Z Erziehungswiss (2022) 25:1431–1453
https://doi.org/10.1007/s11618-022-01121-z

ZfE

Unterrichtsqualität aus der Sicht von externen Raterinnen und Ratern – Analysen zum Reihenfolgeeffekt

Jennifer Igler · Annika Ohle-Peters · Nele McElvany

Eingegangen: 13. April 2021 / Überarbeitet: 18. Mai 2022 / Angenommen: 7. Juni 2022 / Online publiziert: 27. August 2022

Zusammenfassung Unterrichtsqualität wird häufig über Urteile von Externen erfasst. Mit Blick auf die Güte dieser Urteile stellt sich die Frage, ob diese aufgrund von Darbietungsmodalitäten sowie individuellen Faktoren systematisch verzerrt sein können. In der vorliegenden Studie wurden daher der Einfluss der Reihenfolge von Unterrichtsvideos mit unterschiedlichen Qualitätsniveaus sowie von individuellen Faktoren wie Vorerfahrung und mentaler Zustand (Müdigkeit) auf Unterschiede in den Urteilen und Interaktionen untersucht. Hierfür wurden $N = 69$ Studierende zu drei Versuchsgruppen zugeordnet, die zwei zehnminütige Videos in variierender Reihenfolge der Qualitätsniveaus präsentiert bekamen: Gruppe A (niedrig/mittel), B (mittel/hoch) und C (hoch/mittel). Die Analysen zeigten Unterschiede in den mittleren Ratings aufgrund der Reihenfolge sowie der Vorerfahrung und der Müdigkeit. Überdies wurden Interaktionen zwischen der experimentellen Bedingung und der Müdigkeit identifiziert. Die Ergebnisse wiesen auf Verzerrungen der Urteile von Externen hin, besonders im Bereich Motivierung. Konsequenzen für die Interpretation und Durchführung von videogestützten Unterrichtsqualitätsstudien werden diskutiert.

Schlüsselwörter Videorating · Beurteilungsfehler · Unterrichtsqualität · Reihenfolgeeffekt · Unterrichtsforschung

Jennifer Igler (✉) · Dr. Annika Ohle-Peters · Prof. Dr. Nele McElvany
Institut für Schulentwicklungsforschung der Technischen Universität Dortmund, TU Dortmund, Campus Nord (CDI Gebäude), Vogelpothsweg 78, 44227 Dortmund, Deutschland
E-Mail: jennifer.igler@tu-dortmund.de

Dr. Annika Ohle-Peters
E-Mail: annika.ohle-peters@tu-dortmund.de

Prof. Dr. Nele McElvany
E-Mail: nele.mcelvany@tu-dortmund.de

Teaching quality from the perspective of external raters—Analyses of the order effect

Abstract Ratings of external observers are often used for measuring instructional quality. These ratings can be influenced by the order of presentation and individual characteristics such as observers' experience and mental state (tiredness). The present study examined the influence of the presentation order of videos with varying levels of teaching quality and of individual factors such as previous experience and mental state (tiredness) on differences in observer ratings and interactions. Therefore, $N = 69$ students were assigned to three experimental groups. In each group, two videos of ten minutes length were presented in varying order of teaching quality: group A (low/medium), B (medium/high), and C (high/medium). Analyses showed significant differences in the ratings, depending on the sequential order, the observers' experience, and observers' tiredness. An interaction was found between experimental condition and observers' tiredness. Results suggest a bias of external judgements, especially concerning motivational quality. Consequences for interpretation and implementation of video-based studies measuring teaching quality are discussed.

Keywords Order effect · Rater bias · Research on teaching quality · Teaching quality · Video rating

1 Einleitung

Die Bedeutung der Qualität des unterrichtlichen Angebots für den Lernerfolg wird sowohl in theoretischen Modellen als auch in empirischen Studien betont (z. B. Lipowsky 2006). Um Aussagen über die Unterrichtsqualität treffen zu können, wird häufig auf Urteile von Externen zurückgegriffen. Gründe dafür sind die hohe Validität, die prädiktive Kraft für die Schülerleistungen, der emotionale Abstand zum Unterrichtsgeschehen sowie ein höheres Ausmaß an Vergleichsmöglichkeiten (Fauth et al. 2014a; Rakoczy 2008). Im Kontext der Raterurteile anhand von Unterrichtsvideos werden zur Sicherstellung der Objektivität Kodiermanuale entwickelt, welche je nach Inferenz des einzuschätzenden Merkmals ausdifferenziert werden, um Abweichungen in Beurteilungen zur selben Unterrichtseinheit zu minimieren (Seidel und Thiel 2017). Für die Sicherstellung der Validität und Reliabilität der Urteile werden Raterinnen und Rater in der Nutzung dieser Manuale geschult und gemeinsame Trainings durchgeführt (Lotz et al. 2013). Dennoch lassen sich Unterschiede in den Beurteilungen desselben Videos nicht vollständig eliminieren. Messergebnisse können durch Beurteilungsfehler verzerrt werden und somit zu einer Einschränkung der Nutzbarkeit und Interpretierbarkeit der Ergebnisse führen (Pietsch und Tosana 2008). Wenn die Raterinnen und Rater den gleichen Einflussfaktoren unterliegen, können Urteile aufgrund gleicher systematischer Verzerrung trotzdem hohe Interraterreliabilitäten aufweisen. Bisher gibt es nur wenige Erkenntnisse darüber, wie solche Verzerrungen von Einschätzungen der Unterrichtsqualität zustande kommen. Ein möglicher Erklärungsfaktor ist die Reihenfolge, in der Videos beurteilt werden,

da Informationen aus vorherigen Videos die neutrale Betrachtung eines darauffolgenden beeinträchtigen können (Mashburn et al. 2014). Verzerrungen der Urteile aufgrund der Darbietungsreihenfolge der Videos können zusätzlich durch individuelle Merkmale (z. B. geringe Vorerfahrung mit der Thematik und Müdigkeit) verstärkt werden (Gabriel-Busse et al. 2020; Webster et al. 1996). Unterrichtsbeobachtungen und Bewertungen des Lehrkraftverhaltens sowie der Unterrichtsqualität werden zunehmend in der Qualifikation sowie Weiterbildung von Lehrenden und Forschung eingesetzt (Mashburn et al. 2014). Demzufolge ist es bedeutsam, dass Verzerrungen in diesen Bewertungen so gering wie möglich gehalten werden. Die hier vorliegende Studie erweitert die bisherige Befundlage zu Beurteilungsfehlern und untersucht dabei gezielt Effekte der Darbietungsreihenfolge von Unterrichtsvideos sowie der individuellen Merkmale von Raterinnen und Ratern und mögliche Interaktionen.

2 Theoretischer Hintergrund

2.1 Merkmale qualitätsvollen Unterrichts

Qualitätsvoller Unterricht zeichnet sich unter anderem durch eine kognitiv anregende, störungsarme und unterstützende Lernumgebung aus (z. B. Praetorius et al. 2020). Um lernwirksame kognitive Prozesse zu aktivieren, ist das Anknüpfen an methodische sowie inhaltliche Vorkenntnisse der Lernenden und das Stellen herausfordernder Aufgaben bedeutsam (Lipowsky et al. 2009). Für eine erfolgreiche Auseinandersetzung mit dem Lernstoff ist ebenfalls die explizite Benennung von Lernzielen und das Vermitteln von inhaltlichen und methodischen Lernerwartungen von Relevanz (Meyer 2007).

Eine effiziente Klassenführung gilt als Voraussetzung für ein optimales Lernen in der Klasse. Ziel ist es, den Unterricht so zu strukturieren, dass möglichst viel Unterrichtszeit für den Lernstoff verwendet wird und Störungen möglichst vermieden werden (Brophy 1979; Emmer und Stough 2001). Von besonderer Bedeutung sind hierfür die Klarheit und die Einhaltung von Klassenregeln (Emmer und Stough 2001).

Als spezifische Facette der konstruktiven Unterstützung gilt die Motivierungsqualität, welche darauf abzielt, die Lernmotivation von Lernenden anzuregen und aufrecht zu erhalten (Kunter und Trautwein 2013). Für eine hohe Motivierung ist es wichtig, das Erleben von Selbstbestimmung zu fördern. Dies kann durch die Erfüllung psychologischer Grundbedürfnisse nach Autonomie, Kompetenz und sozialer Eingebundenheit erreicht werden (Ryan und Deci 2020). Konkret bedeutet dies, dass den Lernenden Entscheidungsmöglichkeiten geboten werden und ihnen sowie ihren Äußerungen eine wertschätzende Haltung entgegengebracht werden (Hamre und Pianta 2010; Rakoczy 2008).

Die beschriebenen Merkmale lassen sich der Tiefenstruktur von Unterricht zuordnen, welche als besonders relevant für den Lernerfolg gilt (Pauli und Reusser 2003). Allerdings sind die Merkmale der Tiefenstruktur nur indirekt beobachtbar und können nur über die Wahrnehmung von Teilnehmenden des Unterrichts oder die Beurteilung durch Externe erfasst werden (Decristan et al. 2020). Generell übereinstimm-

ten Einschätzungen über Unterrichtsqualitätsmerkmale von Lernenden, Lehrkräften und Externen nur gering (vgl. Clausen 2002; Fauth et al. 2014b). Eine Erklärung dafür ist, dass Unterrichtsqualitätsmerkmale in Abhängigkeit der einzuschätzenden Inhalte (Verhalten der Lehrkraft, Verhalten der Lernenden oder eine Mischung aus beidem) unterschiedlich von den Bewertungsperspektiven (Externe, Lehrkräfte und Lernenden) eingeschätzt werden. Für Unterrichtsqualitätsmerkmale mit geringer Beobachtbarkeit, zum Beispiel Skalen zur konstruktiven Unterstützung der Lehrkräfte, korrelieren die Urteile der Externen nicht stark mit denen der Lernenden- und Lehrkrafteinschätzungen (Fauth et al. 2020). Höhere Zusammenhänge wurden zwischen den drei Perspektiven im Bereich der Klassenführung festgestellt (z. B. Fauth et al. 2014b). Letztlich sind Externe, Lehrkräfte und Lernende nicht in gleicher Weise kompetent das gesamte Spektrum von Unterrichtmerkmalen einzuschätzen (Clausen 2002).

2.2 Unterrichtsbeurteilungen von externen Raterinnen und Ratern

Bei Unterrichtsbeurteilungen von Externen sind die Beobachtungen, die direkt im Klassenraum und die anhand von videografierten Unterrichtssequenzen erfolgen, zu unterscheiden. Beurteilungen auf Basis von Unterrichtsvideos ermöglichen, dass diese differenziert beurteilt und wiederholt angesehen werden und die Bewertungen auch zeitversetzt stattfinden können (Herrle et al. 2016; Mashburn et al. 2014). Urteile der Tiefenstruktur werden über hoch inferente Ratings abgegeben und setzen einen Inferenzschluss seitens der Raterinnen und Rater voraus, wobei von direkt beobachtbaren Merkmalen auf überdauernde Merkmale des Unterrichts geschlossen werden muss (Begrich et al. 2017). Obwohl in Studien zur videogestützten Unterrichtsanalyse Kodiermanuale und Ratertrainings eingesetzt werden, zeigten Untersuchungen, dass teilweise hohe Varianzanteile von bis zu 41,0 % auf Beurteilungsfehler und nicht auf das einzuschätzende Merkmal zurückzuführen sind (Hoyt und Kerns 1999; Pietsch und Tosana 2008; Praetorius 2014). Es bleibt ein Bedarf an theoretischen Begründungen und empirischen Prüfungen für diese Beurteilungsfehler und für den Einfluss von Merkmalen der Studie oder der Beurteilenden auf die Urteile.

2.3 Unterschiede in Unterrichtsbeurteilungen aufgrund des Reihenfolgeeffekts

Ein klassischer Beurteilungsfehler ist der Reihenfolgeeffekt, der bei Gedächtnisprozessen auftritt und auf die Arbeitsweise des Kurzzeitgedächtnisses zurückzuführen ist (Ebbinghaus 1885; Kooken et al. 2017). Demnach hat die Darbietungsreihenfolge, in der Informationen präsentiert werden, einen Einfluss auf die Speicherung und auf die Beurteilung von Informationen (Cushman und Mele 2008). Zum Beispiel können sich Urteile über Information A systematisch danach unterscheiden, ob die Informationen in der Reihenfolge A-B oder B-A vorgelegt werden. Dieser Effekt kann damit erklärt werden, dass die kognitive Beschäftigung mit dem vorigen Bewertungsgegenstand auf die Verfügbarkeit bestimmter Gedächtnisinhalte wirkt (Sudman et al. 1996). Urteile über die Information A können positiver ausfallen, wenn eine negativ bewertete Information B vorausgeht oder negativer ausfallen, wenn eine positiv

bewertete Information B vorausgeht (Miller und Campbell 1959). Folglich kann es zur Auf- oder Abwertung der Qualität eines Unterrichtsvideos kommen, je nachdem in welcher Darbietungsreihenfolge Unterrichtsvideos mit unterschiedlichen Qualitätsniveaus präsentiert werden. Empirisch konnte dies bestätigt werden, indem Ho und Kane (2013) feststellten, dass sich Raterinnen und Rater bei der Bewertung der Unterrichtsqualität eines Videos an dem vorigen Video orientierten und wiesen einen Zusammenhang zwischen einem zweiten Urteil und einem ersten Urteil auf. In einer Studie von Mashburn et al. (2014) konnte gezeigt werden, dass die Varianz, die auf die Urteilsverzerrung der Beurteilenden zurückgeführt werden konnte, minimiert wurde, wenn die Raterinnen und Rater die Reihenfolge von Videos in einer zufälligen Reihenfolge präsentiert bekamen.

Neben der Darbietungsreihenfolge können individuelle Faktoren einerseits die Urteile selbst und andererseits das Auftreten des Reihenfolgeeffekts beeinflussen (Bless et al. 2004).

2.3.1 Effekte der Vorerfahrung

Effekte der Vorerfahrung auf Urteile Unterschiede in Urteilen können aufgrund der Interaktion von Vorerfahrungen mit der Wahrnehmung und Interpretation eines Reizes auftreten (Bless et al. 2004; Myford und Wolfe 2003). Neue Reize werden über eine Bedeutungszuschreibung mittels der im Gedächtnis gespeicherten Kategorien, Schemata und Skripte encodiert und interpretiert (Bless et al. 2004). Demnach unterscheiden sich Urteile in Abhängigkeit der Größe des Speichers an Informationen über und Erfahrungen für das einzuschätzende Thema. Für den Bereich der Unterrichtsforschung fanden Gabriel-Busse et al. (2020) heraus, dass sich Urteilsfindungen über die kognitive Aktivierung und Klassenführung von Unterrichtsvideos sowie Urteilsbegründungen von Lehramtsstudierenden nach einem Zuwachs an Erfahrung vermittelt durch ein Seminar veränderten. Aufgrund einer höheren Erfahrung orientierten sich Studierende bei der Beurteilung stärker an den theoretischen Konstrukten und gewichteten diese unterschiedlich in der Urteilsbegründung. Wolff et al. (2017) zeigten, dass divergierende Erfahrungszeit in der Schule von Lehrkräften zu unterschiedlicher Fokussierung von Unterrichtsqualitätsmerkmalen und zu unterschiedlicher Tiefe der Schlussfolgerungen bei der Beurteilung von Unterrichtsvideos führte. Beispielsweise zogen Lehrkräfte mit weniger Erfahrungszeit in der Schule eher oberflächliche Unterrichtsereignisse für die Urteilsfindung im Vergleich zu Erfahrenen heran.

Effekte der Vorerfahrung auf das Auftreten von Beurteilungsfehlern Überdies kann Erfahrung einen Einfluss auf Verzerrungen in Urteilen haben und geringe Erfahrung kann das Auftreten von Beurteilungsfehlern verstärken (Praetorius 2013). Steht ein größerer Speicher an Informationen für die Verarbeitung zur Verfügung, so können akkuratere Beurteilungen über das Beobachtete getroffen und somit Beurteilungsfehler verringert werden (Bless et al. 2004; Lau und Plessner 2016). Liegt ein Mangel an relevanten Informationen zur Urteilsfindung vor, werden eher irrelevante Informationen berücksichtigt, welche zu systematisch verzerrten Bewertungen führen können (Messner und Schmid 2007). Empirisch konnten Feltz und Cokely

(2011) zeigen, dass Erfahrung mit der Thematik prädiktiv für das Auftreten des Reihenfolgeeffekts ist und eine höhere Erfahrung das Auftreten eines solchen minimieren kann.

2.3.2 Effekte des mentalen Zustands

Effekte des mentalen Zustands auf Urteile Neben Vorerfahrungen kann auch der mentale Zustand einen Einfluss auf die Urteilsfindung haben. Um Urteile zu fällen, werden Situationen zunächst wahrgenommen (Bless et al. 2004). Die Wahrnehmung von Reizen kann von aktuellen Zuständen beeinflusst werden, welche zur Selektion von Informationen führen und demzufolge auch Auswirkungen auf das nachfolgende Urteil haben können (Martin und Wawrinowski 2014). Insbesondere die Müdigkeit kann im Wahrnehmungsprozess dazu führen, dass aufgrund von fehlender Konzentration nicht alle relevanten Informationen beachtet werden können (Bless und Keller 2006). Studien dazu haben gezeigt, dass eine niedrige Konzentrationsfähigkeit oder Müdigkeit zur Selektion der wahrgenommenen Information und anschließend zu einer selektiven Beurteilung führte (Schmidt-Atzert et al. 2004). Eine andere Weise, wie der mentale Zustand auf Urteile wirken kann, ist, dass sich Personen Affektheuristiken (z. B. „Wie fühle ich mich dabei"-Heuristik) bedienen, um von ihrem momentanen Zustand auf ihre Urteile zu schließen (Bless et al. 1990). So können Urteile positiver oder negativer ausfallen, da der eigene mentale Zustand als Informationsquelle dient (Stroebe 2014). In Bezug auf das Beurteilen von Unterrichtsqualität konnte festgestellt werden, dass Raterinnen und Rater mit zunehmender Müdigkeit strengere Urteile vergaben (Mashburn et al. 2014).

Effekte der Müdigkeit auf das Auftreten von Beurteilungsfehlern Die Einschränkung der Informationsverarbeitung durch Müdigkeit kann auch ein höheres Auftreten von Beurteilungsfehlern durch eine mit der Müdigkeit einhergehender Tendenz schnellere Urteile zu fällen und häufigere Nutzung von fehleranfälligen Urteilsheuristiken bewirken (Engle-Friedman et al. 2018; Webster et al. 1996), mitunter etwa ein Auftreten des Reihenfolgeeffekts (Kruglanski und Webster 1996). Epley und Gilovich (2006) konnten zeigen, dass Beurteilungsfehler weniger häufig auftraten, wenn die Probandinnen und Probanden in einem konzentrierten mentalen Zustand waren. Auch in der Unterrichtsbeobachtung kann Müdigkeit Wahrnehmungsverzerrungen verstärken, indem für die Urteilsfindung relevante Unterrichtsereignisse nicht wahrgenommen werden, da offensichtlichere Ereignisse eine höhere Aufmerksamkeit auf sich ziehen (Schwindt 2008).

3 Forschungsfragen und Hypothesen

Ein substanzieller Anteil von Unterschieden in Urteilen kann auf Beurteilungsfehler zurückgeführt werden. Mögliche Erklärungen können Faktoren, wie die Darbietungsreihenfolge sowie individuelle Faktoren (z. B. Vorerfahrung oder mentaler Zustand) der Urteilenden sein. Zudem kann ein Reihenfolgeeffekt durch individuelle Faktoren verringert oder begünstigt werden.

F1 Treten Verzerrungen in den Urteilen der Unterrichtsqualitätsmerkmale kognitive Aktivierung, Klassenführung und Motivierung bei einem Unterrichtsvideo mittleren Qualitätsniveaus aufgrund der sequenziellen Darbietungsreihenfolge auf?

Je nachdem in welcher Reihenfolge Unterrichtsvideos unterschiedlicher Qualitätsniveaus präsentiert werden, können Urteile unterschiedlich ausfallen. Es wird angenommen, dass ein Unterrichtsvideo mittleren Qualitätsniveaus nach einem vorigen Urteil eines Videos niedrigen Qualitätsniveaus positiver bewertet wird, als wenn vorher kein anderes Video bewertet wurde (H1.1). Ebenfalls wird erwartet, dass ein Unterrichtsvideo mittleren Qualitätsniveaus nach einem vorigen Urteil eines Videos eines hohen Qualitätsniveaus negativer bewertet wird, als wenn vorher kein anderes Video bewertet wurde (H1.2).

F2 Stehen die individuellen Faktoren (a) Vorerfahrung (Erfahrungszeit in der Schule) oder (b) mentaler Zustand (Müdigkeit) und die Urteile der Unterrichtsqualitätsmerkmale kognitive Aktivierung, Klassenführung und Motivierung bei einem Unterrichtsvideo mittleren Qualitätsniveaus sowie Unterschiede im Auftreten des Reihenfolgeeffekts in systematischem Zusammenhang?

Aufgrund von unterschiedlich vorgespeicherten Informationen über das einzuschätzende Thema kommt es zu divergierenden Wahrnehmungen und schließlich auch Urteilen (Bless et al. 2004). Dementsprechend wird davon ausgegangen, dass die Urteile der Unterrichtsqualitätsmerkmale kognitive Aktivierung, Klassenführung und Motivierung bei einem Unterrichtsvideo mittleren Qualitätsniveaus durch die Vorerfahrung (Erfahrungszeit in der Schule) der Raterinnen und Rater mitbedingt werden (H2.1). Beurteilungsfehler können durch einen größeren Speicher an Informationen verringert werden (Lau und Plessner 2016). Demnach wird erwartet, dass eine höhere Erfahrungszeit in der Schule das Auftreten des Reihenfolgeeffekts verringert (H2.2). Hinzu wirkt der mentale Zustand (Müdigkeit) auf die Wahrnehmung von Reizen, was wiederum zur Selektion von Informationen führen und Auswirkungen auf das nachfolgende Urteil haben kann (Martin und Wawrinowski 2014). Folglich wird davon ausgegangen, dass die Urteile der Unterrichtsqualitätsmerkmale kognitive Aktivierung, Klassenführung und Motivierung bei einem Unterrichtsvideo mittleren Qualitätsniveaus durch den mentalen Zustand (Müdigkeit) der Raterinnen und Rater mitbedingt werden (H2.3). Aufgrund von Müdigkeit werden häufig schnellere Urteile unter anderem mittels Urteilsheuristiken gefällt, was wiederum zu einem häufigeren Auftreten von Beurteilungsfehlern führen kann (Engle-Friedman et al. 2018). Es wird erwartet, dass eine höhere Müdigkeit das Auftreten des Reihenfolgeeffekts verstärkt (H2.4).

4 Methodisches Vorgehen

4.1 Stichprobe

Die Daten wurden im Frühjahr 2017 im Rahmen eines experimentellen Settings erhoben. Insgesamt nahmen $N = 69$ Studierende an der Studie teil. Die Probandinnen und Probanden wurden zufällig auf drei experimentelle Bedingungen verteilt, die

Tab. 1 Deskriptiva der Gesamt- und Teilstichproben

Stichproben		*n*	M_{Alter} (*SD*)	Weiblich in %	$M_{Semester}$ (*SD*)	Lehramtsstudium in %	Erfahrung in der Schule in %
Alle Probandinnen und Probanden		69	24,91 (4,40)	80,9	7,84 (4,66)	50,7	62,3
Experimentelle Bedingung A		23	24,87 (4,95)	81,8	7,32 (4,03)	47,8	54,5
Video 1 Niedriges Qualitätsniveau	Video 2 Mittleres Qualitätsniveau						
Experimentelle Bedingung B		23	23,78 (3,09)	82,6	6,96 (4,25)	47,8	60,9
Video 1 Mittleres Qualitätsniveau	Video 2 Hohes Qualitätsniveau						
Experimentelle Bedingung C		23	26,14 (4,81)	78,3	9,22 (5,43)	56,5	73,9
Video 1 Hohes Qualitätsniveau	Video 2 Mittleres Qualitätsniveau						

Die höheren Mittelwerte der Semesteranzahl in der Bedingung C können auf zwei einzelne Extremfälle zurückgeführt werden (Semester = 19; Semester = 20)

sich in der Darbietungsreihenfolge der Unterrichtsvideos mit unterschiedlichen Qualitätsniveaus unterschieden: Gruppe A (niedrig/mittel), B (mittel/hoch) und C (hoch/mittel) (vgl. Tab. 1). Eine vorige Poweranalyse ergab, dass bei einer Power von 0,80 und einem Effekt von $f = 0{,}40$ $N = 64$ Personen benötigt werden. Demographische Angaben der Probandinnen und Probanden sowie die Erfahrungszeit in der Schule wurden im Voraus in einem Online-Fragebogen erhoben, damit Studierende des gleichen Studienfachs gleichmäßig auf die Gruppen aufgeteilt werden konnten. Die Probandinnen und Probanden der drei Gruppen unterschieden sich nicht hinsichtlich des Anteils des Geschlechts ($\chi^2(2) = 0{,}16$; $p > 0{,}05$), ihres Alters ($F[2, 65] = 1{,}64$; $p > 0{,}05$), ihrer Semesterzahl ($F[2, 65] = 1{,}58$; $p > 0{,}05$) oder des Anteils des Lehramtsstudiums ($\chi^2(2) = 0{,}46$; $p > 0{,}05$).

4.2 Instrumente und Durchführung

Unterrichtsvideos Die Videosequenzen stammten aus dem DFG geförderten Kooperationsprojekt *Entwicklung und Überprüfung von Kompetenzmodellen zur integrativen Verarbeitung von Texten und Bildern* (BiTe) (vgl. McElvany et al. 2012) und zeigten einen circa zehnminütigen Unterrichtseinstieg in vierten Klassen zur vorgegebenen Thematik „Südamerika“. Die Auswahl der Videosequenzen mit unterschiedlich hohen Qualitätsniveaus für die vorliegende Studie erfolgte in zwei Schritten. Zunächst wurden drei Videos aufgrund ihrer Qualität (niedrig [1] – hoch [4]; *niedrig* $M = 2{,}14$, $SD = 0{,}45$, *mittel* $M = 2{,}24$, $SD = 0{,}46$ und *hoch* $M = 2{,}63$, $SD = 0{,}36$) und des Auftretens der zu untersuchenden Facetten in den ersten 10 min aus einem Pool von schon gerateten Videos des DFG-Projekts BiTe ausgesucht. Um die Auswahl der Videos zu bestätigen, wurden zusätzlich $N = 8$ Expertinnen und Experten aus der Unterrichtsforschung bezüglich der Unterrichtsqualität in den drei ausge-

wählten Videos sowie zum Untersuchungsdesign befragt. Die Zuordnung zu den Kategorien *niedrig, mittel* und *hoch* konnte durch die Expertisen unterstützt werden.

Ratingmanual und Ratingbogen Das Ratingmanual wurde in Anlehnung an Ohle und McElvany (2016) und Praetorius (2014) entwickelt und gliederte sich in einen strukturellen Teil (Informationen zu dem Beurteilungsprozess der Unterrichtssequenzen) sowie einen inhaltlichen Teil (Beschreibung der drei zu beurteilenden Unterrichtsqualitätsmerkmale). Zu den einzelnen Merkmalen *kognitive Aktivierung, Klassenführung* und *Motivierung* wurden zunächst im Ratingmanual eine Definition sowie zwei Facetten mit positiven und negativen Beispielindikatoren gegeben.

Für die Erfassung der Unterrichtsqualität wurden im Ratingbogen drei Items eingesetzt, bei denen jeweils der Gesamteindruck der Qualität der kognitiven Aktivierung, Klassenführung oder Motivierung im gerade gesehenen Video auf einer vierstufigen Antwortskala (niedrig [1] – hoch [4]) eingeschätzt werden sollte (Item „Wir möchten Sie bitten, einen Gesamteindruck der Unterrichtsqualität einzuschätzen. Bitte tragen Sie unten ein, wie hoch Sie die Qualität der Merkmale im gerade gesehenen Video einschätzen.“).

Erfassung individueller Merkmale Zur Erfassung der Vorerfahrungen gaben die Studierenden an, wie viele Wochen Unterrichtserfahrung sie in der Schule gesammelt haben ($M = 16{,}70$, $SD = 31{,}32$; $Min = 0{,}00$, $Max = 200{,}00$). Die *Erfahrungszeit in der Schule* wurde mitunter im Rahmen eines Praktikums ($n = 36$), einer Vertretungsstelle ($n = 7$) oder eines Praxissemesters ($n = 6$) gesammelt. $N = 25$ Probandinnen und Probanden gaben an, keine Erfahrung in der Schule erworben zu haben. Für die Operationalisierung der *Müdigkeit* wurde die „Aktuelle Stimmungsskala“ eingesetzt (Dalbert 1992). Dabei sollte aus einer Liste mit Items, die verschiedene mentale Zustände beschreiben, für jedes Item die Zahl angekreuzt werden, welche den aktuellen Gefühlszustand am besten beschreibt (Vier Items: abgeschlafft, müde, erschöpft und entkräftet; Ratingskala überhaupt nicht [1] – sehr stark [7]; $M = 3{,}12$, $SD = 1{,}25$; $Min = 1{,}00$, $Max = 6{,}75$; $\alpha = 0{,}90$). Bis auf die Variable *Erfahrungszeit in der Schule* (2,9 % fehlend) gab es keine fehlenden Werte.

Durchführung der Untersuchung Vorab wurde den Probandinnen und Probanden die Information gegeben, dass sie an einer Studie zur akkuraten Einschätzung von Unterrichtsqualität anhand von Videosequenzen und der Relevanz des Studienfachhintergrundes teilnahmen. Um die Vergleichbarkeit der Beurteilungen sicherzustellen, wurden alle Probandinnen und Probanden im Bereich *Ratings von Unterrichtsqualität* drei Stunden standardisiert geschult (vgl. Tab. 2). Hierzu erfolgte zuerst eine theoretische Einführung in die Unterrichtsqualität und in die manualbasierte Videoanalyse. Im Anschluss wurden der Inhalt und die Struktur des Ratingmanuals sowie des Ratingbogens erklärt und eine gemeinsame Übung für das Beurteilen einer Unterrichtssequenz anhand eines Videos hohen Qualitätsniveaus angeleitet. Danach folgten zwei Einzelübungen, in denen die Studierenden eigenständig je ein Video niedrigen und mittleren Qualitätsniveaus mit Hilfe des Ratingmanuals und des Ratingbogens (siehe 4.2) einschätzten. Die Urteile wurden jeweils nach jedem Video gemeinsam besprochen, um die Urteilsübereinstimmung zu opti-

Tab. 2 Durchführung der Studie

Zeitpunkt	Schritte		Zeit
Im Voraus	Onlinefragebogen vorab: – Demographische Angaben – Studienfach		5 min
Einführung	Kurze theoretische Einführung in die Dimensionen der Unterrichtsqualität und manualbasierte Videoanalyse		190 min
	Einarbeiten in das Ratingmanual		
	Gemeinsames Raten einer Sequenz „Video mit hohem Qualitätsniveau“ & Besprechung		
	Einzelrating Trainingssequenz 1 „Video mit niedrigem Qualitätsniveau“ & Besprechung		
	Einzelrating Trainingssequenz 2 „Video mit mittlerem Qualitätsniveau“ & Besprechung		
	Pause		
	Skala zur aktuellen Stimmung		
Hauptuntersuchung	Einzelrating der zwei Hauptuntersuchungssequenzen		40 min
	Experimentelle Bedingung A		
	Video 1 Niedriges Qualitätsniveau	Video 2 Mittleres Qualitätsniveau	
	Experimentelle Bedingung B		
	Video 1 Mittleres Qualitätsniveau	Video 2 Hohes Qualitätsniveau	
	Experimentelle Bedingung C		
	Video 1 Hohes Qualitätsniveau	Video 2 Mittleres Qualitätsniveau	
–	Abschluss		5 min
–			240 min (4 h inkl. Pausen)

Tab. 3 Mittelwerte und Standardabweichungen der Unterrichtsvideos

Experimentelle Bedingung	Qualitätsniveau	Kognitive Aktivierung *M* (*SD*)	Klassenführung *M* (*SD*)	Motivierung *M* (*SD*)
A	Niedrig	1,48 (0,51)	2,73 (0,75)	1,65 (0,49)
	Mittel	3,61 (0,50)	2,87 (0,63)	2,83 (0,72)
B	Mittel	3,52 (0,51)	2,61 (0,78)	3,00 (0,74)
	Hoch	3,61 (0,50)	3,26 (0,45)	2,96 (0,71)
C	Hoch	3,70 (0,47)	3,43 (0,59)	2,83 (0,58)
	Mittel	3,09 (0,51)	2,26 (0,75)	2,35 (0,49)

mieren. Nach den zwei Trainingsvideos erreichten die Interraterreliabilitäten, Reliabilitäten des Mittelwertes aller Raterinnen und Rater, zufriedenstellende Werte von $ICC_{2kognitive\ \ Aktivierung} = 0{,}82$, $ICC_{2Klassenführung} = 0{,}86$ und $ICC_{2Motivierung} = 0{,}63$ (Wirtz und Caspar 2002). Anschließend bearbeiteten die Probandinnen und Probanden Fragen zur aktuellen Stimmung, in denen unter anderem die Müdigkeit erfasst wurde.

Zuletzt schätzten die Probandinnen und Probanden je nach experimenteller Bedingung die Unterrichtsqualität von zwei Videos mit einem niedrigen bzw. hohen und mittleren Qualitätsniveau ein (vgl. Tab. 3). Die Urteile erfolgten auch hier direkt

nach jedem der zwei gesehenen Videos, wobei ein erneutes Ansehen des jeweiligen Videos für die Beurteilung der einzelnen Unterrichtsqualitätsmerkmale nicht erlaubt war. Die Gruppe B, die das Referenzvideo mittleren Qualitätsniveaus zuerst sah und daher keiner möglichen Beeinflussung durch ein vorheriges Video unterlag, stellte die Referenzgruppe dar. Für eine vergleichbar lange Durchführung des Experiments erhielten auch die Probandinnen und Probanden der Gruppe B ein zweites Video zum Einschätzen. Den gemeinsamen Abschluss der Sitzung bildete die Aufklärung der Teilnehmenden über den genauen Studienzweck.

4.3 Auswertungsstrategie

Die Analysen zur Skalenreliabilität, Deskriptiva und *t*-Tests wurden mit Hilfe des Programms SPSS 24 durchgeführt. Für die Untersuchung des Reihenfolgeeffekts (F1) wurden *t*-Tests für unabhängige Stichproben gerechnet. Hierbei wurden im ersten Schritt die Unterschiede zwischen den Gruppen A (niedrig/mittel) und B (mittel/hoch) und im zweiten Schritt zwischen den Gruppen B (mittel/hoch) und C (hoch/mittel) jeweils für die Unterrichtsqualitätsmerkmale kognitive Aktivierung, Klassenführung und Motivierung überprüft. Es wurde eine Korrektur nach Bonferroni-Holm vorgenommen, um die Alpha-Fehler-Kumulierung zu berücksichtigen (Holm 1979). Für die Berechnungen des Unterschieds zwischen den Gruppen B (mittel/hoch) und C (hoch/mittel) im Bereich kognitive Aktivierung wurde der Mann-Whitney-U-Test gerechnet, da mit ungleichen Varianzen die Voraussetzung für die Berechnung von *t*-Tests nicht gegeben war. Für eine bessere Interpretierbarkeit der Ergebnisse wurde das Urteil des Videos mittleren Qualitätsniveaus der Gruppe B auf null gesetzt und für die anderen beiden Gruppen die Abweichungen der Urteile zum Mittelwert der Gruppe B bestimmt.

Die Fragestellung 2 zu den Effekten der Vorerfahrung und Müdigkeit wurde mit Hilfe von Regressions- sowie Moderationsanalysen mittels des Mplus-Softwarepakets (Muthén und Muthén 1998–2017) überprüft. In diesen Analysen stellten die Variablen zur Einschätzung der Unterrichtsqualität bei dem Video mittlerer Qualität die abhängigen Variablen dar. Es wurden vorab zwei Dummy-Variablen (Dummy 1: A = 1, B = 0, C = 0 und Dummy 2: A = 0, B = 0, C = 1) für die Gruppenzugehörigkeit der experimentellen Bedingung A und C gebildet. Insgesamt wurden drei Modelle je Unterrichtsqualitätsmerkmal gerechnet. Zur Überprüfung der Hypothese 2.1 und 2.3 wurden in dem Modell I zur Kontrolle des Reihenfolgeeffekts die Dummy-Variablen der Bedingung A und C sowie die Erfahrungszeit oder die Müdigkeit als unabhängige Variable aufgenommen. Für die Überprüfung der Hypothesen 2.2 und 2.4 wurden Moderationsanalysen gerechnet, bei denen Interaktionsterme zwischen den zwei Dummy-Variablen und der Erfahrungszeit oder Müdigkeit definiert wurden. Die unabhängigen Variablen in dem Modell II stellten zur Kontrolle des Reihenfolgeeffekts die Dummy-Variablen der Bedingung A und C sowie die Interaktionsterme da. Ein möglicher Einfluss der individuellen Faktoren auf Unterschiede im Auftreten des Reihenfolgeeffekts würde sich in einer signifikanten Interaktion zwischen den Gruppen A oder C und dem jeweiligen individuellen Faktor zeigen. In dem Modell III wurden schließlich als unabhängige Variablen zur Kontrolle des Reihenfolgeeffekts die Dummy-Variablen der Bedingung A und C, die Erfahrungs-

zeit oder die Müdigkeit und die Interaktionsterme aufgenommen. In allen Modellen wurden die Variablen manifest modelliert und die Variablen Erfahrungszeit und Müdigkeit standardisiert. Da es sich um saturierte Modelle handelt, wurden keine globalen Fitindices angegeben (Geiser 2011). Außerdem wurden Korrelationen zwischen den unabhängigen Variablen zugelassen. Zusätzlich wurde in allen Modellen die in Mplus implementierten Optionen maximum-likelihood-estimator (ML) und full information maximum likelihood (FIML) verwendet.

5 Ergebnisse

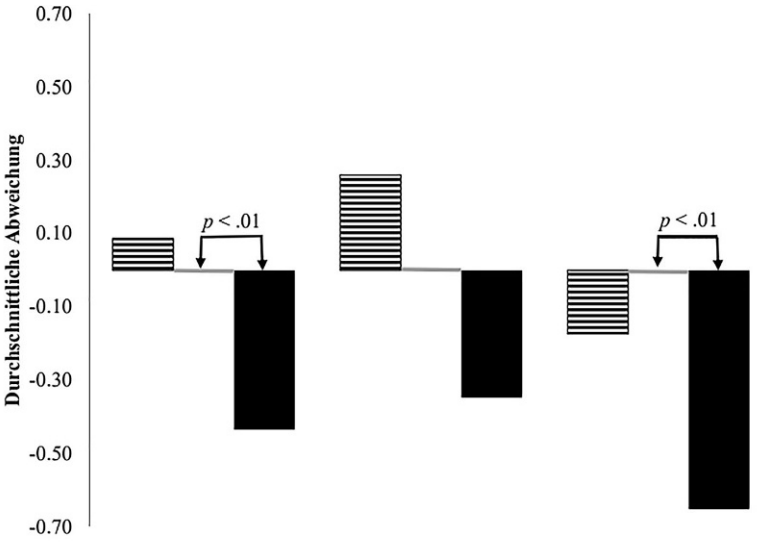
5.1 Unterschiede in den Urteilen aufgrund der Darbietungsordnung

Bezüglich der Fragestellung 1, ob sich Unterschiede in der durchschnittlichen Beurteilung eines Videos mittleren Qualitätsniveaus aufgrund der Darbietungsordnung zeigen, wiesen die Ergebnisse der *t*-Tests auf Unterschiede in den Beurteilungen der Gruppen B und C hinsichtlich den Unterrichtsqualitätsmerkmalen kognitive Aktivierung und Motivierung hin (Abb. 1).[1]

Gruppe A (niedrig/mittel) vs. Gruppe B (mittel/hoch) Zwischen den Urteilen der Gruppen A (niedrig/mittel) und B (mittel/hoch) ließen sich keine Unterschiede bezüglich der kognitiven Aktivierung ($t(44) = 0{,}58$, $p = 0{,}844$, Cohens $d = 0{,}17$), Klassenführung ($t(44) = 1{,}25$, $p = 0{,}654$, Cohens $d = 0{,}37$) und Motivierung ($t(44) = -0{,}81$, $p = 0{,}844$, Cohens $d = -0{,}24$) finden. Raterinnen und Rater, die vor der Beurteilung des Referenzvideos ein Video niedrigen Qualitätsniveaus sahen, beurteilten die Qualität der kognitiven Aktivierung, Klassenführung und Motivierung des Referenzvideos nicht signifikant abweichend. Dementsprechend konnte die Hypothese 1.1 zur Aufwertung der Urteile durch die Reihenfolge nicht bestätigt werden.

Gruppe B (mittel/hoch) vs. Gruppe C (hoch/mittel) Es zeigten sich signifikante Differenzen zwischen den Urteilen der Gruppen B (mittel/hoch) und C (hoch/mittel) hinsichtlich der kognitiven Aktivierung ($U = 161{,}50$, $Z = -2{,}64$, $p = 0{,}048$, Cohens $d = 0{,}84$) sowie der Motivierung ($t(44) = 3{,}54$, $p = 0{,}009$, Cohens $d = 1{,}04$). Demzufolge schätzten die Raterinnen und Rater, die vor dem Referenzvideo ein Video hohen Qualitätsniveaus beurteilten, die Qualität der kognitiven Aktivierung und Motivierung des Referenzvideos mit mittlerer Qualität niedriger ein als die Referenzgruppe. Es wurden keine statistisch signifikanten Differenzen zwischen den Urteilen der Gruppen B und C ($t(44) = 1{,}54$, $p = 0{,}524$, Cohens $d = -0{,}14$) bezüglich der Klassenführung festgestellt. Raterinnen und Rater, die vor der Bewertung des Referenzvideos ein Video mit einem niedrigeren oder höheren Qualitätsniveau sahen, schätzten die Qualität der Klassenführung des Referenzvideos nicht signifikant

[1] Werden für Kontrollanalysen zur Robustheit der Befunde die Semesteranzahl oder das Alter als Kovariaten in Varianzanalysen, bei denen alle drei Gruppen A, B und C verglichen werden, aufgenommen, bleiben die Ergebnisse stabil.



	Kognitive Aktivierung *M* (*SD*)	Klassenführung *M* (*SD*)	Motivierung *M* (*SD*)
Experimentelle Bedingung A	0.09 (0.50)	0.26 (0.63)	-0.17 (0.72)
Experimentelle Bedingung B	0.00 (0.51)	0.00 (0.78)	0.00 (0.74)
Experimentelle Bedingung C	-0.43 (0.51)	-0.35 (0.75)	-0.65 (0.49)

Abb. 1 Durchschnittliche Abweichungen der Urteile des Videos mit mittlerem Qualitätsniveau der experimentellen Bedingungen A (niedrig/mittel) und C (hoch/mittel) und signifikanten Pfaden unterteilt nach den Unterrichtsqualitätsmerkmalen. *Anmerkung*. Zusätzlich zu den hypothesenprüfenden Analysen ließen sich Unterschiede in den Urteilen zwischen den Gruppen A (niedrig/mittel) und C (hoch/mittel) bezüglich der kognitiven Aktivierung ($t(44) = 3{,}50$, $p = 0{,}009$, Cohens $d = 1{,}03$) und der Klassenführung ($t(44) = 2{,}99$, $p = 0{,}035$, Cohens $d = 0{,}88$) finden; jedoch nicht im Bereich Motivierung ($t(44) = 2{,}65$, $p = 0{,}055$, Cohens $d = 0{,}78$)

abweichend ein als die Referenzgruppe. Die vorangegangenen Analysen verweisen darauf, dass die Unterschiede in den Bereichen kognitive Aktivierung und Motivierung auf einen Reihenfolgeeffekt der kritischeren Einschätzung im Sinne einer Abwertung zurückzuführen sind. Die Hypothese 1.2 zur Abwertung kann allerdings nur teilweise empirisch unterstützt werden, da sich nicht für alle Unterrichtsqualitätsmerkmale signifikante Unterschiede finden ließen.

Insgesamt zeigten sich somit eher Abwertungs- als Aufwertungsprozesse durch einen Reihenfolgeeffekt mit mittlerer Ausgangsqualität.

5.2 Prädiktoren für die Urteile und das Auftreten des Reihenfolgeeffekts

Die Ergebnisse zur Beantwortung der Fragestellung 2, ob die individuellen Faktoren Erfahrungszeit oder Müdigkeit prädiktiv für die Urteile der Unterrichtsqua-

Tab. 4 Prädiktoren (Experimentelle Bedingung A und C, Erfahrungszeit und Interaktionsterme Bedingung A oder C × Erfahrungszeit) für die Urteile der Unterrichtsqualitätsmerkmale

	Kognitive Aktivierung						Klassenführung						Motivierung					
	I		II		III		I		II		III		I		II		III	
Prädiktoren	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*
Experimentelle Bedingung A (EB A)	0,08	0,13	0,07	0,13	0,07	0,13	0,16	0,13	0,17	0,13	0,18	0,13	–0,12	0,12	–0,12	0,13	–0,11	0,12
Experimentelle Bedingung C (EB C)	–0,38**	0,12	–0,38**	0,12	–0,38**	0,12	–0,20	0,13	–0,21	0,13	–0,20	0,13	–0,41**	0,12	–0,43**	0,12	–0,41**	0,12
Erfahrungszeit (EZ)	0,01	0,11	–	–	–0,01	0,26	–0,12	0,11	–	–	–0,30	0,23	–0,22*	0,11	–	–	–0,51*	0,21
EB A × EZ	–	–	–0,05	0,11	–0,05	0,14	–	–	0,02	0,12	0,13	0,14	–	–	–0,05	0,11	0,14	0,13
EB C × EZ	–	–	0,05	0,11	0,06	0,21	–	–	–0,07	0,12	0,16	0,21	–	–	–0,10	0,11	0,30	0,20
R^2	0,17*	0,08	0,18*	0,08	0,18*	0,08	0,13	0,08	0,12	0,07	0,14	0,08	0,21*	0,09	0,17*	0,08	0,23**	0,09

Anmerkungen. Saturierte Modelle
*$p < 0{,}05$, **$p < 0{,}01$

litätsmerkmale kognitive Aktivierung, Klassenführung und Motivierung bei dem Unterrichtsvideo mittlerer Qualität als auch für Unterschiede im Auftreten des Reihenfolgeeffekts sind, sind in den Tab. 4 und 5 dargestellt.

Vorerfahrung (Erfahrungszeit in der Schule) Es zeigte sich in allen Modellen, dass die Erfahrungszeit in der Schule nicht prädiktiv für die Urteile der kognitiven Aktivierung und Klassenführung bei dem Unterrichtsvideo mittlerer Qualität war. Hinsichtlich der Urteile der Motivierung bei dem Unterrichtsvideo mittlerer Qualität wies die Erfahrungszeit eine prädiktive Kraft auf; auch unter Berücksichtigung der Interaktionsterme. Raterinnen und Rater, die eine größere Erfahrungszeit in der Schule angaben, schätzten die Qualität der Motivierung des Referenzvideos unabhängig von der Zugehörigkeit der Bedingung niedriger ein als unerfahrene Raterinnen und Rater. Da die Prädiktion nur im Bereich Motivierung zu finden war, konnte die Hypothese 2.1 nur teilweise empirisch unterstützt werden. Die nach der Hypothese 2.2 erwarteten Prädiktionen der Interaktionen zwischen den Gruppen und der Erfahrungszeit auf die Beurteilungen der kognitiven Aktivierung, Klassenführung und Motivierung konnten nicht festgestellt werden. Somit konnte keine Verstärkung oder Verminderung des Reihenfolgeeffekts aufgrund unterschiedlicher Erfahrungszeit in der Schule nachgewiesen werden und die Hypothese 2.2 konnte demnach nicht empirisch gestützt werden.

Mentaler Zustand (Müdigkeit) In Bezug auf die Urteile der kognitiven Aktivierung und Klassenführung bei dem Unterrichtsvideo mittlerer Qualität wies die Müdigkeit in allen Modellen keinen gerichteten Zusammenhang auf. Bezüglich der Motivierung war die Müdigkeit prädiktiv für die Urteile bei dem Unterrichtsvideo mittlerer Qualität. Raterinnen und Rater, die ihre Müdigkeit hoch beurteilten, schätzten unabhängig von der Zugehörigkeit der Bedingung die Qualität der Motivierung des Videos mit mittlerer Qualität höher ein als wachere Raterinnen und Rater. Dieses Ergebnis zeigte sich allerdings nicht unter Berücksichtigung der Interaktionsterme in Modell III. Da sich die Prädiktion nur für den Bereich Motivierung und auch nur in dem Modell I zeigte, konnte die Hypothese 2.3 nur teilweise empirisch unterstützt werden. In Hinblick auf die Urteile in den Bereichen kognitive Aktivierung und Klassenführung ließen sich keine Interaktionseffekte zwischen der experimentellen Bedingung und der Müdigkeit nachweisen, jedoch bezüglich der Urteile im Bereich Motivierung. Raterinnen und Rater, die ihre Müdigkeit hoch beurteilten und der Bedingung A (niedrig/mittel) zugewiesen wurden, schätzten die Qualität der Motivierung des Videos mittlerer Qualität höher ein als wachere Raterinnen und Rater der Bedingung A. Dies bedeutet, dass der Reihenfolgeeffekt durch eine hohe Müdigkeit, also eher müderen Raterinnen und Rater, verstärkt wurde. Dies war allerdings nur für ein positiveres Urteil nach einem Video mit geringerer Unterrichtsqualität zu beobachten und auch nur, wenn die Müdigkeit als Variable nicht einbezogen wurde. Somit konnte die Hypothese 2.4 teilweise empirisch unterstützt werden.

Tab. 5 Prädiktoren (Experimentelle Bedingung A und C, Müdigkeit und Interaktionstherme Bedingung A oder C × Müdigkeit) für die Urteile der Unterrichtsqualitätsmerkmale

	Kognitive Aktivierung						Klassenführung						Motivierung					
	I		II		III		I		II		III		I		II		III	
Prädiktoren	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*
Experimentelle Bedingung A (EB A)	0,08	0,13	0,07	0,13	0,07	0,13	0,16	0,13	0,16	0,13	0,16	0,13	–0,12	0,12	–0,11	0,13	–0,11	0,12
Experimentelle Bedingung C (EB C)	–0,38**	0,12	–0,38**	0,12	–0,38**	0,12	–0,22	0,13	–0,22	0,13	–0,22	0,13	–0,45**	0,12	–0,45**	0,12	–0,45**	0,12
Müdigkeit (M)	0,02	0,11	–	–	–0,06	0,24	–0,02	0,11	–	–	0,00	0,25	0,21*	0,11	–	–	0,08	0,23
EB A × M	–	–	–0,06	0,11	–0,02	0,18	–	–	–0,10	0,11	–0,10	0,18	–	–	0,22*	0,11	0,17	0,17
EB C × M	–	–	0,09	0,11	0,13	0,19	–	–	0,06	0,11	0,06	0,20	–	–	0,10	0,11	0,05	0,19
R^2	0,17*	0,08	0,19*	0,08	0,19*	0,09	0,11	0,07	0,12	0,07	0,12	0,07	0,20*	0,09	0,21*	0,09	0,21**	0,09

Anmerkungen: Saturierte Modelle
*$p < 0{,}05$, **$p < 0{,}01$

6 Diskussion

Die vorliegende Untersuchung ging der Frage nach, ob Urteile von Unterrichtsvideos von der dargebotenen Reihenfolge (F1) sowie von individuellen Merkmalen abhängig sind und welche individuellen Merkmale den Reihenfolgeeffekt verringern oder begünstigen (F2). Die Resultate bezüglich aller drei Unterrichtsqualitätsmerkmale zeigten, dass zwischen den Urteilen der Gruppe A (niedrig/mittel) und B (mittel/hoch) keine bedeutsamen Unterschiede vorlagen. Die Qualität eines Videos mittleren Qualitätsniveaus wurde nicht höher nach einem vorangestellten Urteil eines Videos niedrigen Qualitätsniveaus eingeschätzt, als wenn vorher kein anderes Video beurteilt wurde (H1.1). Die Analysen wiesen keine statistisch bedeutsamen Effekte der experimentellen Bedingung auf, allerdings ließen sich Effekte zwischen den Urteilen der Gruppe A und B von *Cohens d* = 0,17–0,37 finden.

Für die Urteile der kognitiven Aktivierung und Motivierung zeigten sich signifikante Unterschiede zwischen den Gruppen B (mittel/hoch) und C (hoch/mittel). Die kognitive Aktivierung und Motivierung wurden durchschnittlich negativer nach einer vorigen Beurteilung eines Videos hohen Qualitätsniveaus beurteilt, als wenn vorher kein anderes Video bewertet wurde (H1.2). Somit deuteten die Resultate bezüglich der kognitiven Aktivierung und Motivierung eher auf Abwertungs- als auf Aufwertungsprozesse durch einen Reihenfolgeeffekt mit mittlerer Ausgangsqualität hin.

Trotz einer hohen Interraterreliabilität im Bereich der kognitiven Aktivierung, welche für ein gelungenes Training stehen kann, weisen die Ergebnisse auf die Bedeutung der Darbietungsreihenfolge der Videos für das Urteil der kognitiven Aktivierung hin. Somit können Urteile trotz hoher Übereinstimmungen Beurteilungsfehlern unterliegen.

Für die Klassenführung ließen sich keine erwarteten Unterschiede (Gruppe A vs. B und Gruppe B vs. C) aufgrund der sequenziellen Darbietung nachweisen. Die Perspektive der Externen gilt als direktester Weg zur Erfassung der Klassenführung (Clare et al. 2001). Die Klassenführung beinhaltet im Vergleich zu den anderen Merkmalen eindeutig beobachtbare Facetten und setzt somit weniger Interpretationsaufwand voraus (Clausen 2002). Dies kann dazu führen, dass in der vorliegenden Studie keine signifikanten Unterschiede festgestellt wurden und der Grad des Interpretationsaufwandes und das Auftreten eines Reihenfolgeeffekts zusammenhängen.

In Bezug auf Urteile der Motivierung sind die negativen Abweichungen der Urteile im Bereich Motivierung der Gruppe A auffällig. Erwartet wurden positive Abweichungen zu den Urteilen der Referenzgruppe, die sich auch für die kognitive Aktivierung und Klassenführung gezeigt haben. Eine mögliche Erklärung dafür könnte sein, dass die Motivierung als eher schwierig zu beobachtendes Konstrukt gilt und für Urteile das komplexe Zusammenwirken struktureller und inhaltlicher Unterrichtsqualitätsmerkmale sowie auch soziale und persönliche Komponenten der Lehrkräfte und Lernenden berücksichtigt werden muss (Clausen et al. 2003; Rakoczy 2008). Die Komplexität für eine Bewertung dieses Merkmals kann auch zu einer Notwendigkeit von subjektiveren Interpretationen bei hoch inferenten Ratings führen, welche wiederum häufiger von Beurteilungsfehlern betroffen sind als sichtbarere Unterrichtsqualitätsmerkmale (Praetorius 2014). Neben der Komplexität ist

auch der kurze Einblick der Externen in das Unterrichtsgeschehen ein weiterer Faktor, welcher die Urteile verzerren kann. Dieser Stichprobeneffekt, eingeschränkte kurze Beobachtungsstichprobe, wird von der Beobachtbarkeit des Unterrichtsqualitätsmerkmals moderiert und so können Unterrichtsqualitätsmerkmale mit niedriger Beobachtbarkeit den Effekt verstärken (Clausen 2002). Dementsprechend konnten sich bei der Einschätzung der Motivierung eventuell zwei Faktoren gegenseitig bedingen: Komplexität des Unterrichtsqualitätsmerkmals und kurze Beobachtungsstichprobe.

Bezüglich der Fragestellung 2 zeigte sich ein Zusammenhang zwischen der Erfahrungszeit in der Schule und der Urteile des Referenzvideos. Die Studierenden mit einer höheren Erfahrungszeit in der Schule beurteilten die Motivierung des Referenzvideos negativer als Studierende mit weniger Erfahrungszeit. Eine Erklärung können implizite Theorien der erfahrenen Studierenden sein, die auf Grundlage von Erfahrung in der Schule, zum Beispiel im Rahmen von Praktika, entwickelt wurden und auch nicht durch ein Training geändert werden können (Praetorius et al. 2012). Ferner legen erfahrene Raterinnen und Rater häufig ein breiteres Spektrum für die Beurteilung an als Unerfahrene und greifen auch auf Faktoren zurück, die nicht für das Rating vorgegeben waren (Leckie und Baird 2011). Darüber hinaus zeigte sich ein gerichteter Zusammenhang zwischen der Müdigkeit und den Urteilen über die Motivierung bei dem Unterrichtsvideo mittlerer Qualität. Studierende, die ihre Müdigkeit hoch beurteilten, schätzten die Qualität der Motivierung des Videos mit mittlerer Qualität höher ein als wachere Raterinnen und Rater. Dieses Ergebnis ist gegensätzlich zu den Befunden von Mashburn et al. (2014), dass die Probandinnen und Probanden mit zunehmender Müdigkeit strengere Urteile vergaben. Die Wahrnehmung und das Urteilsvermögen werden unter anderem durch kognitive Merkmale wie die Müdigkeit bestimmt (z. B. Feltz und Cokely 2011). So kann es zum Beispiel aufgrund von Müdigkeit zu einer selektiven Informationsaufnahme kommen und demzufolge werden Urteile auf Basis mangelnder Informationen getroffen (Schmidt-Atzert et al. 2004). Dies kann ein Grund sein, warum sich Studierende mit divergierender Ausprägung der Müdigkeit im Urteil über die Motivierungsqualität unterschieden. Die Affektheuristik, dass Urteile auf Basis des eigenen Befindens getroffen werden (Fiske und Taylor 2017), kann an dieser Stelle keine Erklärung für die vorliegenden Resultate sein. Demnach hätten die Studierenden, die aufgrund von Müdigkeit ein schlechteres Befinden hatten, negativer die Motivierungsqualität beurteilt als wachere Studierende mit einem positiveren Befinden.

Ein Interaktionseffekt wurde zwischen der Müdigkeit und der experimentellen Bedingung A (niedrig/mittel) gefunden. Im Einklang mit bisherigen Befunden kann anhand der Ergebnisse vermutet werden, dass die Müdigkeit den Reihenfolgeeffekt verstärkt. Raterinnen und Rater, die ihren Zustand als eher müde einschätzten und zuvor ein Video niedrigen Qualitätsniveaus bewerteten, orientierten sich stärker an diesem und überschätzten die Motivierung des Videos mittlerer Qualität stärker als Personen, die ihre Müdigkeit geringer einschätzten. Auch für die verstärkte Verzerrung bei hoher Müdigkeit kann eine Begründung der Zusammenhang einer zunehmenden Müdigkeit mit sinkender Wahrnehmung der relevanten Lehrkrafthandlungen sein (Schmidt-Atzert und Amelang 2012). So haben eventuell müdere Raterinnen und Rater für die Beurteilung der Qualität bedeutsame Kriterien

aufgrund von niedriger Aufmerksamkeit nicht wahrgenommen. Dass sich die Resultate der Prädiktion der Müdigkeit sowie der Interaktionseffekt nur zeigten, wenn jeweils die andere Variable nicht berücksichtigt wurde, konnte durch die Korrelation von $r = 0{,}58$ ($p = 0{,}000$) zwischen der der Müdigkeit und des Interaktionsterms der Bedingung A × Erfahrungszeit verursacht werden. Alle signifikanten gerichteten Zusammenhänge betrafen den Bereich Motivierung. Diese Resultate bestätigen, dass das Unterrichtsqualitätsmerkmal Motivierung schwierig für Externe einzuschätzen ist und externe Urteile über die Motivierung häufiger durch individuelle Merkmale beeinflusst werden.

6.1 Einschränkungen

Bei der Interpretation der Ergebnisse bleibt zu beachten, dass das Training der Raterinnen und Rater mit drei Stunden vergleichsweise zu anderen Studien kurz war (z. B. Clausen et al. 2003) und besonders die gefundenen Zusammenhänge der Vorerfahrung eventuell geringer ausfallen würden, wenn zum Beispiel eine größere Spannbreite an Videos mit niedriger, mittlerer und hoher Qualität vorab gesichtet worden wäre. Hinzu sollte bei der Einordnung der Resultate berücksichtigt werden, dass in der vorliegenden Studie nur ein Item pro Merkmal eingeschätzt wurde. Dies könnte trotz ausführlicher Beschreibung der Unterrichtsqualitätsmerkmale im Manual und Beispielfacetten zu breiten Interpretationen bei den Raterinnen und Ratern führen.

Für die Vergleiche der Urteile der Qualität des Videos mittleren Qualitätsniveaus wurde als Referenz das Urteil der Gruppe B (mittel/hoch) herangezogen, in der die Studierenden das Referenzvideo zuerst sahen und daher nicht durch ein anderes Video in ihrer Beurteilung beeinflusst wurden. Dieses Vorgehen hat einerseits den Vorteil der Möglichkeit einer Abschätzung der absoluten Größe einer möglichen Urteilsverzerrung aufgrund der wahrscheinlich eher unbeeinflussten Bewertung in Gruppe B. Andererseits können im Vergleich zu den Gruppe A (niedrig/mittel) und C (hoch/mittel), die das Referenzvideo als zweites sahen, zwischen dem zuerst einzuschätzenden Video und dem zweiten Video kognitive Prozesse auftreten (zum Beispiel kognitive Ermüdung), die wiederum zu Beurteilungsfehlern führen können.

Aufgrund des vorigen Urteils eines Videos mit einer entweder negativeren oder positiveren Unterrichtsqualität könnten die Ergebnisse nicht nur auf einen Reihenfolgeeffekt, sondern auch auf einen Kontrasteffekt hinweisen. Hierbei werden Informationen aufgrund der im Kontrast stehenden Vergleichsinformation intensiver wahrgenommen und Urteile in die entgegensetzte Richtung verzerrt (Lenske 2016). Des Weiteren könnte auch der Ankereffekt zutragen gekommen sein, wobei vorige Informationen als Anker für zukünftige Urteile dienen (Tversky und Kahneman 1974). Es ist aber zu beachten, dass die Beurteilungsfehler nicht eindeutig trennbar sind und in einer engen Beziehung stehen (Hermann 2016). Zusätzlich führen sie alle zu dem gleichen unerwünschten Effekt – verzerrte Beurteilungen der Unterrichtsqualität.

6.2 Implikationen für Praxis und Forschung

Die vorliegende Untersuchung kann aufgrund des experimentellen Designs wichtige Hinweise für die Messung der Unterrichtsqualität anhand von Videos liefern. Für weitere Studien zur Unterrichtsqualität, die auf Videoratings zurückgreifen, sollte die Reihenfolge der gezeigten Videos mit unterschiedlichen Qualitätsniveaus beachtet werden. Hierbei kann eine Randomisierung der Videos ein mögliches Auftreten eines Reihenfolgeeffekts verhindern. Mögliche Unterschiede in den Urteilen aufgrund der Vorerfahrung könnten eventuell mit einer umfänglicheren Schulung als in der vorliegenden Studie ausgeglichen werden.

Da der mentale Zustand, wie die Müdigkeit, Beurteilungsfehler verstärken und eine höhere Müdigkeit größere Verzerrungen in den Urteilen hervorrufen kann, ist es bedeutsam dies in zukünftigen Studien zu kontrollieren, zum Beispiel durch eine vorige Erfassung des mentalen Zustandes. Um Fehlattributionen aufgrund des aktuellen mentalen Zustandes zu vermeiden, ist es ebenfalls wichtig, Raterinnen und Rater zu sensibilisieren bei ihren Urteilen nicht ihren aktuellen mentalen Zustand zu berücksichtigen (Stroebe 2014).

Bislang gibt es nur wenige Studien, die mögliche Einschränkungen von Urteilen Externer in der Unterrichtsforschung untersuchten. Folge sind die nicht valide Erfassung von Unterrichtsqualitätsmerkmalen und möglicherweise Fehlinterpretationen von Zusammenhängen zwischen Unterricht und beispielsweise Lernleistung. Insbesondere in dem Bereich der Motivierung sollte neben der Perspektive der Externen auch die Sicht der Lehrkräfte und Lernenden herangezogen werden. Die Komplexität dieses Unterrichtsqualitätsmerkmals, welches sich unter anderem durch die Berücksichtigung der Beziehungen der beteiligten Personen auszeichnet, ist nur eingeschränkt durch die Perspektive von Externen zu erfassen.

Schließlich weist die vorliegende Studie auf mögliche Einschränkungen von Urteilen externer Raterinnen und Rater hin, indem vorige Unterrichtsszenen als mögliche Referenz für das aktuelle Urteil verwendet werden und dies insbesondere zu einer Unterschätzung der Unterrichtsqualität führen kann.

Danksagung Ein besonderer Dank gilt Dr. Matthias Trendtel für die Unterstützung im Bereich der statistischen Methodenwahl.

Förderung Das Institut für Schulentwicklungsforschung förderte und führte die Studie durch.

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt J. Igler, A. Ohle-Peters und N. McElvany geben an, dass kein Interessenkonflikt besteht.

Literatur

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 Erziehungswissenschaften*, *20*(1), 23–47.

Bless, H., & Keller, J. (2006). Urteilsheuristiken. In H.-W. Bierhoff & D. Frey (Hrsg.), *Handbuch der Psychologie: Sozialpsychologie und Kommunikationspsychologie* (S. 294–300). Göttingen: Hogrefe.

Bless, H., Bohner, G., Schwarz, N., & Strack, F. (1990). Mood and persuasion: a cognitive response analysis. *Personality and Social Psychology Bulletin*, *16*, 331–345.

Bless, H., Fiedler, K., & Strack, F. (2004). *Social cognition: how individuals construct social reality*. Philadelphia: Psychology Press.

Brophy, J. (1979). Teacher behavior and its effects. *Journal of Educational Psychology*, *71*, 733–750.

Clare, L., Valdés, R., Pascal, J., & Steinberg, J. R. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools* (CSE Tech. Rep. No. 545). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

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

Cushman, F., & Mele, A. (2008). Intentional action: two-and-a-half folk concepts? In J. Knobe & S. Nichols (Hrsg.), *Experimental philosophy* (S. 171–188). Oxford: Oxford University Press.

Dalbert, C. (1992). Subjektives Wohlbefinden junger Erwachsener: Theoretische und empirische Analysen der Struktur und Stabilität. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *13*(4), 207–220.

Decristan, J., Hess, M., Holzberger, D., & Praetorius, A.-K. (2020). Oberflächen- und Tiefenmerkmale. Eine Reflexion zweier prominenter Begriffe der Unterrichtsforschung. *Zeitschrift für Pädagogik*, *66*, 102–116.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist*, *36*, 103–112.

Engle-Friedman, M., Mathew, G. M., Martinova, A., Armstrong, F., & Konstantinov, V. (2018). The role of sleep deprivation and fatigue in the perception of task difficulty and use of heuristics. *Sleep Science*, *11*(2), 74–84.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: why the adjustments are insufficient. *Psychological Science*, *17*(4), 311–318.

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

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

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.

Feltz, A., & Cokely, E. T. (2011). Individual differences in theory-of-mind judgments: order effects and side effects. *Philosophical Psychology*, *24*(3), 343–355.

Fiske, S. T., & Taylor, S. E. (2017). *Social cognition: from brains to culture* (3. Aufl.). London: SAGE.

Gabriel-Busse, K., Groß-Mlynek, L., Feldhoff, T., & Harring, M. (2020). Eine Unterrichtssequenz – unterschiedliche Einschätzungen. Analyse videografierter Unterrichtssequenzen als Bestandteil einer evidenzbasierten Lehrer/innenausbildung. In I. Gogolin, B. Hannover & A. Scheunpflug (Hrsg.), *Evidenzbasierung in der Lehrkräftebildung* (ZfE-Edition, Bd. 4, S. 291–314). Wiesbaden: Springer VS.

Geiser, C. (2011). *Datenanalyse mit Mplus: Eine anwendungsorientierte Einführung*. Wiesbaden: VS.
Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization, measurement, & improvement. In J. L. Meece & J. S. Eccles (Hrsg.), *Handbook of research on schools, schooling and human development* (S. 25–41). New York: Routledge.
Hermann, H.-P. (2016). *Tourismuspsychologie*. Heidelberg: Springer.
Herrle, M., Rauin, U., & Engartner, T. (2016). Videos als Ressourcen zur Generierung von Wissen über Unterrichtsrealität(en). In U. Rauin & M. Herrle (Hrsg.), *Videoanalysen in der Unterrichtsforschung – Methodische Vorgehensweisen und Anwendungsbeispiele* (S. 8–28). Weinheim: Beltz Juventa.
Ho, A. D., & Kane, T. J. (2013). *The reliability of teaching observations by school personnel*. Seattle: Bill and Melinda Gates Foundation.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.
Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: a metaanalysis. *Psychological Methods*, *4*, 403–424.
Kooken, J., Welsh, M. E., McCoach, D. B., Miller, F. G., Chafouleas, S. M., Riley-Tillmann, T. C., & Fabio, G. (2017). Test order in teacher-rated behavior assessments: is counterbalancing necessary? *Psychological Assessment*, *29*(1), 98–109.
Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing". *Psychological Review*, *103*, 263–283.
Kunter, M., & Trautwein, U. (2013). *Psychologie des Unterrichts*. Paderborn: Ferdinand Schöningh.
Lau, A., & Plessner, H. (2016). *Sozialpsychologie und Sport – Ein Lehrbuch in 12 Lektionen*. Aachen: Meyer & Meyer.
Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: a multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399–418.
Lenske, G. (2016). *Schülerfeedback in der Grundschule. Untersuchungen zur Validität*. Münster: Waxmann.
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 & E. Terhart (Hrsg.), Kompetenz und Kompetenzentwicklung von Lehrerinnen und Lehrern. *Zeitschrift für Pädagogik*, *51*, 47–70.
Lipowsky, F., Rakoczy, K., Drollinger-Vetter, B., Klieme, E., Reusser, K., & Pauli, C. (2009). Quality of geometry instruction and its short-term impact on students? Understanding of Pythagorean Theorem. *Learning and Instruction*, *19*(6), 527–537.
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.
Martin, E., & Wawrinowski, U. (2014). *Beobachtungslehre: Theorie und Praxis reflektierter Beobachtung und Beurteilung* (6. Aufl.). Weinheim: Beltz.
Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pinata, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement*, *74*(3), 400–422.
McElvany, N., Schroeder, S., Richter, T., Hachfeld, A., Baumert, J., Schnotz, W., et al. (2012). Cognitively demanding learning materials with texts and instructional pictures: teachers' diagnostic skills, pedagogical beliefs, and motivation. *European Journal of Psychology of Education*, *27*, 403–420.
Messner, C., & Schmid, B. (2007). Über die Schwierigkeit unparteiische Entscheidungen zu fällen: Schiedsrichter bevorzugen Fußballteams ihrer Kultur. *Zeitschrift für Sozialpsychologie*, *38*(2), 105–110.
Meyer, H. (2007). Zehn Merkmale guten Unterrichts. In W. Endres (Hrsg.), *Lernen lernen – wie stricken ohne Wolle? 13 Experten streiten über Konzepte und Modelle zur Lernmethodik* (S. 167–187). Weinheim: Beltz.
Miller, N., & Campbell, D. T. (1959). Recency and primacy in persuasion as a function of the timing of speeches and measurements. *Journal of Abnormal and Social Psychology*, *59*, 1–9.
Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8. Aufl.). Los Angeles: Muthén & Muthén.
Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Reach measurement: Part I. *Journal of Applied Measurement*, *4*, 386–422.
Ohle, A., & McElvany, N. (2016). Erfassung von Unterrichtsqualität in der Grundschule: Kognitiver Anspruch, Strukturierung und Motivierungsqualität. In N. McElvany, W. Bos, H. G. Holtappels, M. Gebauer & F. Schwabe (Hrsg.), *Bedingungen und Effekte guten Unterrichts. Dortmunder Symposium der Empirischen Bildungsforschung* (S. 117–134). Münster: Waxmann.

Pauli, C., & Reusser, K. (2003). Unterrrichtsskripts im schweizerischen und im deutschen Mathematikunterricht. *Unterrichtswissenschaft*, *31*, 238–272.
Pietsch, M., & Tosana, S. (2008). Beurteilereffekte bei der Messung von Unterrichtsqualität. *Zeitschrift für Erziehungswissenschaft*, *11*(3), 430–452.
Praetorius, A.-K. (2013). Einschätzung von Unterrichtsqualität durch externe Beobachterinnen und Beobachter. Eine kritische Betrachtung der aktuellen Vorgehensweise in der Schulpraxis. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *31*(2), 174–185.
Praetorius, A.-K. (2014). *Messung von Unterrichtsqualität durch Ratings*. Münster: Waxmann.
Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: do they fulfill what they promise? *Learning and Instruction*, *22*, 387–400.
Praetorius, A.-K., 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*, 303–318.
Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht: Unterricht aus der Perspektive von Lernenden und Beobachtern*. Münster: Waxmann.
Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: definitions, theory, practices, and future directions. *Contemporary Educational Psychology*. https://doi.org/10.1016/j.cedpsych.2020.101860.
Schmidt-Atzert, L., & Amelang, M. (2012). *Psychologische Diagnostik* (5. Aufl.). Heidelberg: Springer.
Schmidt-Atzert, L., Büttner, G., & Bühner, M. (2004). Theoretische Aspekte von Aufmerksamkeits-/Konzentrationsdiagnostik. In G. Büttner & L. Schmidt-Atzert (Hrsg.), *Diagnostik von Konzentration und Aufmerksamkeit* (S. 3–22). Göttingen: Hogrefe.
Schwindt, K. (2008). *Lehrpersonen betrachten Unterricht: Kriterien für die kompetente Unterrichtswahrnehmung*. Münster: Waxmann.
Seidel, T., & Thiel, F. (2017). Standards und Trends der videobasierten Lehr-Lernforschung. *Zeitschrift für Erziehungswissenschaften*, *32*, 1–21.
Stroebe, W. (2014). Strategien zur Einstellungs- und Verhaltensänderung. In K. Jonas, W. Stroebe & M. Hewstone (Hrsg.), *Sozialpsychologie* (6. Aufl., S. 231–268). Berlin: Springer.
Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: the application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*, 1124–1131.
Webster, D. M., Richter, L., & Kruglanski, A. W. (1996). On leaping to conclusions when feeling tired: mental fatigue effects on impressional primacy. *Journal of Experimental Social Psychology*, *32*, 181–195.
Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.
Wolff, C. E., Jarodzka, H., & Boshuizen, H. P. A. (2017). See and tell: Differences between expert and novice teachers' interpretations of problematic classroom management events. *Teaching and Teacher Education*, *66*, 295–308.