Beitrag zum Verständnis der Nutzung von Schülerurteilen zur Erfassung von Unterrichtsqualität. Sowohl für Lehr-Lernsettings in Präsenz als auch auf Distanz haben sich Schülerurteile hinsichtlich wesentlicher Eigenschaften als reliable und valide Methode zur Erfassung von Unterrichtsqualität erwiesen. Zudem zeigte sich für den Präsenz- und Distanzunterricht, dass Schülerurteile mit unterschiedlichen Zielkriterien des Unterrichts assoziiert sind und daher wichtige Einblicke in das Lernen von Schülerinnen und Schülern ermöglichen. Im Hinblick auf die Erfassung der Dimension der konstruktiven Unterstützung kann angenommen werden, dass ein Ich-Adressat zu verlässlicheren Urteilen führt. Insgesamt lässt sich festhalten, dass Schülerurteile sowohl auf individueller Ebene als auch auf Klassenebene wichtige Informationen über die Unterrichtsqualität aus Perspektive der Schülerinnen und Schüler bereitstellen können. Jedoch sollten bei der Interpretation von Schülerurteilen mögliche Einflussfaktoren wie die Note eines anderen Faches bedacht werden.

REFERENZEN

Abrami, P. C. (1989). How should we use student ratings to evaluate teaching? *Research in Higher Education, 30*(2), 221–227. https://doi.org/10.1007/BF00992718

Abrami, P. C., Bernard, R. M., Bures, E. M., Borokhovski, E., & Tamim, R. (2011). Interaction in distance education and online learning: Using evidence and theory to improve practice. *Journal of Computing in Higher Education, 23*(2/3), 82–103. https://doi.org/10.1007/s12528-011-9043-x

Abrami P. C., d'Apollonia S., Rosenfield S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In: Perry R. P., & Smart J. C. (Eds), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385–445). Springer. https://doi.org/10.1007/1-4020-5742-3_10

Aebli, H. (1961). *Grundformen des Lehrens.* Ernst Klett.

Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (S. 110–145). Sage.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation In Education, 13*, 153–166. https://doi.org/10.1023/A:1008168421283

Aldrup, K., Klusmann, U., Lüdtke, O., Göllner, R., Trautwein, U. (2018). Social support and classroom management are related to secondary students' general school adjustment: A multilevel structural equation model using student and teacher ratings. *Journal of Educational Psychology, 110*, 1066–1083. https://doi.org/10.1037/edu0000256

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 32, pp. 201–271). Academic Press. https://doi.org/10.1016/S0065-2601(00)80006-4

Arens, A.K. & Möller, J. (2016). Dimensional comparisons in students' perceptions of the learning environment. *Learning and Instruction, 42*, 22–30. https://doi.org/10.1016/j.learninstruc.2015.11.001

Babu, S. & Mendro, R. (2003). *Teacher accountability: HLM-based teacher effectiveness indices in a state assessment program*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, USA.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575–589. https://doi.org/10.1016/S0022-5371(75)80045-4

Barbour, M. K. (2019). The landscape of K-12 online learning: Examining the state of the field. In M. G. Moore & W. C. Diehl (Eds.), *Handbook of distance education* (4th ed., pp. 521–542). New York: Routledge.

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

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. In J. Baumert, P. Stanat & R. Watermann (Hrsg.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* (S. 95–188). Springer VS. https://doi.org/10.1007/978-3-531-90082-7_4

Baumert, J., Gruehn, S., Heyn, S. Köller, O., & Schnabel, K. U. (1997). *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU) – Dokumentation (Band 1).* Berlin: Max-Planck-Institut für Bildungsforschung.

Begrich, L., Fauth, B., Kunter, M., & Klieme, E. (2017). Wie informativ ist der erste Eindruck? Das Thin-Slices-Verfahren zur videobasierten Erfassung des Unterrichts. *Zeitschrift für Erziehungswissenschaft*, 1–25. doi:10.1007/s11618-017-0730-x

Benton, S. L. & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature (IDEA Paper No. 50)*. Kansas State University, Center for Faculty Evaluation and Development.

Benton, S. L., Duchon, D., & Pallett, W. H. (2013). Validity of student self-reported ratings of learning. *Assessment & Evaluation in Higher Education, 38*(4), 377–388. https://doi.org/10.1080/02602938.2011.636799

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56*, 205–213. https://doi.org/10.1177/0022487105275904

Bloom, B. S. (1976). *Human characteristics and school learning*. McGraw-Hill.

Borich, G. D. (1986). Paradigms of teacher effectiveness research. Their relationship to the concept of effective teaching. *Education and Urban Society, 18*, 143–167. https://doi.org/10.1177/0013124586018002002

Bos, W., Strietholt, R., Goy, M., Stubbe, T. C., Tareli, I., & Hornberg, S. (2010). *IGLU 2006: Dokumentation der Erhebungsinstrumente*. Waxmann. http://doi.org/10.5159/IQB_IGLU_2006_v1

Bromme, R., Rheinberg, F., Minsel, B., Winteler, A., & Weidenmann, B. (2006). Die Erziehenden und Lehrenden. In: Krapp, Andreas; Weidenmann, Bernd (Hrsg.): *Pädagogische Psychologie. Ein Lehrbuch*. (5. vollst. überarb. Aufl., S. 269–356). Beltz.

Brophy, J. (2000). *Teaching*. Educational Practices Series-1, International Academy of Education.

Brophy, J. (2006). Observational research on generic aspects of classroom teaching. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 755–780). Lawrence Erlbaum Associates Publishers. http://doi.org/10.4324/9780203874790.ch33

Brophy, J. & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3. Auflage, S. 328–375). Macmillan.

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *Journal für Mathematik-Didaktik, 39*(2), 257–284. https://doi.org/10.1007/s13138-017-0122-z

Brunner, M., Kunter, M., Krauss, S., Baumert, J., Blum, W., Dubberke, T., Jordan, A., Klusmann, U., Tsai, Y.-M., & Neubrand, M. (2006). Welche Zusammenhänge bestehen zwischen dem fachspezifischen Professionswissen von Mathematiklehrkräften und ihrer Ausbildung sowie beruflichen Fortbildung? *Zeitschrift für Erziehungswissenschaft, 9*(4), 521–544. https://doi.org/10.1007/s11618-006-0166-1

Cabooter, E. F. K. (2010). *The impact of situational and dispositional variables on response styles with respect to attitude measures.* Doctoral dissertation, Ghent University, Ghent, Belgium

Carabajal, K., LaPointe, D., & Gunawardena, C. (2007). Group development in online distance learning groups. In M. Moore (Ed.), *Handbook of distance education* (pp. 137–148). Erlbaum.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 723–733.

Carroll, J. B. (1973). Implications of aptitude test research and psycholinguistic theory for foreign-language teaching. *International Journal of Psycholingistics*, *2*, 5–14.

Carroll, J. B. (1989). The Carroll Model. A 25-Year retrospective and prospective view. *Educational Researcher, 18*(1), 26–31.

Centra, J. A. & Gaubatz, N. B. (2000). Is there a gender bias in student evaluations of teaching*? Journal of Higher Education, 70*, 17–33. https://doi.org/10.2307/2649280

Chickering, A. W. & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *AAHE Bulletin, 39*(7), 3–7.

Cho, M.-H. & Kim, B. J. (2013). Students' self-regulation for interaction with others in online learning environments. *Internet and Higher Education, 17,* 69–75. https://doi.org/10.1016/j.iheduc.2012.11.001

Cho, M.-H. & Shen, D. (2013). Self-regulation in online learning. *Distance Education, 34*(3), 290–301. https://doi.org/10.1080/01587919.2013.835770

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

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Government Printing Office.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis, *Review of Educational Research, 77*(1), 113–143. http://dx.doi.org/10.3102/003465430298563

Davis, H. A. (2003). Conceptualizing the role and influence of student-teacher relationships on children's social and cognitive development. *Educational Psychologist, 38*(4), 207–234. http://dx.doi.org/10.1207/S15326985EP3804_2

De Jong, R. & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*, 51–85. https://doi.org/10.1023/A:1011402608575

Deci, E. L. & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality, 19*, 109–134. https://doi.org/10.1016/0092-6566(85)90023-6

Decristan, J., Hess, M., Holzberger, D., & Praetorius, A. K. (2020). Oberflächen- und Tiefenmerkmale: eine Reflexion zweier prominenter Begriffe der Unterrichtsforschung. *Zeitschrift für Pädagogik. Beiheft*, *66*(1), 102–116. https://doi.org/10.3262/ZPB2001102

Den Brok, P., Brekelmans, M., & Wubbels, T. (2006). Multilevel issues in research using students' perceptions of learning environments: The case of the questionnaire on teacher interaction. *Learning environments research*, *9*(3), 199–213. https://doi.org/10.1007/s10984-006-9013-9

DESI-Konsortium (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie*. Beltz

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. https://doi.org/10.1177/0895904808330173

Deutscher Bundestag (2016). *Kurzdarstellung zum Zusammenhang von Schulpflicht und Homeschooling in Deutschland*. https://www.bundestag.de/resource/blob/439052/ae8a7017b058abe1c92853fc9f0e7e33/wd-8-052-16-pdf-data.pdf

Dietrich, J., Dicke, A.-L., Kracke, B., & Noack, P. (2015). Teacher support and its influence on students' intrinsic value and effort: Dimensional comparison effects across subjects. *Learning and Instruction, 39*, 45–54. http://doi.org/10.1016/j.learninstruc.2015.05.007

Ditton, H. (2002). Lehrkräfte und Unterricht aus Schülersicht. Ergebnisse einer Untersuchung im Fach Mathematik. *Zeitschrift für Pädagogik, 48*(2), 262–286.

Doyle, W. (1977). Paradigms for research on teacher effectiveness. *Review of Research in Education, 5*, 163–198. https://doi.org/10.3102/0091732X005001163

Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Ed.), *Handbook of research on teaching. A project of the American Educational Research Association* (pp. 392–431). Macmillan.

Drechsel, B. & Schindler, A. K. (2019). Unterrichtsqualität. In *Psychologie für den Lehrberuf* (S. 353–372). Springer.

Drossel, K., Heldt, M., & Eickelmann, B. (2020). Die Implementation digitaler Medien in den Unterricht gemeinsam gestalten: Lehrer*innenbildung durch medienbezogene Kooperation. *Bildung, Schule, Digitalisierung*, 45.

Dubberke, T., Kunter, M., McElvany, N., Brunner, M., & Baumert, J. (2008). Lerntheoretische Überzeugungen von Mathematiklehrkräften: Einflüsse auf die Unterrichtsgestaltung und den Lernerfolg von Schülerinnen und Schülern. *Zeitschrift für pädagogische Psychologie*, *22*(34), 193–206. https://doi.org/10.1024/1010-0652.22.34.193

Dubs, R. (2009). Leitungsstrukturen in Bildungsorganisationen – Leadership und die Folgen für die Professionalität von Lehrenden. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität: Bedingungen, Genese, Wirkungen und ihre Messung* (S. 503–516). Beltz.

Eickelmann, B., Bos, W., Gerick, J., Goldhammer, F., Schaumburg, H., Schwippert, K., Senkbeil, M., & Vahrenhold, J. (2019). *ICILS 2018 #Deutschland: Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking*. Waxmann.

Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden*. Beltz.

Emmer, E. T. & Strough, L. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*(2), 103–112. https://doi.org/10.1207/S15326985EP3602_5

Eurostat (2017a). *Statistiken über elementare und primare Bildung*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Early_childhood_and_primary_education_statistics/de&oldid=400240

Eurostat (2017b). *Statistiken über sekundare Bildung.* https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Statistiken_%C3%BCber_sekundare_Bildung

Evertson, C., Emmer, E., Sanford, J., & Clements, B. (1983). Improving classroom management: An experiment in elementary classrooms. *Elementary School Journal, 84*, 173–188. https://doi.org/10.1086/461354

Evertson, C. M. & Weade, R. (1989). Classroom management and teaching style: Instructional stability and variability in two junior high English classrooms. *The Elementary School Journal*, *89*(3), 379–393. https://doi.org/10.1086/461581

Fauth, B., Atlay, C., Dumont, H., & Decristan, J. (2021). Does what you get depend on who you are with? Effects of student composition on teaching quality. *Learning and Instruction, 71*, 101355. https://doi.org/10.1016/j.learninstruc.2020.101355

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. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Göllner, R., Lenske, L., Praetorius, A., & Wagner, W. (2020). Who sees what? Theoretical considerations on the measurement of teaching quality from different perspectives. *Zeitschrift für Pädagogik, 66*, 138–155.

Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff, J., Lüdtke, O., Polikoff, M S., Trautwein, U. (2020). Don’t blame the teacher? The need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology*. Advance online publication. https://doi.org/10.1037/edu0000416

Feeley, T. H. F. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education, 51*, 225–236. https://doi.org/10.1080/03634520216519

Feldman, K. A. (1998). Reflections on the study of effective college teaching and student ratings: One continuing quest and two unresolved issues. In J. C. Smart (Ed.), *Higher education; Handbook of theory and research* (Vol. 13, pp. 35–74). Agathon.

Fend, H. (1981). *Theorie der Schule*. (2. Auflage). Urban & Schwarzenberg.

Fend, H. (1982). *Gesamtschule im Vergleich: Bilanz der Ergebnisse des Gesamtschulversuchs*. Beltz.

Fend, H. (1998). *Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistung.* Juventa.

Fend, H. (2008a). *Neue Theorie der Schule. Einführung in das Verstehen von Bildungssystemen* (2. Aufl.). Springer VS.

Fend, H. (2008b). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität*. Springer VS.

Fend, H. (2019). Erklärungen von Unterrichtserträgen im Rahmen des Angebot-Nutzungs-Modells. In U. Steffens & R. Messner (Hrsg), *Unterrichtsqualität. Konzepte und Bilanzen gelingenden Lehrens und Lernens. Grundlagen der Qualität von Schule 3* (S. 91–103). Waxmann.

Fendick, F. (1990). *The correlation between teacher clarity of communication and student achievement gain: A meta-analysis* (Unpublished doctoral dissertation). University of Florida.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28 https://doi.org/10.1177/ 003172171209400306

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117 – 140. https://doi.org/10.1177/001872675400700202

Fisicaro, S. A. & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14*(4), 419–429. http://dx.doi.org/10.1177/014662169001400407

Fraser, B. J., Giddings, G. J., & McRobbie, C. J. (1995). Evolution and validation of a personal form of an instrument for assessing science laboratory classroom environments. *Journal of Research in science Teaching*, *32*(4), 399–422. https://doi.org/10.1002/tea.3660320408

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652-670. https://doi.org/10.1037/0033-295X. 102.4.652

Gage, N. L. & Needels, M. C. (1989). Process-product research on teaching: A review of criticisms. *The Elementary School Journal, 89*, 253–300.

Fischer, C. & Platzbecker, P. (2018). *Auf den Lehrer kommt es an?! Unterstützung für professionelles Handeln angesichts aktueller Herausforderungen*. (Münstersche Gespräche zur Pädagogik, Band 34). Waxmann.

Gewerkschaft für Erziehung und Wissenschaft (2019). *Nur eine 3,8 für die digitale Ausstattung deutscher Schulen*. https://www.gew.de/aktuelles/detailseite/neuigkeiten/drei-minus-fuer-die-digitale-ausstattung-an-deutschen-schulen/

Göllner, R., Fauth, B., Lenske, G., Praetorius, A.K., Wagner, W. (2020). Do student ratings of classroom management tell us more about teachers or classroom composition? *Zeitschrift für Pädagogik, 66*, 156–172.

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

Göllner, R., Wagner, W., Klieme, E., Lüdtke, O., Nagengast, B., & Trautwein, U. (2016). Erfassung der Unterrichtsqualität mithilfe von Schülerurteilen. Chancen, Grenzen und Forschungsperspektiven. In Bundesministerium für Bildung und Forschung (Hrsg.), *Forschungsvorhaben in Ankopplung an Large-Scale-Assessments* (S. 63–82). wbv.

Götz, T., Frenzel, A. C., Hall, N. C., & Pekrun, R. (2008). Antecedents of academic emotions: Testing the internal/external frame of reference model for academic enjoyment. *Contemporary Educational Psychology, 33*, 9–33. https://doi.org/10.1016/j.cedpsych.2006.12.002

Graf, T., Harych, P., Wendt, W., Emmrich, R., & Brunner, M. (2016). Wie gut können VERA-8 Testergebnisse den schulischen Erfolg am Ende der Sekundarstufe I vorhersagen? *Zeitschrift für Pädagogische Psychologie, 30*, 201–211. https://doi.org/10.1024/1010-0652/a000182

Graham, C., Cagiltay, K., Lim, B. R., Craner, J., & Duffy, T. M. (2001). Seven principles of effective teaching: A practical lens for evaluating online courses. *The technology source*, *30*(5), 50.

Gräsel C. & Göbel K. (2015). VII-3 Unterrichtsqualität. In: Reinders H., Ditton H., Gräsel C., Gniewosz B. (Hrsg.) *Empirische Bildungsforschung*. Springer VS. https://doi.org/10.1007/978-3-531-19994-8_8

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182–1186. https://doi.org/10.1037/0003-066X.52.11.1182

Greenwald, A. G. & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209–1217. http://dx.doi.org/10.1037//0003-066x.52.11.1209

Griffin, B. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology, 29*, 410–425. http://dx.doi.org/10.1016/j.cedpsych.2003.11.001

Gruehn, S. (2000). *Unterricht und Lernen*. Waxmann.

Gründler R. (2018). Effektive Nutzung von Elementen klassischer Lehrdidaktik im Fernstudium durch Digitalisierung unter besonderer Berücksichtigung des Lehrenden-Lernenden-Verhältnisses. In: Arnold C., Knödler H. (Hrsg.) *Die informatisierte Service-Ökonomie*. Springer Gabler. https://doi.org/10.1007/978-3-658-21528-6_16

Hamre, B. K. & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. Meece & J. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 25–41). Routledge.

Harnischfeger, A. & Wiley, D. E. (1977). Kernkonzepte des Schullernens. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 9*, 207–228.

Hattie, J. (2003). *Teachers make a difference. What is the research evidence?* URL: https://research.acer.edu.au/cgi/viewcontent.cgi?article=1003&context=research_conference_2003

Hattie, J. (2009). *Visible Learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hawkins, A., Graham, C. R., Sudweeks, R. R., & Barbour, M. K. (2013). Academic performance, course completion rates, and student perception of the quality and

frequency of interaction in a virtual high school. *Distance Education*, *34*(1), 64–83. https://doi.org/10.1080/01587919.2013.770430

Helmke, A. (2007). *Unterrichtsqualität und Unterrichtsentwicklung: Wissenschaftliche Erkenntnisse zur Unterrichtsforschung und Konsequenzen für die Unterrichtsentwicklung*. Bertelsmann.

Helmke, A. (2010). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts*. Klett-Kallmeyer.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (4. Aufl.). Klett-Kallmeyer.

Helmke, A. (2017). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts* (7. Aufl.). Klett-Kallmeyer.

Helmke, A. & Weinert, F. E. (1997). *Bedingungsfaktoren schulischer Leistungen*. Max-Planck-Institut für Psychologische Forschung.

Hochweber, J. & Vieluf, S. (2018). Gender differences in reading achievement and enjoyment of reading: The role of perceived teaching quality. *The Journal of Educational Research*, *111*(3), 268–283. https://doi.org/10.1080/00220671.2016.1253536

Hugener, I., Pauli, C., & Reusser, K. (2006). Videoanalysen. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis*“. Materialien zur Bildungsforschung, Bd. 15. GFPF.

Jaekel, A.-K., Göllner, R., Trautwein, U. (2020). How students' perceptions of teaching quality in one subject are impacted by the grades they receive in another subject — Dimensional comparisons in student evaluations of teaching quality. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000488.

Jäger, R. S. & Helmke, A. (2008). *Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext (MARKUS)* (Version 1) [Datensatz]. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.5159/IQB_MARKUS_v1

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, Gintis, H., Heyns, B., Michaelson, S. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America*. Basic Books.

Krammer, K., Lipowsky, F., Pauli, C., Schnetzler, C. L., & Reusser, K. (2012). Unterrichtsvideos als Medium zur Professionalisierung und als Instrument der Kompetenzerfassung von Lehrpersonen. In M. Kobarg, C. Fischer, I. M. Dalehefte, F. Trepke, & M. Menk (Hrsg.), *Lehrerprofessionalisierung wissenschaftlich begleiten - Strategien und Methoden* (pp. 69–86). Waxmann.

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.

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review, 8*, 265–280. https://doi.org/10.1207/s15327957pspr0803_3

Kleickmann, T., Steffensky, M., & Praetorius, A.-K. (2020). Quality of teaching in science education: More than three basic dimensions? *Zeitschrift für Pädagogik, 66*. Beiheft, 37–55. https://doi.org/10.3262/ZPB2001037

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

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

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: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I. „Aufgabenkultur“ und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.),

TIMSS-Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente (S. 43–57). Bundesministerium für Bildung und Forschung.

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* (S. 231–255). Waxmann.

Köller, O., Trautwein, U., Lüdtke, O., & Baumert, J. (2006). Zum Zusammenspiel von schulischer Leistung, Selbstkonzept und Interesse in der gymnasialen Oberstufe. *Zeitschrift für Pädagogische Psychologie, 20*(1/2), 27–39. https://doi.org/10.1024/1010-0652.20.12.27

Kounin, J. S. (1976). *Techniken der Klassenführung*. Huber.

Kounin, J. S. (2006). *Techniken der Klassenführung* (Orig. der dt. Ausgabe, 1976). Waxmann.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213–236. https://doi.org/10.1002/acp.2350050305

Kuger, S., Klieme, E., Lüdtke, O., Schiepe-Tiska, A., & Reiss, K. (2017). Mathematikunterricht und Schülerleistung in der Sekundarstufe: Zur Validität von Schülerbefragungen in Schulleistungsstudien. *Zeitschrift für Erziehungswissenschaft, Sonderheft 33*, 61–98. Springer VS. https://doi.org/10.1007/s11618-017-0750-6

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the Tripod Student Survey. *Educational Assessment, 22*(4), 253–274. doi:10.1080/10627197.2017.1381555

Kumi-Yeboah, A., Dogbey, J., & Yuan, G. (2017). Exploring factors that promote online learning experiences and academic self-concept of minority high school students. *Journal of Research on Technology in Education, 50*(1), 1–17. https://doi.org/10.1080/15391523.2017.1365669

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. https://doi.org/10.1007/s10984-006-9015-7.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project*. Springer.

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17,* 494–509. https://doi.org/10.1016/j.learninstruc.2007.09.002

Kunter, M. & Ewald, S. (2016). Bedingungen und Effekte von Unterricht: Aktuelle Forschungsperspektiven aus der pädagogischen Psychologie. In N. McElvany, W. Bos, H. G. Holtappels, M. M. Gebauer & F. Schwabe (Hrsg.), *Bedingungen und Effekte guten Unterrichts* (S. 9–31). Waxmann.

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

Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction: motivation for teaching. *Learning and Instruction, 18*(5), 468–482. http://dx.doi.org/10.1016/j.learninstruc.2008.06.008

Kunter, M. & Voss, T. (2011). Das Modell der Unterrichtsqualität COACTIV: Eine multikriteriale Analyse. In: Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., Neubrand, M. (Hrsg.). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 83–113). Waxmann.

Ladenthin V. (2018) Homeschooling. In: Barz H. (Hrsg.) *Handbuch Bildungsreform und Reformpädagogik*. Springer VS. https://doi.org/10.1007/978-3-658-07491-3_49

Lance, C. E., LaPointe, J. A., & Fisicaro, S. A. (1994). Tests of three causal models of halo rater error. *Organizational Behavior and Human Decision Processes, 57,* 83–96. https://doi.org/10.1006/obhd.1994.1005

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

Lenski, A. E., Hecht, M., Penk, C., Milles, F., Mezger, M., Heitmann, P., Stanat, P., & Pant, H. A. (2016). *IQB-Ländervergleich 2012. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente.* Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.20386/HUB-42547

Lipowsky, F. (2006). Auf den Lehrer kommt es an: Empirische Evidenzen für Zusammenhänge zwischen Lehrerkompetenzen, Lehrerhandeln und dem Lernen der Schüler. In C. Allemann-Ghionda (Hrsg.), *Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern* (S. 47–70). Beltz.

Lipowsky, F. (2015). Unterricht. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie.* (S. 69–105). Springer.

Lipowsky, F. & Bleck, V. (2019). Was wissen wir über guten Unterricht? – Ein Update. In U. Steffens & R. Messner (Hrsg.), *Unterrichtsqualität: Konzepte und Bilanzen gelingenden Lehrens und Lernens* (S. 219–249). 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). Klinkhardt.

Lipowsky, F. & Hess, M. (2019). Warum es manchmal hilfreich sein kann, das Lernen schwerer zu machen - Kognitive Aktivierung und die Kraft des Vergleichens. In K. Schöppe & F. Schulz (Hrsg.), *Kreativität & Bildung – Nachhaltiges Lernen* (S. 77–132). kopaed.

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. https://doi.org/10.1016/j.learninstruc.2008.11.001

Liu, F. & Cavanaugh, C. (2012). Factors influencing student academic performance in online high school algebra. *Open Learning, 27*(2), 149–167. https://doi.org/10.1080/02680513.2012.678613

Lotz, M., Gabriel, K., & Lipowsky, F. (2013). Niedrig und hoch inferente Verfahren der Unterrichtsbeobachtung. Analysen zu deren gegenseitiger Validierung. *Zeitschrift für Pädagogik, 59*(3), 357–380.

Lotz, M. & Lipowsky, F. (2015). *Die Hattie-Studie und ihre Bedeutung für den Unterricht. Ein Blick auf ausgewählte Aspekte der Lehrer-Schüler-Interaktion.* http://www.frank-lipowsky.de/wp-content/uploads/Lotz-Lipowsky-2.pdf

Lüders, M. (2012). Der Unterrichtsbegriff in pädagogischen Nachschlagewerken. Ein empirischer Beitrag zur disziplinären Entwicklung der Schulpädagogik. *Zeitschrift für Pädagogik, 58*, 109 – 129.

Lüdtke, O., Robitzsch, A., Trautwein, U., Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings in multilevel modelling. *Contemporary Educational Psychology, 34*, 123–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9*, 215–230. https://doi.org/10.1007/s10984-006-9014-8

Lüdtke, O., Trautwein, U., Schnyder, I., & Niggli, A. (2007). Simultane Analysen auf Schüler- und Klassenebene. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *39*, 1–11. https://doi.org/10.1026/0049-8637.39.1.1

Maier, U., Kleinknecht, M., Metz, K., & Bohl, T. (2010). Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben. *Beiträge zur Lehrerbildung, 28*(l), 84–96.

Mang, J, Ustjanzew, N., Leßke, I., Schiepe-Tiska, A., Reiss, K. (2019). *PISA 2015 Skalenhandbuch: Dokumentation der Erhebungsinstrumente*. Waxmann.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self- concept. *Journal of Educational Psychology, 79*, 280–295. http://dx.doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Lüdtke, O. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*, 106–124. https://doi.org/10.1080/00461520.2012.670488

Marsh, H. W. & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81*, 59–77. http://dx.doi.org/10.1348/ 000709910X503501

Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018). An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over 6 years. *Developmental Psychology, 54*, 263–280. http://dx.doi.org/10.1037/dev0000393

Marsh, H. W. & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92*(1), 202–228. https://doi.org/10.1037/0022-0663.92.1.202

Marsh, H. W. & Seaton, M. (2013). Academic self-concept. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 62–63). Routledge.

Marsh, H. W., Trautwein, U., Lüdtke, O. & Köller, O. (2008). Social comparison and big-fish-little-pond effects on self-concept and efficacy perceptions: Role of generalized and specific others. *Journal of Educational Psychology, 100*, 510–524. https://doi.org/10.1037/0022-0663.100.3.510

Marzano, R. J., Gaddy, B., & Dean, C. (2000). *What works in classroom instruction?* Mid-Continent Research for Education and Learning.

Mayrberger K. (2017). Partizipatives Lernen in der Online-Lehre – Anspruch, Konzept und Ausblick. In: Griesehop H., Bauer E. (Hrsg.) *Lehren und Lernen online*. Springer VS. https://doi.org/10.1007/978-3-658-15797-5_6

McRobbie, C. J., Fisher, D. L., & Wong, A. F. L. (1998). Personal and class forms of classroom environment instruments. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 581–594). Kluwer.

McRobbie, C. J., Fraser, B. J., & Giddings, G. J. (1991). Comparison of personal and class forms of the Science Laboratory Environment Inventory. *Research in Science Education, 1991*(21), 244–252. https://doi.org/10.1007/BF02360478

Merkens, H. (2010). *Unterricht. Eine Einführung*. Springer Verlag.

Messner, R. (2019). Bausteine eines kognitiv aktivierenden Fachunterrichts. In U. Steffens & R. Messner (Hrsg.), *Unterrichtsqualität: Konzepte und Bilanzen gelingenden Lehrens und Lernens* (S. 201–218). Waxmann.

Mohammed, S. S. (2018). Measurement in Emerging Learning Environments. In K. Kennedy & R. E. Ferdig (Eds.), *Handbook of research on K-12 Online and Blended Learning* (pp. 111–120). ETC Press.

Möller, J., Helm, F., Müller-Kalthoff, H., Nagy, N., & Marsh, H. W. (2015). Dimensional comparisons and their consequences for self-concept, motivation, and emotion. In J.

D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (2. Auflage, S. 430–436). Elsevier. http://dx.doi.org/10.1016/B978-0-08-097086-8.26092-3

Möller, J., Müller-Kalthoff, H., Helm, F., Nagy, N., & Marsh, H. W. (2016). The generalized internal/external frame of reference model: An extension to dimensional comparison theory. *Frontline Learning Research, 4*, 1–11. http://dx.doi.org/10.14786/flr.v4i2.169

Möller, J., Pohlmann, B., Köller, O. & Marsh, H.W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79*, 1129–1167. https://doi.org/10.3102/0034654309337522

Möller, J., Streblow, L., & Pohlmann, B (2006). The belief in a negative interdependence of math and verbal abilities as determinant of academic self-concepts. *British Journal of Educational Psychology, 76*, 1–15. https://doi.org/10.1348/000709905X37451

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality and Quantity, 42*(6), 779–794. https://doi.org/10.1007/s11135-006-9067-x

Mullis, I. V. A., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks.* TIMMS & PIRLS International Study Center.

Mussweiler, T. (2003). Comparison processed in social judgment: Mechanisms and consequences. *Psychological Review, 110*, 472-489. https://doi.org/10.1037/0033-295X.110.3.472

National Standards for Quality (2019). *Quality online teaching*. https://www.nsqol.org/the-standards/quality-online-teaching/

Neuweg, G. H. (2011). Das Wissen der Wissensvermittler. Problemstellungen, Befunde und Perspektiven der Forschung zum Lehrerwissen. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 451–471). Waxmann.

Nilsen, T. & Gustafsson, J.-E. (Eds.) (2016). *Teacher quality, instructional quality and student outcomes*. Springer.

Ogrin S., Silber S., Friedrich A., Trautwein U., & Schmitz B. (2017) Entwicklung und empirische Prüfung einer Lehrkräftefortbildung zur Förderung von Selbstregulationskompetenz und mathematischer Kompetenz bei Schülerinnen und Schülern der Haupt- und Werkrealschule („Lernen mit Plan“). In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals* (S. 195–214). Springer VS.

Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology*, *26*, 382–399. https://doi.org/10.1006/ceps.2000.1070

OECD (2019). *PISA - Internationale Schulleistungsstudie der OECD*. https://www.oecd.org/berlin/themen/pisa-studie/

Oser, F. & Baeriswyl, F. J. (2001). Choreographies of teaching: Bridging instruction to learning. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 1031–1065). American Educational Research Association.

Ouzts, K. (2006). Sense of community in online courses. *Quarterly Review of Distance Education, 7*(3), 285–296.

Paulicke, P., Ehmke, T., Pietsch, M., & Schmidt, T. (2019). Wie beeinflusst die Kameraperspektive die Beurteilung der Unterrichtsqualität? *Zeitschrift für Bildungsforschung, 9*, 411–435. https://doi.org/10.1007/s35834-019-00246-2

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS: PreK-3)*. Brookes.

Pohlmann, B., Möller, J., & Streblow, L. (2004). Fremdeinschätzungen von Schülerselbstkonzepten durch Lehrer und Mitschüler. *Zeitschrift für Pädagogische Psychologie, 18*(3/4), 157–169. https://doi.org/10.1024/1010-0652.18.34.157

Polikoff, M.S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education, 121*(2), 183–212. https://doi.org/10.1086/679390

Praetorius, A.-K., Grünkorn, J. & Klieme, E. (2020). Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen. Einleitung in das Beiheft. *Zeitschrift für Pädagogik, 66*, Beiheft 1/20, 9–14.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM Mathematics Education, 50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A.-K., Klieme, E. & Kleickmann, T., Brunner, E., Lindmeier, A., Taut, S., & Charalombous, C. (2020). Towards developing a theory of generic teaching quality. *Zeitschrift für Pädagogik, 66*, Beiheft 1/20, 15–35.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Praetorius, A.-K., Rogh, W., & Kleickmann, T. (2020). Blinde Flecken des Modells der drei Basisdimensionen von Unterrichtsqualität? Das Modell im Spiegel einer internationalen Synthese von Merkmalen der Unterrichtsqualität. *Unterrichtswissenschaft, 48*(3), 303–318. https://doi.org/10.1007/s42010-020-00072-w

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rost, J, & Schiefele, U. (2013). *Programme for International Student Assessment - International Plus 2003, 2004 (PISA-I-Plus 2003, 2004)*. Berlin: IQB - Institute for Educational Quality Improvement.

Rammstedt B. (2010). Reliabilität, Validität, Objektivität. In: Wolf C., Best H. (Hrsg.) *Handbuch der sozialwissenschaftlichen Datenanalyse*. Springer VS. https://doi.org/10.1007/978-3-531-92038-2_11

Rakoczy, K., Klieme, E., Lipowsky, F., & Drollinger-Vetter, B. (2010). Strukturierung, kognitive Aktivität und Leistungsentwicklung im Mathematikunterricht. *Unterrichtswissenschaft, 38*(3), 229–246.

Reddy, L. A., Kettler, R. J., & Kurz, A. (2015). Schoolwide educator evaluation for improving school capacity and student achievement in high-poverty schools: Year 1 of the school system improvement project. *Journal of Educational and Psychological Consutlation, 25*, 90–108. https://doi.org/10.1080/10474412.2014.929961

Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E., & Köller, O. (Hrsg.) (2016). PISA 2015. *Eine Studie zwischen Kontinuität und Innovation.* Waxmann.

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

Reusser, K. (2009). Unterricht. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 881–896). Beltz.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 104*, 700–712. http://dx.doi.org/10.1037/a0027268

Rice, K. L. (2006). A comprehensive look at distance education in the K-12 context. *Journal of Research on Technology in Education, 38(*4), 425–448. http://dx.doi.org/10.1080/15391523.2006.10782468

Rjosk, C., Richter, D., Hochweber, J., Lüdtke, O., Klieme, E., & Stanat, P. (2014). Socioeconomic and language minority classroom composition and individual reading achievement: The mediating role of teaching quality. *Learning and Instruction, 32*, 63–72. https://doi.org/10.1016/j.learninstruc.2014.01.007

Roch, S. G., Paqin, A. R., & Littlejohn, T. W. (2009). Do raters agree more on observable items? *Human Performance, 22*, 391–409. https://doi.org/10.1080/08959280903248344.

Rösler, L., Zimmermann, F., Möller, J., & Retelsdorf, J. (2018). Effects of dimensional comparisons on domain-specific interests in initial teacher education: A validation of the generalized I/E model. *Learning and Individual Differences, 67*, 188–196. https://doi.org/10.1016/j.lindif.2018.08.011

Scheerens, J. & Bosker, J. (1997). *The foundations of educational effectiveness*. Elsevier.

Schenke, K., Ruzek, E., Lam, A. C., Karabenik, S. A., & Eccles, J. S. (2018). To the means and beyond: understanding variation in students' perceptions of teacher emotional support. *Learning and Instruction, 55*, 13–21. https://doi.org/10.1016/j.learninstruc.2018.02.003

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality. *Frontiers in Psychology*, *7*,1–16. https://doi.org/10.3389/fpsyg.2016.00110.

Schön, S. & Ebner, M. (2013). *Gute Lernvideos … so gelingen Web-Videos zum Lernen!* Books on Demand GmbH.

Schulz-Heidorf, K. & Gerick, J. (2017). Aktuelle Perspektiven und Entwicklungen in Large-Scale-Studien: Eine Hinführung. *Tertium Comparationis 23*(1), 1–8.

Seidel, T. (2014). Das Angebot-Nutzungs-Modell in der Unterrichtspsychologie. *Zeitschrift für Pädagogik, 60*(6), 850-866.

Seidel, T. (2015). Klassenführung. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 107–119). Springer.

Seidel, T. & Shavelson, R.J. (2007). Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. https://doi.org/10.3102/0034654307310317

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*, 105–126. https://doi.org/10.1207/s15326977ea1102_2

Simonson, M., Schlosser, C., & Orellana, A. (2011). Distance education research: A review of the literature. *Journal of Computing in Higher Education*, *23*(2-3), 124–142. https://doi.org/10.1007/s12528-011-9045-8

Simonson, M., Zvacek, S. M., & Smaldino, S. (2019). *Teaching and Learning at a Distance: Foundations of Distance Education 7th Edition*. IAP.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Sage.

Stahns, R. & Rieser, S. (2018). Qualität des Leseunterrichts in vierten Klassen in der Grundschule unter den Bedingungen von Mehrsprachigkeit. Ergebnisse der Lehrkräftebefragung von IGLU 2011. *Zeitschrift für Grundschulforschung*, *11*(1), 131–145. https://doi.org/10.1007/s42278-018-0007-3

Staub, F. C. & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology, 94*(2), 344–355. https://doi.org/10.1037/0022-0663.94.2.344

Stigler, J.W. & Hiebert, J. (1999). *The teaching gap. Best ideas from the world's teachers for improving education in the classroom*. The Free Press.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. Jossey-Bass.

Terhart, E. (2006). Kompetenzen von Grundschullehrerinnen und -lehrern: Kontext, Entwicklung, Beurteilung. In P. Hanke (Hg.*), Grundschule in Entwicklung. Herausforderungen und Perspektiven für die Grundschule heute* (S. 233–248). Waxmann.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

UNESCO (2020). *1.37 billion students now home as COVID-19 school closures expand, ministers scale up multimedia approaches to ensure learning continuity*. https://en.unesco.org/news/137-billion-students-now-home-covid-19-school-closures-expand-ministers-scale-multimedia

Van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement, 30*(1), 30–50. https://doi.org/10.1080/09243453.2018.1539015

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98*, 281–300. https://doi.org/10.1037/a0017908

Vieluf, S., Praetorius, A.-K., Rakoczy, K., Kleinknecht, M., & Pietsch, M. (2020). Angebots-Nutzungsmodelle der Wirkweise des Unterrichts: Eine kritische Auseinandersetzung mit ihrer theoretischen Konzeption. *Zeitschrift für Pädagogik, 66*, 63–80. Beiheft.

Wagner, W. (2008). *Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht am Beispiel der Studie DESI (Deutsch Englisch Schülerleistungen International) der Kultusministerkonferenz*. [Dissertation, Universität Koblenz-

Landau]. Deutsche Nationalbibliothek. http://kola.opus.hbz-nrw.de/volltexte/2008/234/pdf/Diss_(Publikation).pdf

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, *28*, 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, *108*(5), 705–721. https://doi.org/10.1037/edu0000075

Waldis, M., Grob, U., Pauli, C., & Reusser, K. (2010). Der schweizerische Mathematik-unterricht 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). Waxmann.

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal, 53*, 1834–1868. https://doi.org/10.3102/0002831216671864

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*(3), 249–294. http://dx.doi.org/10.2307/1170546

Wayne, A. J. & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89–122. https://doi.org/10.3102/00346543073001089

Weidinger, A. F., Spinath, B., & Steinmayr, R. (2015). Zur Bedeutung von Grundschulnoten für die Veränderung von Intrinsischer Motivation und Fähigkeitsselbstkonzept in Deutsch. *Zeitschrift für Pädagogische Psychologie*, *29*, 193–204.

Wilhelm, O. & Kunina-Habenicht, O. (2015). Pädagogisch-psychologische Diagnostik. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (2. Aufl., S. 307–332). Berlin: Springer.