Praetorius, Anna-Katharina [Hrsg.]; Grünkorn, Juliane [Hrsg.]; Klieme, Eckhard [Hrsg.]

Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen

1. Auflage

Weinheim; Basel : Beltz Juventa 2020, 268 S. - (Zeitschrift für Pädagogik, Beiheft; 66)

66. Beiheft April 2020

ZEITSCHRIFT FÜR PÄDAGOGIK

Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen

BELTZ JUVENTA

Zeitschrift für Pädagogik · 66. Beiheft

Zeitschrift für Pädagogik · 66. Beiheft

Empirische Forschung zu Unterrichtsqualität

Theoretische Grundfragen und quantitative Modellierungen

Herausgegeben von
Anna-Katharina Praetorius, Juliane Grünkorn
und Eckhard Klieme

BELTZ JUVENTA

ISSN: 0514-2717
ISBN 978-3-7799-3534-6 Print
ISBN 978-3-7799-3535-3 E-Book (PDF)
Bestellnummer: 443534

1. Auflage 2020

in der Verlagsgruppe Beltz · Weinheim Basel
Werderstraße 10, 69469 Weinheim

Herstellung: Hannelore Molitor
Satz: text plus form, Dresden
Druck und Bindung: Beltz Grafische Betriebe, Bad Langensalza
Printed in Germany

Weitere Informationen zu unseren Autoren und Titeln finden Sie unter: www.beltz.de

Inhaltsverzeichnis

Anna-Katharina Praetorius/Juliane Grünkorn/Eckhard Klieme
Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen. Einleitung in das Beiheft 9

Themenblock I: Dimensionen der Unterrichtsqualität – Theoretische und empirische Grundlagen (englischsprachig)

Anna-Katharina Praetorius/Eckhard Klieme/Thilo Kleickmann/Esther Brunner/Anke Lindmeier/Sandy Taut/Charalambos Charalambous
Towards Developing a Theory of Generic Teaching Quality: Origin, Current Status, and Necessary Next Steps Regarding the Three Basic Dimensions Model 15

Thilo Kleickmann/Mirjam Steffensky/Anna-Katharina Praetorius
Quality of Teaching in Science Education: More Than Three Basic Dimensions? 37

Courtney A. Bell
Commentary Regarding the Section "Dimensions of Teaching Quality – Theoretical and Empirical Foundations" – Using Warrants and Alternative Explanations to Clarify Next Steps for the TBD Model 56

Themenblock II: Angebots-Nutzungs-Modelle als Rahmung (deutschsprachig)

Svenja Vieluf/Anna-Katharina Praetorius/Katrin Rakoczy/Marc Kleinknecht/Marcus Pietsch
Angebots-Nutzungs-Modelle der Wirkweise des Unterrichts: ein kritischer Vergleich verschiedener Modellvarianten 63

Sibylle Meissner/Samuel Merk/Benjamin Fauth/Marc Kleinknecht/Thorsten Bohl
Differenzielle Effekte der Unterrichtsqualität auf die aktive Lernzeit 81

Tina Seidel
Kommentar zum Themenblock „Angebots-Nutzungs-Modelle als Rahmung“ – Quo vadis deutsche Unterrichtsforschung? Modellierung von Angebot und Nutzung im Unterricht 95

Themenblock III: Oberflächen- und Tiefenstruktur des Unterrichts (deutschsprachig)

Jasmin Decristan/Miriam Hess/Doris Holzberger/Anna-Katharina Praetorius
Oberflächen- und Tiefenmerkmale – eine Reflexion zweier prominenter Begriffe der Unterrichtsforschung 102

Miriam Hess/Frank Lipowsky
Zur (Un-)Abhängigkeit von Oberflächen- und Tiefenmerkmalen im Grundschulunterricht – Fragen von Lehrpersonen im öffentlichen Unterricht und in Schülerarbeitsphasen im Vergleich 117

Christine Pauli
Kommentar zum Themenblock „Oberflächen- und Tiefenstruktur des Unterrichts“: Nutzen und Grenzen eines prominenten Begriffspaars für die Unterrichtsforschung – und das Unterrichten 132

Themenblock IV: Zur Bedeutung unterschiedlicher Perspektiven bei der Erfassung von Unterrichtsqualität (englischsprachig)

Benjamin Fauth/Richard Göllner/Gerlinde Lenske/Anna-Katharina Praetorius/Wolfgang Wagner
Who Sees What? Conceptual Considerations on the Measurement of Teaching Quality from Different Perspectives 138

Richard Göllner/Benjamin Fauth/Gerlinde Lenske/Anna-Katharina Praetorius/Wolfgang Wagner
Do Student Ratings of Classroom Management Tell us More About Teachers or About Classroom Composition? 156

Marten Clausen
Commentary Regarding the Section “The Role of Different Perspectives on the Measurement of Teaching Quality” 173

Themenblock V: Modellierung der Wirkungen von Unterrichtsqualität (englischsprachig)

Alexander Naumann/Susanne Kuger/Carmen Köhler/Jan Hochweber
Conceptual and Methodological Challenges in Detecting the Effectiveness of Learning and Teaching 179

Carmen Köhler/Susanne Kuger/Alexander Naumann/Johannes Hartig
Multilevel Models for Evaluating the Effectiveness of Teaching: Conceptual and Methodological Considerations 197

Oliver Lüdtke/Alexander Robitzsch
Commentary Regarding the Section "Modelling the Effectiveness of Teaching Quality" – Methodological Challenges in Assessing the Causal Effects of Teaching 210

Kommentare

Ewald Terhart
Unterrichtsqualität zwischen Theorie und Empirie – Ein Kommentar zur Theoriediskussion in der empirisch-quantitativen Unterrichtsforschung 223

Kurt Reusser
Unterrichtsqualität zwischen empirisch-analytischer Forschung und pädagogisch-didaktischer Theorie – Ein Kommentar 236

Anke Lindmeier/Aiso Heinze
Die fachdidaktische Perspektive in der Unterrichtsqualitätsforschung: (bisher) ignoriert, implizit enthalten oder nicht relevant? 255

Anna-Katharina Praetorius/Juliane Grünkorn/Eckhard Klieme

Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen

Einleitung in das Beiheft

Schulischer Unterricht ist in modernen Gesellschaften eine ubiquitäre Praxis, mit der jede und jeder durch die eigene Schulzeit sehr vertraut ist. Im wissenschaftlichen Kontext existieren vielfältige Vorschläge, diesen Begriff zu definieren. So spricht beispielsweise Reusser (2009) von Unterricht als „einer durch Standards geregelten, planvollen Praxis der intergenerationellen Fähigkeitsübertragung nach definierten Prinzipien und Methoden der Unterweisung innerhalb von festgelegten Organisationsformen und gesellschaftlichen Einrichtungen. Damit verbunden waren seit jeher Ergebniserwartungen […] wie die Vermittlung von Wissen und Fertigkeiten, die erzieherische Formung von Persönlichkeit und des Verhaltens“ (Reusser, 2009, S. 881). Die vielfältigen diesbezüglichen Forschungs- und Theoriebildungsbemühungen werden im Kontext unterschiedlicher disziplinärer und subdisziplinärer Traditionen (Erziehungswissenschaft: Bildungstheorie, Allgemeine Didaktik, Schulpädagogik, empirisch-pädagogische Forschung; Pädagogische Psychologie; Soziologie: Theorie des Erziehungssystems, Professionsforschung u. a. m.) mit ihren spezifischen Begrifflichkeiten, Kategorien und Methoden durchgeführt (Meseth, Proske & Radtke, 2011; Praetorius, Martens & Brinkmann, in Druck; Terhart, 2014). Einige dieser Zugangsweisen zielen auf eine Rekonstruktion des Unterrichtsgeschehens, andere auf normative Vorgaben für professionelles Handeln, wieder andere auf theoretische und empirische Klärungen der Frage, welcher Unterricht Bildung ermöglicht und unterstützt. Einen prominenten Ansatz für die Beantwortung der letztgenannten Frage stellt die quantitative empirische Unterrichtsforschung dar. Seit der TIMSS-Videostudie aus dem Jahr 1995 (Baumert et al., 1997; Stigler & Hiebert, 1999) hat sich dieser Ansatz zu einem der aktivsten Teile der deutschsprachigen Bildungsforschung entwickelt. Zentraler Topos ist Unterrichtsqualität 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. 396). International passt sich diese Forschung in das Educational-Effectiveness-Paradigma ein, das generell nach Merkmalen von Bildungsinstitutionen und -prozessen fragt, die mit der Erreichung von gesetzten pädagogischen Zielen zusammenhängen bzw. – in einer strengeren, kausalanalytisch argumentierenden Variante – diese bewirken (z. B. Muijs et al., 2014; Seidel & Shavelson, 2007). Das aktuelle Beiheft fokussiert auf die quantitative empirische Unterrichtsforschung und thematisiert zentrale Herausforderungen, die mit der Erforschung von Unterrichtsqualität im Sinne dieses Ansatzes einhergehen.

Theoretische und methodische Innovationen in dessen Rahmen ermöglichen es mittlerweile, Aussagen über zentrale Dimensionen von Unterrichtsqualität und das damit verbundene komplexe Bedingungsgefüge zu treffen. Dies schließt eine an Schultheorien anschlussfähige Auswahl von Qualitätsdimensionen ebenso ein wie differenzielle Aussagen über das Zusammenwirken von individuellen Voraussetzungen der Lehrpersonen sowie Lernenden, Kontextmerkmalen wie beispielsweise pädagogischen Traditionen und Qualitätsmerkmalen des Unterrichts. Zugleich lassen sich grundlegende Herausforderungen identifizieren, deren Bearbeitung für eine Weiterentwicklung des Forschungsfeldes von hoher Bedeutung ist. Solche Herausforderungen werden innerhalb eines Forschungsansatzes oft eher randständig thematisiert, weil die Antworten darauf zum paradigmatischen Kern gehören, der in empirischen Originalarbeiten kaum in Frage gestellt wird. Ziel des Beiheftes ist es, basierend auf dem aktuellen Erkenntnisstand, aber auch aus einer Meta-Perspektive heraus, zentrale theoretische und methodologische Herausforderungen des Forschungsfeldes zu identifizieren und eine Plattform für die dezidiert (selbst-)kritische Auseinandersetzung mit diesen zu bieten.

Zeitgleich spiegeln die Herausforderungen, die im Beiheft thematisiert werden, jedoch Grundprobleme einer jeden Unterrichtsforschung (siehe auch Praetorius et al., in Druck). So kann es mit Gruschka (2013) als zentrale Aufgabe bei der Analyse von Unterricht gelten, *Dimensionierungen* und *Skalierungen* des Unterrichtsgeschehens vorzunehmen, um Befunde generalisieren zu können. Dies geschieht in der quantitativen Forschung natürlich mit anderen Methoden, unter Verknüpfung erziehungswissenschaftlicher mit pädagogisch-psychologischen Konzepten, aber die Frage nach grundlegenden Dimensionen von Unterricht (1. Herausforderung) stellt sich paradigmenübergreifend. In der weiteren Ausarbeitung muss jedwede Forschung zu Unterricht auf dessen Komplexität reagieren, wie sie etwa Asbrand und Martens (2018) mit Bezug auf Luhmann herausarbeiten: „Im Kontext der empirischen Bildungsforschung verweist das Angebots-Nutzungs-Modell auf die Kontingenz der Unterrichtsinteraktion“ (Asbrand & Martens, 2018, S. 91); dies ist die 2. im Beiheft thematisierte Herausforderung. Auch dass sich die Bedeutung konkreter Aktivitäten für das Lernen und Verstehen erst auf einer Tiefenebene erschließt (3. Herausforderung) und dass Lehrpersonen und Lernende prinzipiell unterschiedliche Perspektiven auf den Unterricht haben (4. Herausforderung) gehört zu den Grundproblemen jeder Unterrichtsforschung. Schließlich kann es als Ziel von Unterricht benannt werden, „Veränderungen bei den Edukanden herbeizuführen“ (Asbrand & Martens 2018, S. 100). Der quantitative Ansatz beansprucht, solche Veränderungen – die allein aus einer Rekonstruktion des Prozessgeschehens nicht beurteilbar sind – mit Tests und Befragungen explizit zu messen und statistisch als Wirkungen des Unterrichts zu modellieren (5. Herausforderung), auch wenn bildungsphilosophische Konzepte dabei nicht in ihrer Gänze abgedeckt werden können – was einmal mehr zeigt, dass Unterrichtsforschung der Kombination verschiedener Zugänge bedarf.

Das Beiheft ist entsprechend diesen fünf Herausforderungen gegliedert. In jedem Themenblock erfolgt zunächst ein konzeptueller Beitrag, der einen Überblick über den aktuellen Forschungsstand und mit diesem verbundene Herausforderungen sowie Über-

legungen zu möglichen Lösungsansätzen bietet. Anschließend folgt ein empirischer Beitrag, der exemplarisch Überlegungen aus einem Teilbereich des vorhergehenden Beitrags aufgreift und einer empirischen Überprüfung unterzieht. Den Abschluss des Themenblocks bildet eine Diskussion der beiden Beiträge durch ausgewiesene Expertinnen und Experten zu dieser Thematik. Im Anschluss an die fünf Themenblöcke erfolgt eine Diskussion des gesamten Beihefts aus verschiedenen Perspektiven.

Die Autorinnen und Autoren der Beiträge sind mehrheitlich Teil des Leibniz-Netzwerks Unterrichtsforschung, das einen interdisziplinären Kreis von 30 Wissenschaftlerinnen und Wissenschaftlern aus der quantitativen empirischen Unterrichtsforschung mit dem Fokus auf das Fach Mathematik umfasst (www.unterrichtsforschung.dipf.de). Aufgrund der Verankerung der quantitativen empirischen Unterrichtsforschung im internationalen Educational-Effectiveness-Paradigma sind drei der fünf Themenblöcke englischsprachig angelegt. Die Begriffspaare *Angebot und Nutzung* sowie *Oberflächen- und Tiefenstruktur* werden jedoch eher im deutschsprachigen Raum diskutiert; daher sind die zugehörigen Themenblöcke auf Deutsch verfasst.

Themenblock I: Dimensionen der Unterrichtsqualität – Theoretische und empirische Grundlagen (englischsprachig)

Im deutschen Sprachraum dominiert zur Konzeptualisierung von Unterrichtsqualität in quantitativ-empirischen Studien das Modell der drei Basisdimensionen, das Klassenführung, konstruktive Unterstützung sowie kognitive Aktivierung als Komponenten von Unterrichtsqualität unterscheidet.[1] Der konzeptuelle Beitrag des Themenblocks von Praetorius, Klieme, Kleickmann, Brunner, Lindmeier, Taut und Charalambous stellt die Entwicklung und den aktuellen Forschungsstand zu diesem Modell dar sowie eine Reflexion dazu, inwieweit das Modell der drei Basisdimensionen als Theorie im Sinne der analytischen Wissenschaftstheorie angesehen werden kann. Der empirische Beitrag von Kleickmann, Steffensky und Praetorius geht einem der Kritikpunkte an dem Modell nach, inwiefern das Modell einer weiteren Ausdifferenzierung mit zusätzlichen Dimensionen bedarf. In ihrer Diskussion setzt sich Bell unter Bezug auf einen Ansatz von Toulmin mit einer weiteren Möglichkeit auseinander, das Modell der drei Basisdimensionen stärker theoretisch zu fundieren.

1 Klassenführung dient durch den präventiven und intervenierenden Umgang mit Unterbrechungen und Disziplinproblemen dazu, Schülerinnen und Schülern möglichst viel Zeit zum Lernen zur Verfügung zu stellen. Konstruktive Unterstützung fördert deren Erleben von Autonomie, Kompetenz und sozialer Eingebundenheit. Kognitive Aktivierung bezieht sich auf die Anregung kognitiv anspruchsvoller Prozesse des Problemlösens und des Verstehens.

Themenblock II: Angebots-Nutzungs-Modelle als Rahmung (deutschsprachig)

Angebots-Nutzungs-Modelle werden in der quantitativen empirischen Unterrichtsforschung vielfach als Rahmung herangezogen, um zu erklären, warum und wie Unterricht Schülerinnen und Schüler beeinflussen kann. Sie unterscheiden dazu zwischen dem unterrichtlichen Lernangebot und den Lernaktivitäten der Schülerinnen und Schüler. Der konzeptuelle Beitrag des Themenblocks von Vieluf, Praetorius, Rakoczy, Kleinknecht und Pietsch gibt einen Überblick über zentrale Aspekte des Rahmenmodells, reflektiert kritisch die unterschiedlichen Versionen des Modells in der Forschungsliteratur und schlägt abschließend ein integriertes Rahmenmodell vor. Der empirische Beitrag von Meissner, Merk, Fauth, Kleinknecht und Bohl stellt ein Beispiel für die Herausforderung dar, die vermittelnden Effekte der Nutzung des Lernangebots durch die Schülerinnen und Schüler auch empirisch fassbar zu machen. In ihrer Diskussion geht Seidel auf notwendige Schritte für eine Weiterentwicklung der Modellierung von Angebot und Nutzung in der quantitativen empirischen Unterrichtsforschung ein.

Themenblock III: Oberflächen- und Tiefenstruktur des Unterrichts (deutschsprachig)

Mit der Unterscheidung von Oberflächen- versus Tiefenstruktur betont ein wichtiger Teil der deutschsprachigen Fachliteratur im Anschluss an Piaget und Aebli, dass es weniger die einzelnen Lehr-Lern-Aktivitäten, Sozialformen usw. sind, die das Lernen von Schülerinnen und Schülern im Unterricht erklären, sondern vielmehr grundlegende Prozessmerkmale wie etwa die Basisdimensionen oder sogenannte ‚Basismodelle' als Prototypen des Lerngeschehens. Im konzeptuellen Beitrag des Themenblocks von Decristan, Hess, Holzberger und Praetorius wird ein Überblick über aktuelle Sichtweisen und Definitionen des Begriffspaares gegeben. Dabei wird eine begriffliche Schärfung diskutiert – die u. a. in dem Vorschlag mündet, statt von „Strukturen" von „Merkmalen" zu sprechen – und zu einfache Abgrenzungen – z. B. die Annahme, Tiefenmerkmale seien prinzipiell schwerer beobachtbar – werden zurückgewiesen. Hess und Lipowsky zeigen in ihrem empirischen Beitrag, dass die Qualität von Lehrpersonenfragen (als Tiefenmerkmal) nicht unabhängig ist von der Sozialform des Grundschulunterrichts (als Oberflächenmerkmal). In ihrer Diskussion geht Pauli nochmals auf den Ursprung der Unterscheidung von Tiefen- und Oberflächenstrukturen ein und erörtert das Potenzial einer Forschung, die Unterrichtsmerkmale auf diesen beiden Ebenen lokalisiert.

Themenblock IV: Zur Bedeutung unterschiedlicher Perspektiven bei der Erfassung von Unterrichtsqualität (englischsprachig)

Zur empirischen Erfassung von Unterrichtsqualität wird in der quantitativen empirischen Unterrichtsforschung in der Regel auf Einschätzungen von Schülerinnen und Schülern, Lehrpersonen und/oder externen Beobachtenden zurückgegriffen. Der konzeptuelle Beitrag des Themenblocks von Fauth, Göllner, Lenske, Praetorius und Wagner gibt einen Überblick über die oftmals lediglich geringen bis nicht vorhandenen Zusammenhänge der Perspektiven, diskutiert mögliche Gründe und stellt eine Referenten-Perspektiven-Matrix vor, die genutzt werden kann, um Unterschiede zwischen den Perspektiven zu verstehen und Messinstrumente adäquat zu gestalten. Im empirischen Beitrag von Göllner, Fauth, Lenske, Praetorius und Wagner wird der Befund, dass Klassenführung konsistent über viele Studien hinweg mit der Leistungsentwicklung von Schülerinnen und Schülern einhergeht, kritisch in den Blick genommen. Es wird gezeigt, dass die Befunde von der jeweiligen Itemformulierung (Fokus auf die Lehrperson vs. die Schülerinnen und Schüler) beeinflusst werden. In seiner Diskussion reflektiert Clausen zentrale Annahmen der beiden Beiträge und entwickelt Anregungen für eine Weiterentwicklung der Forschung zur Perspektivenspezifität von Unterrichtsqualität.

Themenblock V: Modellierung der Wirkungen von Unterrichtsqualität (englischsprachig)

Zur adäquaten Überprüfung der Wirkungen sind komplexe statistische Modelle notwendig, die verschiedene Setzungen erfordern. Der konzeptuelle Beitrag dieses Themenblocks von Naumann, Kuger, Köhler und Hochweber gibt einen Überblick über die üblichen Vorgehensweisen in der quantitativen Unterrichtsforschung und diskutiert methodische Standards. In dem empirischen Beitrag von Köhler, Kuger, Naumann und Hartig werden verschiedene Mehrebenenmodelle zur Überprüfung der Wirkungen von Unterrichtsqualität verglichen. Die Diskussion von Lüdtke und Robitzsch fokussiert auf methodische Herausforderungen bei der Identifikation kausaler Effekte in nicht-randomisierten Designs.

Um den Austausch zwischen Disziplinen und Paradigmen der Unterrichtsforschung anzuregen, erfolgt die abschließende Gesamtdiskussion des Beiheftes aus drei unterschiedlichen bildungswissenschaftlichen Perspektiven: Allgemeine Erziehungswissenschaft und Didaktik (Terhart), kognitionspsychologisch fundierte Didaktik und Unterrichtsforschung (Reusser) sowie Fachdidaktik (Lindmeier & Heinze).

Literatur

Asbrand, B., & Martens, M. (2018). *Dokumentarische Unterrichtsforschung.* Wiesbaden: Springer VS.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O., & Neubrand, J. (1997). *TIMSS – mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich: deskriptive Befunde.* Opladen: Leske + Budrich.

Gruschka, A. (2013). *Unterrichten – eine pädagogische Theorie auf empirischer Basis.* Opladen: Barbara Budrich.

Klieme, E. (2019). Unterrichtsqualität. In M. Gläser-Zikuda, M. Harring & C. Rohlfs (2018), *Handbuch Schulpädagogik* (S. 393–408). Münster: Waxmann.

Meseth, W., Proske, M., & Radtke, F.-O. (2011). *Unterrichtstheorien in Forschung und Lehre.* Bad Heilbrunn: Klinkhardt.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art – teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*(2), 231–256.

Praetorius, A.-K., Martens, M., & Brinkmann, M. (in Druck). Qualität von Unterricht: Forschungstraditionen, Ergebnisse und Kontroversen. In T. Hascher, W. Helsper & T.-S. Idel (Hrsg.), *Handbuch Schulforschung.* Wiesbaden: Springer VS.

Reusser, K. (2009). Unterricht. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 881–896). Weinheim: Beltz.

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. doi:10.3102/0034654307310317.

Stigler, J., & Hiebert, J. J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* New York: Free Press.

Terhart, E. (2014). Unterrichtstheorie – Einführung in den Thementeil. *Zeitschrift für Pädagogik, 60*(6), 813–816.

Anschrift der Autor_innen

Prof. Dr. Anna-Katharina Praetorius, Universität Zürich,
Lehrstuhl für pädagogisch-psychologische Lehr-Lernforschung und Didaktik,
Institut für Erziehungswissenschaft,
Freiestrasse 36, 8032 Zürich, Schweiz
E-Mail: anna.praetorius@ife.uzh.ch

Dr. Juliane Grünkorn, DIPF | Leibniz-Institut für Bildungsforschung
und Bildungsinformation,
Rostocker Straße 6, 60323 Frankfurt a. M., Deutschland
E-Mail: gruenkorn@dipf.de

Prof. Dr. Eckhard Klieme, DIPF | Leibniz-Institut für Bildungsforschung
und Bildungsinformation,
Rostocker Straße 6, 60323 Frankfurt a. M., Deutschland
E-Mail: klieme@dipf.de

Themenblock I: Dimensionen der Unterrichtsqualität – Theoretische und empirische Grundlagen

Anna-Katharina Praetorius/Eckhard Klieme/Thilo Kleickmann/Esther Brunner/ Anke Lindmeier/Sandy Taut/Charalambos Charalambous

Towards Developing a Theory of Generic Teaching Quality

Origin, Current Status, and Necessary Next Steps Regarding the Three Basic Dimensions Model

Abstract: In this paper we elaborate upon the relevance of theories of teaching (quality) in quantitative empirical research on teaching. First we introduce, the quantitative empirical research approach. Then, we present the origin and current status of research with respect to a model – the Three Basic Dimensions of teaching quality – that is especially popular in quantitative research on teaching quality in German-speaking countries. Next, we reflect on the extent to which the model fulfills criteria for a good theory, before deriving conclusions for future research that focuses on a process of successive theory building.

Keywords: Theory, Teaching Quality, Three Basic Dimensions, Instruction, Model

1. The Relevance of Theories of Teaching and Teaching Quality

Scientific endeavors generally aim to develop theory: namely, general statements that allow for understanding, explaining, predicting, or critically reflecting upon phenomena. According to modern philosophy of science, any acceptable theory should be a coherent, systematic composition of such statements, and research should constantly aim to test and revise those theories. According to Kuhn (1969) and Lakatos (1977), theories are embedded into paradigms and research programs that are grounded in shared a priori assumptions and intended applications. In the present paper we discuss quality criteria for theories of teaching (see Section 3) and apply them to a model of teaching quality, the Three Basic Dimensions (TBD) model (the origins and foundations of which are outlined in Section 2), which has been developed in quantitative research on teaching quality in German-speaking countries.

Introducing a special issue in the journal *Zeitschrift für Pädagogik* on *Unterrichtstheorie* (theory of teaching[1]), Terhart (2014) stated that educational scientists rarely dis-

1 The German concept of *Unterricht* may be translated as teaching or as instruction. In fact, the

cuss teaching from a decisively theoretical, analytical perspective free of practical considerations. For example, while traditional didactics has in some cases been claimed to be a theory of teaching (e. g., Prange, 2005), Oelkers (2000) has emphasized that the 19th and 20th century German approaches to didactics and teaching should be perceived as reflections on professional practice rather than theory. More recently, Lüders (2014) criticized didactics as eclectic, lacking empirical validation, and/or being narrowly focused on curriculum. As examples of appropriate theories of teaching, both Terhart (2014) and Lüders (2014) cite approaches that are based on qualitative methods and conceptual foundations from the social sciences (e. g., practice theory; Reckwitz, 2002).

In the same special issue, Seidel (2014) provided a review of research on teaching based on quantitative methods. She distinguished two broader paradigms. In the first paradigm, psychological theories of cognition, metacognition, or motivation are used to identify features of classroom teaching that may support these aspects of learning. The second paradigm aims to systematize aspects of teaching, studying their structure and their relations with student outcomes. Historically, this paradigm emerged as a derivative and enhancement of the process-product paradigm (Gruehn, 2000; see Section 2). TBD is mentioned as an example of this second paradigm (see Seidel, 2014).[2]

With regard to the status of theories in quantitative research on teaching, Seidel's (2014) review is quite illuminating. Although it was published in a special issue on theories of teaching, all occurrences of the term *theory* were related to psychological theories of learning, cognition, metacognition, or motivation, while the term *model* was always used in discussing aspects of teaching within the second paradigm. Overall, the paper included the term *model* 63 times and the term *theory* just 21 times. Accordingly, Terhart (2014, p. 815) characterized the research reviewed by Seidel as "a process of developing and testing models. These models function as a shared foundation of thinking within a certain context of research, and require stepwise, self-corrective verification" (translated by the authors). What Terhart described is basically the process of theory building and revision, as envisioned in the philosophy of science. But in line with Seidel, he preferred to use the notion of model instead of theory, due perhaps to the sketchy, eclectic, or data-driven nature of some of this work. Theories, we may conclude, compared to models, need to be conceptually richer, more coherent, more robust, and more general.

---

terms teaching and instruction are often used interchangeably. Cohen, Raudenbush, and Ball (2003) consider teaching to be the narrower term, since it might be thought of as "something done by teachers to learners" whereas instruction in their understanding refers to "interactions among teachers and students, around content, in environments" (p. 122). *Unterricht* is very much in line with the latter notion. Nevertheless, we are using the term teaching throughout this paper, as this term meanwhile seems to be used in a much broader sense, encompassing all kinds of classroom and professional activities (cf. e. g., Ball & Forzani, 2009).

2 Although the title of the Seidel (2014) paper refers to psychology of teaching, the second paradigm mentioned could equally be subsumed under educational science. In the present paper, we refrain from establishing borderlines between educational science, educational psychology, and sociology of education, as teaching requires an interdisciplinary approach.

This fuzzy and oftentimes rather pragmatic way of conceptualizing teaching can be found in the international literature as well. For example, the most recent *Handbook of Research on Teaching* (Gitomer & Bell, 2016) does not include any general theory of teaching. The chapter by Cappella, Aber and Kim (2016) integrates research from different areas of psychology and educational science into a complex model which in many respects is similar to the German models of opportunities and uses of instruction (see Section 4). Some prominent current work on teaching in the US, such as Deborah Ball's *high leverage teaching practices* (Ball & Forzani, 2009), is similar to the pragmatic approaches of German didactics, although the link to empirical evidence is much stronger for the former. Other prominent research on teaching evolved in the form of frameworks guiding the development of student surveys or observation protocols (e. g., Danielson, 2013). Frameworks may over time evolve into more elaborated models that also explicate relations among constructs of interest (e. g., the Dynamic Model of Educational Effectiveness; Creemers & Kyriakides, 2008). Eventually, these models might mature into theories (Leplin, 1980) fulfilling the criteria we discuss below (see Section 3). This distinction between frameworks, models, and theories is, however, often not used systematically in research on teaching quality.

The current special issue intends to address the lack of theorizing and systematic revision of theories in quantitative research on teaching. In the present paper, we discuss the core assumption of the second paradigm mentioned by Seidel (2014): namely, the possibility of identifying a limited set of measurable dimensions that are well-founded in theories or models of teaching and that explain effects of teaching on students. As these dimensions are assumed to be the main drivers of teaching effectiveness, they are called quality dimensions. We study this assumption by choosing the TBD model[3] of teaching quality (Klieme, 2019; Kunter & Trautwein, 2013; Praetorius, Klieme, Herbert & Pinger, 2018), which has gained a lot of attention in the German-speaking research literature over the last two decades but which has also been applied internationally, for example in TIMSS (e. g., Nilsen & Gustafsson, 2016) and PISA (e. g., Kuger, Klieme, Lüdtke, Schiepe-Tiska & Reiss, 2017). Thus, studying the theoretical foundations of quantitative research on teaching, using TBD as an example, seems to be a valuable undertaking.

3 Depending on the publication, TBD is called a (theoretical) framework (e. g., Fauth, Decristan, Rieser, Klieme, & Büttner, 2014) or a (theoretical) model (e. g., Klieme, Pauli, & Reusser, 2009). At least in the abstract of an article by Klieme and Rakoczy (2008), TBD is also called a theory.

2. Three Basic Dimensions of Teaching Quality: Origins and Foundations of the Model

Carroll (1963) first introduced the notion of teaching quality to empirical educational research, defining it as the extent to which teaching enables students to learn a task as quickly and effectively as possible. In other words, teaching quality was understood as a moderator variable shaping the relation between student aptitudes, learning time, and learning outcomes. Thus it was not defined substantially, but was deemed to cover any aspect of teaching (i. e., the process) that may help to optimize the student learning outcomes (i. e., the product). In the decades to follow, the so-called process-product paradigm of quantitative empirical research aimed at identifying such aspects. During the late 1980s and the 1990s, the basic notions of the paradigm were refined to include mediator variables, enabling the coverage of cognitive and affective processes that govern student learning. At the turn of the century, a certain substantive consensus had been reached on pivotal aspects of teaching quality. A seminal meta-analysis conducted by Wang, Haertel, and Walberg (1993) showed that classroom management and the quality of student-teacher academic interactions (namely, the intensity as well as the quality of questioning and answering) had about the same mean effect size as cognitive and meta-cognitive student aptitudes, their home environment, and parental support.[4]

Building upon the research conducted in the process-(mediation)-product paradigm, large-scale research on educational trajectories and educational effectiveness was conducted in the 1990s at the Center for Educational Research at the Max Planck Institute for Human Development in Germany. In her dissertation, Gruehn (2000) combined the international research literature with national traditions in didactics (e. g., the notion of Socratic teaching as defined by Wagenschein (1992)) and research on classroom climate (Eder, 1996). Gruehn assessed students' perceptions of teaching in various subjects using a total of 21 questionnaire scales. After dropping nine scales, Gruehn was able to establish five second-order dimensions of teaching quality through Confirmatory Factor Analysis: classroom management, pacing, adaptivity, affective quality, and constructivism (Gruehn, 2000).

At the same time, the Max Planck Institute was operating as the National Center for the Third International Mathematics and Science Study (TIMSS 1995; Baumert et al., 1997). Baumert and colleagues decided to enhance the international TIMSS design by using the same 21 questionnaire scales, formerly used by Gruehn. Furthermore, Germany, the US, and Japan became part of a video study which, for the first time ever, would cover mathematics lessons in nationally representative samples of classrooms, allowing comparisons of teaching in these three countries (Stigler & Knoll, 1999). While publications from the international study were based on codes, ratings, and qualitative

4 Researchers had also set out to integrate findings conceptually (e. g., Rosenshine & Furst, 1973, with their concept of direct instruction). As in the German literature (see Section 1), however, many authors questioned whether these approaches can be called theories (see, e. g., Berliner, 2009; Biddle & Anderson, 1986; Hill & Schrum, 2002; Snow, 1973).

Factor 1 Classroom management	Factor 2 Student support	Factor 3 Cognitive activation
• Dealing effectively with disruptions • Frequency of disruptions (−) • Waste of instructional time (−) • Volatility of the teacher (−) • Clarity of rules • Clarity and structure of teaching • Monitoring • Time on task	• Social orientation • Individual frame of reference • Teachers' diagnostic competence regarding social needs • High interaction speed (−) • Achievement pressure (−)	• Socratic teaching • Challenging practicing • Repetitive practicing (−) • Teachers' ability to motivate students

Note. (−) indicates a reverse-scored scale.

Tab. 1: Scales representing the three basic dimensions in TIMSS-Video 1995

scripts derived from videos only (Stigler & Hiebert, 1999), the German team managed to combine the video recording in Grade 8 with the TIMSS assessment, plus a follow-up test and a student questionnaire implemented one year later. Thus, TIMSS and TIMSS-Video 1995 were combined into a full-size longitudinal study of teaching effectiveness in Germany.

Among others, an attempt was made to turn the 21 scales used by Gruehn (2000) into observation protocols for trained observers (Clausen, 2002). In order to systematize and structure the observational space, Klieme, Schümer, and Knoll (2001) ran an exploratory factor analysis with these high-inference rating scores, resulting in a clear three-dimensional solution (see Tab. 1; English labels provided by the present authors).[5]

Klieme et al. (2001) reported the main findings of this analysis in their non-technical paper, written for teachers interested in professional learning. They provided initial evidence that cognitive activation was positively related to gains in mathematics achievement, while student support was related to a more positive development of students' interest in mathematics. The authors suggested that there might be a non-linear relation between cognitive activation and gains in mathematics achievement (i. e., too much cognitive activation might be suboptimal). Using the same data (but without referring to TBD), Clausen (2002) additionally showed that teaching quality depends on the perspective (i. e., teacher, student, or observer perspective) from which it is evaluated, as these perspectives did not converge in their judgments of teaching quality.

All in all, the foundational work in TIMSS-Video established (a) the three-dimensional structure of teaching quality, (b) its relevance for explaining student outcomes over the course of a school year, also providing hints on (c) non-linear relations with student outcomes, and (d) the perspective-specific nature of judgments of teaching quality. Klieme et al. (2001) introduced TBD as a comprehensive model that was devel-

5 Both the scree test and the Kaiser criterion indicated that three factors, explaining a total of 73 % of the common variance of the scales, could be distinguished. Of the 21 scales, 17 could be unambiguously linked to one of the three factors (criterion: a unique loading > .65). The technical details given here were not reported in the original publication.

oped to reduce the multiplicity of measures of teaching quality available at that time to a smaller set of dimensions, applying the meta-theoretical principle of Occam's razor, which claims that scientific research should avoid redundancy. Instead of the longer lists of facets, aspects, features, factors, components, domains, or dimensions of (effective) teaching or teaching quality that had been published, a reduced set of basic dimensions was believed to be more easily interpreted and analyzed with respect to effects on student learning. The authors stated that the model should apply to all school subjects and grade levels, and potentially even to different countries.

Although they were mainly focused on summarizing the core ideas for a non-research audience, Klieme et al. (2001) established some initial theoretical foundations by linking the three dimensions to a re-conceptualization of traditional (mainly German) didactics published by Diederich and Tenorth (1997), who argued – citing seminal writings from the history of school pedagogy – that classroom teaching requires a certain level of student attentiveness, student motivation, and student understanding. Klieme et al. (2001) interpreted TBD as comprising those aspects of teaching that help teachers achieve these three student outcomes.

Soon after the publication of this foundational, yet informal paper, the research group at the Max Planck Institute became involved in PISA. Indicators for TBD were therefore implemented in national extensions to PISA 2000 and 2003. This is why the model was first explained in research papers published in the context of PISA. Klieme and Rakoczy (2003) provided theoretical explanations as to how and why the dimensions should be linked to student learning (see also Fig. 1) by referring to constructivist theories (as a

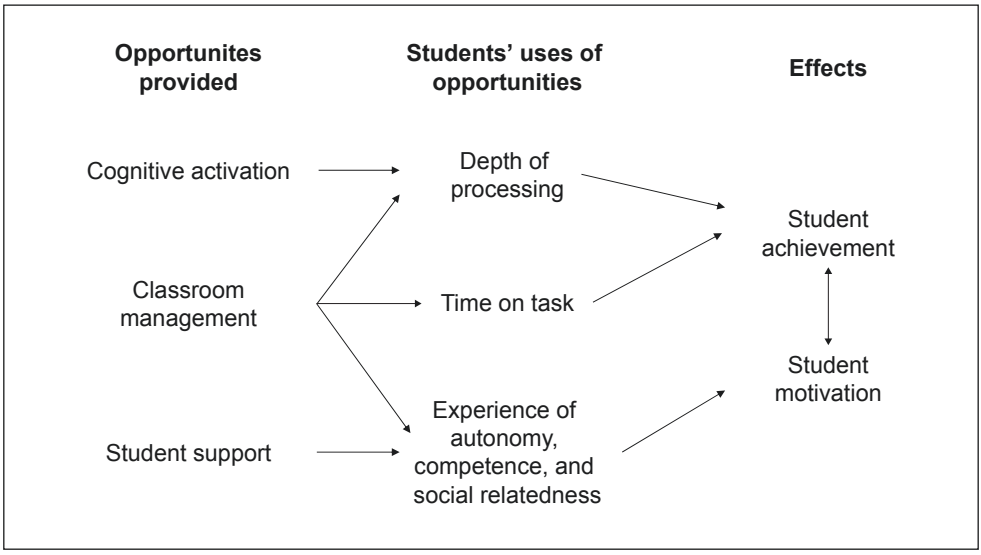

Fig. 1: The assumed relations between the three basic dimensions and student learning (adapted after Klieme et al., 2009, p. 140)

theoretical foundation for cognitive activation) and Self Determination Theory (Ryan & Deci, 2000; as a theoretical foundation for student support). Baumert, Kunter, and colleagues replicated the three dimensions for German data in PISA 2003 (published under the label COACTIV study; see Baumert et al., 2010; Kunter & Voss, 2011).

Internationally, the model of TBD was first presented by Klieme at the annual conference of the American Educational Research Association (AERA) in 2001, but the first scientific papers in English presenting TBD as a theoretical model (Klieme, Pauli & Reusser, 2009) or a theoretical conceptualization (Lipowsky et al., 2009) originated from a Swiss-German video study on teaching the Theorem of Pythagoras. This study also led to enhancements of the model. For example, Rakoczy, Klieme, Lipowsky, and Drollinger-Vetter (2010) discriminated behavioral aspects of classroom management from issues of content structure and clarity. Drollinger-Vetter (2011) and Lipowsky, Drollinger-Vetter, Klieme, Pauli, and Reusser (2018) discriminated generic aspects of cognitive activation from the quality of mathematical subject-matter presented.

Summing up, we can conclude that the TBD model, with its three-dimensional structure of teaching quality, was derived from factor analytical work. Initial empirical evidence showed its hypothesized relation to student achievement and motivation. Some effort has been spent on linking TBD to theoretical considerations, but the theoretical foundations of the TBD could still be improved, as shown in Section 4 of this paper. We envision that reflecting on possible ways to develop TBD into a theory could help to move the field forward by addressing the criticism that progress has so far largely been driven by empirical findings without a proper theoretical basis.

3. Desirable Characteristics of Theories of Teaching Quality

Quantitative empirical research on teaching, as it is addressed in the present special issue, is rooted in the understanding of scientific research that Terhart (2014) characterizes as "working on theories". This refers to a process of developing and testing models as a shared foundation of thinking in a research context that requires self-correcting examination. Despite the unclear delineation of the terms theory and model – and still more fuzzy terms such as shared foundation of thinking, or framework – there is a consensual need: quantitative empirical researchers in the field of teaching should constantly strive towards more mature frameworks/models/theories to gain more sophistication, rigorousness, and validity.

Although there are no clear, agreed-upon criteria for what qualifies as a theory in research, most researchers would probably agree that a theory is a system of concepts, connections, and explanations (e. g., Beck & Krapp, 2006; Kerlinger, 1964; Scriven, 1956/1994; Whetten, 1989). Following Rosenshine (2009), many would likely also agree that a theory on teaching "explains why and how teaching works[6], … explains

6 Others would not necessarily use the notion of "working", which seems to imply a rather technical approach.

why some kinds of teaching work better than others, … can be subjected to attempts at refutation, and … can be modified as new findings emerge" (Rosenshine, 2009, p. 1169).[7] The quantitative empirical paradigm thus clearly intends to link student learning (i. e., changes in dispositions, achievement, and behavior) to the classroom teaching in which students participate. Accordingly, it seems crucial to integrate theories of learning and motivation (e. g., social-constructivist theories of learning, Palincsar, 1998; or the self-determination theory of motivation, Ryan & Deci, 2000) with theories of teaching (e. g., Gage, 2009; Klieme et al., 2006).

In order to guide the development of TBD towards a more mature model, or even to a theory of teaching, criteria for scientific theories are needed. Some authors describe purposes and features of theories (e. g., Scriven, 1956/1994; Westermann, 2000) but do not present a set of crucial criteria on the basis of this description. Others do provide such a set (e. g., Beck & Krapp, 2006; Thagard, 1978; Whetten, 1989), but the criteria vary between authors, show considerable overlap, and are often formulated in a very general way.[8] Some authors, however, have also specified these general criteria for the field of teaching. In the following, we make pragmatic use of Kane and Marsh's (1980, p. 254) *integrated criteria for a general theory of instruction* (see Tab. 2) in order to highlight pivotal aspects regarding TBD. Although this set was proposed almost 40 years ago as a summary and integration of the assumptions of leading researchers, we are not aware of any more up-to-date set of criteria that is equally detailed and specific to theories of teaching.

As sections IA and IIB clearly show, Kane and Marsh (1980) subscribed to the view of analytical philosophy of science, defining theory as a consistent, hierarchical set of statements with empirical support and predictive value. Section IIA, testability, refers to a main aspect of critical rationalism. However, the set of criteria transcends these philosophical views, as limitations and restrictions of intended use are considered (IB), untested hypotheses are taken into account (IIB(3)), and prescriptions for teaching practice are requested (III). An aspect not included in the set of Kane and Marsh (1980) but emphasized in other places (e. g., Kuhn, 1970; Seidel, Prenzel & Krapp, 2014; Westermann, 2000) is a focus on the relation of a newly developed theory to existing approaches: Researchers need to make clear the extent to which a new theory builds upon, extends, refines, or overrules existing theories and observations. Only in doing so can we help in "developing the field by building cumulatively on existing knowledge and theory, rather than constantly attempting to reinvent the wheel" (Muijs et al., 2014, p. 251).

---

7 This view is in line with analytical philosophy of science and critical rationalism (Popper, 2005). Modern philosophy of science (for an overview, see Chalmers, 2007) also offers different perspectives, such as the non-statement view of theory, and scientific anti-realism. Nevertheless, analytical and critical-rationalist views dominate in the thinking of researchers conducting quantitative empirical research.

8 At the same time, such sets also show that in building theories, striking a balance between different criteria is necessary, as some of them are not independent, and some even contradict each other (e. g., comprehensiveness and parsimony in the set by Kane and Marsh, 1980).

I.	**Theoretical Characteristics**
A.	Characteristics and organization of the components. A theory of instruction should consist of a set of:
(1)	logically, and
(2)	theoretically related
(3)	internally consistent statements (axioms, corollaries, postulates), arranged in a
(4)	hierarchical or systematic order, so that
(5)	the higher level constructs integrate the constructs below.
(6)	These statements should be as few as possible to cover all of the theories and findings relevant to the area specified and should be
(7)	clearly defined.
(8)	If possible, these statements should be quantitatively related, as well as
(9)	qualitatively related.
B.	Boundaries
	The boundaries or limitations of concern of the theory should be stated, including such limitations as theories of learning and development subscribed to, philosophies adhered to, characteristics of the students and organizations deemed suitable. The most general theory will have as few such limitations as possible.
II.	**Empirical Characteristics**
	The statements included (except for axiomatic statements and those noted in IIB(3)) should relate to existing empirical evidence in the following manner:
A.	Testability
	The statements should be:
(1)	capable of being easily and clearly restated in the form of hypotheses about which
(2)	evidence can be collected to either verify or refute them.
B.	Support
	The statements should have
(1)	demonstrable empirical support and
(2)	predictive value in similar situations.
(3)	However, at the present time it may be necessary to include as yet untested hypotheses to meet the completeness criteria noted in section IA above.
III.	**Prescriptive Characteristics**
	To be of practical use, a theory of instruction should contain or clearly imply a series of prescriptive statements, specifying how best to obtain given ends, if they are desired. Areas to be covered include strategies, sequencing, materials, reinforcements, motivation.

Tab. 2: Set of integrated criteria for a general theory of instruction (Kane & Marsh, 1980, p. 254)

4. Reflecting on the Theoretical Foundation of TBD

In the following, we apply the core quality criteria mentioned above in an illustrative manner to the current understanding of TBD (see also Section 2). The literature included is based on the TBD review by Praetorius et al. (2018), some reviews available in German (Klieme, 2019; Kunter & Ewald, 2016; Lipowsky & Bleck, 2019) and on the authors' knowledge of the literature.

Our aim in applying the theory quality criteria to TBD is to help researchers in the field of teaching understand how their work can better contribute to theory building by using a model that holds considerable potential to evolve into a theory of teaching. To minimize the risk of subjective interpretations regarding the theoretical and empirical evidence for TBD, we ensured that our author group was diverse by (a) including the main developer of TBD, while at the same time making sure that we included researchers (b) from different disciplines, (c) working in more applied or research-oriented settings, (d) from both German-speaking and other countries, and (e) both working with TBD themselves, and not working with TBD. To ensure a broad and consensual perspective in application to each criterion, the initial drafts of the sections to follow were developed further until a consensus was reached by this interdisciplinary group.

4.1 Are the Statements Logically and Theoretically Related as well as Internally Consistent?

TBD can basically be seen as the combination of a structural component (quality dimensions), and a process component (effectiveness), including different psychological mediators (see Fig. 1).

The structural part of the model, distinguishing the three dimensions, could be elaborated further by including assumed relations between the dimensions (see also Sections 4.2 and 4.6). Conceptually, some overlap exists across dimensions with regard to the specific sub-dimensions used to conceptualize the dimensions (for details, see Praetorius et al., 2018). For example, the sub-dimension of support of competence experience (e. g., constructive approach to errors, differentiation and adaptive support, as well as constructive feedback) may contribute not only to the dimension of student support but also to cognitive activation.

The process part of the model takes up the idea of the model of opportunities and uses of instruction (e. g., Fend, 1998; see also Vieluf, Praetorius, Rakoczy, Kleinknecht & Pietsch, this issue): On the basis of different theories of learning and motivation, assumed mediating processes (i. e., the use of opportunities by students) between the teaching quality dimensions (i. e., the opportunities provided by the teacher) and student outcomes are included. The assumed effects of the three dimensions on student outcomes could, however, be described more specifically in terms of how these effects are assumed to work, at both individual and group levels (see Vieluf et al., this issue).

Additionally, some relations may be missing as, for example, cognitive activation might also have an effect on experience of competence.

4.2 Are the Statements Arranged in a Hierarchical or Systematic Order?

No hierarchy or order has been suggested explicitly for the TBD. In some publications, however, classroom management is described as the prerequisite for other aspects of teaching quality (e. g., Brunner, 2018; Klieme, Schümer & Knoll, 2001). This also fits well with Openshaw and Clarke's (1970) suggestion to distinguish three levels of teaching acts: those that set the stage for learning (i. e., classroom management and student support), those that are at the core of learning (i. e., cognitive activation), and those that appraise the process and the product (this last level is largely missing from the TBD conception).

4.3 How Explicitly is TBD Related to Existing Attempts to Capture a Theory of Teaching?

Klieme et al. (2001) related TBD to the three requirements of classroom teaching distinguished by Diederich and Tenorth (1997): student attentiveness, student motivation, and student understanding. This approach therefore can be seen as an initial theoretical foundation for TBD. In further developments of TBD, motivational and learning theories were added to this understanding (see Section 2).

TBD has also been connected to other frameworks or models that conceptualize teaching quality. This is particularly true for the Classroom Assessment Scoring System (CLASS; see Pianta & Hamre, 2009) with its three teaching dimensions: classroom organization (e. g., behavior management), emotional support (e. g., positive climate), and instructional support (e. g., content understanding). Without explicitly describing similarities and differences between TBD and CLASS, some publications seem to largely equate the dimensions included in both (e. g., Decristan et al., 2015; Fauth, Decristan, Rieser, Klieme & Büttner, 2014; Praetorius et al., 2017; Taut & Rakoczy, 2016). If both are indeed capturing exactly the same dimensions of teaching quality, one would need to critically challenge whether we need both. As the aspects covered in each of the dimensions are not structured and named in the same way, direct comparison of the two approaches is difficult. Through the synthesis of several teaching dimensions that resulted from considering different frameworks and models, including TBD and CLASS, Praetorius and Charalambous (2018), showed differences between the two frameworks/models: whereas TBD covers 10 of the 20 sub-dimensions of teaching quality in this synthesis, CLASS covers 15, and groups some of them differently. We therefore see value in better understanding the degree of overlap between different frameworks or models, as well in comparing the empirical support for the different assumptions (see also Sections 4.8 and 4.9).

4.4 Are the Statements Parsimonious and Comprehensive?

Whereas TBD was described as parsimonious and comprehensive in the initial publication by Klieme et al. (2001), subsequently it has been emphasized that it is parsimonious but likely not comprehensive. Lipowsky and Bleck (2019), for example, have claimed that a fourth dimension, subject matter quality (in their case, mathematics), needs to be added. Nilsen and Gustafsson (2016) enhanced the TBD model with a dimension called clarity, while Kleickmann, Steffensky, and Praetorius (this issue) established a dimension called cognitive support, which they distinguish from motivational support. Such claims receive further support from the synthesis by Praetorius and Charalambous (2018) across TBD and 11 other commonly used observational frameworks or models for describing teaching quality. Of the seven dimensions distinguished in the synthesis, one is not covered at all (i. e., support of practicing); of the 20 sub-dimensions across the seven dimensions, ten are missing in TBD (two generic, e. g., "teacher regularly checks for students' understanding"; three content-specific, e. g., "selecting worthwhile and developmentally appropriate content"; and five correspond to the interaction of generic and content-specific aspects, e. g., "presenting the content in a structured way"[9]). One could argue that at least the generic aspects should be included in TBD, to enable a comprehensive generic view on teaching quality.

4.5 Are the Statements Clearly Defined and Can They Be Tested?

For being able to test statements, explicit and clear hypotheses as well as clearly defined statements are necessary.

For TBD, such explicit and clear hypotheses have been formulated (e. g., the assumed effects of these dimensions on student achievement and student motivation; see Section 2, see also Fig. 1).

Clearly defined statements involve proper and agreed upon definitions and operationalizations of the constructs in which we are interested. However, in the case of TBD, the respective literature does not convey a consistent definition of its main elements and we find great differences in the operationalizations across studies (see Praetorius et al., 2018). This is even the case for classroom management, which is often considered a dimension in which the research community has developed a common understanding over the last decades. Some studies use rather narrow operationalizations (focusing exclusively on disruptions or effective use of time, e. g., Fauth et al., 2014, whereas others are broader (focusing also on aspects such as monitoring or clear rules; Lenske et al., 2016). This variability exists to an even larger extent for cognitive activation and student support. Here, narrow foci on challenging tasks and questions as well

9 In the initial publication on TBD by Klieme et al. (2001), structuredness was included as part of classroom management. In later publications (e. g., Lipowsky et al., 2009), classroom management was focused on time and behavior management.

as on exploration of the students' way of thinking exist for cognitive activation (e. g., Fauth et al., 2014) or on teacher-student relationships and differentiation for student support, respectively (e. g., Praetorius, Vieluf, Saß, Bernholt & Klieme, 2016). Broader foci also include discursive and co-constructive learning or the support of metacognition for cognitive activation (e. g., Korneck et al., 2017) or encompass constructive feedback and choice options for student support, respectively (e. g., Rakoczy & Pauli, 2006). Another question that needs to be addressed more clearly in this context is which perspectives (e. g., observer, student, and teacher ratings) are best suited to capturing each of the (sub)dimensions (see also Fauth, Göllner, Lenske, Praetorius & Wagner, this issue).

4.6 Are the Statements Quantitatively or Qualitatively Related?

The degree to which the statements of a theory are quantitatively and/or qualitatively related is an important prerequisite for the testability of theories (see Section 4.5). For example, although the three basic dimensions are partly positively correlated in empirical studies (e. g., Fauth et al., 2014; Kleickmann et al., this issue; Kunter et al., 2013), publications usually do not include explicit statements on expected relations. Such statements would, however, enhance the possibilities for validating the TBD model.

The non-deterministic relations of specific teaching dimensions and student outcomes are highlighted in the TBD by including students' uses of opportunities (e. g., time on task, depth of processing; see Fig. 1). What could be investigated in more detail is the nature of these relations – for instance, whether they are linear or non-linear (for an initial investigation, see Klieme et al., 2001) or whether an optimum or the maximum level of a teaching characteristic is most conducive for student learning (e. g., Brunner, 2018; Marzano & Marzano, 2003; Puntambekar & Hübscher, 2005).

4.7 Are the Boundaries Stated?

Although TBD is assumed to be broadly applicable to all age levels, school subjects, school forms, and possibly even countries (e. g., Klieme et al., 2001), boundaries can be identified. For example, despite the suggestion that research be value neutral (e. g., Seidel et al., 2014; Reiss & Sprenger, 2017), social research focuses on social phenomena that are per se intertwined with values and socio-cultural aspects in complex ways. International comparative research, for instance, has shown teaching and its effects on student outcomes to be influenced by culture (e. g., Bellens, van Damme, van den Noortgate, Wendt & Nilsen, 2019; Clarke, 2013; Stigler & Hiebert, 1999). It is evident that a theory of teaching cannot evade the complexities arising from cultural-societal contexts. Hammersley (2000) discusses that the only way of dealing with the dilemma is that researchers pay close attention to the values and socio-cultural assumptions implicated in their work and make them explicit. Thus, one could enhance the

explicitness of the TBD with respect to its boundaries by elaborating on the underlying socio-cultural assumptions more explicitly (for an example, see Fischer, Praetorius & Klieme, 2019).

The boundaries of TBD could also be made more explicit with respect to its capability to model content-specific teaching aspects. In general, TBD aims to cover generic aspects of teaching quality. Content-specific aspects therefore are explicitly excluded (e. g., accuracy of the content taught; see for example Charalambous & Litke, 2018). However, due to the fact that most studies on TBD have been tied to mathematics (see Section 4.9), as well as the challenge that cognitive activation is closely intertwined with the content taught, the focus on only generic aspects of teaching quality is not as clear-cut as one might think.

4.8 Are the Statements Supported Empirically?

Models of teaching quality are often highly complex, considering a large number of aspects that are interrelated. Hence it is not possible to generate empirical data in a single study that is suited to support such models as a whole. Empirical support therefore typically pertains to sub-parts, and researchers aim to accumulate evidence across several studies to eventually come close to a complete picture. The structural assumption that three dimensions can be distinguished has been supported by several studies using factor analysis (for an overview, see Praetorius et al., 2018). The fact that these studies differed in regard to the instruments and operationalizations used, could be seen as additional support for the model, as it is robust in this way. However, in most studies the decision-making with respect to selecting or developing the specific scales used for measuring TBD, is not explained. Therefore, one could question whether such decisions were conceptually-driven prior to data collection, or whether they resulted from a data-informed a posteriori decision process – a problem that, of course applies not only to research on TBD but to empirical studies in general without an open science approach (see e. g., Open Science Collaboration, 2017).

An important assumption in the initial publication on TBD by Klieme et al. (2001) – one that is, however, not explicitly stated in later publications – was that the three dimensions are comprehensive for capturing teaching quality. One way of testing this assumption is by testing the increment in predictive value of student outcomes when taking into account additional teaching aspects. Existing studies indicate the need to include aspects of teaching quality that go beyond TBD (see Section 4.4).

Furthermore, the assumption that all three dimensions have effects on student outcomes could only be confirmed for approximately half of the findings, according to a review of the current empirical literature that included only multi-level longitudinal designs, as these allow methodologically proper testing of such effects (see Praetorius et al., 2018). These inconsistent findings should be discussed, and should inform a revision of TBD in future publications. Such revisions have been suggested by Klieme in several conference presentations, although these suggestions have not as yet been

published. In 2012 for example, Klieme proposed a six-factor model, splitting each of the basic dimensions into a behavioral/interactive factor and a content-related factor (Klieme, 2012).

4.9 Are There Yet Untested Main Statements?

One of the assumptions of TBD is that the three dimensions are relevant for all subjects, grade levels, school forms and, potentially, different countries. Of the 21 empirical studies provided in the overview by Praetorius et al. (2018), the majority were conducted in mathematics (12 studies), with the remaining studies being distributed between German (5 studies), science education/physics (4 studies), English as a Foreign Language (2 studies), accounting (1 study), or across all subjects (3 studies).[10] This points to the need to test TBD more frequently in subjects other than mathematics and science education (e. g., language, social sciences, or musical/aesthetic subjects).

These studies mainly took place in lower secondary grades (9 studies), some additionally covering the lower part of upper secondary school (6 studies), whereas studies in primary school (4 studies) and those spread across all grades (2 studies) were rather the exception. This variation shows that evidence is needed on TBD in upper secondary grades and primary school.

TBD has been implemented in international studies such as PISA and TIMSS (see Section 2), and its dimensional structure, as well as the cross-sectional relations with outcomes, provides initial support for its applicability on an international level. It is important to note, however, that these studies also partly indicate that the dimensions might not be interpretable in the same way across countries (e. g., Fischer et al., 2019; see also Section 4.7).

Finally, the more complex assumptions on mediating processes between TBD and student outcomes – such as students' time on task for classroom management, students' experiences of autonomy, competence, and social relatedness for student support, as well as students' depth of processing for cognitive activation – have been stated theoretically (Klieme, Lipowsky, Rakoczy & Ratzka, 2006), but so far have rarely been investigated empirically (exceptions are, for example, Helm, 2016, as well as Rakoczy et al., 2019). The reasons for this, among others, include the lack of clarity with respect to how to conceptualize the mediating processes (see Vieluf et al., this issue), and how to develop reliable and valid measures capturing student thinking and attentiveness (e. g., using experience sampling methods).

10 The sum of the named studies exceeds 21, as some studies covered several subject matters.

4.10 Are There Yet Untested Main Statements?

According to Kane and Marsh (1980), theories of teaching quality should not only fulfill certain criteria for research purposes, but they should also contain, or clearly imply, statements that are useful for guiding and improving teaching practice. The purpose of TBD has been to serve both research and professional practice by providing an analytic tool on how to conceptualize teaching quality and its effects on student outcomes. Although the model was not developed with the goal of offering concrete directions for practice, it was meant to stimulate professional learning: TBD provides a general heuristic that may guide feedback and reflection on teaching. In fact, even the initial publication on TBD by Klieme et al. (2001) was accompanied by a CD-ROM providing material for professional training. Further research projects using TBD have produced professional development materials on cognitive activation for teachers (see Krammer et al., 2006). Furthermore, in educational policy (e. g., school evaluation and inspection systems in Germany), TBD has been used as an input in the development of standards describing quality teaching, and for the validation of classroom observation protocols (Taut & Rakoczy, 2016).

5. Conclusion

"[P]leas for better theory fall on receptive ears but recalcitrant hands. Everyone agrees that our theories should be stronger, as long as it does not require us to do anything differently" (Sutton & Staw, 1995, p. 378). Although developing stronger theories represents an arduous task, we believe it is time to take action so as to bring the field of teaching quality forward. The answers to the questions above provide some initial ideas on how to foster theory development for TBD specifically; yet the analytic process outlined above can also be transferred to other teaching models and frameworks. A summary of the most important needs for action might read as follows: (a) defining and summarizing the main statements of a model or framework in a testable, parsimonious but comprehensive way; (b) providing an overview of the extent to which the statements are empirically supported, and revising them in cases where they are not; (c) relating these single statements logically, theoretically, and internally consistently to each other, as well as arranging them hierarchically or systematically; (d) clearly stating the boundaries of the model or framework; (e) being more explicit on how the model or framework of interest is related to other existing approaches of conceptualizing teaching quality; and eventually (f) ensuring its relevance and potential for improving teaching practice.

It is hoped that in the years to come scholars will undertake this work more consistently and collaboratively, in producing stronger theories of teaching quality that, in turn, can help us better understand how teaching contributes to student learning.

References

Ball, D. L., & Forzani, F. M. (2009). The work on teaching and the challenge for teacher education. *Journal of Teacher Education, 60*(5), 497–511.

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,* 133–180.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O., & Neubrand, J. (1997). *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde.* Wiesbaden: Springer.

Beck, K., & Krapp, A. (2006). Wissenschaftstheoretische Grundfragen der Pädagogischen Psychologie. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie* (S. 33–73). Weinheim: Beltz.

Bellens, K., van Damme, J., van den Noortgate, W., Wendt, H., & Nilsen, T. (2019). Instructional quality: Catalyst or pitfall in educational systems' aim for high achievement and equity? An answer based on multilevel SEM analyses of TIMSS 2015 data in Flanders (Belgium), Germany, and Norway. *Large-Scale Assessments in Education, 7,* 1–27.

Berliner, D. C. (2009). Review regarding the book. In N. L. Gage (Eds.), *A conception of teaching* (see back of the book). New York: Springer Science and + Business Media.

Biddle, B. J., & Anderson, D. S. (1986). Theory, methods, knowledge, and research on teaching. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 230–252). New York: MacMillan.

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *Journal für Mathematik-Didaktik, 39*(2), 257–284.

Cappella, E., Aber, J. L., & Kim, H. K. (2016). Teaching beyond achievement tests: Perspectives from developmental and education science. In D. H. Gitomer, & C. A. Bell (Eds.), *Handbook of research on teaching* (pp. 249–348). Washington, DC: American Education Research Association.

Carroll, J. B. (1963). A primer of programmed instruction in foreign language teaching. *International Review of Applied Linguistics in Language Teaching, 1*(2), 115–141.

Chalmers, A. F. (2007). *Wege der Wissenschaft: Einführung in die Wissenschaftstheorie* (6. Band). Berlin: Springer.

Charalambous, C. Y., & Litke, E. (2018). Studying instructional quality by using a content-specific lens: The case of the Mathematical Quality of Instruction framework. *ZDM Mathematics Education, 50,* 445–460.

Clarke, D. J. (2013). International comparative research into educational interaction: Constructing and concealing difference. In K. Tirri & E. Kuusisto (Eds.), *Interaction in educational settings* (pp. 5–22). Rotterdam: Sense Publishers.

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität.* Münster: Waxmann.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119–142.

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools.* London/New York: Routledge.

Danielson, C. (2013) *The framework for teaching evaluation instrument.* www.danielsongroup.org [24. 09. 2019].

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

Diederich, J., & Tenorth, H.-E. (1997). *Theorie der Schule: Ein Studienbuch zu Geschichte, Funktionen und Gestaltung.* Berlin: Cornelsen Scriptor.

Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit.* Münster: Waxmann.

Eder, F. (1996). *Schul- und Klassenklima. Ausprägung, Determinanten und Wirkungen des Klimas an höheren Schulen.* Innsbruck: Studien Verlag.

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.

Fend, H. (1998). *Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistungen.* Weinheim/München: Juventa.

Fischer, J., Praetorius, A.-K., & Klieme, E. (2019). *The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality. Educational Assessment, Evaluation and Accountability.* https://www.springerprofessional.de/the-impact-of-linguistic-similarity-on-cross-cultural-comparabil/16658412 [24. 09. 2019].

Gage, N. L. (2009). *A conception of teaching.* New York: Springer Science and + Business Media.

Gitomer, D. H., & Bell, C. A. (2016) (Eds.). *Handbook of research on teaching* (5th ed.). Washington: American Educational Research Association.

Gruehn, S. (2000). *Unterricht und schulisches Lernen. Schüler als Quellen der Unterrichtsbeschreibung.* Münster: Waxmann.

Hammersley, M. (2000). Varieties of social research. *International Journal of Social Research Methodology, 3*(3), 221–229.

Helm, C. (2016). Zentrale Qualitätsdimensionen von Unterricht und ihre Effekte auf Schüleroutcomes im Fach Rechnungswesen. *Zeitschrift für Bildungsforschung, 6*(2), 101–119.

Hill, J. R., & Schrum, L. (2002). Theories on teaching: Why are they so hard to find?!? *TechTrends, 46*(5), 22–26.

Kane, R., & Marsh, C. J. (1980). Progress toward a general theory of instruction? *Educational Leadership,* 253–255.

Kerlinger, F. N. (1964). *Foundations of behavioral research: Educational and psychological inquiry.* New York: Holt, Rinehart and Winston.

Klieme, E. (2012, August). Qualities and effects of teaching. Integrating findings across subjects and cultures. *Keynote lecture at the EARLI SIG 18 conference,* Zurich.

Klieme, E. (2019). Unterrichtsqualität. In M. Gläser-Zikuda, M. Harring & C. Rohlfs (2018). *Handbuch Schulpädagogik* (S. 393–408). Münster: Waxmann.

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

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). Münster: Waxmann.

Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann & M. Weiss (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 333–359). Wiesbaden: VS Verlag für Sozialwissenschaften.

Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik: Outcomeorientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik, 54*(2), 222–227.

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. Forschunsbefunde, Refominitiativen, Prasixberichte und Video-Dokumente* (S. 43–57). Bonn: Bundesministerium für Bildung und Forschung.

Korneck, F., Krüger, M., & Szogs, M. (2017). Professionswissen, Lehrerüberzeugungen und Unterrichtsqualität angehender Physiklehrkräfte unterschiedlicher Schulformen. In E. Sumfleth & H. Fischler (Hrsg.), *Professionelle Kompetenzen von Lehrkräften der Chemie und Physik. Studien zum Physik- und Chemielernen Bd. 200* (S. 113–133). Berlin: Logos.

Krammer, K., Ratzka, N., Klieme, E., Lipowsky, F., Pauli, C., & Reusser, K. (2006). Learning with classroom videos: Conception and first results of an online teacher-training program. *ZDM Mathematics Education, 38*(5), 422–432.

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

Kuhn, T.S. (1969/1997). *Die Struktur wissenschaftlicher Revolutionen* (14. Aufl.). Frankfurt a.M.: Suhrkamp.

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). Münster: Waxmann.

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

Kunter, M., & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 85–113). Münster: Waxmann.

Lakatos, I. (1977). *The methodology of scientific research programmes: Philosophical papers Volume 1.* Cambridge: Cambridge University Press.

Lenske, G., Wagner, W., Wirth, J., Thillmann, H., Cauet, E., Liepertz, S., & Leutner, S. (2016). Die Bedeutung des pädagogisch-psychologischen Wissens für die Qualität der Klassenführung und den Lernzuwachs der Schüler/innen im Physikunterricht. *Zeitschrift für Erziehungswissenschaft, 19,* 211–233.

Leplin, J. (1980). The role of models in theory construction. In T. Nickles (Ed.), *Scientific discovery, logic, and rationality* (pp. 267–283). Dordrecht: Springer.

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.

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

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. Grundlagen der Qualität von Schule* (Bd. 3, S. 219–249). Münster: Waxmann.

Lüders, M. (2014). Erziehungswissenschaftliche Unterrichtstheorien. *Zeitschrift für Pädagogik, 60*(6), 832–849.

Marzano, R. J., & Marzano, J. S. (2003). The key to classroom management. *Educational Leadership, 61*(1), 6–13.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art – teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*(2), 231–256.

Nilsen, T., & Gustafsson, J.-E. (Eds.) (2016). *Teacher quality, instructional quality and student outcomes*. Heidelberg: Springer.

Oelkers, (2000). Anmerkungen zur Reflexion von "Unterricht" in der deutschsprachigen Pädagogik des 20. Jahrhundert. In D. Benner & H. E. Tenorth (Hrsg.), *Bildungsprozesse und Erziehungsverhältnisse im 20. Jahrhundert* (42. Beiheft der Zeitschrift für Pädagogik S. 166–185). Weinheim: Beltz.

Open Science Collaboration (2017). Maximizing the reproducibility of your research. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 1–21). New York: Wiley.

Openshaw, K., & Clarke, S. C. T. (1970). General teaching theory. *Journal of Teacher Education, 21*(3), 403–416.

Palincsar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review of Psychology, 49,* 345–375.

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

Popper, K. R. (2005). *Logik der Forschung. Gesammelte Werke* (Band 3). Tübingen: Mohr Siebeck.

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

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM Mathematics Education, 50*(3), 407–426.

Praetorius, A.-K., Lauermann, F., Klassen, R. M., Dickhäuser, O., Janke, S., & Dresel, M. (2017). Longitudinal relations between teaching-related motivations and student-reported teaching quality. *Teaching and Teacher Education, 65,* 241–254.

Praetorius, A.-K., & Charalambous, C. Y. (2018) Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM Mathematics Education, 50*(3), 535–553.

Prange, K. A. (2005). Didaktik. In J. Kade, W. Helsper, C. Lüders, B. Egloff, F.-O. Radtke & W. Thole (Hrsg.), *Pädagogisches Wissen* (S. 183–188). Stuttgart: Kohlhammer.

Puntambekar, S., & Hübscher, R. (2005). Tools for scaffolding students in a complex environment: What have we gained and what have we missed? *Educational Psychologist, 40,* 1–12.

Rakoczy, K., Klieme, E., Lipowsky, F., & Drollinger-Vetter, B. (2010). Strukturierung, kognitive Aktivität und Leistungsentwicklung im Mathematikunterricht. *Unterrichtswissenschaft, 38*(3), 229–246.

Rakoczy, K., & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". Teil 3: Videoanalysen. Materialien zur Bildungsforschung* (S. 206–233). Frankfurt a. M.: GFPF.

Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., & Besser, M. (2019). Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy. *Learning and Instruction, 60,* 154–165.

Reckwitz, A. (2002) Toward a theory of social practices. A development in culturalist theorizing. *European Journal of Social Theory, 52,* 245–265.

Reiss, J., & Sprenger, J. (2017). *Scientific objectivity. The Stanford encyclopedia of philosophy* (Winter 2017 Edition). https://plato.stanford.edu/archives/win2017/entries/scientific-objectivity/ [24.09.2019].

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. In R.M.W. Travers (Eds.), *Second handbook of research on teaching* (pp. 122–183). Chicago: Rand McNally.

Rosenshine, B. (2009). A conception of teaching. Review of N.L. Gage. Springer, Norwell, MA (2009). *Teaching and Teacher Education, 25,* 1169–1171.

Ryan, R.M., & Deci, E.L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and wellbeing. *American Psychologist, 55*(1), 68–78.

Scriven, M. (1994). A possible distinction between traditional scientific disciplines and the study of human behavior. In M. Martin & L.C. McIntyre (Eds.), *Readings in the philosophy of social science* (pp. 71–78). Cambridge: The MIT Press.

Seidel, T. (2014). Angebots-Nutzungs-Modelle in der Unterrichtspsychologie: Integration von Struktur- und Prozessparadigma. *Zeitschrift für Pädagogik, 60*(6), 850–866.

Seidel, T., Prenzel, M., & Krapp, A. (2014). Grundlagen der Pädagogischen Psychologie. In T. Seidel & A. Krapp (Hrsg.), *Pädagogische Psychologie* (S. 21–36). Weinheim: Beltz.

Snow, R.E. (1973). Theory construction for research on teaching. In R.M.W. Travers (Ed.), *Second handbook of research on teaching* (pp. 77–112). Chicago, IL: Rand McNally.

Stigler, J.W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* New York: Free Press.

Stigler, J.W., & Knoll, S. (1999). The TIMSS Videotape Classroom Study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States. *Education Statistics Quarterly, 1,* 109–112.

Sutton, R.I., & Staw, B.M. (1995). What theory is not. *Administrative Science Quarterly, 40*(3), 371–384.

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction, 46,* 45–60.

Terhart, E. (2014). Unterrichtstheorie: Einführung in den Thementeil. *Zeitschrift für Pädagogik, 60*(6), 813–816.

Thagard, P.R. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy, 75*(2), 76–92.

Wagenschein, M. (1992). *Verstehen lehren. Genetisch – Sokratisch – Exemplarisch* (9. ed.). Weinheim: Beltz.

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.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik. Ein Lehrbuch zur Psychologischen Methodenlehre.* Göttingen: Hogrefe.

Whetten, D.A. (1989). What constitutes a theoretical contribution? *Academy of Management Review, 14*(4), 490–495.

Zusammenfassung: Der vorliegende Beitrag fokussiert auf die Bedeutung von Theorien zu Unterricht und dessen Qualität für die quantitative empirische Unterrichtsforschung. Dabei wird zunächst der quantitative empirische Forschungsansatz vorgestellt. Anschließend stellen wir die Herkunft und den aktuellen Status eines populären Modells in der deutschsprachigen quantitativen Unterrichtsforschung dar, dem Modell der drei Basisdimensionen von Unterrichtsqualität. Es folgt eine Reflexion in welchem Ausmaß dieses Modell Kriterien guter Theorien erfüllt. Abschließend werden Schlussfolgerungen für zukünftige Forschung mit einem Fokus auf die Entwicklung von Theorien gezogen.

Schlagworte: Theorie, Unterrichtsqualität, drei Basisdimensionen, Unterricht, Modell

Contact

Prof. Dr. Anna-Katharina Praetorius, Universität Zürich,
Lehrstuhl für pädagogisch-psychologische Lehr-Lernforschung und Didaktik,
Institut für Erziehungswissenschaft,
Freiestrasse 36, 8032 Zürich, Schweiz
E-Mail: anna.praetorius@ife.uzh.ch

Prof. Dr. Eckhard Klieme, DIPF | Leibniz-Institut für Bildungsforschung
und Bildungsinformation,
Rostocker Straße 6, 60323 Frankfurt a. M., Deutschland
E-Mail: klieme@dipf.de

Prof. Dr. Thilo Kleickmann, Christian-Albrechts-Universität zu Kiel,
Institut für Pädagogik, Abteilung Schulpädagogik,
Olshausenstr. 75, 24118 Kiel, Deutschland
E-Mail: kleickmann@paedagogik.uni-kiel.de

Prof. Dr. Esther Brunner, Pädagogische Hochschule Thurgau
Unterer Schulweg 3, 8280 Kreuzlingen, Schweiz
E-Mail: esther.brunner@phtg.ch

Prof. Dr. Anke Lindmeier, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften
und Mathematik,
Olshausenstr. 62, 24118 Kiel, Deutschland
E-Mail: lindmeier@ipn.uni-kiel.de

Prof. Dr. Sandy Taut, Bayerisches Landesamt für Schule,
Qualitätsagentur,
Stuttgarter Straße 1, 91710 Gunzenhausen, Deutschland
E-Mail: sandy.taut@isb.bayern.de

Assistant Prof. Dr. Charalambos Charalambous, University of Cyprus,
Department of Education,
Theophanides Building,11 – 13 Dramas Str., 1077 Nicosia, Cyprus
E-Mail: cycharal@ucy.ac.cy

Thilo Kleickmann/Mirjam Steffensky/Anna-Katharina Praetorius

Quality of Teaching in Science Education

More Than Three Basic Dimensions?

Abstract: The three basic dimensions framework for assessing quality teaching distinguishes the dimensions of cognitive activation, student support, and classroom management. Research from various disciplines suggests, however, that cognitive support is not sufficiently represented in the framework as yet. In the present study, two-level factor analyses based on student ratings of teaching quality in science education (2,659 students, grades 4 and 6) suggest four dimensions of teaching quality, with cognitive support being a separate dimension. Moreover, cognitive support predicted student achievement. The results suggest that including cognitive support as a separate dimension contributes to a more comprehensive, yet parsimonious framework of teaching quality.

Keywords: Quality of Teaching, Three Basic Dimensions Framework, Student Ratings, Science Education, Factor Analysis

1. Introduction

The basic dimensions framework a generic framework used to describe teaching quality (e. g., Klieme, Schümer & Knoll, 2001; Kunter & Voss, 2013; Praetorius, Klieme, Herbert & Pinger, 2018b) that is particularly prominent in German-speaking countries. In the basic dimensions approach, teaching quality is considered as consisting of three basic dimensions: cognitive activation, student support, and classroom management. Cognitive activation aims to involve students in higher order thinking processes and to engage them in knowledge construction and revision (e. g., through challenging tasks and exploring prior knowledge). Student support, which originated from research on classroom climate, primarily aims to foster student motivation (e. g., by supporting experiences of autonomy and social relatedness). Classroom management aims to organize the complex and dynamic teaching situation and, ultimately, to use instructional time in a productive way (e. g., through monitoring, clear rules and routines, and other strategies to prevent disruptions).

In the basic dimensions framework, cognitive activation and classroom management are considered to foster student achievement, while student support and classroom management likewise are supposed to stimulate student motivation (Klieme & Rakoczy, 2008; Kunter & Voss, 2013; Praetorius et al., 2018b). However, evidence on the predictive validity of the three basic dimensions is mixed. While some studies have shown that basic dimensions predict student learning and motivation in the hypothesized way, others did not (Praetorius et al., 2018b).

The three basic dimensions are clearly a parsimonious framework, as they reduce the complexities of high quality teaching to three core dimensions. However, is this framework sufficiently comprehensive? In this article, we argue that another feature of quality teaching – cognitive support – is not sufficiently represented in the basic dimensions framework. Cognitive support aims to reduce cognitive demands in challenging learning environments, so that students can master them (e. g., Kirschner, Sweller & Clark, 2006; Puntambekar & Hübscher, 2005). We assume that cognitive support represents a separate dimension of quality teaching. Moreover, we assume that prediction of student progress becomes more stringent – theoretically and empirically – when cognitive support is included as a separate dimension in the framework.

2. Cognitive Support as an Important Feature of Teaching Quality

Research from different disciplines highlights the role of cognitive support provided by teachers when students are involved in challenging learning environments. Social-constructivist theories emphasize the role of scaffolding student learning (Puntambekar & Hübscher, 2005; van de Pol, Volman & Beishuizen, 2010). While the notion of scaffolding originally referred to the adaptive support of students during student-teacher interactions, based on on-going diagnosis of the individual learner's progression, later versions included "blanket" scaffolding, where support is the same for all students and can relate to larger instructional units (e. g., through clarity of goals or conceptual coherence of the content presented; for an overview, see Puntambekar & Hübscher, 2005; van de Pol et al., 2010).

Similarly, theories from cognitive psychology, and cognitive load theory in particular, point to the critical role of guidance in complex learning environments. They suggest that learners' working memory capacities clearly restrict unguided learning (e. g., Kirschner et al., 2006). Therefore, teachers need to reduce the complexity of a learning environment through cognitive support by, for example, modeling, explaining, and structuring. In this tradition, the bulk of evidence demonstrates that learners, and novice learners in particular, should be provided with cognitive support while learning the concepts and procedures of a particular domain, and should not be left to discover those concepts or procedures by themselves (e. g., Alfieri, Brooks, Aldrich & Tenenbaum, 2011; Kirschner et al., 2006).

Those theories have influenced research in various content domains – for instance, in science education. In this domain, the concept of guided inquiry underscores the role of guidance, structure, and focused goals in reducing the complexity of the inquiry process (e. g., Hardy, Jonen, Möller & Stern, 2006; Steffensky, Gold, Holodynski & Möller, 2015).

In sum, cognitive support aims to reduce complexity and cognitive demands by means of structuring content and promoting clarity, so that students can master the respective tasks and successfully gain understanding. Cognitive support includes both adjusted (also referred to as differentiated or calibrated) support during individual stu-

dent-teacher interactions (e. g., by modeling, explaining, highlighting, giving analogies, and informative feedback) and blanket support, which latter is the same for groups of students or the entire class. Ultimately, blanket cognitive support includes structuring and clarity in larger instructional units (e. g., clarity of goals, coherence of the content covered in relation to student activities, reduction of task difficulty, visualizations and representations used in instructional materials; Puntambekar & Hübscher, 2007). Cognitive support can be realized in different classroom settings, such as teacher-centered or student-centered settings, or individualized teaching (Hardy et al., 2006; van de Pol et al., 2010).

The need to adapt support to learners' prerequisites (as highlighted in the first component of cognitive support) is also evinced in the concept of adaptive teaching (e. g., Hardy et al., 2011). However, cognitive support does not equal adaptive teaching. First, cognitive support also includes blanket scaffolding, and second, adaptive teaching is not relevant to cognitive support exclusively, but also to other dimensions of teaching quality, such as cognitive activation and motivational support (e. g., Kyriakides, Creemers & Panayiotou, 2018).

3. Cognitive Support and Basic Dimensions of Teaching Quality

Although the basic dimensions framework is quite commonly used in instructional research in German-speaking countries, there is no common operationalization of the three dimensions (Praetorius et al., 2018b). Accordingly, cognitive support has been considered differently across studies. Roughly four types how cognitive support has been considered can be distinguished.

In the first type (type-1 operationalization), cognitive support is not or only rudimentarily considered in the operationalization of the three basic dimensions: If considered at all, it is only rudimentarily included in student support, which is primarily focused on a supportive climate providing students with experiences of social relatedness and autonomy. Examples of this type can be found in the studies of Decristan et al. (2015), Fauth, Decristan, Rieser, Klieme and Büttner (2014), and Helm (2016).

In the second type (type-2 operationalization), studies consider cognitive support as part of student support. In such studies, student support includes both cognitive and motivational support. While cognitive support refers to the reduction of cognitive demands through structuring, as noted in the previous section, motivational support refers to support of autonomy and social relatedness. In contrast to the first type, features of cognitive support are included in the assessment of student support more comprehensively. Examples of this second type are the studies of Hochweber and Vieluf (2016), Klieme and Rakoczy (2008), Kunter and Voss (2013), and Praetorius, Lenske, and Helmke (2012).

In the third type (type-3 operationalization), specific aspects of cognitive support are included in the classroom management dimension. In particular, specific aspects of lesson clarity or structure (e. g., excursiveness [recoded]) are included as features of class-

room management. The studies of Klieme et al. (2001) as well as Taut and Rakoczy (2016) are examples of this type.

The fourth type (type-4 operationalization), considers cognitive activation and cognitive support as highly interconnected aspects of one dimension of quality teaching. Cognitive structuring (Einsiedler & Hardy, 2010) and instructional support (Pianta & Hamre, 2009) are examples of such merged dimensions. Instructional support comprises key features of cognitive support (e. g., clear presentation of material, quality of feedback), but also of cognitive activation (e. g. fostering higher-level thinking, provision of engaging lessons and materials; Pianta & Hamre, 2009).

In sum, type one and three do not represent the construct of cognitive support in a comprehensive way. They include only parts of the construct, at best. Type two does consider cognitive support. However, in this type, the hypothesized way in which the three basic dimensions predict student learning and motivation is not straightforward from a theoretical point of view. In particular, it does not take into account the important role of cognitive support in student learning, because student support is supposed to foster student motivation only (Praetorius et al., 2018b). Explicitly differentiating between cognitive and motivational support would allow the setting up of differential hypotheses of student progress, with cognitive support predicting student learning and motivational support predicting student motivation. Type four points to the interconnectedness of cognitive activation and cognitive support.

4. The Present Study

In the present study, we used student ratings to assess teaching quality in upper elementary and lower secondary science education. Recent studies have demonstrated that elementary school children can already differentiate between cognitive activation, supportive climate (with a focus on motivational support), and classroom management in science lessons (Decristan et al., 2015; Fauth et al., 2014). In the present study, we included items explicitly designed to assess cognitive support, to test our assumption of four basic dimensions, with cognitive support being a separate dimension of teaching quality.

First, we investigated the factor structure of student ratings of teaching quality, aiming thereby to test whether a four-factor model including the dimensions of cognitive activation, cognitive support, motivational support, and classroom management fits the data better than alternative, more parsimonious models that include cognitive support in a common factor with motivational support (as suggested in the aforementioned type-2 operationalization), classroom management (type-3 operationalization), or cognitive activation (type-4 operationalization).

Second, we investigated the predictive validity of the hypothesized four dimensions of high quality teaching for students' conceptual understanding of the water cycle as well as for students' interest. We hypothesized that cognitive activation, cognitive support, and classroom management predict student understanding, while motivational support and classroom management predict student interest.

5. Method

5.1 Participants and Design

Our analyses were based on data from a study on science education in German fourth- and sixth-grade classes (Kauertz et al., 2011). The sample used to test the factor structure consisted of 60 fourth-grade classrooms with 1,326 students (age: $M = 10.3$ years, $SD = 0.6$ years; 53 % male) and 54 sixth-grade classrooms with 1,333 students (age: $M = 12.2$ years, $SD = .7$ years; 54 % male). While grade 4 is comprehensive schooling in Germany, the sixth-grade subsample included 28 classes from non-academic track schools (Hauptschule) and 26 classes from academic track schools (Gymnasium). All classes were located in North Rhine-Westphalia in the west of Germany.

For research question 1 (factor structure), we used the full data set of 2,659 students in 114 classrooms. For research question 2 (predictive validity), we used the subsample of 60 classrooms with 1,326 fourth-graders, as valid information on student achievement and interest were available for this subsample.

Students rated teaching quality at the end of the teaching unit on the topic of the water cycle. All participating 114 teachers taught this topic in their science classes. States of matter, condensation, evaporation, and conditions affecting these processes were the key physics concepts addressed in this unit. The topic of the water cycle, and the respective physics concepts, are valid for the science curricula of grades 4 and 6 in North Rhine-Westphalia.

Students' understanding of the water cycle and related physics concepts was assessed using repeated measures directly before and after the teaching unit. Students' interest in the teaching unit was also assessed directly after the teaching unit. As a covariate, students' interest in physics topics was assessed before the teaching unit.

5.2 Measures

Students rated teaching quality along a set of 20 items assigned to four scales: cognitive activation (five items), cognitive support (five items), motivational support (five items), and classroom management (five items; see Table A1 in the Appendix for item wording). Cognitive activation included provoking cognitive conflict, testing of hypotheses, justification of beliefs, and application of newly constructed knowledge in everyday situations. Two items, in contrast to the commonly used generic items in previous studies (e. g. Fauth et al., 2014), specifically addressed cognitive activation in the context of science inquiry. Cognitive support comprised clarity of goals and procedures, highlighting of important aspects, adequate reduction of complexity, and the absence of incomprehensible terms.

Motivational support included teacher sensitivity to student problems, positive feedback, and autonomy support. As three items addressed teacher sensitivity, further aspects of motivational support, such as support of student autonomy, were addressed to

a small degree only. Classroom management was operationalized by the absence of disruptions and no wasting of time. Consequently, this measure did not assess teacher actions to prevent disruptions (for a similar operationalization see Fauth et al., 2014; for an operationalization focusing on teacher actions see Kunter, Baumert & Köller, 2007). All items were rated on a four-point scale ranging from "strongly disagree" (1) to "strongly agree" (4). Hence, the scale mean was 2.5.

In respect of students rating the quality of teaching, variance within classes can be distinguished from variance between classes. In the present study, we were interested in the shared perceptions of students reflected in between-class variance. Intra-class correlations indicated substantial variance between classes *(ICC1, see Tab. 1)* and the reliability of the class-level measures *(ICC2):* For cognitive activation, cognitive support, motivational support, and classroom management *ICC2* were .80, .80, .87, and .84 respectively. Cronbach's alphas were .71, .61, .83, and .83 respectively.

Student understanding of the water cycle was assessed by a test consisting of 26 multiple-choice items covering physics concepts such as states of matter, condensation, evaporation, and conditions affecting related processes. The distractors included the typical alternative conceptions of students, derived from literature on student conceptions (e. g., Amin, Smith & Wiser, 2017). Students' pre- and posttest scores (WLE scores) were scaled concurrently using the Rasch model. EAP/PV-reliability was .74 at pretest and .82 at posttest. The values for *ICC2* were .78 and .85 for pre- and posttest respectively.

Student interest in the teaching unit was measured using a scale of six items (sample item: I always looked forward to the lessons on the water cycle). Cronbach's alpha and *ICC2* were .82 and .81 respectively. We used a slightly modified version of the scale to obtain a covariate of students' prior interest. In this scale, the items did not refer to the topic of the water cycle, as this had not been taught yet, but more generally to physics topics in elementary science education. The Cronbach's alpha and *ICC2* of that scale were .80 and .81 respectively.

The wording of all items for teaching quality, student understanding, and student interest was simple, and all items were read out loud by a trained data collector in order to minimize language and reading problems.

In the analyses on the prediction of student outcomes, we included measures of cognitive ability using the CFT 20-R (Cronbach's alpha =.72; Weiß, 2006), a German version of the Culture Fair Intelligence Tests, and socio-economic status was operationalized by the sum of the International Socio-Economic Index assigned to father and mother (Ganzeboom, de Graaf & Treiman, 1992) as covariates.

5.3 Data Analyses

We used two-level confirmatory factor analyses (CFA) to test our assumptions on the factor structure of student ratings of teaching quality. In particular, we used a doubly-latent model (see Figure 1) according to the framework suggested by Marsh et al. (2009).

Our model comparisons were based on several goodness-of-fit indices (Hu & Bentler, 1998), such as the Akaike Information Criterion (AIC). For all models, we specified cross-level invariant factor loadings. In order to control for school type differences, we included two dummy variables as between-level covariates in the CFA models.[1]

In order to investigate the prediction of student outcomes, we used two-level regression analyses. In these analyses, we used the manifest-latent approach (Marsh et al., 2009), with latent aggregation of within-level constructs to between-level constructs, to model student ratings of teaching quality. We introduced students' general cognitive abilities (CFT), socio-economic status (SES), and gender as manifest covariates on the within-class level. Student understanding and interest (pre- and posttest scores) were also included, using the manifest-latent approach. On the between level, posttest student understanding respectively interest were regressed on teaching quality, pretest understanding respectively interest. We conducted separate models for student understanding and interest, as well as for the dimensions of teaching quality. We used full information maximum likelihood (FIML) estimation with robust standard errors implemented in M*plus* (Muthén & Muthén, 1998–2012) to deal with missing data.

6. Results

6.1 Descriptive Results

Table 1 shows descriptive results of the study variables on the between level. The means for cognitive activation, cognitive support, and motivational support were high (M = 3.16, 3.12, and 3.20) and those for classroom management medium (M = 2.61). The correlation between cognitive activation and motivational support was particularly high (r = .82), while the other correlations were low to medium. The proportion of variance between classes was descriptively higher for motivational support and classroom management ($ICC1$ = .24 and .22 respectively) compared to cognitive activation and cognitive support ($ICC1$ = .15 and .16).

The mean values for the achievement test indicate that the items were rather difficult for elementary school children (M = 6.61 at pretest and 9.64 at posttest; theoretical maximum = 22). However, the test differentiated sufficiently between classes ($ICC1$ = .15 and .22 for pre- and posttest).

1 Following the recommendation of a reviewer, we did not include further covariates in the CFA models.

	2	3	4	5	6	7	8	9	10	11	M	SD	ICC1
1. Cognitive activation (1–4)	.38	.82	.19	−.01	−.03	.38	.68	−.11	−.13	.03	3.16	0.28	0.15
2. Cognitive support (1–4)		.50	.41	.34	.49	.46	.58	.02	.34	.45	3.12	0.28	0.16
3. Motivational support (1–4)			.18	.18	.04	.36	.72	.02	−.04	.08	3.20	0.40	0.24
4. Classroom management (1–4)				.15	.35	.20	.25	.09	.19	.38	2.61	0.39	0.22
5. Achievement, T1 (0–24)					.60	.28	.11	−.05	.41	.35	6.61	1.30	0.14
6. Achievement, T2 (0–24)						.24	.11	−.12	.35	.45	9.64	1.89	0.20
7. Interest, T1 (1–4)							.70	.04	.24	.12	3.26	0.28	0.16
8. Interest, T2 (1–4)								.07	.12	.04	3.14	0.39	0.16
9. Gender (male = 1, female = 0)									−.04	−.01	0.53	0.09	0.00
10. SES (mother and father; 32–180)										.23	69.06	13.25	0.07
11. Cognitive ability (0–30)											14.93	1.17	0.06

Note. Statistics are based on the original metric of the variables (scale range in brackets). The intra-class correlation (ICC1) indicates the proportion of variance between classes.

Tab. 1: Descriptive statistics for study variables on the between level: Zero order correlations, means, standard deviations, and intra-class correlations

6.2 Research Question 1: Factor Structure

Research question 1 addressed the factor structure of student ratings of teaching quality in the units on the water cycle. To test our hypotheses, we compared the fit of five CFA models (Tab. 2): Model 1 featured four factors (cognitive activation, cognitive support, motivational support, and classroom management) on each level. This was the model we expected to show the best model fit. Model 2 featured three factors and represented the second type of basic dimensions framework, as outlined above. In this model, cognitive and motivational support formed one common factor ("support-factor"). Model 3 also featured three factors, and represented the third type of basic dimensions framework. In this model, cognitive support and classroom management formed one common factor. In Model 4, another three-factor model, cognitive activation and cognitive support formed one common factor, as suggested in the type-4 operationalization. Due to the high zero-order correlation between cognitive activation and motivational support on the between level, we additionally tested another three-factor model variant: In this model, cognitive activation and motivational support formed one common factor

Model	Features	AIC	BIC	RMSEA	CFI	TLI	SRMR within	SRMR between
1	4 factors	**122434**	**123045**	**0.03**	**0.91**	**0.90**	**0.05**	**0.11**
2	3 factors (CS and MS merged)	123140	123699	0.04	0.86	0.85	0.06	0.14
3	3 factors (CS and CM merged)	123599	124157	0.05	0.82	0.81	0.08	0.19
4	3 factors (CA and CS merged)	123172	123731	0.04	0.85	0.84	0.06	0.12
5	3 factors (CA and MS merged)	123102	123660	0.04	0.86	0.85	**0.05**	0.12
6	1 factor	127464	127952	0.08	0.53	0.51	0.11	0.21

Note. CA = cognitive activation, CS = cognitive support, MS = motivational support. Values indicating the best model fit are in bold type.

Tab. 2: Results of the two-level confirmatory factor analyses: model-fit indices

(Model 5). Model 6, finally, included only one global factor for teaching quality. Table 1 shows the model fit indices for the six models. All fit indices consistently supported Model 1.

Standardized factor loadings on the between level in the best fitting model (model 1) ranged from $\lambda = .69$ to .92 for cognitive activation, from $\lambda = .66$ to .94 for cognitive support, from $\lambda = .81$ to 1.00 for motivational support, and from $\lambda = .86$ to .98 for classroom management. Figure 1 shows that factor correlations on the between level were medium-sized to high.

The latent correlations between the four factors on the between level indicated that cognitive activation and motivational support were particularly highly correlated ($r = .84$) while the other correlations were of medium size ($r = .39$ to .54).

6.3 Research Question 2: Prediction of Student Understanding and Interest

Our second research question addressed the predictive validity of the four factors of teaching quality. Table 3 shows the results for the prediction of students' conceptual understanding of the water cycle. Models 1–4 included only one dimension of teaching quality, whereas model 5 incorporated all four dimensions simultaneously. The results from models 1–4 showed that cognitive support and classroom management predicted student understanding, while cognitive activation and motivational support did not. In model 5, only cognitive support was significantly related to student understanding. The predictors in model 5 explained 54% of between-level variance in students' posttest understanding.

Table 4 shows the respective results for prediction of student interest. In the models 1–4, cognitive activation, cognitive support, and motivational support predicted stu-

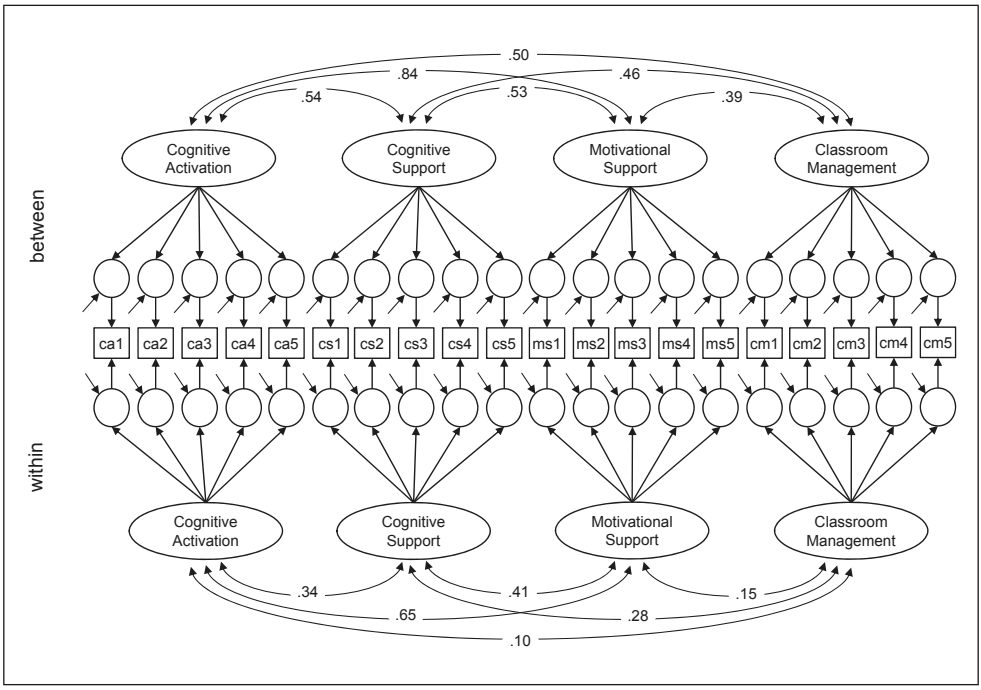

Fig. 1: Two-level confirmatory factor analysis model featuring each of four factors on the within and the between levels. Factor loadings are constrained to be equal across levels. On the between level, two dummy variables coding the type of school were included as covariates.

dent interest, but classroom management did not. In model 5, none of the dimensions of teaching quality was significantly related to student interest. The predictors in model 5 explained 86 % of between-level variance in students' posttest interest.

	M0		M1		M2		M3		M4		M5	
	B	**SE**	**B**	**SE**	**B**	**SE**	**B**	**SE**	**B**	**SE**	**B**	**SE**
Within												
Understanding, pretest	**.54**	.03	**.54**	.03	**.53**	.03	**.54**	.03	**.54**	.03	**.53**	.03
Cognitive ability	**.13**	.03	**.13**	.03	**.11**	.03	**.12**	.03	**.12**	.03	**.11**	.03
SES	**.07**	.03	**.07**	.03	**.06**	.03	**.07**	.03	**.07**	.03	**.06**	.03
Gender (male)	−.02	.03	−.02	.03	.00	.02	−.02	.02	−.02	.03	.00	.02
Cognitive activation			.01	.03							.01	.03
Cognitive support					**.10**	.03					**.13**	.03
Motivational support							−.02	.04			−.07	.04
Classroom management									.02	.02	.00	.02
R^2	*.37*		*.37*		*.38*		*.37*		*.37*		*.38*	
Between												
Understanding, pretest	**.53**	.10	**.53**	.10	**.44**	.12	**.55**	.10	**.52**	.11	**.43**	.16
Cognitive activation			−.04	.14							.04	.38
Cognitive support					**.35**	.15					**.59**	.18
Motivational support							−.07	.13			−.48	.36
Classroom management									**.30**	.15	.13	.16
R^2	*.29*		*.29*		*.40*		*.29*		*.38*		*.54*	

Note. Significant coefficients ($p < .05$) in bold.

Tab. 3: Results of two-level regression analyses predicting students' conceptual understanding of the water cycle and related physics concepts

	M0		M1		M2		M3		M4		M5	
	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE
Within												
Interest, pretest	**.39**	.03	**.31**	.03	**.33**	.03	**.31**	.03	**.38**	.03	**.26**	.03
Cognitive ability	.01	.02	.01	.02	−.04	.03	.02	.02	.00	.02	−.01	.02
SES	−.02	.03	.00	.03	−.04	.03	−.02	.03	−.02	.03	−.04	.03
Gender (male)	**−.07**	.03	−.04	.03	−.03	.03	−.04	.03	**−.07**	.03	−.02	.03
Cognitive activation			**.32**	.03							**.14**	.03
Cognitive support					**.29**	.03					**.20**	.03
Motivational support							**.40**	.03			**.26**	.03
Classroom management									**.09**	.03	.02	.03
R^2	*.16*		*.26*		*.23*		*.31*		*.17*		*.36*	
Between												
Interest, pretest	**.77**	.07	**.54**	.10	**.57**	.10	**.56**	.10	**.75**	.08	**.50**	.10
Cognitive activation			**.51**	.11							.17	.29
Cognitive support					**.37**	.11					.14	.13
Motivational support							**.55**	.11			.34	.31
Classroom management									.08	.11	−.02	.10
R^2	*.60*		*.81*		*.70*		*.85*		*.60*		*.86*	

Note. Significant coefficients ($p < .05$) in bold.

Tab 4: Results of two-level regression analyses predicting students' interest in the teaching unit on the water cycle

7. Discussion

Research from different disciplines suggested that cognitive support plays a crucial role in high quality teaching. As previous conceptions of the basic dimensions framework represented cognitive support insufficiently, or included it in other dimensions (see operationalization types 2, 3, and 4 as outlined above), we tested a modification of the original framework in the present study. In particular, we suggested a four-dimensional framework, with cognitive support forming a factor separate from cognitive activation, motivational support, and classroom management. We tested this framework in relation to its factor structure and the prediction of student outcomes in a teaching unit on the topic of the water cycle in the domain of science education.

Our factor-analytic results supported the notion that cognitive activation, cognitive support, motivational support, and classroom management form four separate, yet cor-

related dimensions of student perceived teaching quality in science education (for similar conceptions of teaching quality, see Korneck, Krüger & Szogs, 2017; Lipowsky, 2015). Alternative models featuring only three factors yielded poorer model fit. These models included models representing the second, third, and fourth type of the basic dimensions framework noted in the introduction. In particular, a model with only one factor representing general teaching quality, showed insufficient model fit with regard to common threshold values. The medium to high factor correlations might (also) reflect functional dependencies among the dimensions (e. g., classroom management as a prerequisite for effective classroom teaching).

The results on the predictive validity of the four dimensions of teaching quality were only partly in line with our hypotheses. The two-level regression models including the dimensions of teaching quality separately as predictors, showed that cognitive support and classroom management predicted student understanding as hypothesized, but cognitive activation did not. Considering the four dimensions of teaching quality simultaneously, only cognitive support was significantly related to student understanding. This finding underscores the importance of cognitive support as rated by students, for student learning. Concerning student interest, the models including the dimensions of teaching quality separately showed that cognitive activation, cognitive support, and motivational support were predictive, but classroom management was not. Considering the four dimensions of teaching quality simultaneously, none of the dimensions of teaching quality was significantly related to student interest. This finding suggests that cognitive activation, cognitive support, and motivational support share common variance with student interest and lack incremental predictive validity.

The result that student perceived cognitive activation did not predict student learning as well as the very high correlation between cognitive activation and motivational support has also been found in other studies (e. g., Fauth et al., 2014). This pattern may challenge the assumption that student ratings – or at least young children's ratings – are suitable for measuring cognitive activation: Students might not be sufficiently capable of separating cognitive activation from their own abilities. Cognitive interviews (e. g., Lenske, 2016) might be one way of determining to what extent this is the case.

Aside from the non-expected results concerning cognitive activation and classroom management, cognitive and motivational support yielded the hypothesized pattern of results: while cognitive support predicted student achievement (and interest), motivational support predicted student interest only. Thus, the separate assessment and modeling of the two dimensions might be promising for future research based on student ratings.

While previous studies using student ratings to assess teaching quality did not find significant effects of student support on student achievement gains (e. g., Fauth et al., 2014; Pinger, Rakoczy, Besser & Klieme, 2017), cognitive support was significantly related to student understanding in the present study, and indeed, was the strongest predictor of student understanding among the four dimensions of teaching quality. These results underscore the relevance of cognitive support as rated by students.

In essence, cognitive support serves to reduce cognitive demands in challenging, cognitively activating learning environments so that students can master them. Future

research might address the specific roles of adaptive support during student-teacher interactions, and also blanket scaffolding, where support is the same for all students and can refer to larger instructional units (Puntambekar & Hübscher, 2005). Moreover, the presumably different mediating mechanisms through which cognitive and motivational support, as well as classroom management, affect student outcomes would be a highly relevant field for future research (Praetorius et al., 2018b).

8. Limitations

The present study complements recent findings on student perceived teaching quality in science education (Decristan et al., 2015; Fauth et al., 2014). As in these previous studies, we assessed teaching quality and student outcomes in the context of a specific teaching unit (here, on the water cycle). Long-term effects therefore were not in the scope of the present research. The design allowed, however, a close alignment of outcome measures and teaching unit.

The operationalizations of the three basic dimensions have been heterogeneous in previous studies (Praetorius et al., 2018b). In the present study, available items and operationalizations were restricted, due to our re-analysis of student ratings of teaching quality. Our measure of cognitive support did not cover feedback, for instance, and our measure of classroom management was limited to items on disruptions, and did not consider teacher actions to prevent disruptions (such as monitoring). Our motivational support measure mainly addressed teacher sensitivity and assessed autonomy support, for instance, only marginally. Finally, our measure of cognitive activation included two items addressing cognitive activation in the context of science inquiry. These do not consistently follow a generic approach to teaching quality. Another issue seems to be item formulations that did not exclusively refer to the class (i. e., the level of our analysis) but to individual students (Marsh et al., 2012). Finally, the sample size did not allow us to test factor structure and measurement invariance between grades 4 and 6.

Future research then should test the generalizability of our results across content domains and age groups in short- and long-term designs, and with more comprehensive measures of cognitive activation, cognitive support, motivational support, and classroom management, including other perspectives (teachers and external observers in particular).

9. Conclusion

The three basic dimensions framework is a parsimonious, generic model for teaching quality that is prominent in German-speaking countries in particular. It has provided an important step in establishing common ground for the description of quality teaching. A next step might now be to test potentially relevant modifications of the framework in order to achieve greater comprehensiveness.

The present study suggests such a modification of the basic dimensions approach. On the basis of student ratings of teaching quality in science education, it provides preliminary evidence on the factorial and predictive validity of a four-factor model differentiating cognitive activation, cognitive support, motivational support, and classroom management.

Future research may test this framework in varying settings, as well as including further increments (see, for instance, Praetorius & Charalambous, 2018a) by means of factor analyses and tests of predictive validity. Such research could enable exploration of the theoretical and empirical value added of more comprehensive over more parsimonious frameworks of teaching quality.

References

Alfieri, L., Brooks, P.J., Aldrich, N.J., & Tenenbaum, H.R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103,* 1–18. doi:10.1037/a00210171.

Amin, T.G., Smith, C.L., & Wiser, M. (2014). Student conceptions and conceptual change. In N.G. Lederman, & S.K. Abell (Eds.), *Handbook of research on science education* (Vol. 2, pp. 57–81). New York, NY: Routledge.

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., & Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal, 52,* 1133–1159. doi: 10.3102/0002831215596412.

Einsiedler, W., & Hardy, I. (2010). Kognitive Strukturierung im Unterricht: Einführung und Begriffsklärungen. *Unterrichtswissenschaft, 38,* 194–209. doi:10.3262/UW1003194.

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. doi:10.1016/j.learninstruc.2013.07.001.

Ganzeboom, H.B.G., de Graaf, P.M., & Treiman, D.J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21,* 1–56. doi:10.1016/0049-089X(92)90017-B.

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "Floating and Sinking". *Journal of Educational Psychology, 98,* 307–326. doi:10.1037/0022-0663.98.2.307.

Hardy, I., Hertel, S., Kunter, M., Klieme, E., Warwas, J., Büttner, G., & Lühken, A. (2011). Adaptive Lerngegebenheiten in der Grundschule: Merkmale, methodisch-didaktische Schwerpunktsetzung und erforderliche Lehrerkompetenzen. In W. Helsper & R. Tippelt (Hrsg.), *Pädagogische Professionalität* (57. Beiheft der Zeitschrift für Pädagogik, S. 819–833).

Helm, C. (2016). Zentrale Qualitätsdimensionen von Unterricht und ihre Effekte auf Schüleroutcomes im Fach Rechnungswesen. *Zeitschrift für Bildungsforschung, 6,* 101–119. doi:10.1007/s35834-016-0154-3.

Hochweber, J., & Vieluf, S. (2016). Gender differences in reading achievement and enjoyment of reading: The role of perceived teaching quality. *Journal of Educational Research, 111,* 268–283. doi:10.1080/00220671.2016.1253536.

Hu, L., & Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453. doi:10.1037/1082-989X.3.4.424.

Kauertz, A., Kleickmann, T., Ewerhardy, A., Fricke, K., Lange, K., Ohle, A., Pollmeier, K., Tröbst, S., Walper, L., Fischer, H., & Möller, K. (2011). *Dokumentation der Erhebungsinstrumente im Projekt PLUS.* Essen: Forschergruppe und Graduiertenkolleg nwu-essen.

Kirschner, P.A., Sweller, J., & Clark, R.E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41,* 75–86. doi:10.1207/s15326985ep4102_1.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: Aufgabenkultur und Unterrichtsgestaltung. In Bundesministerium für Bildung und Forschung (Hrsg.), *TIMSS – Impulse für Schule und Unterricht* (S. 43–57). München: Medienhaus Biering.

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

Korneck, F., Krüger, M., & Szogs, M. (2017). Professionswissen, Lehrerüberzeugungen und Unterrichtsqualität angehender Physiklehrkräfte unterschiedlicher Schulformen. In E. Sumfleth & H. Fischler (Eds.), *Professionelle Kompetenzen von Lehrkräften der Chemie und Physik. Studien zum Physik- und Chemielernen Bd. 200* (S. 113–133). Berlin: Logos.

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17,* 494–509. doi: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). New York: Springer. doi:10.1007/978-1-4614-5149-5_6.

Kyriakides, L., Creemers, B., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: the contribution of the dynamic model. *ZDM: The International Journal on Mathematics Education.* doi:10.1007/s11858-018-0919-3.

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

Lipowsky, F. (2015). Unterricht. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 69–105) (2. überarb. Aufl.). Heidelberg: Springer.

Marsh, H.W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44,* 764–802. doi:10.1080/00273170903333665.

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

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

Pianta, R., & Hamre, B.K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38,* 109–119. doi:10.3102/0013189X09332374.

Pinger, P., Rakoczy, K., Besser, M., & Klieme, E. (2017). Interplay of formative assessment and instructional quality – Interactive effects on students' mathematics achievement. *Learning Environment Research, 21,* 61–79. doi:10.1007/s10984-017-9240-2.

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. doi:10.1016/j.learninstruc.2012.03.002.

Praetorius, A.-K., & Charalambous, C. (2018a). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM Mathematics Education.* doi:10.1007/s11858-018-0946-0.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018b). Generic dimensions of teaching quality: The German framework of the three basic dimensions. *ZDM: The International Journal on Mathematics Education.* doi:10.1007/s11858-018-0918-4.

Puntambekar, S., & Hübscher, R. (2005). Tools for scaffolding students in a complex environment: What have we gained and what have we missed? *Educational Psychologist, 40,* 1–12. doi:10.1207/s15326985ep4001_1.

Steffensky, M., Gold, B., Holodynski, M., & Möller, K. (2015). Professional vision of classroom management and learning support in science classrooms: Does professional vision differ across general and content-specific classroom interactions? *International Journal of Science and Mathematics Education, 13,* 351–368. doi: 10.1007/s10763-015-9627-4.

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction, 46,* 45–60. doi:10.1016/j.learninstruc.2016.08.003.

Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review, 22,* 271–297. doi:10.1007/s10648-010-9127-6.

Weiß, R. H. (2006). *CFT 20-R. Grundintelligenztest Skala 2e Revision [Culture fair test]*. Göttingen, Germany: Hogrefe.

Item	λ_w	λ_b
Cognitive activation		
Our teacher demonstrates with an experiment that our explanations are not correct yet.	.57	.89
We often observe things that astonish us.	.64	.92
We can test our assumptions with experiments.	.65	.93
Our teacher asks us to give reasons for our assumptions.	.41	.69
Our teacher asks us to use what we have learned to explain observations from everyday life.	.44	.74
Cognitive support		
In the classroom, often too many questions are treated at the same time. (reverse coded)	.58	.89
In the classroom, I often do not know what we are talking about. (reverse coded)	.63	.94
I know exactly what to do when I am working.	.44	.66
Our teacher often uses foreign words we do not understand. (reverse coded)	.40	.70
Our teacher rarely helps us during classroom talk to find a solution. (reverse coded)	.32	.74
Motivational support		
Our teacher has always time if I want to talk with her/him about something particular.	.77	.97
Our teacher pays attention to my problems.	.81	1.00
If I do not like something, I can talk to our teacher.	.73	.99
If we try hard, we get complimented.	.59	.87
During instruction, we are supported to work autonomously.	.51	.81
Classroom management		
Students are fooling around in the lessons. (reverse coded)	.65	.90
The teacher has to be loud quite often. (reverse coded)	.68	.86
After asking us to be quiet, our teacher has to wait a long time until everybody is actually quiet. (reverse coded)	.66	.91
During instruction, it is often turbulent and loud. (reverse coded)	.76	.98
At the beginning of a lesson, it takes a long time until the students are quiet. (reverse coded)	.61	.87

Tab. A1: Items used to assess teaching quality and standardized factor loadings on the within (λw) and between (λb) level

Zusammenfassung: Das Modell der drei Basisdimensionen von Unterrichtsqualität unterscheidet die Dimensionen kognitive Aktivierung, konstruktive Unterstützung und Klassenführung. Forschung aus unterschiedlichen Disziplinen legt nahe, dass kognitive Unterstützung in dem Modell nicht hinreichend berücksichtigt ist. Faktorenanalysen auf der Basis von Schülereinschätzungen zum naturwissenschaftlichen Unterricht (2,659 Schüler*innen, Klasse 4 und 6) weisen in der vorliegenden Studie auf ein vierdimensionales Modell mit kognitiver Unterstützung als separater Dimension hin. Kognitive Unterstützung sagte zudem den Lernerfolg der Schüler*innen voraus. Die Ergebnisse unterstreichen, dass die Ergänzung um diese Dimension zu einem vollständigeren, aber dennoch sparsamen Modell von Unterrichtsqualität beiträgt.

Schlagworte: Unterrichtsqualität, drei Basisdimensionen, Schülerratings, naturwissenschaftlicher Unterricht, Faktorenanalyse

Contact

Prof. Dr. Thilo Kleickmann, Kiel University,
Olshausenstr. 75, 24118 Kiel, Germany
E-Mail: kleickmann@paedagogik.uni-kiel.de

Prof. Dr. Mirjam Steffensky, Leibniz-Institute for Science and Mathematics Education,
Olshausenstr. 62, 24118 Kiel, Germany
E-Mail: steffensky@ipn.uni-kiel.de

Prof. Dr. Anna-Katharina Praetorius, University of Zurich,
Institute of Education,
Freiestrasse 36, 8032 Zürich, Switzerland
E-Mail: anna.praetorius@ife.uzh.ch

Courtney A. Bell

Commentary Regarding the Section "Dimensions of Teaching Quality – Theoretical and Empirical Foundations"

Using Warrants and Alternative Explanations to Clarify Next Steps for the TBD Model

Abstract: The field has not reached consistent findings regarding the structure of teaching, despite many studies using the Three Basic Dimensions (TBD) model. This section's author offers two productive pathways for the field – one criterion-based and the other empirically focused. This discussion offers a third pathway: the use of the Toulmin's (1958) argument structure. That pathway will focus on an analysis of the most threatening alternative explanations to existing TBD claims and a renewed commitment to specifying and interrogating the warrants implicit to the model's existing claims. Toulmin's argument structure is briefly explained and examples of its application to TBD model findings are considered.

Keywords: Validity, Validity Argument, Toulmin, Teaching Effectiveness, Classroom Interactions

1. Introduction

The field has struggled to come to consistent findings regarding the structure of teaching, despite many studies using the Three Basic Dimensions (TBD) model (Praetorius, Klieme, Kleickmann, Brunner, Lindmeier, Taut & Charalambous, in this issue). The TBD model hypothesizes three factors of teaching quality as cognitive activation, student support, and classroom management; and coheres with the well-researched three factor Teaching Through Interactions (TTI) framework (Hamre et al., 2013), the framework operationalized in the CLASS observation system.

Despite surface similarities between the TBD and TTI models, there are important differences that warrant the TBD be considered on its own structure and merits. Specifically, the TBD model has been operationalized in both student questionnaires and observation systems, allowing for potential testing of robustness to measurement mode. The model has factor analytic evidence that supports its hypothesized structure, but as Kleickmann, Steffensky, and Praetorius (in this issue) carefully document, studies do not operationalize the three factors in the same ways so it is unclear how to interpret such findings. Finally, predictive results from the model are uneven (Praetorius et al., in this issue). Together, these realities demand we pay careful attention to what may and may not generalize across studies.

Globally, we face similar concerns around consistency of instrumentation and findings across studies. We all must build systematic understandings across studies of teaching models. In the TBD context we can ask: 'How should the field move forward to develop systematic understanding of the TBD model?' The two papers in this section offer different approaches. Praetorius et al. (in this issue) offer a criterion approach. Specifically, they compare the validation evidence for the TBD model against general criteria of a good theory pointing out where additional evidence is needed. Kleickmann and colleagues (in this issue) take an empirical approach that reconsiders the specific constructs being measured in the model, ultimately presenting a new four-factor structure. Each approach offers productive foci along which the field might progress. Each has strengths and weaknesses. To encourage an even broader perspective, I suggest a third approach.

I argue that future work on the TBD model would benefit from a review of studies focused on the most threatening alternative explanations to existing claims and a renewed commitment to specifying and interrogating the warrants implicit to the model's existing claims. To begin, I specify three assumptions upon which my suggestions rest. Then I use Toulmin's model of argumentation (1958) to show how a heuristic based on the structure of arguments might be used at the field-level to identify future directions for TBD research.

2. Assumptions About Teaching and Research Knowledge of Teaching

There are three assumptions foundational to how knowledge of teaching models accumulates. First, teaching is intertwined with learning. If learning is situated (it occurs in specific schools with specific students at a moment in time), teaching is definitionally situated in social-historical contexts (Lave & Wenger, 1991). Teaching is also a complex performance that reflects interactions of specific subjects, students, and teachers (Cohen, Raudenbush, & Ball, 2003).

These facts imply that the interpretation of TBD-based scores must initially be understood to apply only to the context in which the scores were gathered. This applies to the measurement system (see Liu, Bell, Jones & McCaffrey, 2019), the specific intersection of teachers, students, and subject (Cohen, Ruzek & Sandilos, 2018; Campbell & Ronfeldt, 2018), and the purpose for which scores were created (Vitiello, Bassok, Hamre, Player & Williford, 2018). For example, a factor analysis of ratings created for a research study of gymnasium classrooms should not be understood to generalize to realschule classrooms unless there is evidence of stability across those school types.

Finally, data do not "speak for themselves" (AERA/APA/NCME, 2014); they are not evidence until they are linked to a specific purpose and assembled into an evidentiary argument. For example, the TTI framework was originally developed in pre-Kindergarden classrooms and a three-factor model best fit the data (Hamre et al, 2013) with some support for a bifactor model (Hamre, Hatfield, Pianta, & Jamil, 2014). Research on the associated secondary-level CLASS observation system produced evidence that the TTI framework could be described by a one factor model (BMGF, 2012). The argu-

ment for a three-factor model of the secondary system ultimately considered data across multiple studies; and referred to factor analytic findings at other grade levels, how the instrument had been used in professional development, and the theoretical predictions from the TTI framework (Hafen et al., 2014).

Given these assumptions, understanding the generalizability of the TBD model may be advanced through identification of the major claims and warrants in need of attention and the subsequent application of a general heuristic. Toulmin's (1958) explanation of the structure of informal arguments is one such heuristic. Assessment scholars have made use of Toulmin's heuristic, arguing that it helps researchers make sense of assessment data in the form of validation arguments and metaphors (Kane, 2006; Mislevy, 2018). The approach is also useful at the level of a field. I demonstrate this by using it to consider hypothetical claims across studies on the TBD model.

3. Using the Structure of Arguments to Identify Assumptions and Alternative Explanations

In Toulmin's terms (Fig. 1), arguments are structured around data, claims, warrants, and alternative explanations. There are patterns in datum (D) that lead us to make certain claims (C). Those claims are supported because we have warrants (W) or justifications for the claims; warrants are supported by logical, theoretical, empirical or other types of backing (B). Claims may be undermined if we can identify alternative explanations (A) for the relationships between the data and the claim and support the explanations with

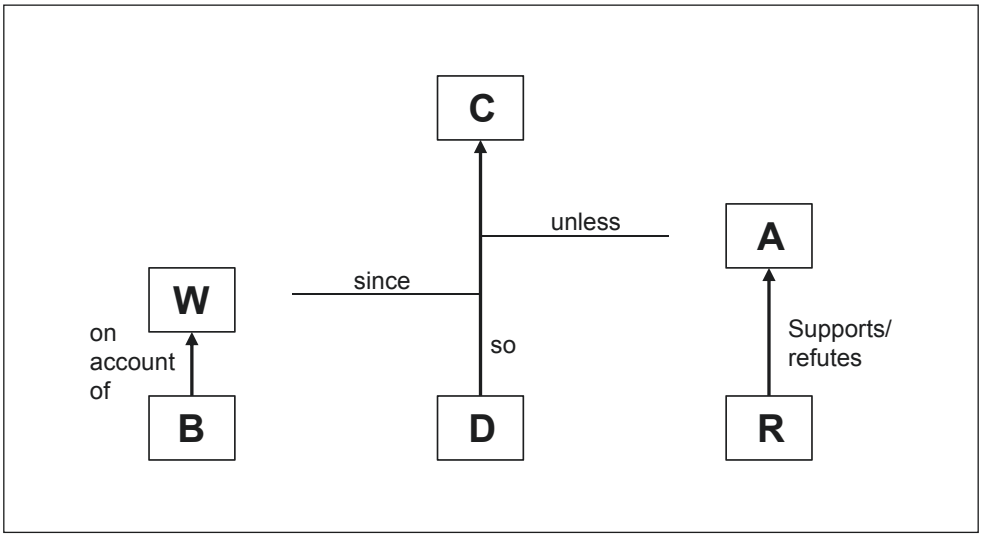

Fig. 1: Toulmin's (1958) structure of informal arguments (figure modified from Toulmin 1958, p. 97)

refuting evidence (R). If evidence supports the alternative explanation, the claim is undermined. If the alternative explanation is refuted, the claim is strengthened.

Using this structure, we can better identify and consider potential research trajectories for the TBD model. We begin with the foundational assumption that context matters. For example, U. S. researchers have found that for the same teacher, higher classroom observation ratings are associated with higher incoming achievement levels of the students taught (Campbell & Ronfeldt, 2018). Additionally, some evidence suggests that when students' racial background matches that of the teacher, students learn more (Yarnell & Bohrnstedt, 2018).

4. Claim 1

Focusing on context, Toulmin's logic can be used to generate multiple additional research foci for the TBD model. Here is one example of a hypothetical claim of interest:

Assume that in two studies of gymnasium mathematics classrooms, a three-factor model fits the student questionnaire data (D). Since factor analyses provide evidence of latent factors (W) and the specific type of factor analysis carried out in both studies appropriately models student questionnaire data (B), the TBD model of teaching is supported (C). An alternative explanation that the three-factor model is supported only in mathematics gymnasium classrooms (A) is refuted in two additional studies that carry out similar factor analyses finding the best fitting model is the same three factor model in German and history Gesamtschule classrooms (R).

The act of writing out this single hypothetical claim specification immediately demands we clarify the warrants upon which our research enterprise rests. It also forces the articulation of alternative explanations.

Continuing with the example, in stating the backing for the use of factor analysis, two relevant research findings come into focus. First, recent research on factor analytic approaches in observation instruments measuring teaching quality suggests that factor analysis should take account of measurement error. Not accounting for measurement error can lead to erroneous conclusions about the factor structure (McCaffrey, Yuan, Savitsky, Lockwood & Edelen, 2015). This suggests a line of inquiry focused on various factoring methods that can be applied to existing and new data.

Second, aggregating evidence across different studies' factor analyses traditionally assumes the constructs have been operationalized into the same instruments, otherwise the results may reflect different instruments, not different factor structures. As previously noted, there are various instruments used across studies of the TBD model (Praetorius et al., in this issue), thus pointing to a second potential line of inquiry.

This single claim specification also prompts us to consider alternative explanations. In addition to the articulated alternative explanation, which focuses on the role of subject matter, there are other explanations. For example: 1) the TBD model is only supported in ratings from secondary classrooms, 2) the TBD model is only supported in ratings from gymnasium classrooms, or 3) the TBD model is only supported in mod-

els that are measured by ratings from student questionnaires. The alternative explanations consider aspects of the context – grade level, school type, and measurement mode. By specifying these alternative explanations, researchers can consider which are most threatening.

5. Claim 2

Validity is best conceived of as a set of related claims moving from the scoring inferences to implication inferences (Kane, 2006). To develop a persuasive TBD validity argument many types of evidence are needed. Evidence about predictive validity and utility are particularly important to the improvement of teaching. Here is a second hypothetical claim that speaks directly to the utility of the TBD model.

Assume that in an experiment that coaches gymnasium science teachers using the TBD model, the students of coached teachers show no difference in motivation but significantly different academic growth compared to the comparison teachers' students (D). Because teacher practices changed in response to guidance around the TBD model (W), the TBD model supports effective coaching of teachers (C). An alternative explanation that coaching around any other three factor model would support changes in teacher practice better than the TBD model (A) is supported in another experiment carried out in science gymnasium classrooms. It shows that students of teachers coached in a different three factor model had higher levels of both motivation and academic growth compared with the comparison teachers' students (R).

As with the first hypothetical claim, this claim draws our attention to warrants. The warrant raises questions about the utility of the TBD model. Are teachers able to understand the model? To what degree does coaching around the model support teacher learning? What aspects of student learning should be related to improvements in the model's factors?

It also draws our attention to alternative explanations for the claim that the TBD model supports effective coaching of teachers. For example, the following alternative explanations could also apply: 1) the TBD model supports teacher learning among gymnasium teachers, 2) the TBD model helps teachers of traditional academic subjects learn, 3) coaching around any one of the factors in the model would help teachers learn. Again, the alternative explanations consider contextual features – school type and subject matter – as well as the relationships among the factors in the TBD model.

6. Developing Research Priorities From Inquires of Warrants and Alternative Explanations

While the Toulmin heuristic can be used to generate many claims, warrants, and alternative explanations, research is limited by constraints such as funding and time. Researchers must make choices about what is worthy of resources. Such choices are informed

by technical considerations, but they are also informed by researchers' values. What is most important – specifying the structure of teaching? Improving teaching? Demonstrating relationships between teaching and student outcomes?

While each of these values is important, individual researchers and teams must make decisions about which alternative hypotheses and warrants to interrogate given the specific values of the research team. If teams explicitly consider the relationships and details of the TBD model's claims, warrants, and alternative explanations, each team can pursue increasingly nuanced lines of inquiry. This could result in progress toward a robust and persuasive validity argument for the model.

References

American Educational Research Association (AERA), American Psychological Association (APA) & National Council of Measurement in Education (NCME) (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association. http://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition [15. 10. 2019].

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*(6), 1233–1267. doi:10.3102/0002831218776216.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119–142. doi:10.3102/01623737025002119.

Cohen, J., Ruzek, E., & Sandilos, L. (2018). Does teaching quality cross subjects? Exploring consistency in elementary teacher practice across subjects. *AERA Open, 4*(3), 1–16. doi:10.1177/2332858418794492.

BMGF = The Bill and Melinda Gates Foundation (2012). Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains. http://k12education.gatesfoundation.org/resource/gathering-feedback-on-teaching-combining-high-quality-observations-with-student-surveys-and-achievement-gains-3/ [15. 10. 2019].

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms revisiting the factor structure and practical application of the classroom assessment scoring system – secondary. *The Journal of Early Adolescence, 35*(5-6), 651–680.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & Hamagami, A. (2013). Teaching through Interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal, 113*(4), 461–487. doi:10.1086/669616.

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher-child interactions: Associations with preschool children's development. *Child Development, 85*(3), 1257–1274. doi:10.1111/cdev.12184.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). New York: Praeger Publishers.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge: Cambridge University Press.

Liu, S., Bell, C. A., Jones, N., McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability, 31*(1), 31–61. https://doi.org/10.1007/s11092-018-09291-3.

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice, 34*(2), 34–46. doi:10.1111/emip.12061.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement.* New York: Routledge.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Vitiello, V. E., Bassok, D., Hamre, B. K., Player, D., & Williford, A. P. (2018). Measuring the quality of teacher–child interactions at scale: Comparing research-based and state observation approaches. *Early Childhood Research Quarterly, 44,* 161–169. doi:https://doi.org/10.1016/j.ecresq.2018.03.003.

Yarnell, L. M., & Bohrnstedt, G. W. (2018). Student-teacher racial match and its association with black student achievement: An exploration using multilevel structural equation modeling. *American Educational Research Journal, 55*(2), 287–324. doi:10.3102/0002831217734804.

Zusammenfassung: Das Forschungsfeld hat trotz vieler Studien zum Modell der drei Basisdimensionen keine konsistenten Erkenntnisse über die Struktur von Unterrichtsmerkmalen gewonnen. Die Autorinnen und Autoren dieses Themenblocks bieten zwei produktive Wege für die weitere Forschung an – einen kriterienbasierten und einen empirisch fokussierten. Die vorliegende Diskussion stellt einen dritten Weg vor: die Verwendung der Argumentationstheorie nach Toulmin (1958). Dieser Weg würde darin bestehen, möglichst starke alternative Erklärungen für Aussagen des Modells der drei Basisdimensionen aufzustellen und zu analysieren, sowie verstärkt die impliziten Annahmen des Modells zu spezifizieren und zu hinterfragen. Toulmins Theorie der Struktur von Argumenten wird kurz erläutert und es werden Beispiele auf ihre Anwendung der Befunde zum Modell der drei Basisdimensionen hin untersucht.

Schlagworte: Validität, Argumentation, Toulmin, Wirkung von Unterricht, Interaktionen im Klassenzimmer

Contact

Courtney A. Bell, Educational Testing Service | ETS,
Center for Global Assessment,
Rosedale Road, MS 13-E, Princeton, NJ 08541,
E-Mail: cbell@ets.org

Themenblock II: Angebots-Nutzungs-Modelle als Rahmung

Svenja Vieluf/Anna-Katharina Praetorius/Katrin Rakoczy/Marc Kleinknecht/ Marcus Pietsch

Angebots-Nutzungs-Modelle der Wirkweise des Unterrichts

Ein kritischer Vergleich verschiedener Modellvarianten

Zusammenfassung: Dieser Beitrag widmet sich Angebots-Nutzungs-Modellen der Wirkweise des Unterrichts. Konkreter hat er zum Ziel, verschiedene Angebots-Nutzungs-Modelle zu vergleichen und durch diesen Vergleich konzeptuelle Unschärfe innerhalb des Ansatzes aufzudecken. Kritisch diskutiert werden sollen Unterschiede hinsichtlich a) dem zugrundeliegenden Verständnis von Unterricht, Angebot und Nutzung, b) Zusammenhängen zwischen Angebot und Nutzung, c) der Bedeutung von Wahrnehmung und Interpretation, d) der Verortung von Angebot und Nutzung im Mehrebenensystem, e) der Bedeutung von Kontexten und f) der aufgeführten Kriterien unterrichtlicher Wirksamkeit. Der Beitrag schließt mit einem Fazit, in dessen Rahmen ein integriertes Angebots-Nutzungs-Modell vorgestellt wird.

Schlagworte: Angebots-Nutzungs-Modell, Unterrichtsqualität, (Ko-)Konstruktion, Mehrebenenmodell, Kriterien unterrichtlicher Wirksamkeit

1. Einführung

Angebots-Nutzungs-Modelle der Wirkweise des Unterrichts, die vor allem auf die Arbeiten von Fend (1981, 1982, 1998, 2008a, 2008b) sowie Helmke und Weinert (1997) und Helmke (2003) zurückgehen, bilden schematisch und auf hohem Abstraktionsniveau ab, welche Struktur- und Prozessmerkmale im Zusammenspiel auf mehreren Ebenen beeinflussen können, wie effektiv Schüler*innen im Unterricht lernen. Zugrunde liegt ihnen die konstruktivistische Vorstellung, dass Lernende aktive (Ko-)Konstrukteure ihres eigenen Lernfortschritts sind und deshalb die Effektivität des Unterrichts nicht nur davon abhängt, welche Lerngelegenheiten sich den Lernenden in der Schule bieten, sondern auch davon, ob und wie sie diese nutzen. Neben dem Zusammenspiel von Angebot und Nutzung berücksichtigt das Modell auch Einflüsse von Merkmalen der individuellen Akteure (Lehrende und Lernende) sowie der Bildungskontexte (des Bildungs-

systems, der Einzelschule und der Schulklasse; z. B. Helmke, 2010; Lipowsky, 2006; Reusser & Pauli, 2010; Seidel, 2014).

Angebots-Nutzungs-Modelle haben im deutschsprachigen Raum breite Akzeptanz gefunden und liegen einer Vielzahl empirisch-quantitativer Studien zugrunde (z. B. DESI-Konsortium, 2008; Hardy et al., 2011; Kunter et al., 2011; Reusser & Pauli, 2010). Kohler und Wacker (2013) bezeichneten sie gar als das „derzeit prominenteste Wirkmodell innerhalb der Schul- und Unterrichtsforschung" (S. 241). Ihr Erfolg dürfte auch damit zusammenhängen, dass sie das komplexe Zusammenspiel vieler Faktoren auf verschiedenen Ebenen übersichtlich darzustellen vermögen. So eignen sie sich als Grundlage für die Verständigung über Unterricht in Forschung, Lehre und Praxis, für die Integration verschiedener Befunde aus empirischen Untersuchungen und für die Planung neuer Studien (Kohler & Wacker, 2013; Seidel, 2014). Mittlerweile sind allerdings eine Vielzahl unterschiedlicher Angebots-Nutzungs-Modelle formuliert worden (Brühwiler & Blatchford, 2011; Helmke, 2003, 2007, 2010; Klieme, Lipowsky, Rakoczy & Ratzka, 2006; Kunter & Trautwein, 2013; Lipowsky, 2006; Reusser & Pauli, 2003, 2010; Seidel, 2014), welche signifikante Unterschiede aufweisen. Diese wollen wir im Folgenden kritisch diskutieren – auch mit Blick auf die Implikationen dieser Unterschiede für Forschung und Praxis.

2. Konzeptionelle Grundlagen der Angebots-Nutzungs-Modelle der Wirkungsweise des Unterrichts

Angebots-Nutzungs-Modelle der Wirkungsweise des Unterrichts basieren auf Überlegungen von Fend (1981) und finden sich in grafischer Form erstmalig bei Fend (1982, S. 215). Es handelt sich dabei um eine deutschsprachige Weiterentwicklung des – vor allem durch US-amerikanische Forschung geprägten – Prozess-Produkt-Paradigmas. Letzterem liegt die Annahme zugrunde, Unterschiede zwischen Schüler*innen in Bezug auf Lernresultate (meist die in Tests gezeigten fachlichen Leistungen) ließen sich mit Unterschieden in Unterrichtsprozessen (meist isolierbarem Verhalten der Lehrenden und Lernenden, aber auch beschreibbaren Mustern der Interaktion zwischen diesen Akteuren) erklären. Als Reaktion auf Kritik an der Unterkomplexität dieser Annahme (zusammengefasst z. B. in Gage & Needles, 1989) wurde das Paradigma verschiedentlich erweitert. So wurden kognitive Informationsverarbeitungsprozesse als Mediatoren des Effekts von Unterrichtsmerkmalen auf Lernergebnisse im sog. „kognitiven Mediationsparadigma" aufgenommen (z. B. Borich, 1986; Doyle, 1977; Rothkopf, 1976; Winne, 1987). Zudem erhielten auch lernbegleitende Prozesse wie die Lernmotivation (z. B. Pintrich, 2003) und Emotionen (z. B. Mayring & Rhoeneck, 2003) und Kontextmerkmale wie die Klassenstufe und das Fach (siehe z. B. Dunkin & Biddle, 1974) immer mehr Beachtung.

An diese Überlegungen knüpfte Fend (1981, 1982, 1998, 2008a, 2008b) an und verband sie mit Luhmanns Systemtheorie und handlungstheoretischen Überlegungen. Den Unterschied zur Prozess-Produkt-Tradition erklärte er selber wie folgt:

> Neben […] Erweiterungen bei den *administrativen, organisatorischen und inhaltlichen Stützen der Qualität des Angebotes* würde ich einen grundsätzlich anderen Akzent als den der ‚Produktion' setzen. Ich meine, dass Unterricht und Lernen als interaktiver Prozess anzusehen ist, der handlungstheoretisch besser abzubilden ist, als dies beim Denken in Kategorien von Produktionsfaktoren der Fall ist. (Fend, 1998, S. 323; Hervorh. d. Verf.)

Neu ist bei ihm vor allem die Konzeption von Angebot und Nutzung, die ihre Wurzeln in Luhmanns Systemtheorie hat (Fend, 2008a). Luhmann (2002 u. a.) beschrieb Systeme (Individuen, aber auch Klassen oder Schulen) als operativ geschlossen, sodass sie einander nicht direkt beeinflussen, sondern bloß durch strukturelle Koppelung zur Bildung neuer Strukturen anregen können. Mit Bezug darauf argumentierte Fend (2008a): „Die Angebotsseite des Bildungsprozesses kann nicht beinhalten, die Eigendynamik der Nutzungsseite auszuschalten und damit die volle Verantwortung für die Ergebnisse des Bildungsprozesses zu übernehmen. Die Verantwortung liegt angesichts unhintergehbarer Autopoiesis und Eigenintentionalität auch auf der Nutzungsseite" (S. 130). Lernen kann zwar durch pädagogische Kommunikation im Unterricht angeregt und unterstützt, aber nicht determiniert werden. Lernende bleiben „Produzent der eigenen Entwicklung" (Fend, 2008a, S. 132). Luhmanns Systemtheorie wird auch dem operativen Konstruktivismus zugerechnet (De Haan & Rülcker, 2009). Die konstruktivistische Grundlage wird bei Fend (2008a) ebenfalls deutlich, etwa wenn er die Unvorhersehbarkeit und Unbeherrschbarkeit des Unterrichtsgeschehens, die Ambivalenz und Missverständnisanfälligkeit jeder Kommunikation im Klassenzimmer, die sog. „doppelte Kontingenz", betont (Fend, 2008a, S. 323). So verortete Fend die Verantwortung für Momente des Scheiterns im Unterricht nicht ausschließlich bei den Lehrenden. Diese werden entlastet und Schüler*innen in ihrem Recht auf Selbstbestimmung ernst genommen. Anders als Luhmann, dessen Fokus auf abstrakten Systemen liegt, versuchte Fend (2008a) aber auch intentionales soziales Handeln von Individuen im Schulsystem zu verstehen. Hierfür bezog er sich u. a. auf den akteurszentrierten Institutionalismus (Scharpf, 2000) und argumentierte, Institutionen (also Regelsysteme) strukturierten die Chancen und Restriktionen des Handelns vor, die individuellen Wahrnehmungen, Präferenzen und Ressourcen der Akteure hätten aber ebenfalls einen Einfluss darauf, wie das Handeln innerhalb dieses Rahmens ausgestaltet würde (Fend, 2008a, S. 152). Die Regelungen auf höheren Ebenen würden von den Akteuren auf darunterliegenden Ebenen rekontextualisiert (Fend, 2008a, 2008b).

Helmke (2003) griff die Idee einer Trennung von Angebot und Nutzung auf, weniger jedoch die von Fend beschriebenen system- und handlungstheoretischen Grundlagen (er zitierte auch nur Fend, 1981, nicht aber Fend, 1982 und 1998). Stattdessen verband er den Angebots-Nutzungs-Gedanken mit eigenen Vorarbeiten (Helmke & Weinert, 1997). Sein Angebots-Nutzungs-Modell, das er später noch mehrmals modifiziert hat (z. B. Helmke, 2007, 2010), lässt sich als erweitertes kognitives Mediationsparadigma beschreiben. Ähnlich wie bei Fend hat er neben Angebot und Nutzung auch Einflüsse individueller Eingangsvoraussetzungen auf die Nutzung sowie Einflüsse des

Klassenkontextes und des fachlichen Kontextes auf Angebot und Nutzung dargestellt. Zusätzlich nahm Helmke ein Feld „Lehrerpersönlichkeit“ in das Modell auf, dessen theoretische Grundlage im sog. „Expertenparadigma“ zu finden ist, das sich mit der Frage beschäftigt, welches berufsbezogene Wissen und Können Lehrpersonen von Novizen unterscheidet bzw. welches Wissen und Können notwendige Voraussetzung dafür ist, die Anforderungen des Lehrerberufes erfolgreich zu bewältigen (u. a. Berliner, 1992; Bromme, 1992). Außerdem zog Helmke explizit lern- und motivationspsychologische Grundlagen heran und ergänzte die kognitive Mediation um motivationale und emotionale Vermittlungsprozesse zwischen Unterrichtsangebot und Nutzung durch die Schüler*innen. Schließlich hat er – anders als Fend – auch die Wirkungen des Unterrichts im Modell dargestellt und führte hier als Wirkungen neben fachlichen Effekten auch überfachliche Effekte auf.

Beide Angebots-Nutzungs-Modelle der Wirkungsweise des Unterrichts sind im deutschsprachigen Raum breit rezipiert und zitiert worden und sie fungieren als Grundlage für eine Vielzahl empirischer Studien. Zudem sind weitere Varianten des Angebots-Nutzungs-Modell entwickelt worden, die sich mal stärker auf Fend, mal stärker auf Helmke beziehen, deren Modelle aber nie eins zu eins übernehmen (u. a. Brühwiler & Blatchford, 2011; Klieme et al., 2006; Kunter & Trautwein, 2013; Lipowsky, 2006; Reusser & Pauli, 2003, 2010; Seidel, 2014). Vereinzelt sind die Modelle auch in englischsprachigen Publikationen zitiert worden (z. B. Brühwiler & Blatchford, 2011; Lipowsky et al., 2009) – primär handelt es sich jedoch um einen deutschsprachigen Diskurs.

3. Unterschiede zwischen verschiedenen Angebots-Nutzungs-Modellen der Wirkungsweise des Unterrichts und ihre theoretische Bedeutung

Vergleicht man sämtliche Angebots-Nutzungs-Modelle der Wirkungsweise des Unterrichts, wird deutlich, dass diese zwar ähnliche Felder umfassen – nämlich stets ein Angebots- und ein Nutzungsfeld, mehrere Felder für Kontexte sowie (außer bei Fend, 1982, 1998) je ein Wirkungsfeld –, sie sich sonst aber in Bezug auf nahezu alle Aspekte unterscheiden. Augenfällig sind zunächst Unterschiede bezüglich der innerhalb der Felder aufgeführten Faktoren. Allerdings sind die Modelle lediglich als Rahmen zu verstehen, der immer wieder neu „mit spezifischeren Konstrukten und theoriegeleiteten Hypothesen „gefüllt“ werden muss“ (Lipowsky, 2015, S. 76), wobei die konkreten Faktoren innerhalb der Felder unter anderem vom Zielkriterium abhängen dürften (Helmke, 2002; Klieme, 2006; Reusser, 2009). Insofern sind die in den Feldern aufgeführten Faktoren exemplarisch zu verstehen. Dies wird nicht immer berücksichtigt, wenn Angebots-Nutzungs-Modelle als Grundlage für empirische Studien herangezogen werden, lässt sich aber prinzipiell als Rechtfertigung für die Unterschiede anführen. Von noch grundsätzlicherer Relevanz für das Design neuer Studien, aber auch für die Integration der Ergebnisse verschiedener Studien und die Kommunikation über unterricht-

liche Wirkungen, erscheinen insofern andere Unterschiede, welche im Folgenden ausführlicher diskutiert werden sollen.

Zur Illustration dieser Unterschiede wird jeweils die folgende Unterrichtsszene herangezogen: Eine Lehrerin zeichnet ein Koordinatensystem und zwei Punkte an eine Tafel und fordert ihre Schüler*innen auf, eine Parabel durch diese Punkte zu legen. Nachdem mehrere Schüler*innen Parabeln an die Tafel gezeichnet haben, fragt sie: „Wie lange können wir das jetzt machen hier?“ Ein Schüler antwortet: „Irgendwie war jetzt die Parabel von [Schülerin 1] und [Schülerin 2] fast nur ein anderer, ich sag mal ‚Winkel‘. Es gibt praktisch nur zwei ‚Ausgangsparabeln‘ und die kann man dann ja beliebig stauchen und strecken“. Im Folgenden werden jeweils zunächst theoretisch Unterschiede zwischen den Angebots-Nutzungs-Modellen diskutiert und dann werden die Gedanken mit dem Beispiel verknüpft.

3.1 Das zugrundeliegende Verständnis von Unterricht

Allgemein herrscht in den Bildungswissenschaften keine hinreichende Klarheit darüber, was Unterricht als pädagogische Form auszeichnet (z. B. Lüders, 2012). In der Literatur zu Angebots-Nutzungs-Modellen wird häufig der sozial-interaktive Charakter des Unterrichts hervorgehoben. So bezeichnete etwa Fend (1998) Unterricht als „interaktiven Prozess“ (S. 323), Reusser (2009) als „interaktionales Geschehen“ (S. 893), und Helmke, Piskol, Pikowsky und Wagner (2009) als „eine hoch komplexe gemeinsame Aktivität von Lehrerinnen und Lehrern sowie Schülerinnen und Schülern“ (S. 98). Ähnlich argumentierte Klieme (2006), „Unterricht als sozialer Prozess wie auch das darin verhandelte Wissen“ stelle „eine Ko-‚Produktion‘ der beteiligten Personen“ dar (S. 765). Letztere Definition deutet auch an, dass Unterricht in diesem Diskurs nicht allein als Interaktion zwischen Lehrenden und Lernenden betrachtet wird, sondern dass zusätzlich die Beziehung der Akteure mit dem fachlichen Inhalt im Fokus steht, wie dies auch Reusser (2009) betont hat. Diese Überlegungen sind bereits im didaktischen Dreieck enthalten (vgl. Reusser, 2009) und machen deutlich, dass Unterricht und die darin stattfindenden Interaktions- und Lernprozesse stets fach- und inhaltsspezifisch konzipiert werden müssen. Unterschiedliche Unterrichtsinhalte können unterschiedliche fachdidaktische Bearbeitungen erfordern, damit Lernende zu fachspezifischen Lernaktivitäten angeregt werden (Seidel & Shavelson, 2007).

Lipowsky (2015) sowie Reusser und Pauli (2003, 2010) stellten Unterricht außerdem als ein Gesamtgeschehen dar, in dem Angebot und Nutzung zusammenspielen. Dagegen wurde in anderen Modellen der Unterricht mit dem Angebot gleichgesetzt. Besonders offensichtlich ist dies bei Helmke (2003, 2007, 2010) sowie bei Kunter und Trautwein (2013), in deren Modellen jeweils das Angebotsfeld die Überschrift „Unterricht“ trägt. Doch auch bei Seidel (2014) heißt es: „In diesem Sinne werden Unterricht und das Lehren bzw. Gestalten von Lernumgebungen als eine *Angebotsstruktur* aufgefasst, die von den Lernenden für sich genutzt werden will (soll)“ (S. 857). Die Autor*innen von Angebots-Nutzungs-Modellen sind sich also augenscheinlich uneins, in-

wieweit die Nutzung von Lernangeboten Teil des Unterrichts ist oder ob der Unterricht bloß mit dem Angebot gleichzusetzen ist. Dies wirft die Frage auf, wie denn eigentlich „Angebot" und „Nutzung" jeweils definiert werden.

Die Definition des Angebots

Wie schon in Bezug auf den Unterrichtsbegriff, so herrscht auch hinsichtlich der Bestimmung des Angebotsbegriffs keine Einigkeit zwischen den Autor*innen verschiedener Angebots-Nutzungs-Modelle. Von manchen wird das Angebot mit dem Verhalten der Lehrenden gleichgesetzt. So steht im Angebots-Nutzungs-Modell von Brühwiler und Blatchford (2011) hinter den „classroom processes" in Klammern „teacher acting" (S. 97). Auch bei Lipowsky (2015) findet sich die Formulierung „Lernangebote des Lehrers" (S. 76) und bei Fend (2008b) stehen im Feld „Angebotshandeln" nur die Lehrenden. Im Kontrast dazu betonte Helmke (2010): „Selbstverständlich können auch Schülerinnen und Schüler Angebote machen, und sie tun es unaufhörlich" (S. 76). Als Beispiele nannte er „reciprocal teaching" sowie innoffizielle Parallelkommunikation und heimliches Helfen unter Schüler*innen (Helmke, 2007, 2010). In der Abbildung seines Modells wird dieser Gedanke jedoch auch nicht deutlich.

Nimmt man den Gedanken der interaktiven Ko-Produktion von Unterricht, den Klieme (2006), aber auch Fend (2008b), formuliert haben, ernst, so ist davon auszugehen, dass Schüler*innen noch viel grundsätzlicher an der Gestaltung der inhaltlichen Auseinandersetzung mit dem Unterrichtsgegenstand beteiligt sind. Dies kann etwa anhand des Klassengesprächs verdeutlicht werden, einer Sozialform die in deutschen Schulen vielfach den Unterricht dominiert (Jurik, Seidel & Gröschner, 2012; Seidel, 2011). Ein gelingendes Klassengespräch kann nämlich nicht allein durch anregende Gesprächsimpulse der Lehrenden realisiert werden, sondern erfordert auch eine aktive Beteiligung der Schüler*innen. In unserem Unterrichtsbeispiel kann nicht nur die Aufforderung der Lehrerin, eine Parabel durch zwei Punkte zu legen, ein Lernangebot für die Klasse darstellen, sondern auch etwa der Schülerbeitrag: „Es gibt praktisch nur zwei ‚Ausgangsparabeln' und die kann man dann ja beliebig stauchen und strecken", da dieser ebenfalls bei Mitschüler*innen Denkprozesse anregen könnte. Eben dies könnte als „Ko-Produktion" aufgefasst werden. Theoretisch kohärent wäre demnach davon auszugehen, dass das Angebot von Lehrenden und Schüler*innen in ihrer Interaktion gemeinsam gestaltet wird, wie es Helmke (2007, 2009, 2010) formuliert hat.

Die Definition der Nutzung

Wenn nun aber das Handeln der Lernenden im Unterricht auch als Teil des Angebots zu verstehen wäre, was ist dann noch die Nutzung? In allen Angebots-Nutzungs-Modellen, in denen das Nutzungsfeld überhaupt genauer spezifiziert ist (Helmke, 2003, 2007, 2010; Klieme et al., 2006; Kunter & Trautwein, 2013; Seidel, 2014), werden *kognitive Lernprozesse* in diesem Feld genannt. Hier wird deutlich, dass das Angebots-Nutzungs-Modell eine Weiterentwicklung des kognitiven Mediationsparadigmas ist (s. o.). Als theoretische Grundlage für die Konzeptualisierung der kognitiven Lernprozesse werden von Helmke (2003) sowie von Kunter und Trautwein (2013) sowohl Informationsver-

arbeitungstheorien als auch sozialkonstruktivistische Theorien angeführt. In den Beschreibungen anderer Angebots-Nutzungs-Modelle wird diese Frage weniger explizit beantwortet. Zumindest Reusser (2009) scheint sich stärker auf kognitiv-konstruktivistische Theorien zu beziehen.

Zusätzlich werden in manchen Angebots-Nutzungs-Modellen auch *Motivation* und *Emotionen* dem Feld der Nutzung zugeordnet (Klieme et al., 2006; Seidel, 2014), wobei diese Prozesse jedoch in anderen Modellen als Mediatoren zwischen Angebot und Nutzung verstanden werden (Helmke, 2003; Kunter & Trautwein, 2013). Zusätzlich tauchen Kognitionen, Motivation und Emotionen in vielen Modellen auch im Feld des Lernpotenzials bzw. der Lernvoraussetzungen auf (Helmke, 2007, 2010; Lipowsky, 2006; Reusser & Pauli 2003, 2010; Brühwiler & Blatchford, 2011; Seidel, 2014), teilweise aber auch bei den Wirkungen (Seidel, 2014; Lipowsky, 2006; Reusser & Pauli, 2003, 2010).

Um diese Mehrfachnennung von Motivation und Emotionen in den Modellen zu verstehen, ist es zunächst nützlich, zwischen Prädispositionen (Traits) und aktuellem Erleben (States) zu unterscheiden. Motivationale und affektive Prädispositionen wie z. B. motivationale Orientierungen (Krapp, 1999, S. 392), individuelle Interessen und Fähigkeitsselbstkonzepte (Kunter & Trautwein, 2013, S. 49), oder Gefühlstendenzen wie die Neigung zu Angst oder Ärger können in Interaktion mit situativen Merkmalen (bzw. dem Angebot) bestimmte Zustände bzw. Formen des motivationalen und emotionalen Erlebens wahrscheinlicher machen als andere. Insofern lassen sich die Prädispositionen (Traits) tatsächlich in plausibler Weise dem Feld des Lernpotenzials bzw. der Lernvoraussetzungen zuordnen – wie es in einigen Modellen der Fall ist – und die States, also aktuelle motivationale und emotionale Zustände, als Teil des Nutzungsgeschehens im Klassenzimmer verstehen. Es wäre allerdings hilfreich, wenn diese Unterscheidung zwischen Traits und States in Angebots-Nutzungs-Modellen noch expliziter vorgenommen würde.

Darüber hinaus stellt sich die Frage, ob situative Zustände motivationalen und emotionalen Erlebens (States) eher als Teil der Nutzung konzeptualisiert werden sollten oder als Mediatoren des Effekts des Angebotes auf die Nutzung. Pekrun und Jerusalem (1996, S. 7–8) argumentierten, dass sich lern- und leistungsbezogene Kognitionen (z. B. Wunsch und Absichtskognitionen) zunächst auf die Emotions- und Motivationsbildung auswirkten und letztere dann kognitive Lernprozesse beeinflussten, dass der Lernprozess aber auch wieder auf emotionale und motivationale Zustände zurückwirke und in dem Prozess zudem kontinuierlich neue lern- und leistungsbezogene Kognitionen ausgelöst würden. In unserem Unterrichtsbeispiel ist möglicherweise der erste Gedanke eines Schülers mit geringer Selbstwirksamkeit: „Ohje, das verstehe ich schon wieder nicht“. Diese leistungsbezogene Kognition wirkt sich negativ auf sein emotionales Erleben aus; er fühlt sich frustriert und folgt den Erläuterungen und der weiteren Diskussion nicht, sondern beschäftigt sich mit etwas anderem. So verpasst er die Lerngelegenheit in der Klasse. Der Schüler, den wir eingangs mit seinem Unterrichtsbeitrag zitiert haben, erlebt dagegen Freude an seiner Erkenntnis und hat eine hohe Selbstwirksamkeitserwartung. Deshalb meldet er sich und wird schließlich auch aufgerufen. Er ist zufrieden mit

seinem Beitrag und erlebt in Folge Stolz und Freude, was ihn wiederum motiviert, der anschließenden Diskussion aufmerksam zu folgen und sich erneut mit einer Wortmeldung einzubringen. Dass Kognitionen, Emotionen und Motivation also – wie anhand des Beispiels verdeutlicht – kontinuierlich reziprok miteinander interagieren, könnte dafür sprechen, motivationales und emotionales Erleben als Teil der Nutzung zu konzeptualisieren (wie es bei Klieme et al., 2006, und Seidel, 2014, der Fall ist), da es nur schwer von den kognitiven Prozessen zu trennen ist[1]. Dabei sollte aber deutlich gemacht werden, dass Emotionen und Motivationen nicht als alternative Nutzungsprozesse aufzufassen sind (was hieße, dass entweder Kognitionen oder Emotionen oder Motivation den Effekt auf Lernergebnisse vermitteln würden), sondern dass Nutzung vielmehr erst aus der reziproken Wechselwirkung all dieser Prozesse miteinander resultiert und Lernen ohne Kognitionen nicht stattfinden kann.

Neben Kognitionen, Emotionen und Motivation werden in manchen Angebots-Nutzungs-Modellen auch noch *äußere Aktivitäten* im Feld der Nutzung aufgeführt, zum Beispiel „sozialer Austausch" bei Kunter und Trautwein (2013, S. 17) oder „Zuhören", „Ausprobieren", „in Gruppen arbeiten" und „sich an Gesprächen beteiligen" bei Seidel (2014, S. 858). Diese Handlungen von Schüler*innen können allerdings z. T. auch eine Lerngelegenheit für andere Schüler*innen im Klassenraum darstellen und somit als Teil des Angebots aufgefasst werden. In unserem Unterrichtsbeispiel macht die Beteiligung des zitierten Schülers am Unterrichtsgespräch mit dem Beitrag, „es gibt praktisch nur ‚Ausgangsparabeln' und die kann man dann ja beliebig stauchen und strecken", natürlich seine eigenen kognitiven Lernprozesse sichtbar, er könnte damit aber eben auch bei Mitschüler*innen Lernprozesse anregen (s. o.). Zudem kann auch Schüler*innenverhalten in der Klasse, welches selbst keine Lerngelegenheit für andere darstellt, trotzdem die Lerngelegenheiten aller beeinflussen und insofern das Angebot mitgestalten: Ein im Klassenraum herumirrender Schüler bietet vermutlich keine Lerngelegenheit im Sinne eines von der Lehrkraft intendierten Lernziels für andere Schüler*innen, aber er lenkt sie möglicherweise ab. Oder die Lehrperson ermahnt ihn lautstark und unterbricht damit die Lernaktivität anderer Schüler*innen. Damit hat das Herumirren einen zentralen Einfluss auf das Angebot. Wenngleich äußere Aktivitäten häufig als Indikatoren für Nutzung verwendet werden, spricht viel dafür, sie eher als Teil des Angebots aufzufassen: Als Indikatoren für innere Aktivitäten sind sie nicht immer eindeutig (beispielsweise kann es gut sein, dass der im Klassenraum herumirrende Schüler selber abgelenkt ist, es ist aber genauso denkbar, dass er bloß das Bedürfnis hatte, sich beim Nachgrübeln über eine fachliche Frage zu bewegen). Vor allem aber sind äußere Aktivitäten immer als Teil der Interaktion aufzufassen, in der der Unterricht ko-produziert wird, und können selber auch ‚Lernangebot' sein oder dieses zumindest beeinflussen. Um Ambivalenz zu vermeiden, könnte es folglich hilfreich sein, nur die inneren kognitiven, motivationa-

1 Die Verknüpfung zwischen Emotionen und Kognitionen ist bei „hot cognitions" stärker als bei „cold cognitions" (Abelson, 1963), doch auch das Auftreten letzterer ist im Lernprozess nicht ganz unabhängig von emotionalen und motivationalen Prozessen.

len und emotionalen Prozesse als Nutzung aufzufassen und äußere Aktivitäten zwar als Resultat von inneren Nutzungsprozessen, aber nicht als Nutzung selber zu konzeptualisieren.

3.2 Zusammenhänge zwischen Angebot und Nutzung

Neben der Frage nach einer Klärung der Begriffe Unterricht, Angebot und Nutzung, stellt sich auch die Frage, wie Angebot und Nutzung zusammenhängen. Selbstverständlich ist, dass das Angebot einen Effekt auf die Nutzung hat. Nutzung kann aber auch auf das Angebot zurückwirken: Es stellt die Grundlage für das Handeln der Schüler*innen im Unterricht dar und dieses beeinflusst wiederum das Lehrendenhandeln in der Interaktion. Hätten die Schüler*innen in unserem Beispiel nicht mitgedacht, dann hätten sie auch keine Parabeln an die Tafel zeichnen und der Lehrerin nicht antworten können. Dann hätte die Lehrerin ihre Frage vielleicht nochmal anders gestellt oder selber die Antwort gegeben. Auf alle Fälle wäre das Lernangebot in dieser Stunde ein anderes gewesen, wenn die Nutzung anders gewesen wäre. Entsprechend finden sich auch in den meisten Angebots-Nutzungs-Modellen – außer jenen von Helmke (2003, 2007, 2010) sowie Kunter und Trautwein (2013) – in beide Richtungen weisende Pfeile zwischen Angebot und Nutzung, die eine reziproke Beziehung zwischen Angebot und Nutzung illustrieren.

In einigen Publikationen wird zudem die Annahme formuliert, dass die Wirkung des Angebots auf die Nutzung in Abhängigkeit von Merkmalen der Lernenden variieren kann bzw. dass Angebot und Schülermerkmale in Bezug auf ihren Effekt auf die Nutzung interagieren können (Brühwiler & Blatchford, 2011, S. 97; Helmke & Weinert, 1997, S. 140; Helmke, 2010, S. 75; Reusser, 2009, S. 892). In unserem Unterrichtsbeispiel dürfte das Potenzial der Eingangsfrage auch davon abhängen, ob ein*e Schüler*in weiß, wie eine Parabel mathematisch definiert ist. In einem Oberstufenkurs hätte die Frage weniger Interesse geweckt, in einer Grundschulstunde hätte sie die Schüler*innen vermutlich überfordert. Solche Interaktionseffekte werden allerdings in keinem der Modelle grafisch dargestellt. Zugegebenermaßen kann das Modell dadurch unübersichtlich werden. Die bisherige Darstellung lässt jedoch den Eindruck entstehen, das Angebot habe bei alle Schüler*innen dieselbe Wirkung auf die Nutzung.

3.3 Die Bedeutung subjektiver Wahrnehmung und Interpretation

In manchen Varianten des Angebots-Nutzungs-Modells von Helmke (2003, 2010) wird der Zusammenhang zwischen Angebot und Nutzung nicht bzw. nicht nur durch Motivation und Emotionen vermittelt, sondern (auch) durch die Wahrnehmung und Interpretation des Angebots durch die Schüler*innen. Eine solche Differenzierung zwischen dem Angebot und dessen Wahrnehmung wurde bereits im Prozess-Mediations-Produkt-Paradigma der 1970er Jahre vorgenommen (z. B. bei Rothkopf, 1976). Die Wahrnehmung

beruht unmittelbar auf sensorischen Informationen und wird unter Einbezug von Vorerfahrungen und Wissen sowie in Abhängigkeit von den eigenen Prädispositionen interpretiert (Pekrun, 1988). Die konstruktivistische Annahme, dass dabei nicht einfach ein objektives Abbild der Welt entsteht, sondern eine subjektiv geprägte Interpretation konstruiert wird, ist eine Grundlage für die von Fend (1981) mit Bezug zu Luhmann erwähnte „doppelte Kontingenz" des Unterrichtsgeschehens. Sich den Zwischenschritt der Wahrnehmung und Interpretation einer stets uneindeutigen Situation bewusst zu machen, kann helfen zu verstehen, warum Kommunikation im Unterricht manchmal misslingt. In unserem Unterrichtsbeispiel wäre genaugenommen die korrekte Antwort der Schüler*innen auf die Frage der Lehrerin, wie lange sie jetzt weiter Parabeln an die Tafel zeichnen können, „bis in alle Unendlichkeit", gewesen. Der antwortende Schüler hat wohl weitergedacht und vermutet, dass die Lehrerin auf die Konzepte „Streckung" und „Stauchung" hinaus will, von denen er offensichtlich schon mal gehört hatte. In einer anderen Klasse hätte vielleicht niemand verstanden, worauf die Lehrerin abzielte und es wäre zu einer Unterbrechung im Unterrichtsfluss gekommen. So haben Wahrnehmung und Interpretation durch individuelle Akteure stets einen großen Einfluss auf den Verlauf der Unterrichtsinteraktion. Allerdings interpretieren natürlich nicht nur Schüler*innen die Geschehnisse im Klassenzimmer, sondern auch Lehrende. Z. B. wird die Lehrerin in unserem Beispiel eine Vorstellung davon entwickelt haben, was der Schüler mit seiner Formulierung „Winkel" einer Parabel" gemeint haben könnte, und seine Aussage je nach Interpretation als „richtig" oder „falsch" klassifiziert haben (in unserem Beispiel als „richtig"). Folglich wäre es unseres Erachtens sinnvoll, neben der Wahrnehmung und Interpretation der Unterrichtssituation durch die Lernenden auch ein Feld für die Wahrnehmung und Interpretation der Unterrichtssituation durch die Lehrenden im Angebots-Nutzungs-Modell zu ergänzen.[2]

3.4 Die Verortung von Angebot und Nutzung im Mehrebenensystem

Die Frage nach der Verortung von Angebot und Nutzung im Mehrebenensystem wird nur in zwei Modellen thematisiert: Brühwiler und Blatchford (2011) ordnen das Angebot der Klassenebene und die Nutzung der Individualebene zu, während Reusser und Pauli (2003, 2010) beide Ebenen nicht trennen.

Gegen eine ausschließliche Verortung des Angebotes auf Klassenebene, wie sie bei Brühwiler und Blatchford zu finden ist, lässt sich einwenden, dass das Angebot in vielerlei Hinsicht innerhalb von Klassen variiert: Besonders offensichtlich wird dies, wenn systematisch Methoden der Binnendifferenzierung und Individualisierung eingesetzt werden, also kein Klassengespräch wie in unserem Beispiel stattfindet, sondern Schüler*innen beispielsweise unterschiedliche Aufgaben bearbeiten (z. B. Bönsch, 1995).

2 Weiterhin stellt diese Überlegung in Frage, ob es eine objektive Messung des Angebotes überhaupt geben kann, oder ob nicht jede Beurteilung (auch jene durch Forschende) eine subjektive Konstruktion des Unterrichts reflektiert.

Individuelle Unterschiede hinsichtlich der Lern-, Beziehungs- und Unterstützungsangebote können sich aber auch unbeabsichtigt ergeben, etwa wenn Lehrende einen wärmeren emotionalen Umgang mit ihren „Lieblingsschüler*innen“ pflegen (Rosenthal, 1994), anders auf Störungen reagieren je nachdem wer deren Urheber*in war (Skiba, Michael, Nardo & Peterson, 2002), oder im Unterrichtsgespräch vorwiegend mit wenigen leistungsstarken Schüler*innen interagieren (z. B. Lipowsky, Rakoczy, Pauli, Reusser & Klieme, 2007). Folglich spricht viel dafür, das Unterrichtsangebot auf beiden Ebenen – der Klassen- und der Individualebene – zu konzeptualisieren.

Die Nutzung beschreibt dagegen eindeutig einen individuellen Prozess. Kognitionen, motivationales und emotionales Erleben hängen nicht nur vom Unterrichtsangebot, sondern auch von dessen Wahrnehmung und Interpretation durch die Lernenden sowie von ihren individuellen Lernvoraussetzungen ab. Folglich variieren sie interindividuell in derselben Unterrichtssituation. Dies gilt auch für äußere Aktivitäten (z. B. Beiträge zum Plenumsgespräch oder zu einer Gruppenarbeit leisten) – sofern diese als Nutzung verstanden werden. Allerdings können äußere Aktivitäten auch Lernangebote für die Mitlernende darstellen und als solche dann wiederum (auch) auf der Klassenebene konzeptualisiert werden (s. o.).

3.5 Die Bedeutung der Kontexte

Fend (2008a, 2008b) beschreibt die Beziehung zwischen den Ebenen mit dem Konzept der Rekontextualisierung: Auf höheren Ebenen des Schulsystems werden Handlungsrahmen vorgegeben, welche jedoch nicht eins zu eins umgesetzt, sondern von den Akteuren aktiv an die lokalen Handlungsbedingungen adaptiert werden. Dabei spielen auch ihre individuelle Wahrnehmung und Interpretation der Situation sowie ihre Ressourcen (Kompetenzen u. a.) eine Rolle. Gelingt die Adaptation nicht auf zufriedenstellende Weise, so können die Akteure zudem Einfluss auf höhere Ebenen des Systems ausüben, damit der Handlungsrahmen besser an die Bedingungen in der Praxis angepasst wird. Insofern können Beziehungen zwischen Ebenen reziprok sein: Kontexte beeinflussen nicht nur das Handeln von Akteuren, sondern diese können auch handelnd Kontexte verändern. Allerdings wird diese Reziprozität in der Publikation von Fend (2008b) grafisch nicht dargestellt. Auch in den meisten anderen Angebots-Nutzungs-Modellen finden sich unidirektionale Pfeile. Eine Ausnahme stellen nur Brühwiler und Blatchford (2011) sowie Seidel (2014) dar, in deren Modellen alle Effekte reziprok sind (außer jenen des Bildungssystems). In keinem der Modelle sind zudem Interaktionen zwischen Vorgaben auf höheren Ebenen und Merkmalen von Akteuren dargestellt, die Fend (2008b) ebenfalls nahelegt, wenn er die Bedeutung individueller Ressourcen für die Umsetzung von Vorgaben beschreibt. Beide Ergänzungen erscheinen insofern theoretisch sinnvoll.

3.6 Aufgeführte Kriterien unterrichtlicher Wirksamkeit

Angebots-Nutzungs-Modelle sind multikriterial, das heißt, sie führen jeweils mehrere unterrichtliche Wirkungen auf. In allen Modellen wurden fachliche Leistungen und Kompetenzen genannt, in den meisten Modellen auch überfachliche Kompetenzen, wie Problemlösefähigkeit oder soziale Kompetenz. Zudem wurden Interesse und Motivation als Zielkriterien aufgeführt (nämlich bei Lipowsky, 2006; Reusser & Pauli, 2003; Seidel, 2014) sowie erzieherische Wirkungen (bei Helmke, 2007, 2010; Reusser & Pauli, 2010). Folglich scheint das in der Vergangenheit, beispielsweise von Oser, Dick und Patry (1992) aufgeworfene Problem einer einseitigen Konzentration der Effektivitätsforschung auf das ‚Produkt' Leistung in Angebots-Nutzungs-Modellen überwunden. Allerdings unterscheiden sich die Modelle in Bezug auf die konkrete Auswahl der Zielkriterien und diese erscheint insgesamt eher willkürlich – zumindest wird ihre theoretische und normative Grundlage für keines der Modelle expliziert. Dies kann insofern problematisiert werden als einerseits die Inhalte aller anderen Felder von der Auswahl dieser Kriterien abhängen dürften (Helmke, 2002; Klieme, 2006; Reusser, 2009) und andererseits diese Auswahl eine stark normative Komponente hat. Zwar kann durchaus auch empirisch bestimmt werden, welche Funktionen Schulunterricht für die Gesellschaft und/oder für Individuen erfüllt, wie es etwa Fend (1981) getan hat. Die Frage, ob er auch genau diese Funktionen erfüllen *sollte,* die implizit im Effektivitätsbegriff steckt (siehe z. B. Klieme & Tippelt, 2008), erfordert aber trotzdem eine normative Setzung. Deshalb wäre eine explizitere Begründung der Auswahl von Wirkungskriterien im Angebots-Nutzungs-Modell wünschenswert.

4. Fazit und Vorschlag eines integrierten Modells

Fasst man die Überlegungen der vorangegangenen Abschnitte zusammen, so lässt sich zunächst feststellen, dass verschiedene Angebots-Nutzungs-Modelle grundlegende Unterschiede aufweisen. Was bedeuten diese Unterschiede für Forschung und Praxis?

In der Forschung hat die Wahl des Angebots-Nutzungsmodells vor allem Implikationen für das Studiendesign: Soll zur Beschreibung des Angebotes Lehrendenhandeln beobachtet werden oder die gesamte Interaktion im Klassenzimmer? Lassen sich Angebot und Nutzung überhaupt getrennt erfassen? Wird dabei ausschließlich die Klassenebene betrachtet oder wird berücksichtigt, dass Lehrende mit jeder*m Schüler*in anders interagieren? Macht es Sinn, die Bedeutung von Kognitionen, Emotionen oder Motivation der Schüler*innen isoliert zu untersuchen? Sollen die Effekte unidirektional oder reziprok modelliert werden und werden auch Interaktionen untersucht? Welche Wirkungen werden betrachtet? Auf all diese Fragen liefern die existierenden Angebots-Nutzungs-Modelle unterschiedliche Antworten. Ein Anliegen unseres Artikels ist es, Forschenden zu helfen, sich bewusst und aufgrund theoretischer Erwägungen für das eine oder andere Angebots-Nutzungs-Modell zu entscheiden. Zudem schlagen wir im Folgenden ein eigenes integriertes Modell vor.

Für Lehrende sind die Unterschiede zwischen Modellen ebenfalls relevant: Angebots-Nutzungs-Modelle implizieren generell eine ausgeglichene Ergebnisverantwortung; diese liegt weder alleine bei den Lehrenden noch alleine bei den Schüler*innen. Allerdings implizieren Modelle, in denen Unterricht und Angebot sowie Angebot und Lehrerhandeln gleichgesetzt werden, eine stärkere Verantwortung der Lehrenden. Schüler*innen wird in diesem Fall nur Verantwortung für die eigenen Nutzungsprozesse zugeschrieben. Wenn dagegen das Angebot und damit auch der Unterricht im Modell als ko-konstruiert konzeptualisiert werden, wenn also angenommen wird, dass der Verlauf des Unterrichts davon beeinflusst wird, wie sich Schüler*innen am Unterricht beteiligen, dann haben die Schüler*innen auch eine Mitverantwortung für das Angebot. Weiterhin kann die Aufnahme von Feldern der Wahrnehmung und Interpretation des Angebotes darauf aufmerksam machen, dass unerwartetes und unerwünschtes Schüler*innen-Handeln nicht absichtliche Verweigerung bedeuten muss, sondern das Resultat von Missverständnissen und/oder einer Adaptation der Vorgaben an die individuellen Handlungsbedingungen der Schüler*innen sein kann. Die Aufnahme der beiden Felder verdeutlicht zudem, dass auch Lehrende die Schüler*innen missverstehen können. So könnten Lehrende ermutigt werden, genauer hinzuhören und nachzufragen, was die Schüler*innen meinen oder warum sie etwas tun. Die Darstellung von Interaktionseffekten im Modell kann schließlich auf die Bedeutung der (Makro- und Mikro-)Adaptivität von Unterricht hinweisen.

Folglich können die beschriebenen Unterschiede zwischen verschiedenen Angebots-Nutzungs-Modellen grundsätzlich andere Forschungsdesigns und Strategien zur Verbesserung von Unterricht in der Praxis nahelegen. Deshalb wird im Folgenden ein integriertes Angebots-Nutzungs-Modell präsentiert, das die im vorangegangenen Text dargestellten Überlegungen aufgreift (Abb. 1). Dabei fokussieren wir auf Prozesse des Unterrichts und vernachlässigen eine genauere Beschreibung der Prozesse in den Kontexten, da das Modell sonst zu umfangreich würde. Das Modell bleibt zudem abstrakt, um je nach Fach, Wirkungskriterium und Zweck mit spezifischen Konstrukten und theoriegeleiteten Hypothesen bzw. empirisch fundierten Erkenntnissen gefüllt werden zu können.

Unterricht wird in unserem Modell konzeptualisiert als Interaktion von Lehrenden, Lernenden und einem Unterrichtsgegenstand. Wir gehen nicht davon aus, dass es nur die Lehrenden sind, die das Lernangebot für Lernende gestalten, sondern wir betrachten Unterricht als fachspezifische Ko-Konstruktion. Entsprechend nehmen wir eine reziproke Beziehung zwischen dem Lernangebot und dessen Nutzung durch die Schüler*innen an, wobei diese Beziehung vermittelt wird über die Wahrnehmung und Interpretation der Geschehnisse durch die Schüler*innen. Nutzung verstehen wir als komplexen Prozess, in dem sich Kognitionen, emotionales und motivationales Erleben ständig gegenseitig beeinflussen. Das Angebot wird weiterhin auch von den Lehrenden wahrgenommen und interpretiert, wodurch auch bei ihnen kognitive Prozesse, motivationales und emotionales Erleben ausgelöst werden, was dann wiederum die Basis für ihr Handeln, also ihre Reaktionen auf die Geschehnisse im Klassenzimmer bilden. Merkmale der Lehrenden und der Schüler*innen haben einen Einfluss auf Angebot und Nutzung, aber

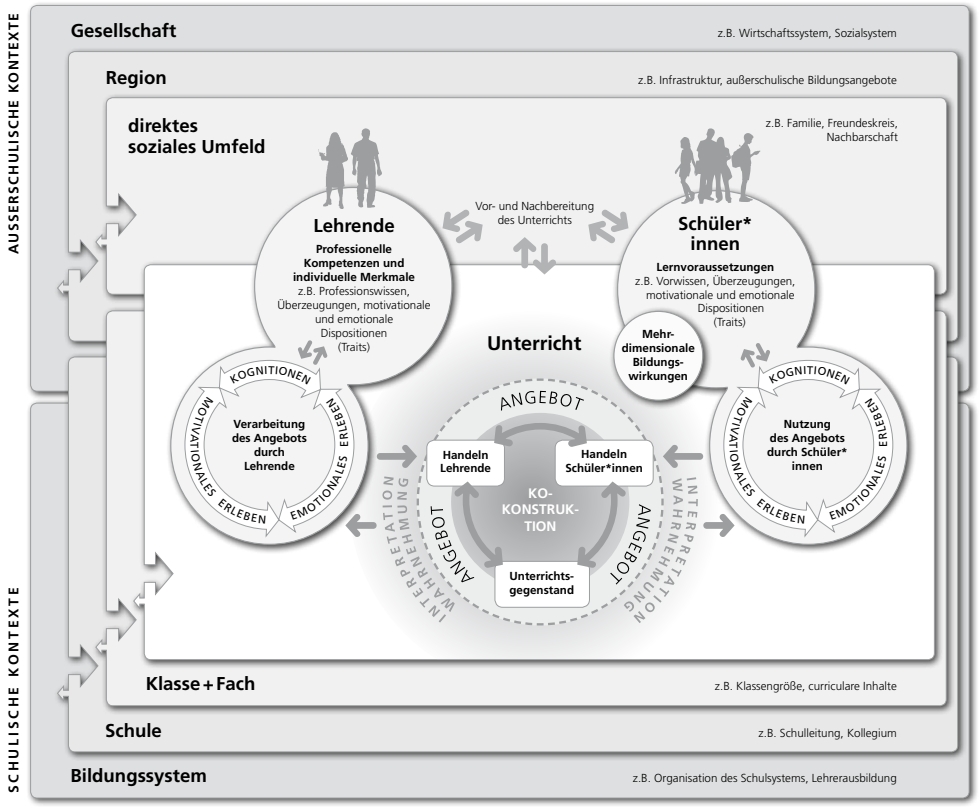

Abb. 1: Integriertes Angebots-Nutzungs-Modell der Wirkweise des Unterrichts

die Unterrichtsprozesse wirken auch wieder auf diese individuellen Merkmale zurück. Der Unterricht wird zudem auch außerhalb der Schule vor- und nachbereitet. Gerahmt wird er durch Kontexte auf verschiedenen Ebenen: die individuellen Lebenskontexten der Schüler*innen und Lehrenden, der Kontext der Klasse und des Faches, der Kontext der Schule und der Kontext des Bildungssystems. Wir gehen davon aus, dass auf höheren Ebenen jeweils Vorgaben gemacht und Handlungsrahmen gesteckt werden, dass diese jedoch von den Akteuren rekontextualisiert werden. Effekte zwischen Ebenen sind weiterhin potenziell reziprok (Akteure können auch Kontexte beeinflussen) und komplexe Interaktionen zwischen Feldern sind wahrscheinlich, weshalb die Kontexte als Rahmen gezeichnet werden, statt spezifische Effekte dieser Kontexte darzustellen.

Literatur

Abelson, R. P. (1963). Computer simulation of „hot cognition". In S. S. Tomkins & S. Messick, (Eds.), *Computer simulation of personality: Frontier of psychological theory* (pp. 277–302). New York: Wiley.

Berliner, D. C. (1992). The nature of expertise in teaching. In F. Oser, A. Dick & J.-L. Patry, (Eds.), *Effective and responsible teaching: The new synthesis* (S. 227–248). San Francisco: Jossey-Bass.

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

Bönsch, M. (1995). *Differenzierung in Schule und Unterricht.* München: Ehrenwirth Verlag.

Bromme, R. (1992). *Der Lehrer als Experte: zur Psychologie des professionellen Wissens.* Bern: Huber.

Brühwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction, 21,* 95–108.

De Haan, G., & Rülcker, T. (2009). *Der Konstruktivismus als Grundlage für die Pädagogik* (Berliner Beiträge Bd. 7). Frankfurt a. M.: Peter Lang.

DESI-Konsortium (Hrsg.) (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie.* Weinheim: Beltz.

Doyle, W. (1977). Paradigms for research on teacher effectiveness. *Review of Research in Education, 5,* 163–198.

Dunkin, M., & Biddle, B. (1974). *The study of teaching.* New York: Holt, Rinehart, & Winston.

Fend, H. (1981). *Theorie der Schule.* München: Urban & Schwarzenberg.

Fend, H. (1982). *Gesamtschule im Vergleich. Bilanz der Ergebnisse des Gesamtschulversuchs.* Weinheim/Basel: Beltz.

Fend, H. (1998). *Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistungen.* Weinheim/München: Juventa.

Fend, H. (2008a). *Neue Theorie der Schule. Einführung in das Verstehen von Bildungssystemen* (2. Aufl.). Wiesbaden: VS Verlag.

Fend, H. (2008b). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität.* Wiesbaden: VS Verlag.

Gage, N. L., & Needles, M. C. (1989). Process-product research on teaching: a review of criticisms. *The Elementary School Journal, 89,* 253–300.

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

Helmke, A. (2002). Kommentar: Unterrichtsqualität und Unterrichtsklima – Perspektiven und Sackgassen. *Unterrichtswissenschaft, 30*(3), 261–277.

Helmke, A. (2003). *Unterrichtsqualität: Erfassen, Bewerten, Verbessern.* Seelze: Kallmeyersche Verlagsbuchhandlung.

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.* Seelze: Klett-Kallmeyer.

Helmke, A., Piskol, K., Pikowsky, B., & Wagner, W. (2009). Schüler als Experten von Unterricht. Unterrichtsqualität aus Schülerperspektive. *Lernende Schule, 46-47,* 98–105.

Helmke, A., & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Hrsg.), *Enzyklopädie der Psychologie. Band 3: Psychologie der Schule und des Unterrichts* (S. 71–176). Göttingen: Hogrefe-Verlag.

Jurik, V., Seidel, T., & Gröschner, A. (2012). Was wissen wir über Lehrerhandeln im Unterricht? *Pädagogik, 64*(2), 42–45.

Klieme, E. (2006). Empirische Unterrichtsforschung: aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde. Einführung in den Thementeil. *Zeitschrift für Pädagogik, 52,* 765–773

Klieme, E., & Tippelt, R. (Hrsg.) (2008). *Qualitätssicherung im Bildungswesen. Eine aktuelle Zwischenbilanz* (53. Beiheft der Zeitschrift für Pädagogik). Weinheim u. a.: Beltz.

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. 128–146). Münster: Waxmann.

Kohler, B., & Wacker, A. (2013). Das Angebots-Nutzungs-Modell. Überlegungen zu Chancen und Grenzen des derzeit prominentesten Wirkmodells der Schul- und Unterrichtsforschung. *Die Deutsche Schule, 105,* 241–257.

Krapp, A. (1999). Intrinsische Lernmotivation und Interesse. Forschungsansätze und konzeptuelle Überlegungen. *Zeitschrift für Pädagogik, 45,* 387–406

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Hrsg.) (2011). *Professionelle Kompetenz von Lehrkräften – Ergebnisse des Forschungsprogramms COACTIV.* Münster: Waxmann.

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

Lipowsky, F. (2015). Unterricht. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (2. überarb. Auflage, S. 69–105). Heidelberg: Springer.

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* (51. Beiheft der Zeitschrift für Pädagogik, S. 47–70). Weinheim/Basel: Beltz.

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 the Pythagorean Theorem, *Learning and Instruction, 19*(6), 527–537.

Lipowsky, F., Rakoczy, K., Pauli, C., Reusser, K., & Klieme, E. (2007). Gleicher Unterricht – gleiche Chancen für alle? Die Verteilung von Schülerbeiträgen im Klassenunterricht. *Unterrichtswissenschaft, 35,* 125–147.

Lüders, M. (2012). Der Unterrichtsbegriff in pädagogischen Nachschlagewerken. Ein empirischer Beitrag zur disziplinären Entwicklung der Schulpädagogik. *Zeitschrift für Pädagogik, 58,* 109–129.

Luhmann, N. (2002). *Das Erziehungssystem der Gesellschaft.* Frankfurt a. M.: Suhrkamp Verlag.

Mayring, P., & Rhoeneck, C. v. (Hrsg.) (2003). *Learning emotions.* Bern u. a.: Lang.

Oser, F., Dick, A., & Patry, J.-L. (Hrsg.) (1992). *Effective and responsible teaching. The new synthesis.* San Francisco: Jossey-Bass.

Pekrun, R. (1988). *Emotion, Motivation und Persönlichkeit.* München/Weinheim: Psychologie Verlags Union.

Pekrun, R., & Jerusalem, M. (1996). Leistungsbezogenes Denken und Fühlen: Eine Übersicht zur psychologischen Forschung. In J. Möller & O. Köller (Hrsg.), *Emotionen, Kognitionen und Schulleistungen.* Weinheim: Beltz.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95,* 667–686.

Reusser, K. (2009). Unterricht. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 881–896). Weinheim: Beltz.

Reusser, K., & Pauli, C. (2003). *Mathematikunterricht in der Schweiz und in weiteren sechs Ländern. Bericht über die Ergebnisse einer internationalen und schweizerischen Video-Unterrichtsstunde. Doppel-CD-Rom (unter Mitarbeit der Video-Projektgruppe des Pädagogischen Instituts der Universität Zürich).* Zürich: Universität Zürich, Pädagogisches Institut.

Reusser, K., & Pauli, C. (2010). Unterrichtsgestaltung und Unterrichtsqualität – Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht: Einleitung und Überblick. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 9–32). Münster: Waxmann.

Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science, 3,* 176–179.

Rothkopf, E. Z. (1976). Writing to teach and reading to learn: A perspective on the psychology of written instruction. In N. L. Gage (Ed.), *The psychology of teaching methods.* Chicago: University of Chicago Press.

Scharpf, F. W. (2000). *Interaktionsformen: Akteurzentrierter Institutionalismus in der Politikforschung.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Seidel, T. (2011). Lehrerhandeln im Unterricht. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 605–629). Münster: Waxmann.

Seidel, T. (2014). Angebots-Nutzungs-Modelle in der Unterrichtspsychologie: Integration von Struktur- und Prozessparadigma. *Zeitschrift für Pädagogik, 60*(6), 850–866.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the last decade: Role of theory and research design in disentangling meta-analysis results. *Review of Educational Research. 77*(4), 454–499.

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *The Urban Review, 34,* 317–342.

Winne, P. H. (1987). Why process-product research cannot explain process-product findings and a proposed remedy: The cognitive mediational paradigm. *Teaching and Teacher Education, 3,* 333–356.

Abstract: This paper discusses so-called ‘opportunity-use models of the effects of teaching’. More specifically, it compares different ‘opportunity-use-models’ with the aim to unveil existing conceptual fuzziness. The following issues will be critically discussed: a) the underlying conception of ‘teaching’, ‘opportunity’, and ‘use’, b) relations between ‘opportunity’ and ‘use’, c) the role of perception and interpretation, d) the localization of ‘opportunity’ and ‘use’ in a multilevel system, e) the role of the context, and f) the selection of teaching effectiveness criteria. In the conclusion an integrated opportunity-use model is developed.

Keywords: Opportunity-use Models of Teaching, Teaching Quality, (Co-)Construction, Multilevel Model, Criteria of Teaching Effectiveness

Anschrift der Autor_innen

Dr. Svenja Vieluf, DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation,
Rostocker Str. 6, 60323 Frankfurt a. M., Deutschland
E-Mail: vieluf@dipf.de

Prof. Dr. Anna-Katharina Praetorius, Universität Zürich,
Lehrstuhl für pädagogisch-psychologische Lehr-Lernforschung und Didaktik,
Institut für Erziehungswissenschaft,
Freiestrasse 36, 8032 Zürich, Schweiz
E-Mail: anna.praetorius@ife.uzh.ch

Prof. Dr. Katrin Rakoczy, HSD Hochschule Döpfer GmbH,
Waidmarkt 3 und 9, 50676 Köln, Deutschland
und DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation,
Rostocker Str. 6, 60323 Frankfurt a. M., Deutschland
E-Mail: rakoczy@dipf.de

Prof. Dr. Marc Kleinknecht, Leuphana Universität Lüneburg,
Institut für Bildungswissenschaft, Schulpädagogik und Schulentwicklung,
Universitätsallee 1, 21335 Lüneburg, Deutschland
E-Mail: marc.kleinknecht@leuphana.de

PD Dr. Marcus Pietsch, Leuphana Universität Lüneburg,
Universitätsallee 1, 21335 Lüneburg, Deutschland
E-Mail: marcus.pietsch@leuphana.de

Sibylle Meissner/Samuel Merk*/Benjamin Fauth/Marc Kleinknecht/ Thorsten Bohl*

Differenzielle Effekte der Unterrichtsqualität auf die aktive Lernzeit

Zusammenfassung: Angebot-Nutzungs-Modelle weisen Angebots- und Prozessmerkmale des Unterrichts sowie Prädispositionen der Lernenden aus, um Lernerfolge zu erklären. Zumeist werden hier direkte Effekte dargestellt, wohingegen differenzielle Effekte von Angebotsvariablen auf die Nutzung nicht abgebildet sind, obschon diese theoretisch häufig angenommen werden. Anhand des Konstrukts der aktiven Lernzeit wurde in dieser Studie untersucht, ob solche Interaktionen zwischen Unterrichtsqualität und Schüler*innenmerkmalen nachweisbar sind. Zwar zeigten sich in den Daten direkte Effekte von Klassenmanagement und dem eingeschätzten Leistungsniveau auf die aktive Lernzeit, ein entsprechender Nachweis für differenzielle Effekte konnte jedoch nicht erbracht werden.

Schlagworte: Aktive Lernzeit, Instruktionsqualität, individuelle Lernausgangslagen, Differenzielle Effekte, Angebot-Nutzungs-Modell

1. Einführung

Seit den 1980er Jahren dominieren in der deutschsprachigen Unterrichtsforschung sog. Angebot-Nutzungs-Modelle, welche Determinanten von Schulerfolg und deren Relationen untereinander ausweisen. Konzeptuell wird dabei zwischen unterrichtlichem Angebot, dessen Nutzung sowie daraus resultierender Erträge unterschieden (Überblick bei Vieluf, Praetorius, Rakoczy, Kleinknecht & Pietsch, in diesem Heft). Eine solche Systematisierung erlaubt eine Verständigung über das komplexe Wirkungsgefüge innerhalb von Bildungsprozessen und macht deutlich, dass neben Kompetenzen der Lehrkraft und Prozessmerkmalen des Unterrichts eine Vielzahl an individuellen und kontextuellen Faktoren Einfluss auf das schulische Lehren und Lernen nimmt. Zwar liegt es nahe anzunehmen, dass Lernende aufgrund ihrer individuellen Lernausgangslagen Unterrichtsqualität in unterschiedlicher Weise zu nutzen vermögen und sich daher auch differenzielle Effekte ergeben können (Vieluf et al., in diesem Heft; Kunter & Ewald, 2016), doch wurde diesem anzunehmenden Interaktionseffekt bislang vergleichsweise wenig Beachtung geschenkt. Die hier vorliegende Studie greift dieses Desiderat auf und untersucht anhand des Kriteriums der aktiven Lernzeit moderierende Einflüsse individueller Schüler*innenmerkmale auf die Nutzung des Unterrichtsangebotes.

* Es besteht eine geteilte Erstautorenschaft.

2. Theoretischer Hintergrund

2.1 Aktive Lernzeit als Konstrukt

Als aktive Lernzeit (auch *time on task* genannt) wird in der Literatur jenes Zeitmaß bezeichnet, das angibt, wie lange Lernende sich bewusst konkreten Lerninhalten widmen (Berliner, 1990; Carroll, 1963). Sie stellt folglich eine quantitativ-individuelle Größe dar, ist auf der Nutzungsseite zu verorten und von der *accounted time* und der *instructional time* auf Angebotsseite zu unterscheiden. Im Unterschied zur sog. *academic learning time* ist mit der aktiven Lernzeit zwar noch kein Qualitätsmoment verbunden (Berliner, 1990), gleichwohl ist sich die empirische Forschung darin einig, dass ein Mehr an aktiver Lernzeit durch die längere Beschäftigung mit dem Lerngegenstand den Lernfortschritt der Schüler*innen deutlich begünstigen kann (Berliner, 1990; Carroll, 1963; Helmke, 2012). Aktive Lernzeit gilt somit als zentraler Prädiktor für Lernerfolge (Klieme, 2018) und geht in signifikanter Weise mit Tiefenverarbeitung einher (Everaert, Opdecam & Maussen, 2017). Wie hoch das Maß an aktiver Lernzeit ausfällt, hängt von vielerlei Faktoren ab: Sowohl unterrichtliche und kontextuelle Faktoren als auch individuelle Merkmale der Schüler*innen beeinflussen das Ausmaß ihrer aktiven Lernzeit (Bloom, 1974; Romero & Barberà, 2011). Im Folgenden wird dargestellt, welche Merkmale sich auf unterrichtlicher Seite (2.2) und mit Blick auf die individuellen Lernausgangslagen (2.3) als einflussreich auf die aktive Lernzeit (im Sinne direkter Effekte) erwiesen haben. Weiterhin wird skizziert, welche differenziellen Effekte von Unterricht in der Forschung bislang bekannt sind (2.4).

2.2 Bedeutung unterrichtlicher Merkmale für die aktive Lernzeit

Innerhalb des deutschen Fachdiskurses haben sich drei Dimensionen von Unterrichtsqualität etabliert, die das Nutzungsverhalten der Schüler*innen positiv beeinflussen: kognitive Aktivierung, konstruktive Unterstützung von Schüler*innen sowie strukturierte Klassenführung (Klieme & Rakoczy, 2008; Fauth, Decristan, Rieser, Klieme & Büttner, 2014). Mit Blick auf die aktive Lernzeit ist der begünstigende Einfluss einer strukturierten Klassenführung sehr gut belegt, da eine störungspräventive Unterrichtsführung die tatsächliche Instruktionszeit erhöht, die wiederum Voraussetzung für das Maß an aktiver Lernzeit ist (Klieme & Rakoczy, 2008; Seidel & Shavelson, 2007; Marzano & Marzano, 2003). Auch die positive Wirkung der Dimension konstruktive Unterstützung ist gut belegt (z. B. Seiz, Decristan, Kunter & Baumert, 2016). Sie beschreibt jene Handlungen von Lehrkräften, die darauf abzielen, die Lernprozesse ihrer Schüler*innen zu befördern – sei dies durch soziale Aspekte (z. B. wertschätzende Beziehung, Ermunterung, Motivierung) oder durch Maßnahmen der Individualisierung (Anpassung des Schwierigkeitsgrads/Lerntempos, positive Fehlerkultur usw.; Kunter & Trautwein, 2013). Interessanterweise sind im US-amerikanischen Raum drei sehr ähnliche Dimensionen gängig: emotional support (Klassenklima u. Schülerorientierung), classroom organisa-

tion (Verhaltensregulation, Unterrichtsregie u. Produktivität) und instructional support (Feedback, Sprachmodellierung u. mentale Konzeptentwicklung) (z. B. Hamre et al., 2013).

2.3 Bedeutung individueller und familiärer Merkmale für die aktive Lernzeit

Individuelle Lernausgangslagen werden in allen etablierten Angebot-Nutzungs-Modellen als auf die Nutzung maßgeblich einwirkende Variable ausgewiesen (vgl. Vieluf et al., in diesem Heft). Der Einfluss kognitiver (z. B. Intelligenz, Vorwissen), motivational-volitionaler (z. B. Neugier, Erfolgszuversicht) sowie sozial-emotionaler Faktoren (z. B. Gewissenhaftigkeit, Impulskontrolle) auf Lernzuwächse ist empirisch sehr gut belegt (Überblick bei Hasselhorn et al., 2014). Individuelle Lernausgangslagen lassen sich folglich als differenzielles Lernpotenzial betrachten, das wiederum über familiäre Faktoren (z. B. sozioökonomischer Hintergrund, Ethnie) beeinflusst wird (Kunter & Trautwein, 2013). Dabei gilt als empirisch abgesichert, dass Kinder aus höheren Sozialschichten mit höherem Bildungshintergrund hinsichtlich ihrer schulischen Leistungen begünstigt sind (z. B. Conger, Conger & Martin, 2010).

In Carrolls ‚Model of school learning' (1963) wird der Einfluss des Lernpotenzials auf die aktive Lernzeit ebenfalls dargestellt, indem die Determinanten von tatsächlich aufgewendeter Lernzeit und individuell benötigter Zeit (Lerngeschwindigkeit) in Relation gesetzt werden: Die Ausdauer und Lernbereitschaft an einer Aufgabe zu arbeiten sowie die im Unterricht dafür individuell zugestandene Zeit nehmen Einfluss auf die aufgewendete Lernzeit eines Schülers/einer Schülerin. Benötigt ein*e Schüler*in mehr Zeit für die Aufgabenbewältigung als ihm/ihr zugestanden wird, kommt es zu Lernschwierigkeiten; wird ihm/ihr deutlich mehr Zeit zur Verfügung gestellt als er/sie aufgrund seiner/ihrer Lernausgangslagen benötigt, kommt es zu Überdruss. Idealerweise wenden Schüler*innen also für die Bearbeitung einer Aufgabe exakt so viel Zeit auf, wie sie aufgrund ihrer individuellen Lernausgangslagen benötigen (Hasselhorn & Gold, 2009). Ein solches Prinzip der individuellen Passung liegt dem Ansatz des adaptiven Unterrichts zugrunde (Klieme & Warwas, 2011). Maßnahmen der Individualisierung können somit zu einer höheren aktiven Lernzeit führen.

2.4 Differenzielle Wirkungen von Unterricht

Im Rahmen der Aptitude-Treatment-Interaction-Forschung erhielt die Frage der Passung zwischen Unterrichtsangebot (treatment) und den individuellen Lernausgangslagen (aptitudes) erstmals Aufmerksamkeit (Snow, 1991). Während in den ersten Jahren die Erforschung der Wirksamkeit von Sichtstrukturen im Vordergrund standen, liegt das Augenmerk der aktuellen ATI-Forschung auf tiefenstrukturellen Merkmalen (Seiz et al., 2016; Kunter & Voss, 2011). Es zeigte sich, dass insbesondere Schüler*innen mit bildungsrelevanten Risiken (Migration, niedriger SES, geringe Vorleistungen) von einer

hohen Unterrichtsqualität profitieren: Durch hohe Steuerung und Strukturierung können kompensatorische Wirkungen für Risikoschüler*innen eintreten (z. B. Cadima, Leal & Burchinal, 2010; Hamre & Pianta, 2005; Kunter & Ewald, 2016). Für Schüler*innen mit günstigen Lernausgangslagen sind hingegen erhöhte Freiheitsgrade von Vorteil; zudem lassen sie sich in ihrer Performanz durch niedrige Unterrichtsqualität weniger beeinflussen (Snow, 1991; Hahn, Rohlfs, Wacker & Bohl, 2016).

3. Forschungsfragen

Empirisch lassen sich mit Blick auf die Bedingtheit von aktiver Lernzeit somit Faktoren identifizieren, die sich sowohl der Angebots- als auch der Nutzungsseite zuordnen lassen. Ein Nachweis der theoretisch plausiblen differenziellen Effekte auf das Nutzungsverhalten (Brühwiler & Blatchford, 2011, S. 97; Helmke, 2012, S. 71; Reusser, 2009, S. 892) steht im Kontext der aktiven Lernzeit indes noch aus. Die hier dargelegte Studie nimmt sich dieser Forschungslücke an und widmet sich folgenden Forschungsfragen:

1) Inwiefern kann das Ausmaß an zu beobachtender aktiver Lernzeit durch Merkmale der Unterrichtsqualität erklärt werden?
   Hypothese:
   Wir erwarten, dass insbesondere die Dimensionen *Klassenführung/Strukturierung* und *Individualisierung* die aktive Lernzeit vorhersagen können. Auch von der Dimension *Motivierung* erwarten wir prädiktive Kraft.
2) Inwiefern bedingen individuelle Merkmale auf Seiten der Nutzer*innen das Ausmaß ihrer zu beobachtenden aktiven Lernzeit?
   Hypothese:
   Schüler*innen mit günstigen Lernvoraussetzungen (leistungsstark, hoher SES) werden höhere Ausprägungen der aktiven Lernzeit aufweisen.
3) Interagieren Effekte der Unterrichtsqualität auf die zu beobachtende aktive Lernzeit mit den individuellen Lernausgangslagen?
   Hypothese:
   Gerade Schüler*innen mit ungünstigen Voraussetzungen (leistungsschwach, niedriger SES) werden hinsichtlich ihrer aktiven Lernzeit von einer hohen Unterrichtsqualität profitieren.

4. Methoden

4.1 Design und Stichprobe

Zur Beantwortung der Forschungsfragen wird auf einen Teildatensatz des in Baden-Württemberg (D) durchgeführten Forschungsprojektes WissGem (Bohl & Wacker, 2016) zurückgegriffen und eine Reanalyse der Daten vorgenommen. In einem mess-

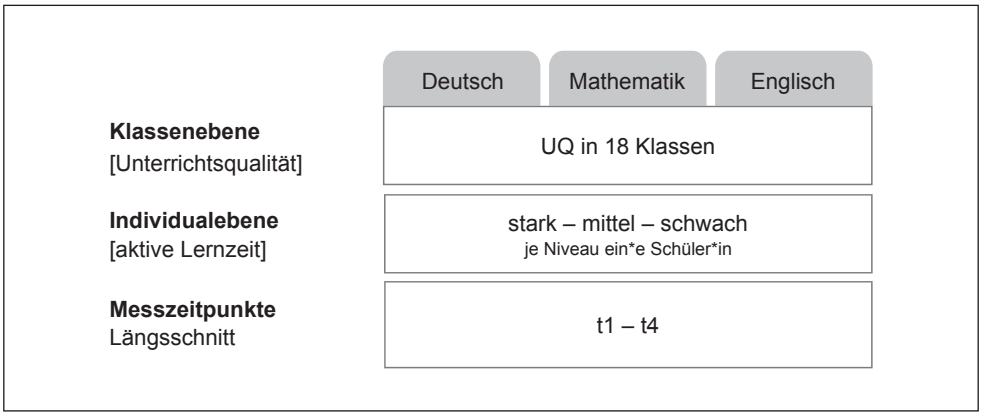

*Abb. 1: Erhebungsdesign. Klassenebene: In jedem Fach wurde in 18 Klassen die Unterrichtsqualität gemessen; Individualebene: Je Klasse und Fach wurden bei Schüler*innen unterschiedlicher Leistungsniveaus (stark, mittel, schwach) zusätzlich drei Individualbeobachtungen bzgl. ihrer aktiven Lernzeit vorgenommen; Messzeitpunkte: Beide Variablen (UQ u. aktive Lernzeit) wurden im Längsschnitt zu vier Messzeitpunkten erhoben*

wiederholten Design wurden in N = 18 Klassen im Deutsch-, Mathematik- und Englischunterricht der Klassenstufe 6 über vier Erhebungszeiträume hinweg sowohl die Unterrichtsqualität als auch die aktive Lernzeit ausgewählter Schüler*innen über Unterrichtsbeobachtung im Feld erfasst. Um der Frage nachgehen zu können, inwiefern die Unterrichtsqualität *differenziell* mit dem Nutzungsverhalten verschiedener Schüler*innen zusammenhängt, wurde in einer Klasse für jedes Fach jeweils ein*e Schüler*in randomisiert aus drei verschiedenen Leistungsniveaugruppen gezogen (Hahn et al., 2016, S. 261) und hinsichtlich ihrer aktiven Lernzeit beobachtet (vgl. Abb. 1). Der Datensatz enthält damit für jeden Erhebungszeitraum (pro beobachteter Klasse und Fach) Individualbeobachtungen dreier Schüler*innen sowie Ratings der Unterrichtsqualität. Eingang in die Stichprobe der aktuellen Studie fanden jene Schüler*innen, die mindestens zu zwei Erhebungszeitpunkten beobachtet werden konnten ($N_{\text{Schüler*innen}} = 107$, $N_{\text{Klassen}} = 18$).

4.2 Instrumente

Aktive Lernzeit

Die ausgewählten Schüler*innen wurden hinsichtlich ihrer aktiven Lernzeit (v. a. in selbstregulierten Arbeitsphasen) im Feld beobachtet und mittels eines hochinferenten Ratings des Single-Items („Der Schüler bzw. die Schülerin beschäftigt sich die ganze Zeit mit dem Unterrichtsgegenstand“) auf einer vierstufigen Likert-Skala (Kategorien von 1 = *trifft nicht zu* bis 4 = *trifft zu*) durch geschulte Rater*innen eingeschätzt. Die Schulung der Rater*innen erfolgte zunächst auf Basis von Unterrichtsvideos. Dabei

	Rater*in							
Beobachtung	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Rating 1	×	×						
Rating 2		×	×					
Rating 3			×	×				
Rating 4				×	×			
Rating 5					×	×		
Rating 6						×	×	
Rating 7							×	×
Rating 8	×							×

Abb. 2: Muster der rotierten Doppelbeobachtungen (Incomplete Connected Design) zur Überprüfung der Interraterreliabilität

wurde von allen Rater*innen das Nutzungsverhalten eines vorher bestimmten Lernenden unter Zuhilfenahme eines Ratingmanuals unabhängig voneinander eingeschätzt. Die zu ratenden Videosequenzen entsprachen in der Länge exakt dem für die Datenerhebung festgelegten zeitlichen Beobachtungsumfang von 40 Minuten. Abweichende Urteile in den Videoschulungen wurden im Nachgang ausführlich diskutiert, um ein gemeinsames Verständnis des Items zu erzielen. Die hierbei gewonnenen Konkretionen (z. B. Indikatoren für die Reduktion der zu beobachtenden aktiven Lernzeit) wurden im Sinne der künftigen Handhabung im Ratingmanual festgehalten. Um die innerhalb der Videoschulungen erreichte Interraterreliabilität während der längsschnittlichen Datenerhebung aufrechtzuerhalten, wurden zwischen den Messzeitpunkten rotierte Doppelbeobachtungen in einem sog. Incomplete Connected Design (Eckes, 2011) durchgeführt (siehe Abb. 2). Das Verfahren wies sehr gute Interraterkonsistenzmaße auf (Krippendorffs α: *Min* = .56, *Max* = 1; *MW* = .89; *Med* = 1; vgl. Hahn et al., 2016, S. 263).

Unterrichtsqualität

Auch die Unterrichtsqualität wurde über Beobachtung unter Rückgriff auf ein bereits validiertes Instrument erfasst (Pietsch, 2010; Müller, Pietsch & Bos, 2011). Nach 40 Minuten Unterrichtsbeobachtung im Feld wurden 30 Items auf einer vierstufigen Likert-Skala (Kategorien von 1 = *trifft nicht zu* bis 4 = *trifft zu*) hochinferent geratet (Meissner, Merk, Pietsch & Bohl, 2016). Das eingesetzte Instrument wurde im Zuge der Hamburger Schulevaluation mithilfe von Multifacettenraschmodellen ursprünglich eindimensional modelliert, um Entwicklungsstufen der Unterrichtsqualität rückmelden zu können, auch wenn die Items des Instruments sechs inhaltlich unterscheidbaren Dimensionen von Unterrichtsqualität zugeordnet werden können. Diese sechsfaktorielle Struktur ließ sich in der vorliegenden Studie jedoch nicht bestätigen (Meissner & Merk, 2019).

Eine explorative Faktorenanalyse resultierte in drei Faktoren, die starke Korrespondenz zu drei Faktoren des Originalinstrumentes aufweisen: *Klassenmanagement und Strukturierung* (KMS), *Individualisierung* (IDV) sowie *Motivierung* (MOT). Detaillierte Ergebnisse der explorativen Faktorenanalyse, Wortlaut der Items sowie ihre jeweilige Zuordnung zu den Dimensionen können ebenfalls der Analysendokumentation entnommen werden (Meissner & Merk, 2019). Die Reliabilität der explorativ generierten Skalen wurde mit dem Koeffizient McDonald ω (Dunn, Baguley & Brunsden, 2013) geschätzt (KMS: .909; 95 %-KI [.896, .921]; IDV: .801; 95 %-KI [.774, .828]; MOT: .748; 95 %-KI [.713, .784]).

Kontrollvariablen

Zur Analyse der jeweiligen Lernausgangslagen wurden über eine zusätzliche schriftliche Befragung folgende Individualdaten erhoben, die als Kontrollvariablen in den Analysen Berücksichtigung fanden:

Sozioökonomischer Hintergrund (SES): Dieser wurde mithilfe des Highest International Socio-Economic Index of Occupational Status (HISEI), basierend auf dreistelligen ISCO 08 Berufsklassifikationen quantifiziert (Ganzeboom, De Graaf & Treiman, 1992).

Kognitive Grundfähigkeiten (KFT): Hierbei wurde auf den non-verbalen Subtest N2 des revidierten Tests kognitiver Fähigkeiten für 4.–12. Klassen von Heller und Perleth (2000) zurückgegriffen. Die enthaltenen 25 Items zeigen in einer konfirmatorischen Faktorenanalyse für binäre Outcomes, welche anhand des Three-Stages-Weighted-Least-Squares-Ansatzes (Muthén, 1984) geschätzt wurden, eine sehr gute Modellanpassung ($\chi^2 = 1624.8$, df = 275, Tucker-Lewis Index [TLI] = 0.992, Comparative Fit Index [CFI] = 0.992, Root Mean Square Error of Approximation [RMSEA] = 0.040, Standardized Root Mean Square Residual [SRMR] = 0.050). Die Schätzung der internen Konsistenz erfolgte aufgrund der Operationalisierung (dichotome Items) anhand der polychorischen Korrelationsmatrix (Gadermann, Guhn & Zumbo, 2012; ordinales $\alpha = .970$).

Leistungsniveau: Die Information über das fachgebundene Leistungsniveau der Schüler*innen wurde über die Einschätzung der jeweils unterrichtenden Lehrkraft gewonnen, die alle Schüler*innen ihrer Klasse einem von drei Leistungsniveaus (stark/mittel/schwach) zuordnete. Um die Zuverlässigkeit dieser Einschätzung empirisch zu untermauern, wurde im Rahmen der Analysen der KFT herangezogen. In einem Regressionsmodell für ordinale Variablen mit Random Intercepts (Cumulative Link Mixed Model; Agresti, 2010) konnte das eingeschätzte Leistungsniveau im Fach durch den KFT prädiziert werden ($\beta = .64$, $p = .017$ bei hochsignifikanter Überlegenheit gegenüber einem Nullmodell, das nur das Random Intercept enthielt; Meissner & Merk, 2019).

Statistische Analysen

Der vorliegenden Studie liegen geclusterte Daten mit einer sog. kreuzklassifizierten Struktur (Snijders & Bosker, 2012; siehe Abb. 3) zugrunde: Jeder Datenpunkt der Variable aktive Lernzeit kann zum einen genau einem Lernenden zugeordnet werden, zum anderen können mehrere Beobachtungen der aktiven Lernzeit einer Erfassung von Unterrichtsqualität (und so zu genau einer Klasse in einer Erhebungsphase und einem Fach) zugeordnet werden.

Innerhalb der ersten Cluster (Schüler*innen) variiert etwa die Variable Klassenmanagement, die Variable HISEI jedoch nicht. Innerhalb der zweiten Cluster (Unterrichtsstunden) ist es umgekehrt. Eine Möglichkeit, trotz dieser komplexen Struktur unverzerrte Parameterschätzungen und korrekte Standardfehler für entsprechend spezifizierte Mehrebenen-Regressionsmodelle zu erhalten, bieten die im Paket „lme4“ (Bates, Mächler, Bolker & Walker, 2014) enthaltenen Maximum-Likelihood-Schätzer. Sie bieten jedoch keine Möglichkeit des modellimanenten Umgangs mit fehlenden Werten, weshalb diese vor der Analyse unter Berücksichtigung der Datenstruktur multipel (20-fach; Bodner, 2008) imputiert wurden. Dazu wurde das Paket „pan“ (Zhao & Schafer, 2016) eingesetzt, welches einen sogenannten joint modelling approach (Grund, Lüdtke & Robitzsch, 2016) verfolgt, nach dem alle Variablen simultan mit einem einzigen statistischen Modell imputiert werden (Schafer & Yucel, 2002). Anschließend wurden die Mehrebenen-Regressionsmodelle einzeln für jeden Datensatz geschätzt und die Ergebnisse (sowohl Fixed als auch Random Effects) nach den Regeln von Rubin (1987) kombiniert.

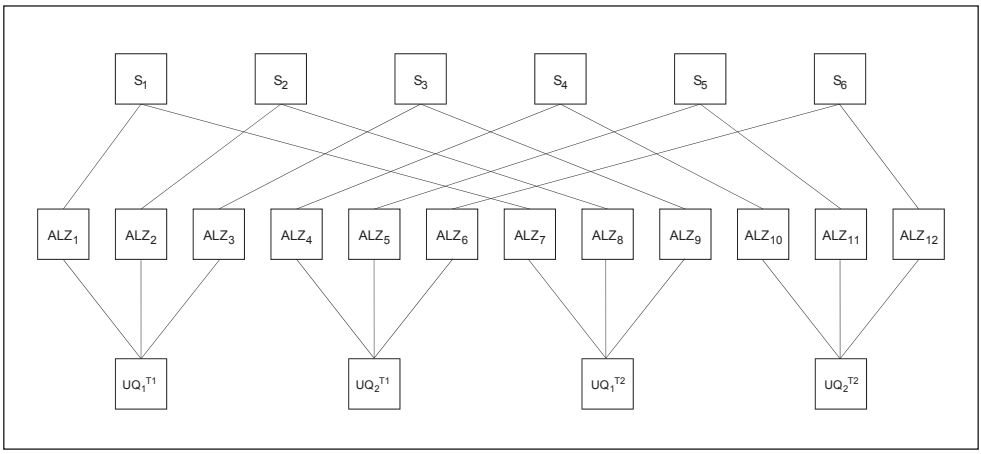

*Abb. 3: Exemplarische Veranschaulichung der kreuzklassifizierten Datenstruktur (S_i = beobachtete Schüler*innen, ALZ_i = beobachtete aktive Lernzeit; UQ_1^{T2} = beobachtete Unterrichtsqualität zum 2. Messzeitpunkt (T2) in Klasse 1)*

5. Ergebnisse

Im ersten Mehrebenenmodell mit gekreuzten Random Faktoren wurden die Unterrichtsqualitätsmaße als Prädiktoren für die zu beobachtende aktive Lernzeit aufgenommen (Mod 1). Von diesen zeigte lediglich die Dimension des *Klassenmanagements* einen signifikanten Effekt auf die zu beobachtende aktive Lernzeit der Schüler*innen. Je besser also die Klassenführung ausgeprägt war, desto höher fiel die durchschnittliche aktive Lernzeit der beobachteten Schüler*innen aus. Die beiden weiteren Dimensionen *Individualisierung* und *Motivierung* konnten das Ausmaß an zu beobachtender aktiver Lernzeit hingegen nicht vorhersagen.

Im zweiten Modell (Mod 2) wurden die Variablen als Prädiktoren aufgenommen, welche innerhalb der Schüler*innen invariant waren (KFT, SES, Leistungsniveau). Von diesen zeigte das von der Lehrkraft eingeschätzte Leistungsniveau der Schüler*innen (zwei Dummyvariablen: mittleres und hohes Leistungsniveau) signifikante Effekte. Wie Tabelle 1 entnommen werden kann, sind diese Effekte durchaus als bedeutsam einzuschätzen: So weisen Schüler*innen mit hohem Leistungsniveau eine 0.650 Likertstu-

	Mod 1			**Mod 2**		
	B	**SE**	**FMI**	**B**	**SE**	**FMI**
Intercept	2.266	0.317	0.251	2.617	0.233	0.087
KFT				− 0.004	0.011	0.172
HISEI				0.003	0.004	0.053
$I^{\text{mittleres Leistungsniveau}}$				0.421*	0.168	0.008
$I^{\text{hohes Leistungsniveau}}$				0.650***	0.179	0.014
KMS	0.249*	0.122	0.383			
IDV	− 0.024	0.099	0.372			
MOT	0.033	0.099	0.360			
	Random Effects					
	s^2			**s^2**		
$\text{Intercept}_{\text{Schüler*innen}}$	0.361			0.295		
$\text{Intercept}_{\text{Unterrichtsstunde}}$	0.108			0.142		
Residuen	0.550			0.538		

*Tab. 1: Modellschätzungen (*p < .05, **p < .01, ***p < .001. FMI = Fraction of Missing Information. KFT = Kognitiver Fähigkeitstest. HISEI = Highest International Socio-Economic Index of Occupational Status. $I^{\text{mittleres Leistungsniveau}}$ = Individuelles fachgebundenes Leistungsniveau. KMS = Klassenmanagement u. Strukturierung. IDV = Individualisierung. MOT = Motivierung)*

fen höhere zu beobachtende aktive Lernzeit auf als Schüler*innen mit niedrigem Leistungsniveau.

Schließlich wurde in einem dritten Modell (Mod 3) untersucht, inwiefern Merkmale der Unterrichtsqualität mit Merkmalen der Schüler*innen interagieren und sich auf diese Weise differenzielle Nutzungseffekte erkennen lassen. Dazu wurden Interaktionseffekte zwischen den Within-Cluster zentrierten Variablen (Enders & Tofighi, 2007) der Unterrichtsqualität und den Dummyvariablen des Leistungsniveaus spezifiziert. Diese Interaktionseffekte zeigten sich jedoch als nicht signifikant (Koeffizienten siehe Meissner & Merk, 2019).

6. Diskussion und Ausblick

Angebot-Nutzungs-Modelle dienen als Heuristik, um sich über die verschiedenen Einflussfaktoren auf schulische Lernprozesse zu verständigen. In der Literatur finden sich wiederholt Formulierungen, die auf theoretischer Ebene andeuten, dass das unterrichtliche Angebot abhängig von den individuellen Lernausgangslagen ganz unterschiedliche Wirkungen zeigen kann, womit differenzielle Effekte angedeutet werden (z. B. Vieluf et al., in diesem Heft; Kunter & Ewald, 2016). Ziel der hier vorliegenden Studie war es, Prädiktoren der aktiven Lernzeit auf Angebotsseite zu ermitteln und diese auf differenzielle Effekte hin zu überprüfen. Im Rahmen unserer Studie ließ sich der Befund vorangegangener Studien bestätigen, dass die Dimension *Klassenmanagement und Strukturierung* auf Angebotsseite positiv mit der zu beobachtenden aktiven Lernzeit auf der Nutzungsseite assoziiert ist. Erwartungswidrig zeigten die Dimensionen *Individualisierung* und *Motivierung* indes keine Vorhersagekraft, obschon sich gemäß der Theorie beide begünstigend auf die aktive Lernzeit auswirken könnten. Ein Grund für diesen ausbleibenden Effekt könnte in dem Umstand liegen, dass sich ein erfolgreiches Klassenmanagement zwar für alle Schüler*innen gleichermaßen positiv auszuwirken scheint, Motivierung und Individualisierung hingegen womöglich eher auf individueller Ebene relevant werden und daher differenziell operationalisiert werden müssten. Zu vermuten wäre überdies, dass das Konstrukt des Klassenmanagements leichter der Beobachtung zugänglich ist als die Dimensionen Motivierung und Individualisierung, weshalb die beiden letztgenannten Unterrichtsmerkmale ggf. über Fragebögen leichter zu erfassen sind – gelten doch gerade Schülerurteile mit Blick auf Unterrichtsqualität als sehr verlässlich (z. B. Fauth et al., 2014). Eine weitere Erklärung könnte im Erhebungsdesign liegen, gemäß welchem Unterrichtsqualität und die zu beobachtende aktive Lernzeit zwar innerhalb des gleichen Erhebungszeitraums, jedoch nicht zum identischen Zeitpunkt gemessen wurden. Anzunehmen wäre, dass das Konstrukt Klassenmanagement zeitlich stabiler ist als die anderen beiden Dimensionen, weshalb die gleichzeitige Erhebung der zu beobachtenden aktiven Lernzeit und der Unterrichtsqualitätsmaße vermutlich eine eindeutigere Prädiktion erlauben würde.

Im Weiteren konnte gezeigt werden, dass das von der Lehrkraft eingeschätzte fachgebundene Leistungsniveau hohe prädiktive Kraft aufweist, nicht aber der SES. Auch

der KFT zeigte keine prädiktive Kraft, obschon über diese Variable die Leistungseinschätzung der Lehrkraft validiert werden konnte. Möglicherweise ist jedoch die Leistungseinschätzung der Lehrkraft bereits durch das Nutzungsverhalten konfundiert. In dieser Variable könnte sich also die enge Verwobenheit von Angebot und Nutzung zeigen. Zwar können Lehrkräfte die Leistungsreihung innerhalb von Klassen zuverlässig einschätzen (Ingenkamp, 1995), doch würde der Einbezug von standardisierten Leistungstests sicherlich eine objektivere Datenbasis liefern.

Womöglich sind bereits diese Limitationen der Studie Grund dafür, dass der hypothetisierte Interaktionseffekt nicht gefunden wurde. Gleichwohl erscheint es aus theoretischer Sicht lohnenswert, weiter nach Interaktionseffekten zu suchen und dabei über den Einbezug weiterer Variablen nachzudenken. So könnten sich gerade motivational-affektive Variablen, selbstregulative Fähigkeiten der Lernenden sowie die Beziehung zwischen Lehrkraft und Schüler*innen oder aber die Beziehungsqualität der Schüler*innen untereinander in künftigen Studien zu differenziellen Effekten als ertragreich erweisen.

Literatur

Agresti, A. (2010). *Analysis of ordinal categorical data* (2. Auflage). Hoboken, N. J.: Wiley.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Berliner, D. (1990). What's all the fuss about instructional time? In M. Ben-Peretz & R. Bromme (Hrsg.), *The Nature of time in schools: Theoretical concepts, practitioner perceptions* (S. 3–35). New York: Teachers College Press.

Bloom, B. S. (1974). Time and learning. *American Psychologist, 29*(9), 682–688.

Bodner, T. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: a Multidisciplinary Journal, 15*(4), 651–675.

Bohl, T., & Wacker, A. (Hrsg.) (2016). *Die Einführung der Gemeinschaftsschule in Baden-Württemberg: Abschlussbericht der Wissenschaftlichen Begleitforschung (WissGem)*. Münster: Waxmann Verlag.

Brühwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction, 21*(1), 95–108.

Cadima, J., Leal, T., & Burchinal, M. (2010). The quality of teacher-student interactions: Associations with first graders' academic and behavioral outcomes. *Journal of School Psychology, 48*(6), 457–482.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723–733.

Conger, R., Conger, K., & Martin, M. (2010). Socioeconomic status, family processes, and individual development. *Journal of Marriage and the Family, 72*(3), 685–704.

Dunn, T., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), S. 399–412.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments. Language testing and evaluation: Vol. 22.* Frankfurt a. M./Berlin/Bern/Bruxelles/New York/Oxford/Wien: Lang.

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

Everaert, P., Opdecam, E., & Maussen, S. (2017). The relationship between motivation, learning approaches, academic performance and time spent. *Accounting Education, 26*(1), 78–107.

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), 1–9.

Gadermann, A., Guhn, M., & Zumbo, B. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. Practical Assessment, *Research & Evaluation, 17*(3), 1–13.

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 2*(1), 1–56.

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of multilevel missing data: An introduction to the R Package pan. *SAGE Open, 6*(4)

Hahn, E., Rohlfs, C., Wacker, A., & Bohl, T. (2016). Umgang mit Heterogenität: Eine quantitative Beobachtungsstudie zur aktiven Lernzeit von Schülerinnen und Schülern unterschiedlicher Leistungsniveaus. In T. Bohl & A. Wacker (Hrsg.), *Die Einführung der Gemeinschaftsschule in Baden-Württemberg: Abschlussbericht der Wissenschaftlichen Begleitforschung (WissGem)*. (S. 255–274). Münster: Waxmann Verlag.

Hamre, B., & Pianta, R. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949–967.

Hamre, B., Pianta, R., Downer, J., DeCoster, J., Mashburn, A., Jones, S., & Hamagami, A. (2013). Teaching through interactions. *The Elementary School Journal, 113*(4), 461–487.

Hasselhorn, M., & Gold, A. (2009). *Pädagogische Psychologie: Erfolgreiches Lernen und Lehren* (2., durchges. Aufl.). Stuttgart: Kohlhammer.

Hasselhorn, M., Andresen, S., Becker, B., Betz, T., Leuzinger-Bohleber, M., & Schmid, J. (2014). Children at risk of poor educational outcomes: Theoretical concepts and empirical results. *Child Indicators Research, 7*(4), 695–697.

Heller, K. A., & Perleth, C. (2000). *KFT 4–12 + R: Kognitiver Fähigkeitstest für 4. bis 12. Klassen. Revision.* Göttingen: Beltz.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (4. aktual. Auflage). Seelze-Velber: Kallmeyer u. a.

Ingenkamp, K. (1995). *Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte* (9., überarb. u. erw. Aufl.). Weinheim: Beltz.

Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik, 54*(2), 222–237.

Klieme, E., & Warwas, J. (2011). Konzepte der individuellen Förderung. *Zeitschrift für Pädagogik, 57*(6), 805–818.

Klieme, E. (2018). Unterrichtsqualität. In M. Gläser-Zikuda, M. Harring & C. Rohlfs (Hrsg.), *Handbuch Schulpädagogik* (S. 1–24). Stuttgart: UTB; Waxmann.

Kunter, M., & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 85–113). Münster u. a.: Waxmann.

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

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). Münster/New York: Waxmann.

Marzano, R., & Marzano, J. (2003). The key to classroom management. *Educational Leadership, 61*(1), 6–13

Meissner, S., & Merk, S. (2019). *Differential effects of instructional quality on time on task. Documentation of analyses.* https://doi.org/10.17605/OSF.IO/Z9EFY.

Meissner, S., Merk, S., Pietsch, M., & Bohl, T. (2016). Unterrichtsqualität an Gemeinschaftsschulen. Ergebnisse quantitativer und qualitativer Untersuchungen zur Qualität unterrichtlicher Prozesse. In T. Bohl & A. Wacker (Hrsg.), *Die Einführung der Gemeinschaftsschule in Baden-Württemberg: Abschlussbericht der Wissenschaftlichen Begleitforschung (WissGem)* (S. 193–212). Münster: Waxmann.

Müller, S., Pietsch, M., & Bos, W. (2011). *Schulinspektion in Deutschland: Eine Zwischenbilanz aus empirischer Sicht.* Münster: Waxmann.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), S. 115–132.

Pietsch, M. (2010). Evaluation von Unterrichtsstandards. *Zeitschrift für Erziehungswissenschaft, 13*(1), 121–148.

Reusser, K. (2009). Von der Bildungs- und Unterrichtsforschung zur Unterrichtsentwicklung. Probleme, Strategien, Werkzeuge und Bedingungen. *Beiträge zur Lehrerbildung, 27*(3), 295–312.

Romero, M., & Barberà, E. (2011). Quality of e-learners' time and learning performance beyond quantitative time-on-task. *The International Review of Research in Open and Distributed Learning, 12*(5), 125–137.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. Harvard University, Hoboken: John Wiley & Sons Inc.

Schafer, J., & Yucel, R. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics, 11*(2), 437–457.

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*(4), 454–499.

Seiz, J., Decristan, J., Kunter, M., & Baumert, J. (2016). Differenzielle Effekte von Klassenführung und Unterstützung für Schülerinnen und Schüler mit Migrationshintergrund. *Zeitschrift Für Pädagogische Psychologie, 30*(4), 237–249.

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.)*. Los Angeles: Sage.

Snow, R. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology, 59*(2), 205–216.

Zhao, J., & Schafer, J. (2016). *pan: Multiple imputation for multivariate panel or clustered data.* [Computer software].

Abstract: Supply-use models serve as heuristics, as they display influencing teacher and classroom characteristics, as well as individual preconditions that determine learning outcomes. However, most of the established supply-use models do not include any interactions among those variables – even though differential effects of instructional quality on usage are often assumed in theory, especially when taking into account the varying learning preconditions of the students. The present study investigated whether interactions between instructional quality and student learning characteristics can be found in regard to time on task. While direct effects from classroom management and estimated subject-related student achievement to time on task can be proven from the data, there was no evidence for differential effects.

Keywords: Time on Task, Instructional Quality, Individual Learning Preconditions, Differential Effects, Supply-Use-Model

Anschrift der Autor*innen

Sibylle Meissner, Eberhard Karls Universität Tübingen,
Tübingen School of Education,
Wilhelmstraße 31, 72074 Tübingen, Deutschland
E-Mail: Sibylle.Meissner@uni-tuebingen.de

Jun.-Prof. Dr. Samuel Merk, Eberhard Karls Universität Tübingen,
Institut für Erziehungswissenschaft, Abteilung Schulpädagogik,
Münzgasse 22, 72070 Tübingen, Deutschland
E-Mail: Samuel.Merk@ife.uni-tuebingen.de

Prof. Dr. Benjamin Fauth, Institut für Bildungsanalysen Baden-Württemberg (IBBW),
Heilbronner Str. 172, 70191 Stuttgart, Deutschland
E-Mail: benjamin.fauth@ibbw.kv.bwl.de

Prof. Dr. Marc Kleinknecht, Leuphana Universität Lüneburg,
Institut für Bildungswissenschaft, Schulpädagogik und Schulentwicklung,
Universitätsallee 1, 21335 Lüneburg, Deutschland
E-Mail: marc.kleinknecht@leuphana.de

Prof. Dr. Thorsten Bohl, Eberhard Karls Universität Tübingen,
Institut für Erziehungswissenschaft, Abteilung Schulpädagogik,
Münzgasse 22, 72070 Tübingen, Deutschland
E-Mail: Thorsten.Bohl@ife.uni-tuebingen.de

Tina Seidel

Kommentar zum Themenblock „Angebots-Nutzungs-Modelle als Rahmung“

Quo vadis deutsche Unterrichtsforschung? Modellierung von Angebot und Nutzung im Unterricht

Zusammenfassung: Der Beitrag diskutiert die Modellierung von Angebot und Nutzung aus einer psychologischen Perspektive. Drei Punkte werden herausgestellt: Erstens bedarf die Angebotsseite einer weiteren Ausdifferenzierung und Vereinheitlichung, insbesondere im Hinblick auf die Integration allgemein-didaktischer, fachdidaktischer und pädagogisch-psychologischer Konzepte. Zweitens hat die Modellierung der Nutzungsseite stark von der Integration prozessorientierter psychologischer Theorien profitiert, erfordert aber weitere Abgrenzungen bei der Einordnung der Lernaktivitäten aufseiten der Lernenden. Drittens orientiert sich der internationale Forschungsstand bislang wenig an den im deutschsprachigen Raum entwickelten Angebots-Nutzungs-Modellen.

Schlagworte: Modelle des Lehrens und Lernens, Angebots-Nutzungs-Modell, Unterrichtsforschung, Lernaktivitäten, Prozessorientierung

1. Einleitung

Die Weiterentwicklung der Prozesse-Produkt Modelle der 80er Jahre und die Ausarbeitung von Angebots-Nutzungs-Modellen stellen – wie in diesem Themenheft von Vieluf, Praetorius, Rakoczy und Kleinknecht in überzeugender Weise aus einer rückblickenden Perspektive zusammengefasst – eine der zentralen Durchbrüche der Unterrichtsforschung in den vergangenen 30 Jahren dar. Einer der wesentlichen Kritikpunkte an den Prozess-Produkt-Modellen lag in der vereinfachten Modellierung einer direkten Wirkung von Unterrichtshandlungen auf Lernergebnisse, die als zu mechanistisch und wenig angepasst an das komplexe Wirkungsgefüge von sozialen Ko-Konstruktionen innerhalb eines Klassenverbands galt (Gage & Needles, 1989). Aufbauend auf der in den 90er Jahren zunehmenden Lehr-Lern-Forschung und der Modellierung von Lernprozessen (Reusser, 1995) wurden die Unterrichtsmodelle vor allem im deutschsprachigen Raum durch Fend (2008) und Helmke (2015) systematisch erweitert und dabei die ‚Nutzungsseite‘ eingeführt. Die grundlegende Differenzierung zwischen einer Angebots- und einer Nutzungsseite und den damit verbundenen Implikationen zur Erklärung unterrichtlicher Lehr-Lern-Prozesse war somit zentraler Fortschritt der deutschsprachigen Unterrichtsforschung. In der nun folgenden Diskussion möchte ich einige Überlegungen zur Angebots- und zur Nutzungsseite in der Unterrichtsmodellierung zusammenfassen und mit einem kurzen Ausblick auf den internationalen Diskurs schließen.

2. Die Angebotsseite von Unterricht erfordert weitere konzeptuelle Arbeit

Obwohl weitgehend Konsens darüber herrscht, dass Unterricht sowohl eine Angebots- als auch eine Nutzungsseite umfasst, bestehen recht große Differenzen hinsichtlich der theoretischen Ausdifferenzierung (Vieluf et al., in diesem Heft), insbesondere im Hinblick auf die inhaltliche Konzeptualisierung des unterrichtlichen Angebots. Ob die Angebotsseite stärker aus einer allgemein-didaktischen, fachdidaktischen oder pädagogisch-psychologischen Perspektive betrachtet wird, hängt oft auch von den individuellen Forschungsinteressen ab (Seidel, 2014b). Diese Vielfalt kann einerseits dazu beitragen, in einem breiter orientierten Ansatz verschiedene Perspektiven anzuwenden und sich gegenseitig zu ergänzen. Andererseits besteht aber auch die Gefahr, dass man in einem Forschungsfeld keine präzise und einheitliche begriffliche Determination der Angebotsstrukturen im Unterricht findet.

Dennoch gab es in den vergangenen Jahren auch zentrale Entwicklungen, die entscheidend zur theoretischen Ausdifferenzierung der Angebotsseite beigetragen haben. Erstens gibt es bei vielen Autorinnen und Autoren die Differenzierung zwischen Oberflächen- und Tiefenmerkmalen von Unterricht, die in erheblicher Weise durch die theoretischen Überlegungen zu Choreographien des Unterrichts angestoßen wurde (Oser & Baeriswyl, 2001). Zweitens beziehen sich viele Forschungsarbeiten auf generische, fächerübergreifende Tiefenmerkmale von Unterricht, wie sie beispielsweise in den Basisdimensionen von Klassenführung, kognitiver Aktivierung und Lernunterstützung zum Ausdruck kommen (Klieme, Lipowsky, Rakoczy & Ratzka, 2006). Ein empirisches Beispiel für die Modellierung und empirische Prüfung der inhaltlichen Angebotsseite stellt der Beitrag von Meissner, Merk, Fauth, Kleinknecht und Bohl (in diesem Heft) dar, indem auf das generische Tiefenmerkmal von Klassenführung fokussiert wird.

Trotz dieser hilfreichen Vereinheitlichungen würde allerdings die Angebotsseite von Unterricht davon profitieren, weitere Theoriearbeit aus unterschiedlichen Perspektiven (allgemein-didaktisch, fachdidaktisch, pädagogisch-psychologisch) anzuwenden. In der deutschen Unterrichtsforschung fokussieren wir derzeit stark auf die empirische Prüfung differenzierbarer Angebotsdimensionen, z. B. den Basisdimensionen guten Unterrichts. Es fehlen aber weiterhin theoretische Modellierungen vor allem zu den Prozessen beim Ablauf unterrichtlicher Angebote (Seidel, 2014a). Beispielsweise müsste berücksichtigt werden, in welcher Beziehung das situationale Unterrichtsangebot zu den überdauernden, stabilen Angebotsstrukturen steht. Erste empirische Hinweise deuten darauf hin, welche Angebotsstrukturen anscheinend als stärker variabel zwischen einzelnen Stunden zu sehen sind (z. B. kognitive Aktivierung) und welche eher längerfristige Angebotsstrukturen abbilden (z. B. Klassenführung) (Praetorius, Pauli, Reusser, Rakoczy & Klieme, 2014).

Die Ausdifferenzierung der Angebotsseite sollte zudem stärker die Rolle der Lernenden berücksichtigen. Beispielsweise wird von Vieluf et al. (in diesem Heft) die berechtigte Frage aufgeworfen, wie die im öffentlichen Klassenverband für alle Lernenden zugänglichen Handlungen der einzelnen Schülerinnen und Schüler zu werten sind. Diese individuellen Schülerhandlungen stellen durchaus ein Lernangebot für die ge-

samte Klasse dar, beispielsweise in dem ein wichtiger inhaltlicher Beitrag eingebracht oder eine zentrale Frage gestellt wird. Bislang wurden die Lernaktivitäten der Schülerinnen und Schüler meist ausschließlich auf der Nutzungsseite in Form von äußeren und inneren Lernaktivitäten modelliert. Der Vorschlag von Vieluf et al. (in diesem Heft), gerade die äußeren, von anderen beobachtbaren Aktivitäten der Lernenden auf der Angebotsseite zu modellieren, hat meines Erachtens ein hohes Potential. Auf diese Weise könnte es gelingen, auf der Angebotsseite die Interaktivität zwischen Lehrenden und Lernenden stärker zu berücksichtigen.

3. Die Nutzungsseite von Unterricht: Integration psychologischer Theorien

Der Kern der Weiterentwicklung der Prozess-Produkt-Modelle aus den 80er Jahren bestand in der näheren Betrachtung der Nutzungsprozesse aufseiten der Lernenden (vgl. Helmke, 2015). Dabei wurde die starre Auffassung aufgegeben, dass Lehrhandlungen direkt und unmittelbar Lernergebnisse beeinflussen. Nach den Angebots-Nutzungs-Modellen werden Wirkungsweisen von Lehrhandlungen nun moderiert durch individuelle Nutzungsprozesse aufseiten der Lernenden. Diese Nutzungsprozesse beinhalten kognitive, motivationale und emotionale Prozesse der Aufnahme, Verarbeitung und Integration von Lerninhalten (Winne, 1987). Diese sind wiederum in erheblichem Maße von längerfristig aufgebauten kognitiven und motivational-affektiven Lerndispositionen der Schülerinnen und Schüler determiniert, können aber auch durch die situative Wahrnehmung der Angebotsstrukturen in der Lernumgebung beeinflusst werden. Bei der Integration der Nutzungsprozesse in aktuelle Unterrichtsmodelle waren vor allem psychologische Theorien hilfreich (Seidel, 2014a). Mithilfe solcher Ansätze können Nutzungsprozesse der Lernenden aus kognitiver, motivationaler und emotionaler Sicht erklärt werden. Arbeiten aus dem Bereich der Lernstrategieforschung und die damit verbundenen Prozesse bei der Verarbeitung von Lerninhalten – der Aufnahme, Elaboration und Organisation – waren insbesondere zur Erklärung kognitiver Lernprozesse im Unterricht hilfreich (Schiefele, Wild & Winteler, 1995; Winne, 1987). Einen weiteren wichtigen Stellenwert nahm die Selbstbestimmungstheorie der Motivation ein (Deci & Ryan, 1993). Diese half zu modellieren, wie auf der Basis der Wahrnehmung individueller psychologischer Bedürfnisse qualitative Unterschiede in der Lernmotivation erklärbar sind, die sich wiederum auf die Tiefe der kognitiven Verarbeitung von Lerninhalten auswirken können (Prenzel, Krapp & Schiefele, 1986). Durch die Integration dieser psychologischen Theorien ist somit die Nutzungsseite der aktuellen Unterrichtsmodelle vergleichsweise ausdifferenziert und dessen Bedeutung durch vielfältige empirische Studien auch belegt (vgl. Seidel, 2014a).

Die derzeitige methodische Herangehensweise in der empirischen Prüfung der Angebots-Nutzungs-Modelle liegt vorrangig in der Anwendung von Mehrebenenmodellen, bei denen klassischerweise Angebotsstrukturen auf der Klassenebene und Nutzungsprozesse auf der Individualebene modelliert werden. Diese Art der Modellierung

hat einen engen Bezug zur Lernpsychologie, die Lernen als situativen, aktiven und individuellen Prozess auffasst und davon ausgeht, dass diese Prozesse sowohl von individuellen Lerndispositionen (z. B. Vorwissen, Interesse, Selbstkonzept) als auch von der Lernumgebung (z. B. den Angebotsstrukturen) beeinflusst werden (Kunter & Trautwein, 2013). Jüngere Forschungsarbeiten weisen nun darauf hin, dass sich trotz dieser interindividuellen Unterschiede auch Gruppen von Lernenden identifizieren lassen, die sich in ihren Dispositionsmustern durchaus ähnlich sind (Seidel, 2006; Südkamp, Praetorius & Spinath, 2018). Die Identifizierung solcher Schülergruppen könnte die Komplexität der Nutzungsseite erheblich reduzieren. Für die Weiterentwicklung der Unterrichtsmodelle stellt sich in diesem Zusammenhang die Frage, ob man nicht neben der Betrachtung der Individualebene diese Schülergruppen als eine weitere Ebene modellieren kann. Die Modellierung solcher Gruppen könnte Aufschluss darüber geben, ob ähnliche Nutzungsprozesse durchlaufen werden, die wiederum vergleichbare Auswirkungen auf Lernergebnisse erklären. Ein Beispiel für einen Schritt in diese Richtung stellt der Beitrag von Meissner et al. (in diesem Heft) dar, in dem Schülergruppen mit ähnlichen Dispositionsmustern (Differenzierung zwischen hohem und mittlerem Leistungsniveau, ähnliche kognitive Grundfähigkeiten und sozialer Hintergrund) betrachtet wurden.

Abschließend bleibt noch ein Punkt zu diskutieren: Inwiefern kann die weitere Ausdifferenzierung der Angebots-Nutzungs-Modelle dazu beitragen, klare und einheitliche Trennungslinien zwischen der Angebots- und der Nutzungsseite zu modellieren? Ein Schlüsselmoment liegt hier in der näheren Bestimmung der Lernaktivitäten der Schülerinnen und Schüler und deren Rolle für die Bereitstellung (Angebotsseite) bzw. Nutzung von Angeboten (Nutzungsseite). Wie diffizil dieser Punkt ist, soll an einem Beispiel der beiden Beiträge im Themenheft zur Trennung zwischen Angebot und Nutzung deutlich gemacht werden. Im Beitrag von Meissner et al. (in diesem Heft) wird die aktive Lernzeit gemessen über Videobeobachtungen im Unterricht und über ein Rating mit der Aussage „Der Schüler bzw. die Schülerin beschäftigt sich die ganze Zeit mit dem Unterrichtsgegenstand“. In diesem empirischen Beitrag wird die aktive Lernzeit als abhängige Variable der Nutzungsseite von Unterricht zugeschlagen. Im Beitrag von Vieluf et al. (in diesem Heft) wird dagegen vorgeschlagen, gerade die äußeren, beobachtbaren Lernaktivitäten der Schülerinnen und Schüler der Angebotsseite zuzuordnen, da diese Schüleraktivitäten wieder die Angebotsstruktur beeinflussen und dazu beitragen, dass Diskussionen und Schülerbeteiligungen am Unterricht im Klassenverband entstehen. Folgt man dieser Argumentation, müsste man die aktive Lernzeit anstelle der Nutzungsseite (Meissner et al., in diesem Heft) wiederum der Angebotsseite (Vieluf et al., in diesem Heft) zuordnen. Dieser Vergleich ist ein schönes Beispiel dafür, dass vor allem im Bereich der Angebotsseite wichtige theoretische Ausdifferenzierungen vollzogen werden müssen, die vor allem das Wechselspiel zwischen Handlungen der Lehrenden und der Lernenden abbilden. Auf der Basis gilt es zu diskutieren, mit welchen Instrumenten und Messverfahren die Angebots- und die Nutzungsseite am besten erfasst werden kann.

4. Quo vadis deutsche Unterrichtsforschung? Die deutsche Unterrichtsforschung in der internationalen Diskussion

Die im deutschsprachigen Raum erarbeiteten, durchaus als komplex wahrgenommenen Angebots-Nutzungs-Modelle werden im internationalen Diskurs bislang erstaunlich wenig aufgegriffen und diskutiert (vgl. Seidel, Hetmanek, Mok & Knogler, 2017). Das mag daran liegen, dass eine Vielzahl der aktuellen Publikationen deutscher Forscherinnen und Forscher in international zugänglichen Zeitschriftenorganen vorrangig empirischer Natur sind. In vielen Fällen werden häufig Ergebnisse methodisch anspruchsvoller Studien präsentiert, die es folglich bei dieser Form der Publikation dann nur noch eingeschränkt erlauben, die ebenfalls als sehr komplex zu betrachtenden theoretischen Modellierungen näher auszuführen. Dies mag unter anderem dazu geführt haben, dass es im Prinzip fast keine aktuelleren international veröffentlichten Publikationen mit einem expliziten Bezug zu einem Angebots-Nutzungs-Modell gibt (eine Ausnahme: Brühwiler & Blatchford, 2011). Ausführlichere konzeptuelle Beiträge in der englischsprachigen Literatur, wie beispielsweise die Darlegung der Choreographien des Lehrens und Lernens von Oser & Baeriswyl (2001) im „Handbook of Research on Teaching“, sind bislang eher die Ausnahme.

Betrachtet man die internationale Ausrichtung der Unterrichtsforschung, ist es interessant, dass man beispielsweise im US-amerikanischen Raum gegenwärtig stark daran interessiert ist, die Verbesserung der Angebotsseite von Unterricht zu erforschen. Dieses Interesse spiegelt sich unter anderem im so breit wie umfassend angelegten MET-Projekt (Measuring Effective Teaching) der Gates Stiftung wieder, in dem eine Reihe an führenden Unterrichtsforscherinnen und -forscher involviert ist (Kane, Kerr & Pianta, 2014). Ein weiteres Beispiel stellen die jüngeren Forschungsarbeiten zur inhaltlichen Konzeptualisierung sogenannter „high leverage“, auch „ambitious practices“ genannt, dar (Cohen, 2015; Cohen, Schuldt, Brown & Grossman, 2016). Diese Beispiele verdeutlichen den im US-amerikanischen Forschungsraum verbreiteten Fokus auf die genauere Bestimmung der Angebotsseite von Unterricht. Und zwar auf eine Weise, bei der man modelliert, wie die Angebote in einer möglichst ‚optimalen‘ Form aussehen können. Die Definition der Angebotsseite ergibt sich demnach über die Festlegung eines zu erreichenden Zielzustands. Dies ist meines Erachtens auch für die deutsche Unterrichtsforschung eine interessante Perspektive, die allerdings auch den Mut erfordert, konkret zu benennen, wie die Angebotsseite in optimaler Weise aussehen kann.

Literatur

Brühwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction, 21*(1), 95–108. doi:10.1016/j.learninstruc.2009.11.004.

Cohen, J. (2015). Challenges in identifying high-leverage practices. *Teachers College Record, 117*(7), 1–41.

Cohen, J., Schuldt, L. C., Brown, L., & Grossman, P. (2016). Leveraging observation tools for instructional improvement: Exploring variability in uptake of ambitious instructional practices. *Teachers College Record, 118*(11), 1–36.

Deci, E., & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik, 39,* 223–238.

Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität.* Wiesbaden: VS Verlag.

Gage, N. L., & Needles, M. C. (1989). Process-product research on teaching: A review of criticisms. *The Elementary School Journal, 89*(3), 253–300.

Helmke, A. (2015). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts* (4. Auflage). Seelze-Velber: Kallmeyer.

Kane, T. J., Kerr, K. A., & Pianta, R. C. (Hrsg.). (2014). *Designing teacher evaulation systems.* San Francisco, CA: Jossey-Bass.

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. 128–146). Münster: Waxmann.

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

Oser, F., & Baeriswyl, F. J. (2001). Choreographies of teaching: Bridging instruction to learning. In V. Richardson (Hrsg.), *Handbook of research on teaching* (S. 1031–1065). Washington, D. C.: American Educational Research Association.

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. doi:10.1016/j.learninstruc.2013.12.002.

Prenzel, M., Krapp, A., & Schiefele, H. (1986). Grundzüge einer pädagogischen Interessentheorie. *Zeitschrift für Pädagogik, 32,* 163–173.

Reusser, K. (1995). Lehr-Lernkultur im Wandel: Zur Neuorientierung in der kognitiven Lernforschung. In R. Dubs & R. Dörig (Eds.), *Dialog Wissenschaft und Praxis* (S. 164–190). St. Gallen: IWP.

Schiefele, U., Wild, K. P., & Winteler, A. (1995). Lernaufwand und Elaborationsstrategien als Mediatoren der Beziehung von Studieninteresse und Studienleistung. *Zeitschrift für Pädogogische Psychologie, 9*(3/4), 181–188.

Seidel, T. (2006). The role of student characteristics in studying micro teaching-learning environments. *Learning Environments Research, 9*(3), 253–271.

Seidel, T. (2014a). Angebots-Nutzungs-Modelle in der Unterrichtspsychologie: Integration von Struktur- und Prozessparadigma. *Zeitschrift für Pädagogik, 60*(6), 828–844.

Seidel, T. (2014b). Lehrerhandeln im Unterricht. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (2. überarb. und erw. Auflage, S. 781–806). Münster: Waxmann.

Seidel, T., Hetmanek, A., Mok, S. Y., & Knogler, M. (2017). Meta-Analysen zur Unterrichtsforschung und ihr Beitrag für die Realisierung eines Clearing House Unterricht für die Lehrerbildung. *Zeitschrift für Bildungsforschung, 7*(3), 311–325. doi:10.1007/s35834-017-0191-6.

Südkamp, A., Praetorius, A.-K., & Spinath, B. (2018). Teachers' judgment accuracy concerning consistent and inconsistent student profiles. *Teaching and Teacher Education, 76,* 201–213. doi:10.1016/j.tate.2017.09.016.

Winne, P. H. (1987). Why process-product research cannot explain process-product findings and a proposed remedy: The cognitive mediational paradigm. *Teaching and Teacher Education, 3*(4), 333–356. doi:10.1016/0742-051x(87)90025-4.

Abstract: This paper discusses current developments in opportunity-usage-models from a psychological perspective. Three points are made: first, opportunity structures in teaching need to be defined more clearly and coherently, particularly with regard to the integration of teaching concepts from a general-didactical, content-didactical and educational-psychological perspective. Second, teaching models have profited from integrating process-oriented psychological theories in order to explain the usage of opportunities by individual students. However, further steps have to be taken with regard to student activities and their role in being either part of the provision of learning opportunities for others, or being part of internal usages. Third, opportunity-usage models have not so far reached high international visibility.

Keywords: Models of Teaching and Learning, Opportunity-Use Model, Teaching Research, Learning Activities, Process Orientation

Anschrift der Autorin

Prof. Dr. Tina Seidel, Technische Universität München,
TUM School of Education,
Friedl Schöller-Stiftungslehrstuhl für Pädagogische Psychologie,
Arcisstraße 21, 80333 München, Deutschland
E-Mail: tina.seidel@tum.de

Themenblock III: Oberflächen- und Tiefenstruktur des Unterrichts

Jasmin Decristan/Miriam Hess/Doris Holzberger/Anna-Katharina Praetorius

Oberflächen- und Tiefenmerkmale

Eine Reflexion zweier prominenter Begriffe der Unterrichtsforschung

Zusammenfassung: Das Begriffspaar ‚Oberflächenmerkmale'/‚Oberflächenstrukturen' und ‚Tiefenmerkmale'/‚Tiefenstrukturen' hat zunehmend Einzug in die Forschungsliteratur erhalten, soll es doch eine Brücke zur Verknüpfung von Lehren und Lernen schlagen. Im vorliegenden Beitrag wird nach einem Überblick zu den historischen Wurzeln das Verständnis der jeweiligen Begriffe zusammengefasst. Anhand von Arbeiten aus der (Fach-)Didaktik und der pädagogisch-psychologischen Unterrichtsforschung werden unterschiedliche Konzeptualisierung von Tiefenmerkmalen aufgezeigt und Befunde zur Lernwirksamkeit von Tiefenmerkmalen angeführt. Der Beitrag schließt mit Vorschlägen zur begrifflichen Schärfungen und diskutiert theoretische und empirische Desiderata bezogen auf das Begriffspaar.

Schlagworte: Oberflächenmerkmale des Unterrichts, Sichtstruktur des Unterrichts, Tiefenmerkmale des Unterrichts, Tiefenstruktur des Unterrichts, Unterrichtsqualität

1. Oberflächen- und Tiefenmerkmale als prominentes Begriffspaar der Unterrichtsforschung[1]

Das zentrale Erkenntnisinteresse der Unterrichtsforschung bezieht sich auf die Frage ‚Wie kann Unterricht das Lernen der Schülerinnen und Schüler unterstützen?'. In diesem Zuge wurde immer wieder betont, dass es darum gehe, ‚teaching', und ‚learning' zu verknüpfen (vgl. Gage, 1963; Oser & Baeriswyl, 2001). Dabei wird Lernen als individueller, aktiver und/oder sozialer Prozess betrachtet, der durch Unterricht bzw. Lehren lediglich angeregt und begleitet werden kann.[2]

1 Wir danken insbesondere Eckhard Klieme, Anke Lindmeier und Christine Pauli für die zahlreichen fruchtbaren Impulse, die diesen Beitrag maßgeblich mit geprägt haben.

2 Wir beziehen uns an dieser Stelle – im Einklang insbesondere mit der neueren englischsprachigen Literatur – auf ein Verständnis von Lernen, das nicht outputbezogen, also auf Wissens- oder Kompetenzerwerb eingegrenzt ist, sondern prozessbezogen in einem (sozial-)konstruktivistischen Sinne.

Eine klassische Antwort auf die genannte Kernfrage besteht in der Nennung einzelner ‚teaching methods' (Wallen & Travers, 1963), Elementen von ‚teacher performance' (Rosenshine & Fürst, 1971) oder ‚teaching practices' (Walberg & Paik, 2000), die sich in empirischen Studien als förderlich für das Erreichen bestimmter Lernziele erwiesen haben (mit Bezug auf ein in den entsprechenden Überblicksarbeiten meist nicht näher spezifiziertes kognitives Ziel). Aktuelle Varianten solcher Merkmalslisten stellen unter anderem die prominente und kontrovers diskutierte Meta-Metaanalyse von Hattie (2009) oder die im angloamerikanischen Raum in Forschung und Praxis populäre Konzeption sogenannter ‚high leverage teaching practices' dar (vgl. Ball & Forzani, 2011).

Sowohl im internationalen wissenschaftlichen Diskurs (z. B. Snook, Clark, Harker, O'Neill & O'Neill, 2010) als auch in der deutschsprachigen Forschungslandschaft (z. B. Gruschka, 2007) sind solche Listen vielfältig kritisiert worden, auch weil sie theoretische Vorstellungen von Unterricht, insbesondere einer Verknüpfung von Lehren und Lern*prozessen,* vermissen lassen. Gleichzeitig gibt es Vorschläge, um dieses Theoriedefizit zu bearbeiten. Hierzu gehört die Metapher der Unterscheidung von ‚Oberfläche' und ‚Tiefe'. Die Grundidee ist es, ‚hinter' oder ‚unterhalb' des an der Oberfläche beobachtbaren Verhaltens eine Logik des pädagogischen Handelns zu erkennen, welche theoretisch beschreiben und empirisch prüfen kann, warum der entsprechende Unterricht lernen anregen und unterstützen kann. Damit soll die Unterscheidung in Oberfläche und Tiefe eine systematische Brücke schlagen zwischen Unterricht und Lernen – „Bridging Instruction and Learning", wie es Oser und Baeriswyl in ihrem wichtigen Aufsatz aus dem Jahr 2001 bezeichneten.

Tatsächlich hat in den letzten Jahren vor allem in der deutschsprachigen Unterrichtsforschung zunehmend eine Unterscheidung sogenannter Oberflächen- und Tiefenmerkmale für die Beschreibung und Analyse von Unterricht Einzug gehalten. Diese Unterteilung geht prinzipiell mit unbestreitbaren Vorteilen einher: Sie soll zum einen eine prägnantere, eingängigere und vor allem sparsamere Antwort liefern auf die eingangs gestellte Frage, wie Unterricht das Lernen unterstützen kann. Zum anderen soll sie die Aufmerksamkeit weg lenken vom Disput darüber, welche Unterrichtsmethode die effektivste ist – jenem Disput, der weite Teile der frühen (quantitativ-)empirischen Unterrichtsforschung dominierte. Anstelle einer bloßen Propagierung oder eines reinen Effektivitätsvergleichs bestimmter Methoden (z. B. Wochenplan) oder spezifischer Ablaufmuster (z. B. Direkte Instruktion) wird der Fokus darauf gelenkt, sich mit der Frage auseinanderzusetzen, *warum* der entsprechende Unterricht (oder eine bestimmte Methode) mehr oder weniger förderlich ist für Lernprozesse und deren Ergebnisse und *wie* Unterricht in diesem Sinne ‚lernwirksam' gestaltet werden kann. Ungeachtet dieser Chancen und der damit einhergehenden Popularität des Begriffspaars entsteht bei einer detaillierteren Betrachtung der gegenwärtigen Literatur der Eindruck, dass die Differenz von Oberflächen- und Tiefenmerkmalen je nach Forschungskontext sehr unterschiedlich verstanden wird. Vor diesem Hintergrund verfolgt der vorliegende Beitrag das Anliegen, Gemeinsamkeiten und Unterschiede zwischen verschiedenen Zugängen zu Oberflächen- und Tiefenmerkmalen herauszuarbeiten, sowie anhand eines detaillierteren Blicks auf Tiefenmerkmale zu diskutieren, inwieweit es das Begriffspaar ermög-

licht, tatsächlich eine theoretische Brücke zwischen Lehren und Lernen zu schlagen. Dazu wird im Folgenden zunächst auf historische Wurzeln der begrifflichen Unterscheidung von Oberfläche und Tiefe eingegangen.

2. Historische Wurzeln der begrifflichen Unterscheidung von Oberflächen- und Tiefenmerkmalen

Eine konzeptionelle Unterscheidung zwischen etwas ‚Greifbarem' und etwas ‚Dahinterliegendem', zwischen Erscheinung und Wesen, durchzieht die gesamte Philosophie und Erkenntnistheorie seit Platons Höhlengleichnis. Aber auch wenn es um pädagogisches Handeln geht, wird die genannte Gegenüberstellung bemüht. Beispielsweise wurde in der Allgemeinen Didaktik schon früh eine analoge Unterscheidung vorgenommen: So spricht Klingberg (1972) unter Bezugnahme auf Diesterweg von einer äußeren und einer inneren Seite der Unterrichtsmethodik. Während sich die äußere Seite auf methodische Grundformen und Sozialformen bezieht, ist die innere Seite auf den ‚methodischen Gang' bestimmt, welcher nicht auf Anhieb erfasst werden kann.

Roth (1965) und Aebli (1983) verknüpfen pädagogische Überlegungen zur Vorbereitung und Gestaltung von Unterricht mit lerntheoretischen Grundlagen und stellen im Gegensatz zu Klingberg somit das Lernen in den Fokus. Aus einer handlungstheoretischen Perspektive heraus wird unter Rückgriff auf Dewey und Piaget der Zusammenhang von Denken und Handeln und somit der operatorische Charakter des Denkens herausgestellt. Aebli (1983) betont, dass Lernen eine innere Aktivität darstellt, die es gilt durch Lehren anzuregen und zu unterstützen. In der Konsequenz arbeitet er zwölf Grundformen eines handlungsorientierten Unterrichts heraus. Auch Roth (1965) betrachtet Lernen als einen inneren Prozess und unterscheidet zwischen verschiedenen (Lern-)Ausgangslagen, Lernschritten und Lernzielen. Um Erkenntnis- und Denkprozesse anzuregen, führt Roth auf Basis empirischer Ergebnisse spezifische Lernhilfen an (z. B. Strukturierungen, Übungen, Lob und Tadel).

Die Begrifflichkeiten Sicht-/Oberflächen- und Tiefenstrukturen selbst wurden unserer Kenntnis nach erstmals von Chomsky (1965) im Kontext der sogenannten generativen Transformationsgrammatik verwendet, um den Prozess der Sprachproduktion und -rezeption zu beschreiben: Auf der Tiefenebene enthält ein Satz abstrakte syntaktische Informationen und semantische Relationen. Durch sprachliche Transformationen lassen sich verschiedene Sätze auf der Oberflächenebene bilden, die aber eine identische Bedeutung haben. Im unterrichtlichen Kontext wurde das Begriffspaar von Oser und Patry (1990) eingeführt: „Betrachtet man den Verlauf eines einfachen Lernprozesses (…), so kann man einerseits sichtbares Aufeinanderreihen von Handlungen feststellen (Sichtstruktur), und man muss zugleich annehmen, dass diesen sichtbaren Abläufen verborgene gesetzmäßige Verkettungen von Lernschritten (Basisstruktur) zugrunde liegen" (S. 3). Die Autoren sprechen von Basismodellen (teils synonym für Tiefenstrukturen, s. u.), die jeweils aus einer festgelegten Abfolge bestimmter Lernschritte bestehen:

> Die Basisstruktur besteht aus einer für jeden Lernenden absolut notwendigen, feststehenden Kette von Operationen, die nicht durch etwas anderes ersetzt werden kann. Der ganzheitliche Charakter dieser jeweiligen Kette wird bestimmt durch lernpsychologische Gesetzmäßigkeiten einerseits und durch den Typ des Ziels bzw. die Inhalte andererseits. (Oser & Patry, 1990. S. 3)

Seitdem haben die Begrifflichkeiten in viele Publikationen der Unterrichtsforschung im deutschsprachigen Raum Eingang gefunden. Ausgehend von dem Eindruck, dass das Begriffspaar – vor allem aber der Begriff Tiefenmerkmale – in einigen Publikationen teils in recht verkürzter Form und teils mit unterschiedlichen Grundannahmen verwendet wurde, werden im Folgenden auf Basis einer Literaturrecherche in Fachdatenbanken[3] wesentliche Charakteristika gegenwärtiger Verständnisse von Oberflächen- und Tiefenmerkmalen herausgearbeitet. Rahmend sei dabei erwähnt, dass sich diese Unterscheidung exklusiv im Kontext der deutschsprachigen Unterrichtsforschung finden lässt und in englischsprachigen Publikationen lediglich in Form von Übersetzungen („visible or sight structure of a lesson“ und „deep structure of learning“, Klieme, Pauli & Reusser, 2009, S. 139; Oser & Baeriswyl, 2001, S. 1032) Einzug erhalten hat.

3. Oberflächen- bzw. Sichtmerkmale

Oberflächenmerkmale – häufig auch bezeichnet als *Sichtstrukturen* oder Sichtmerkmale – werden oft als für Außenstehende leicht erschließbare, „verhaltensnahe, gut beobachtbare und abgrenzbare Merkmale“ des Unterrichts (Pauli & Reusser, 2006, S. 783) charakterisiert. Gleichermaßen wird auch auf deren Vielfältigkeit und Austauschbarkeit verwiesen: Auf der „sichtbaren Oberfläche“ des Unterrichts (Reyer, 2004, S. 21) befinden sich methodische und organisatorische Gestaltungsmerkmale (Pauli, 2012, S. 14), Sozial- und Inszenierungsformen und Medien (Reusser, 2008, S. 231), strukturelle Rahmenbedingungen (Kunter & Trautwein, 2013, S. 65), Aufgabenmaterial und gut beobachtbares Lehrer- und Schülerverhalten (Pauli & Reusser, 2006, S. 784). „Die Sichtstruktur ist also das Wechselhafte, Austauschbare, das an den Lernenden und vom Lernenden immer neu Adaptierbare“ (Oser & Patry, 1990, S. 3). Die Sichtstruktur stellt somit das freie und frei zu gestaltende Moment des Lernverlaufs dar, sodass Unterricht auf der Ebene von Oberflächenmerkmalen „prinzipiell in fast grenzenloser Vielfalt gestaltet werden“ kann (Fischer et al., 2003, S. 182). In der Unterrichtspraxis spiegelt sich diese Vielfalt jedoch nicht wider, vielmehr sind Oberflächenmerkmale hochgradig typisiert und es zeigt sich ein eingeschränktes, fach- und schulformspezifisches Methodenrepertoire (z. B. Hage, Bischoff & Dichanz, 1985; Oser & Baeriswyl, 2001).

3 Hierfür wurde per Schlagwortsuche nach den Begriffen Oberfläche, Sicht und Tiefe gesucht. Anschließend wurden 40 Beiträge aus den Jahren 1990 bis 2018 gesichtet.

4. Tiefenmerkmale

Tiefenmerkmale gelten dagegen allenfalls als indirekt beobachtbar (z. B. Kunter & Trautwein, 2013; Pauli & Reusser, 2006; Reyer, 2004) und ihnen wird eine zentrale Rolle für das Lernen der Schülerinnen und Schülern beigemessen (z. B. Fischer, Reyer, Wirz, Bos & Höllrich, 2002; Kunter & Trautwein, 2013; Reusser, 2008). Jenseits dieses gemeinsamen Nenners lassen sich bei einer näheren Betrachtung der Literatur jedoch ganz unterschiedliche Konzeptualisierungen von Tiefenmerkmalen ausmachen. Zwei prominente Zugänge sollen hier exemplarisch weiter vertieft werden, um die Unterschiedlichkeit in den Annahmen zu Tiefenmerkmalen zu verdeutlichen, nämlich (fach-) didaktische Arbeiten unter Bezugnahme auf die Theorie der Basismodelle und pädagogisch-psychologische Arbeiten unter Bezugnahme auf Unterrichtsqualitätsdimensionen.

4.1 Theorie der Basismodelle

Oser und Kollegen (z. B. Oser & Baeriswyl, 2001; Oser & Patry, 1990) spezifizieren sogenannte ‚Basismodelle', welche die sich hinter einer Sichtstruktur befindende Ebene darstellen. Für jedes spezifische Lernziel (z. B. Problemlösen oder Werte- und Identitätsaufbau) wird ein Basismodell definiert. Jedem der zwölf Basismodelle liegt eine gesetzmäßige Verkettung von Lernschritten zugrunde, die einen Lernweg zum spezifischen Lernziel beschreiben (vgl. Fischer et al., 2003, S. 183). Erst diese spezifische Verknüpfung der einzelnen Lernschritte als Handlungskettenelemente bildet insgesamt einen Lernweg und bietet dadurch die Grundlage einer Tiefenstrukturierung des Unterrichts (z. B. Krumbacher, 2016; Reyer, 2004). Die gewählte Terminologie (Lernschritte und -wege, Handlungskettenelemente) lässt sich als Ausdruck einer handlungstheoretischen Perspektive auf Lernen interpretieren. In den Basismodellen wird eine Zielperspektive des Lernens eingenommen. Anschließend wird das Lernen der Schülerinnen und Schüler in den Fokus gestellt und Lernprozesse werden strukturiert. Entsprechend lassen sich Parallelen zur bildungstheoretischen Didaktik von Klafki (1985) und der lerntheoretischen Didaktik von Aebli (1983) und Roth (1965) sowie Herbarts Formalstufentheorie finden (einen Überblick geben Oser & Baeriswyl, 2001). In fachdidaktischen Operationalisierungen und Erweiterungen der Theorie der Basismodelle wird teil sowohl für die Lehrenden als auch für die Lernenden jeweils eine separate Sichtstruktur und ein Basismodell spezifiziert (z. B. Fischer et al., 2002; Krumbacher, 2016; Reyer, 2004). Dem an der Oberfläche verorteten Verhalten von Lehrkräften und von Schülerinnen und Schülern liegt dann jeweils eine separate Tiefenebene zugrunde. Bei den Lehrkräften ist dies ein entsprechender Lehrzieltyp (z. B. Typ Erfahrungswissen oder Problemlösen), bei den Schülerinnen und Schülern sind es die argumentativen Denkschritte bzw. Handlungskettenschritte.

Wichtig zu erwähnen ist, dass es kein einheitliches Verständnis dazu gibt, ob Basismodelle mit Tiefenstrukturen gleichzusetzen sind. In späteren Arbeiten findet sich

oftmals eine Gleichsetzung (z. B. Fischer et al., 2003; Krumbacher, 2016; Reusser, 2009): So bezeichnet beispielsweise Krumbacher (2016) die Tiefenstrukturierung als „bewusste Sequenzierung mentaler Verarbeitungsprozesse" (S. 31). Und auch Reusser (2009, S. 888) formuliert, dass „sich die Tiefenstruktur auf dessen [des Unterrichts] invariante, psychologisch notwendige Basisprozesse und Elemente" bezieht. Oser und Baeriswyl (2001) selbst setzen im Aufsatz jedoch nur an einer Stelle beide Begriffe gleich („deep structure of learning (basis-model)"; S. 1032). Und auch Oser und Sarasin (1995) bezeichnen lediglich allgemein Sichtstrukturen als „Ausdruck von Tiefenstrukturen" (S. 2), ohne den expliziten Bezug zu Basismodellen herzustellen. Reyer (2004) wiederum unterscheidet Basismodelle von Tiefenstrukturen und sieht letztere als „die syntaktischen Regeln, Abläufe und Verknüpfungen zwischen den als Oberflächenmerkmalen sichtbaren Aspekten von Unterricht" (S. 60) an. Aus dieser Formulierung lassen sich Tiefenmerkmale (auch) als Verknüpfung von Basis-Modell und Oberflächenmerkmalen deuten. Diese Verknüpfung aus Lehren und Lernen, aus Methodenfreiheit einerseits und festen Lernschritten andererseits wird von Oser und Baeriswyl (2001) wiederum als ‚Choreographie' des Unterrichts bezeichnet.

4.2 Pädagogisch-psychologische Unterrichtsqualitätsforschung

In der pädagogisch-psychologischen Unterrichtsforschung sind Tiefenmerkmale mit Lehren *und* Lernen verbunden. Reusser und Pauli (2010) verstehen unter Bezug auf Aeblis kognitionspsychologische Didaktik Tiefenmerkmale als „jene psychologischen Prozesse und Merkmale des Lehrens und Lernens, welche dem Unterricht als psychologisch-didaktische Qualitätsdimensionen zugrunde liegen" (S. 19; vgl. auch Hasselhorn & Gold, 2013; Kunter & Trautwein, 2013; Reusser, 2008). Im Gegensatz zur Theorie der Basismodelle beziehen sich Tiefenmerkmale nicht auf die Unterrichts*planung,* d. h. die didaktische Strukturierung des Unterrichts passend zum angestrebten Lernziel (als generative Funktion der Tiefenstruktur, s. u.), sondern ausschließlich auf die Unterrichts*durchführung.* Tiefenmerkmale sind auf der „Mikroebene des unterrichtlichen Handelns" (Hasselhorn & Gold, 2013, S. 375) verortet und beziehen sich dabei auf die Interaktionen zwischen Lehrenden und Lernenden sowie deren Auseinandersetzung mit dem Lerninhalt (z. B. Kunter & Trautwein, 2013). Von dieser Perspektive ausgehend werden Tiefenmerkmale in der pädagogisch-psychologischen Literatur oft mit der Qualität[4] unterrichtlicher Prozesse gleichgesetzt (vgl. Hasselhorn & Gold, 2013; Klieme, Schümer & Knoll, 2001; Kunter & Trautwein, 2013; Pauli, 2012;

4 Nach Berliner (2005) sind Qualitätsurteile sowohl ein Ergebnis normativer Setzungen wie gesellschaftlichen, kulturellen und historischen Vorstellungen von gutem Unterricht (‚good teaching') als auch von Wirksamkeitsprüfungen (‚effective teaching'). Diesem Verständnis zufolge ist der Qualitätsbegriff streng genommen breiter zu fassen und nicht mit Tiefenmerkmalen gleichzusetzen.

Reusser & Pauli, 2010). Als Unterrichtsqualitätsmerkmale werden häufig die im deutschen Sprachraum weit verbreiteten und an internationale Konzeptionen anschlussfähigen ‚Basisdimensionen' angeführt (vgl. Klieme et al., 2001, S. 50): kognitive Aktvierung (im Sinne eines zum vertieften Nachdenken anregenden Unterrichts), konstruktive Unterstützung (u. a. als wertschätzende Lehrer-Schüler-Interaktion und konstruktives Feedback) und Klassenführung (u. a. als effiziente Zeitnutzung, reibungslose Übergänge sowie effizienter Umgang mit Disziplinstörungen). Weitere Beiträge in diesem Heft (Kleickmann, Steffensky & Praetorius; Praetorius et al.) gehen genauer auf diese Dimensionen ein.

Abweichend von Oser wird nicht von einem spezifischen Lernziel ausgegangen, sondern es wird der Anspruch einer fach- und lernzielunabhängigen Gültigkeit erhoben (Klieme et al., 2001). Die zu den Qualitätsdimensionen spezifizierten Vorstellungen zur erfolgreichen Verknüpfung von Lehren und Lernen werden größtenteils nicht aus der Didaktik gespeist, sondern primär aus pädagogisch-psychologischen Theorien. Anstelle einer Spezifikation und fixen Verkettung kognitiver Schritte werden unter Verweis auf Konzepte und Theorien wie ‚Verarbeitungstiefe', ‚time on task' und ‚Selbstbestimmung' vielmehr allgemeine kognitive und motivationale Wirkmechanismen angenommen (vgl. auch Klieme, Lipowsky, Rakoczy & Ratzka, 2006). Zur Verknüpfung von Lehren und Lernen wird oft auf Angebot-Nutzungs-Modelle (z. B. Fend, 1980) verwiesen und dabei zwischen Lehren als Angebot und Nutzung als kognitive und motivationale Prozesse des Lernens unterschieden, ohne dass diese Prozesse bislang lerntheoretisch ausgearbeitet wurden. Eine Verortung von Tiefenmerkmalen rein auf der Angebotsseite lässt sich zwar aus der Forschungstradition ableiten, ist aber vor dem Hintergrund der Konzeptualisierung von Tiefenmerkmalen als Prozesse des Lehrens und Lernens kritisch zu hinterfragen (für weitergehende kritische Auseinandersetzungen siehe Vieluf, Praetorius, Rakoczy, Kleinknecht & Pietsch, in diesem Heft).

5. Empirische Zugänge zur Erfassung von Tiefenmerkmalen

In Anlehnung an Reyer (2004, S. 60) lassen sich zwei Funktionen von Tiefenmerkmalen unterscheiden: Die ‚generative Funktion' von Tiefenmerkmalen ist es, (als Planungsgrundlage für Lehrkräfte) die Oberfläche zu erzeugen (vgl. auch Oser & Baeriswyl, 2001). Die „rezeptive Funktion" besteht wiederum darin, auf Basis der Oberfläche Schlussfolgerungen auf die darunterliegende latente Tiefenebene zu ziehen. Insofern lassen sich für theoretische Konzeptualisierungen und empirische Zugänge unterschiedliche Wege beschreiten.

Mit Bezug auf die Basismodelle von Oser und Kollegen wird empirisch versucht, mittels der Kodierung von ‚Inhaltshandlungen' oder ‚Lehrzieltypen' interpretative Schlussfolgerungen zu den hinter einer Oberfläche liegenden Tiefenmerkmalen zu ziehen (z. B. Reyer, 2004). Es ist jedoch festzuhalten, dass die theoretischen Annahmen zu einer Tiefenstrukturierung von Unterricht einer empirischen Prüfung bislang nicht standhalten können. So fasst Krumbacher (2016, S. 34) zusammen:

> Obwohl inzwischen etliche empirische Studien zu den Basismodellen durchgeführt wurden (…), kann noch keine Kausalbeziehung zwischen einem basismodell-konformen Unterrichtsangebot und dem Lernzuwachs der Schülerinnen und Schüler hergestellt werden. Insofern lässt sich eine Lernwirksamkeit der Basismodelle bis heute nicht eindeutig empirisch belegen.

In der pädagogisch-psychologischen Unterrichtsforschung werden meist Unterrichtsqualitätsmerkmale herangezogen, um Tiefenmerkmale und somit Prozesse des Lehrens und Lernens zu beschreiben. Den Operationalisierungen liegen jedoch ganz unterschiedliche methodische Herangehensweisen zugrunde, die sich sowohl hinsichtlich des Ausmaßes an nötigen Inferenzen als auch bezüglich des Einbezugs von Indikatoren auf Seiten der Lehrenden und/oder Lernenden deutlich unterscheiden. Insofern muss stets kritisch hinterfragt werden, inwieweit die jeweilige Operationalisierung noch mit den theoretischen Grundannahmen zu Tiefenmerkmalen vereinbar ist. Beispielsweise wird durch eine ausschließliche oder separate Operationalisierungen über das Verhalten von Lehrkräften eine Trennung von Lehren und Lernen vorgenommen, die oft nur unzureichend mit den theoretischen Vorstellungen von Tiefenmerkmalen verknüpft ist. Berücksichtigt man diese Vielfalt der Operationalisierungen, überrascht nicht, dass Praetorius, Klieme, Herbert und Pinger (2018) in ihrer Zusammenschau bisheriger längsschnittlicher Mehrebenen-Studien zum Modell der drei Basisdimensionen zu dem Schluss kommen, dass sich positive Effekte dieser Tiefenmerkmale auf fachliche Leistungen nur partiell bestätigen lassen.

In Teilen der Literatur wird begrifflich vom Anspruch Abstand genommen, Lernprozesse explizit mit in den Blick zu nehmen. Daher schlagen beispielsweise Kunter und Trautwein (2013) sowie Lotz (2015) vor, vom Potenzial zur kognitiven Aktivierung zu sprechen. Auch Seidel (2003) bringt den Umstand, dass Lernprozesse nicht direkt erfassbar sind, mit dem Begriff Gelegenheitsstrukturen zum Ausdruck.

Als empirische Hinweise für eine höhere Lernwirksamkeit von Tiefenmerkmalen im Vergleich zu Oberflächenmerkmalen werden auch Arbeiten aus der internationalen Forschung zu Schul- und Unterrichtseffektivität herangezogen. Im vorliegenden Beitrag soll exemplarisch die prominente Meta-Metaanalyse von Hattie (2009) eingehender betrachtet werden. So fasst Hattie beispielsweise einzelne Unterrichtsmerkmale zu Kategorien wie ‚working conditions' oder ‚teaching' zusammen (S. 244). Die „working conditions" können entsprechend der obigen Beschreibung recht eindeutig den Oberflächenmerkmalen zugeordnet werden. Entsprechend der theoretischen Annahmen zeigen sich auch empirisch konsistent geringe Effektstärken (z. B. ‚within-class grouping' mit $d = 0.16$, ‚ability grouping' mit $d = 0.12$. Die unter „teaching" angeführten Merkmale weisen dagegen allesamt bedeutsame Effekte auf (z. B. ‚teacher-student relationships' mit $d = 0.72$). Allerdings können nicht alle dort angeführten Merkmale zweifelsfrei der Tiefenebene zugeordnet werden. Zum Teil handelt sich um (gut beobachtbare) Unterrichtsmethoden, die ebenso der Oberfläche hätten zugeordnet werden können (z. B. ‚reciprocal teaching' mit $d = 0.74$ oder ‚direct instruction' mit $d = 0.59$). Erst ein genauerer Blick auf die jeweils implizierten theoretischen Wirkmechanismen ermöglicht erste

theoretische Brückenbausteine zur Tiefenebene des Unterrichts: Reziprokes Lernen stellt beispielsweise ein stark strukturiertes Setting zur Instruktion von Lernstrategien dar (vgl. Rosenshine & Meister, 1993) und auch die Direkte Instruktion ist ein Maßnahmenbündel, das transparente Lernziele, eine fortlaufende Lerndiagnostik bei (korrektivem) Feedback und ein hohes Maß an Strukturierung umfasst (vgl. Rosenshine & Stevens, 1986). Und auch mit Bezug auf ein pädagogisch-psychologisches Verständnis von Tiefenmerkmalen lassen sich im Rahmen der Umsetzung der Methode deutliche Unterschiede erwarten. Vor diesem Hintergrund liefern Befunde der Effektivitätsforschung allenfalls Hinweise für eine höhere Lernwirksamkeit von Tiefenmerkmalen.

6. Reflexion von Oberflächen- und Tiefenmerkmalen des Unterrichts

In der abschließenden Reflexion von Oberflächen- und Tiefenmerkmalen werden zunächst Vorschläge zur begrifflichen Schärfung unterbreitet. Anschließend wird auf die Lernwirksamkeit von Oberflächen- und Tiefenmerkmalen eingegangen und schließlich diskutiert, inwieweit das Begriffspaar eine theoretische Brücke zur Verknüpfung von Lehren und Lernen bieten kann.

6.1 Begriffliche Schärfungen

Die bisherigen Anführungen haben verdeutlicht, dass es einer Schärfung der Begriffe Sichtstrukturen und Tiefenstrukturen bedarf. Als erstes schlagen wir vor, den Begriff ‚Sicht' durch ‚Oberfläche' zu ersetzen. Die Terminologie Sichtstrukturen wird oft damit begründet, dass diese im Vergleich zu Tiefenmerkmalen besonders gut sichtbar seien (z. B. Kunter & Trautwein, 2013; Pauli & Reusser, 2006; Reyer, 2004). Bei einer genaueren Betrachtung des Begriffspaars lässt sich jedoch schlussfolgern, dass die Sichtbarkeit kein zentrales Kriterium sein kann. So liegt in der (Unterrichts-)Forschung stets eine zentrale Herausforderung darin, nicht direkt sichtbare (latente) Merkmale, wie psychologische Prozesse des Lehrens und Lernens, zu erfassen. Diese müssen letztlich für eine empirische Prüfung stets mit Hilfe von Operationalisierungen über (Verhaltens-) Indikatoren zugänglich gemacht und somit immer auf eine sicht- und interpretierbare Ebene gehoben werden – etwa um die ‚rezeptive Funktion' von Tiefenmerkmalen abzubilden. Dies bedeutet jedoch nicht, wie bereits Oser und Baeriswyl (2001, S. 1048) angemerkt haben, dass sich auch theoretisch die Tiefenmerkmale aus Oberflächenmerkmalen ableiten lassen und es bedeutet auch nicht, dass alle Oberflächenmerkmale gleichermaßen Ausdruck von Tiefenmerkmalen sind. Erst durch theoretische Vorstellungen zur Konzeptualisierung von Tiefenmerkmalen und Auseinandersetzungen damit, wie und warum Unterricht Lernprozesse anregen kann, lassen sich Tiefenmerkmale spezifizieren.

Ein weiteres Argument für die Unterscheidung von Oberflächen- und Tiefenmerkmalen ist, dass es der Unterrichtsforschung, wenn sie das Ziel der empirisch abgesicherten

Theoriebildung verfolgt, inhaltlich nicht um etwas ‚Sichtbares' geht, sondern vielmehr um die Frage, wie auf einer ‚tieferen' Ebene Unterricht und Lernen zusammenhängen. Warum aber ist immer wieder von ‚Sichtbarkeit' die Rede? Eine banale, aber im Forschungsprozess wichtige Erklärung könnte der oftmals vorgenommene Zugang durch Analyse von Unterrichtsvideos bieten. Bei anderen Zugängen, wie fragebogenbasierten Erhebungen, sind Sichtstrukturen als Begriff eher noch problematischer, denn die Antworten, auch auf standardisierte Befragungen, beruhen immer auf individuell gefilterten und interpretierten Wahrnehmungen (vgl. Fauth, Göllner, Lenske, Praetorius & Wagner, in diesem Heft). Zusammengenommen bestimmen im Wesentlichen die theoretischen Grundannahmen und nicht die Sichtbarkeit, ob ein Unterrichtsmerkmal theoretisch auf der Oberfläche oder in der Tiefenebene verortet wird.

Eine zweite Schärfung bezieht sich auf den Begriff ‚Strukturen'. Er wurde maßgeblich durch die Basismodelle geprägt und hat in diesem Kontext eine zentrale Bedeutung – geht es doch um spezifische Strukturierungen kognitiver Denkschritte. Oberflächenmerkmale hingegen sind dadurch charakterisiert, dass sie gerade nicht mit spezifischen Sequenzierungen oder Strukturierungen verbunden sind, sondern eine Vielzahl an Kategorisierungen und Beschreibungen zulassen und als austauschbar gelten. Bereits Reyer (2004, S. 59) schlägt vor, stattdessen den allgemeineren Begriff ‚Oberflächenmerkmale' zu verwenden. Für Tiefenmerkmale gilt hingegen, dass stets genauer reflektiert und begründet werden sollte, ob der Strukturbegriff in dem verwendeten Zusammenhang angemessen ist. So geht es in der Forschung zu Unterrichtsqualitätsdimensionen primär um die Unterscheidbarkeit von Merkmalen und allenfalls implizit um die (zeitliche, inhaltliche, soziale) Strukturierung von Unterrichtsprozessen. Insofern empfehlen wir, zunächst allgemeiner von Oberflächen- oder Tiefenmerkmalen des Unterrichts oder von verschiedenen „Ebene(n) des Unterrichtsgeschehens" (Reusser & Pauli, 2010, S. 19) zu sprechen.

6.2 Zur Unterstützung von Lernprozessen durch Oberflächen- und Tiefenmerkmale

Die theoretischen Annahmen zu Tiefenmerkmalen legen – unabhängig von der jeweiligen Konzeptualisierung – nahe, dass Tiefenmerkmale besonderes Potenzial zur Unterstützung der Lernprozesse von Schülerinnen und Schülern haben. Dennoch können empirische Studien diese Annahme nicht konsistent bestätigen (vgl. Abschnitt 5). Bislang lässt sich nicht hinreichend klären, inwiefern die theoretischen Annahmen zu den Tiefenmerkmalen modifiziert werden müssen oder inwiefern die jeweiligen Operationalisierungen nicht hinreichend valide sind.

Bislang zu wenig Beachtung hat in diesem Zuge der Fachinhalt erfahren. Nicht zuletzt die Meta-Analyse von Seidel und Shavelson (2007) liefert Belege für die Relevanz der Fachlichkeit für Lernprozesse. Wie sich allerdings Fachlichkeit und Tiefenstrukturen zueinander verhalten, bleibt gegenwärtig ein zu klärendes Forschungsfeld. Beispielsweise liefern Lipowsky, Drollinger-Vetter, Klieme, Pauli und Reusser (2018) Hin-

weise, dass sich neben generischen auch fachliche Dimensionen von Unterrichtsqualität identifizieren lassen.

Auch das Verhältnis von Oberflächen- und Tiefenmerkmalen an sich und in ihrem Zusammenspiel mit Bezug auf die Unterstützung von Lernprozessen ist bislang nicht hinreichend geklärt (vgl. auch Reusser & Pauli, 2010, S. 19). Folgt man theoretischen Vorstellungen, sollten Oberflächen- und Tiefenmerkmale unabhängig voneinander sein. Bereits Oser und Baeriswyl (2001, S. 1043) nehmen im Kontext ihrer Choreographie-Metapher an, dass ein bestimmter Lernschritt durch ganz unterschiedliche Ausgestaltungen des Unterrichts an der Oberfläche adressiert werden kann. Und auch im Rahmen einer pädagogisch-psychologischen Konzeptualisierung von Tiefenmerkmalen lässt sich zunächst einmal annehmen, dass die Qualität unterrichtlicher Interaktionen nicht mit der Ausgestaltung von Unterricht an der Oberfläche korrespondiert. Andererseits legen theoretische Überlegungen zu den Wirkmechanismen verschiedener Methoden, Materialien und Sozialformen nahe, dass Oberflächenmerkmale ein ‚spezifisches Anregungspotenzial' bieten, also unterschiedlich gut geeignet sein könnten, um ein bestimmtes Lernziel zu erreichen. Oberflächenmerkmale können somit unterschiedlich gut geeignete Werkzeuge zur Anregung kognitiver und motivationaler Prozesse sein (vgl. bereits Bransford, Brophy & Williams, 2000, S. 61). Auch Befunde der Effektivitätsforschung zu Direkter Instruktion und zum Reziproken Lehren deuten darauf hin, dass es einige Methoden gibt, die zwar primär auf der Oberflächenebene verortet sind, aber durch ihre theoretischen Wirkmechanismen eng mit der Unterstützung kognitiver Prozesse verbunden sind. Dies lässt sich in den Wirkungsmodellen der quantitativ-empirischen Unterrichtsforschung (vgl. auch Köhler, Kuger, Naumann & Hartig, in diesem Heft; Naumann, Kuger, Köhler & Hochweber, in diesem Heft) abbilden, indem Tiefenmerkmale als vermittelnde oder moderierende Variablen mit Bezug auf die Wirksamkeit von Oberflächenmerkmalen eingeführt werden (vgl. auch Decristan et al., 2015).

6.3 Zur Ausgangsfrage: Verknüpfung von Lehren und Lernen

Wie eingangs formuliert, besteht eine zentrale Herausforderung der Unterrichtsforschung darin, Lehren und Lernen zu verknüpfen und somit eine Antwort auf die Frage zu geben, wie Unterricht das Lernen der Schülerinnen und Schüler unterstützen kann. Die Unterscheidung in Oberflächen- und Tiefenmerkmale des Unterrichts lässt sich als ein Versuch bezeichnen, eine theoretische Brücke zwischen Lehren und Lernen zu schlagen (vgl. Oser & Baeriswyl, 2001).

In der Theorie der Basismodelle von Oser und Kollegen (z. B. Oser & Patry, 1990) wird die Verknüpfung von Lehren und Lernen als Choreographie des Unterrichts bezeichnet (vgl. Oser & Baeriswyl, 2001) und damit das Zusammenspiel aus Freiheit bei der Ausgestaltung des Unterrichts an der Oberfläche bei gleichzeitigen Restriktionen bezüglich des Lernweges. Jeder Lernweg hängt vom jeweils verfolgten Lernziel ab. Diese Zieldimension des Lernens ist zentraler Bestandteil der bildungstheoretischen Didaktik (Klafki, 1985) und scheint ein besonderes Potenzial zur Verknüpfung von Lehren

und Lernen zu haben, da stets die Frage mit berücksichtigt wird, welche Zielsetzung der Unterricht bzw. die Lehrkraft in der jeweiligen Unterrichtssituation konkret verfolgt. Wenn die empirische Unterrichtsforschung systematisch die Zielperspektiven berücksichtigen würde, könnte sie möglicherweise nicht mehr von allgemeingültigen, generischen Qualitätsdimensionen sprechen.

Zudem wird in der Theorie der Basismodelle entsprechend lerntheoretischer Ansätze (z. B. Aebli, 1983; Roth, 1965) explizit vom Lernen der Schülerinnen und Schüler aus gedacht. Auch liefern didaktische Vorstellungen, die von ‚Lernen als Bezugspunkt für die Gestaltung von Unterricht' ausgehen, kognitionspsychologische Grundlagen des Lernens berücksichtigen und Ergebnisse der empirischen Lehr-Lern-Forschung einbeziehen (z. B. Tulodziecki, Herzig & Blömeke, 2017), eine wichtige Grundlage für die Verknüpfung von Lehren und Lernen. Wie oben mit Bezug auf Aebli und Roth argumentiert, bedarf eine didaktisch informierte und empirisch gestützte Unterrichttheorie der engen Verbindung mit Lerntheorien.

Die pädagogisch-psychologische Unterrichtsqualitätsforschung liefert einen anderen Baustein, um die Brücke zwischen Lehren und Lernen zu schlagen: Hierbei werden vor allem die Phase der Unterrichtsgestaltung sowie Prozesse des Lehrens und Lernens gemeinsam in den Blick genommen. Während die Theorie der Basismodelle vor allem eine Antwort auf die Frage danach, *warum* eine Methode eingesetzt wird, bietet, können die über Unterrichtsqualitätsdimensionen abgeleiteten Erkenntnisse Antworten auf die Frage geben, *wie* die Inhalte und Methoden umgesetzt werden sollten. Lernen wird hierbei deutlicher aus einer konstruktivistischen Perspektive betrachtet. Mittels Angebot-Nutzungs-Modellen wird veranschaulicht, dass Lernen stets ein eigenständiger aktiver Prozess ist, der durch Unterricht lediglich unterstützt oder angeregt werden kann. Um die theoretische Brücke zum Lernen (als Ergebnis) zu schlagen, werden als ‚Nutzungsprozesse' nicht nur kognitive, sondern auch motivationale Mechanismen des Lernens betrachtet. Diese Prozesse bilden in der pädagogisch-psychologischen Unterrichtsqualitätsforschung allerdings weder einen zentralen Ausgangspunkt für die Konzeptualisierung von Tiefenmerkmalen noch haben sie wesentliche Implikationen für deren Erforschung. Entsprechend sind sie gegenwärtig nicht weiter theoretisch spezifiziert. Vielmehr werden als empirische Evidenz für die postulierten Annahmen oftmals die über Tests erfassten fachlichen Lernergebnisse herangezogen.

Zusammengenommen können die verschiedenen Konzeptualisierungen des Begriffspaars aus Oberfläche und Tiefe unterschiedliche theoretische und empirische Brückenbausteine liefern, um Lehren und Lernen zu verknüpfen. Diese – technisch gesehen mediierenden und moderierenden – Brücken gilt es jedoch in Zukunft präziser zu erforschen und hieraus überzeugendere theoretische Grundlagen zur Verknüpfung von Lehren und Lernen auszuarbeiten. Die Zugänge aus (Fach-)Didaktik und pädagogisch-psychologischer Forschung können hierfür, trotz unterschiedlicher Konzeptualisierungen, wichtige sich ergänzende Bausteine liefern.

Literatur

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

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education, 60,* 497–511.

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56,* 205–213.

Bransford, J., Brophy, S., & Williams, S. (2000). When computer technologies meet the learning sciences: Issues and opportunities. *Journal of Applied Developmental Psychology, 21*, 59–84.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, Massachusetts: MIT Press.

Decristan, J., Hondrich, A. L., Büttner, G., Hertel, S., Klieme, E., Kunter, M., Lühken, A., Adl-Amini, K., Djakovic, S.-K., Mannel, S., & Hardy, I. (2015). Impact of additional guidance in science education on primary students' conceptual understanding. *The Journal of Educational Research, 108,* 358–370.

Fend, H. (1980). *Theorie der Schule.* München: Urban & Schwarzenberg.

Fischer, H. E., Reyer, T., Wirz, C., Bos, W., & Höllrich, N. (2002). Unterrichtsgestaltung und Lernerfolg im Physikunterricht. In M. Prenzel & J. Doll (Hrsg.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen* (45. Beiheft der Zeitschrift für Pädagogik, S. 124–138).

Fischer, H. E., Klemm, K., Leutner, D., Sumfleth, E., Tiemann, R., & Wirth, J. (2003). Naturwissenschaftsdidaktische Lehr-Lernforschung: Defizite und Desiderata. *Zeitschrift für Didaktik der Naturwissenschaften, 9,* 179–209.

Gage, N. L. (Hrsg.) (1963). *Handbook of research on teaching.* Chicago, IL: Rand McNally.

Gruschka, A. (2007). „Was ist guter Unterricht?". Über neue Allgemein-Modellierungen aus dem Geiste der empirischen Unterrichtsforschung. *Pädagogische Korrespondenz, 36,* 10–43.

Hage, K., Bischoff, H., & Dichanz, H. (1985). *Das Methoden-Repertoire von Lehrern. Eine Untersuchung zum Unterrichtsalltag in der Sekundarstufe I.* Opladen: Leske und Budrich.

Hasselhorn, M., & Gold, A. (2013). *Pädagogische Psychologie: Erfolgreiches Lernen und Lehren* (3. Auflage). Stuttgart: Kohlhammer.

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

Klafki, W. (1985). *Neue Studien zur Bildungstheorie und Didaktik: Beiträge zur kritisch-konstruktiven Didaktik.* Beltz: Weinheim.

Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule* (S. 127–146). Münster: Waxmann.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In J. Tomáš & T. Seidel (Hrsg.), *The power of video studies in investigating teaching and learning in the classroom* (S. 137–160). Münster: Waxmann.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur" und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.), *TIMSS – Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 43–58). Bonn: Bundesministerium für Bildung und Forschung.

Klingberg, L. (1972). *Einführung in die Allgemeine Didaktik.* Berlin: Volk und Wissen.

Krumbacher, C. (2016). *Die Relevanz lernprozessorientierter Sequenzierung im physikbezogenen Sachunterricht – eine Videostudie zur Berücksichtigung von Tiefenstrukturen beim Experimentieren.* Duisburg-Essen: Universitätsbibliothek Duisburg-Essen.

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

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

Lotz, M. (2015). *Kognitive Aktivierung im Leseunterricht der Grundschule. Eine Videostudie zur Gestaltung und Qualität von Leseübungen im ersten Schuljahr.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Oser, F., & Baeriswyl, F. (2001). Choreographies of teaching. Bridging instruction to learning. In V. Richardson (Hrsg.), *AERA's Handbook of Research on Teaching* (S. 1031–1065). Washington: American Educational Research Association.

Oser, F., & Patry, J.-L. (1990). *Choreographien unterrichtlichen Lernens: Basismodelle des Unterrichts.* Berichte zur Erziehungswissenschaft (Nr. 89). Freiburg (CH): Pädagogisches Institut der Universität Freiburg.

Oser, F., & Sarasin, S. (1995). *Basismodelle des Unterrichts: von der Sequenzierung als Lernerleichterung.* https://publishup.uni-potsdam.de/files/410/OSERSARA.pdf [02. 10. 2019].

Pauli, C., & Reusser, K. (2006). Von international vergleichenden Video Surveys zur videobasierten Unterrichtsforschung und -entwicklung. *Zeitschrift für Pädagogik, 52*, 774–798.

Pauli, C. (2012). Merkmale guter Unterrichtsqualität im mathematisch-naturwissenschaftlichen Unterricht aus der Perspektive von Lernenden und Lehrpersonen. In R. Lazarides & A. Ittel (Hrsg.), *Differenzierung im mathematisch-naturwissenschaftlichen Unterricht: Implikationen für Theorie und Praxis* (S. 13–34). Bad Heilbrunn: Klinkhardt.

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*, 407–426.

Reusser, K. (2008). Empirisch fundierte Didaktik – didaktisch fundierte Unterrichtsforschung. In M. A. Meyer, M. Prenzel & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik* (9. Sonderheft der Zeitschrift für Erziehungswissenschaft, S. 219–237). Wiesbaden: VS Verlag für Sozialwissenschaften.

Reusser, K. (2009). Unterricht. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 881–896). Weinheim: Beltz.

Reusser, K., & Pauli, C. (2010). Unterrichtsgestaltung und Unterrichtsqualität – Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht: Einleitung und Überblick. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 9–32). Münster: Waxmann.

Reyer, T. (2004). *Oberflächenmerkmale und Tiefenstrukturen im Unterricht.* Berlin: Logos.

Rosenshine, B., & Fürst, N. (1971). Research on teacher performance criteria. In B. O. Smith (Hrsg.), *Research in teacher education* (S. 37–72). Englewood Cliffs, NJ: Prentice-Hall.

Rosenshine, B., & Meister, C. E. (1993). *Reciprocal teaching: A review of 19 experimental studies* (Technical Report No. 574). Urbana, IL: Center for the Study of Reading, University of Illinois.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. C. Wittrock (Hrsg.), *Handbook of research on teaching* (S. 376–391). New York: Macmillan.

Roth, H. (1965). *Pädagogische Psychologie des Lehrens und Lernens.* Hannover: Schröder.

Seidel, T. (2003). *Lehr-Lernskripts im Unterricht.* Münster: Waxmann.

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,* 454–499.

Snook, I., Clark, J., Harker, R., O'Neill, A.-M., & O'Neill, J. (2010). Critic and conscience of society: A reply to John Hattie. *Journal of Educational Studies, 45,* 93–98.
Tulodziecki, G., Herzig, B., & Blömeke, S. (2017). *Gestaltung von Unterricht: eine Einführung in die Didaktik.* Bad Heilbrunn: UTB.
Walberg, H. J., & Paik, S. J. (2000). *Effective educational practices.* Brüssel: International Academy of Education & International Bureau of Education.
Wallen, N. E., & Travers, R. M. W. (1963). Analyses and investigation of teaching methods. In N. L. Gage (Hrsg.), *Handbook of research on teaching* (S. 448–505). Chicago, IL: Rand McNally.

Abstract: 'Surface characteristics' (also 'sight structures') and 'deep characteristics' (also 'deep structures') of teaching are prominent terms in current research because they provide opportunities for bridging teaching and learning. This paper provides an overview of the origins of the terms and briefly summarizes current understandings of them. The paper also points to literature from didactics and psychology of education research on teaching that implies different conceptualizations of deep characteristics of teaching. After examining empirical research on surface characteristics and deep characteristics, the paper provides suggestions for sharpening the definitions of these terms and discusses further directions for theoretical and empirical research on this conceptual pairing.

Keywords: Surface Characteristics of Teaching, Sight Structure of Teaching, Deep Characteristics of Teaching, Deep Structure of Teaching, Teaching Quality

Anschrift der Autor_innen

Prof. Dr. Jasmin Decristan, Bergische Universität Wuppertal,
Institut für Bildungsforschung in der School of Education,
Gaußstr. 20, 42119 Wuppertal, Germany.
E-Mail: decristan@uni-wuppertal.de

Dr. Miriam Hess, Friedrich-Alexander-Universität Erlangen-Nürnberg,
Institut für Grundschulforschung,
Regensburger Str. 160, 90478 Nürnberg, Germany.
E-Mail: miriam.hess@fau.de

Prof. Dr. Doris Holzberger, Technische Universität München,
School of Education,
Zentrum für internationale Bildungsvergleichsstudien,
Arcisstraße 21, 80333 München, Germany.
E-Mail: doris.holzberger@tum.de

Prof. Dr. Anna-Katharina Praetorius, Universität Zürich,
Lehrstuhl für pädagogisch-psychologische Lehr-Lernforschung und Didaktik,
Institut für Erziehungswissenschaft,
Freiestrasse 36, 8032 Zürich, Schwitzerland (CH).
E-Mail: anna.praetorius@ife.uzh.ch

Miriam Hess/Frank Lipowsky

Zur (Un-)Abhängigkeit von Oberflächen- und Tiefenmerkmalen im Grundschulunterricht

Fragen von Lehrpersonen im öffentlichen Unterricht und in Schülerarbeitsphasen im Vergleich

Zusammenfassung: In der Literatur zur Unterrichtsqualität wird davon ausgegangen, dass die Qualität von Unterricht vor allem von dessen Tiefenstruktur abhängt, während die Oberflächenstruktur hierfür einen Rahmen bildet. Um diesen Zusammenhang empirisch zu untersuchen, wird anhand von Videodaten aus dem Leseunterricht des ersten Schuljahres (N = 47 Videos) analysiert, ob Oberflächenmerkmale (hier: öffentlicher Unterricht vs. Schülerarbeitsphasen) systematisch mit dem kognitiven Niveau der Fragen von Lehrpersonen als ausgewähltem Aspekt der Tiefenstruktur zusammenhängen. Chi-Quadrat-Tests zeigen, dass sich die im Unterricht gestellten Fragen teilweise deutlich zwischen öffentlichem Unterricht und Schülerarbeitsphasen unterscheiden, woraus Implikationen für die (videobasierte) Unterrichtsbeobachtung abgeleitet werden.

Schlagworte: Oberflächenstruktur des Unterrichts, Tiefenstruktur des Unterrichts, Unterrichtsqualität, Fragen von Lehrpersonen, Grundschule

1. Einleitung

Im vorangegangenen Beitrag arbeiteten Decristan, Hess, Holzberger und Praetorius heraus, dass in der aktuellen Forschung zur Beurteilung der Unterrichtsqualität vor allem Tiefenmerkmale[1] des Unterrichts berücksichtigt werden. Gleichzeitig kann aber angenommen werden, dass Oberflächenmerkmale des Unterrichts einen gewissen Rahmen bereitstellen, sodass die Qualität von Unterricht bzw. deren Beurteilung im Rahmen von (videobasierten) Unterrichtsbeobachtungen auch von Aspekten auf Ebene der Oberfläche abhängig sein dürfte. So könnten beispielsweise verschiedene Sozialformen im Unterricht ein unterschiedlich hohes Potenzial zur Ermöglichung einer hohen Qualität der Interaktionen zwischen Lehrpersonen und Lernenden bieten (vgl. Abschnitt 2.2).

Anhand von Beobachtungsdaten aus dem Leseunterricht des ersten Schuljahres, welche aus dem PERLE-Projekt stammen, wird daher der Frage nachgegangen, ob die Oberflächenmerkmale des Unterrichts (hier operationalisiert anhand der Unterscheidung zwischen öffentlichem Unterricht und Schülerarbeitsphasen) systematisch mit dem kognitiven Niveau der Fragen von Lehrpersonen zusammenhängen. Fragen kön-

1 In Anlehnung an den Beitrag von Decristan, Hess, Holzberger und Praetorius (in diesem Heft) werden im Beitrag im Folgenden die Begriffe Oberflächen- und Tiefenstruktur weitgehend durch „Oberflächen-/Tiefenmerkmale“ oder „Oberflächen-/Tiefenebene“ ersetzt.

nen als ein Aspekt kognitiver Aktivierung zu den Tiefenmerkmalen des Unterrichts gezählt werden, da sie kognitive Prozesse anregen können (vgl. Abschnitt 2.3).

2. Theoretischer Hintergrund und Forschungsstand

2.1 Oberflächen- und Tiefenmerkmale

Bezugnehmend auf den Beitrag von Decristan et al. (in diesem Heft) werden Oberflächenstrukturen als methodisch-organisatorische Gestaltungsmerkmale des Unterrichts verstanden, die nicht auf direktem Weg kognitive Prozesse anregen, aber einen mehr oder weniger geeigneten Rahmen schaffen können, um die Voraussetzungen für die Anregung von Denkprozessen bereitzustellen. Tiefenmerkmale hingegen dürften direkt die mentalen Verarbeitungsprozesse der Lernenden adressieren.[2]

Decristan et al. (in diesem Heft) formulieren als ein Desiderat der aktuellen Forschung die Frage, inwiefern sich Oberflächenmerkmale des Unterrichts auf dessen Tiefenstruktur auswirken bzw. damit in Zusammenhang stehen. Dieser Frage soll daher im vorliegenden Beitrag nachgegangen werden. Als klassische Oberflächenmerkmale gelten die Sozialformen des Unterrichts. Daher werden hier zwei Sozialformen, der öffentliche Unterricht und die Schülerarbeitsphasen (Einzel-, Parter- und Gruppenarbeit), betrachtet. Im Folgenden wird zunächst darauf eingegangen, was diese Phasen grundlegend kennzeichnet.

2.2 Unterrichtsqualität in Schülerarbeitsphasen und im öffentlichen Unterricht

Unterricht kann in verschiedenen Sozialformen stattfinden, wobei grundlegend zwischen eher schülerzentrierten Formen wie Einzelarbeit, Partnerarbeit und Gruppenarbeit (Schülerarbeitsphasen) und lehrerzentrierten Formen (öffentlicher Unterricht) unterschieden werden kann. Innerhalb des öffentlichen Unterrichts können Methoden wie das Unterrichtsgespräch oder der Vortrag von Lehrpersonen zum Einsatz kommen. Durch die Sozialformen „wird der Unterricht methodisch-organisatorisch strukturiert“ (Lotz, 2013, S. 123). Da die Wahl der Sozialform allein aber noch keinen direkten Einfluss auf die Anregung kognitiver Prozesse haben dürfte, besteht in der aktuellen Literatur Konsens darüber, sie zu den Oberflächenmerkmalen des Unterrichts zu zählen (z. B. Denn, Hess & Lipowsky, 2017; Hugener, 2008).

Es kann aber angenommen werden, dass gewisse Oberflächenmerkmale eine höhere kognitive Aktivierung der Lernenden begünstigen können, da sich die Funktionen von öffentlichen Phasen und Schülerarbeitsphasen unterscheiden. Während im öffentlichen

2 Für eine umfassende Auseinandersetzung mit den verschiedenen Begriffsverständnissen wird an dieser Stelle auf den vorangegangenen Beitrag von Decristan et al. (in diesem Heft) verwiesen.

Unterricht beispielsweise häufig neue Inhalte eingeführt werden, dienen Schülerarbeitsphasen oftmals dem Üben, Festigen und Automatisieren (z. B. Krammer, 2009). Schülerarbeitsphasen stellen außerdem besonders hohe Anforderungen an die Lehrperson, welche die Arbeitsprozesse von vielen Schülern parallel im Blick haben müsste, um alle Lernenden mit passgenauen und auf dem jeweiligen Niveau kognitiv herausfordernden Instruktionen, Hilfen und Feedback versorgen zu können. Insbesondere wenn die Lernenden an unterschiedlichen Aufgaben arbeiten, lässt sich annehmen, dass im Sinne der Cognitive Load Theorie (Sweller, 1994) „die Organisation und die Steuerung einer so hohen Anzahl unterschiedlicher Prozesse bereits so viele kognitive Kapazitäten der Lehrperson beanspruchen, dass für eine anspruchsvollere inhaltliche Unterstützung keine freien Kapazitäten mehr zur Verfügung stehen“ (Lipowsky & Lotz, 2015, S. 179). In Schülerarbeitsphasen kommt der Lehrperson auch die Funktion eines Tutors zu. Effektive Tutoren können nach Lepper und Woolverton (2002) durch sieben Merkmale charakterisiert werden, die die Autoren im sogenannten Inspire-Modell zusammenfassen (intelligent, nurturant, socratic, progressive, indirect, reflective, encouraging). Gute Lernbegleitung ist demnach u. a. gekennzeichnet durch fundiertes, intelligentes Wissen über das Fach und dessen Didaktik, die Schaffung eines ‚nahrhaften‘, guten Lernklimas sowie ein sokratisches, also durch Fragen und Impulse aktivierendes Lehrverhalten. Damit verdeutlicht das Modell, dass die Ansprüche an das Verhalten von Lehrpersonen in Schülerarbeitsphasen sehr hoch sind.

Dass es möglich ist, Lehrpersonen zu trainieren, den hohen Ansprüchen, die Schülerarbeitsphasen an das Verhalten der Lehrpersonen stellen, besser gerecht zu werden, zeigten Galton, Hargreaves und Pell (2008). Sie beobachteten in Gruppenarbeitsphasen nachhaltigere Interaktionen auf einem höheren kognitiven Niveau als in Klassengesprächen, nachdem die teilnehmenden Lehrpersonen vorab an einem Training teilgenommen hatten, in dem es darum ging, die Fähigkeiten der Lernenden für Gruppenarbeiten zu verbessern.

Bislang gibt es nur wenige Studien, die Unterschiede in der Unterrichtsqualität zwischen öffentlichem Unterricht und Schülerarbeitsphasen systematisch untersuchen. Oft wird in Beobachtungsstudien nur eine der beiden Phasen fokussiert, sodass Vergleiche innerhalb einer Unterrichtsstunde nicht möglich sind. Auch wird die Unterrichtsqualität häufig global über den Verlauf des gesamten Unterrichts beurteilt, sodass keine direkten Vergleiche zwischen den Phasen gezogen werden können. Ergebnisse verschiedener Studien miteinander zu vergleichen, die entweder Schülerarbeitsphasen oder öffentlichen Unterricht in den Blick genommen haben, ist oftmals schwierig, da in diesem Fall meist noch sehr viele weitere Merkmale variieren (z. B. Klassenstufe, Fach, Unterrichtsinhalt etc.).

Hugener, Rakoczy, Pauli und Reusser (2006) schlagen daher vor, durch die gleichzeitige Erfassung von Oberflächen- und Tiefenmerkmalen innerhalb einer Studie zu überprüfen, „ob gewisse Lernaktivitäten mit ausgewählten Qualitätseinschätzungen zusammenhängen“, wobei zwei grundlegende Möglichkeiten bestehen: „(1) Lektionen mit hohen Qualitätseinschätzungen können daraufhin analysiert werden, ob sie sich in Bezug auf die tatsächlich inszenierten Lehr-Lernaktivitäten unterscheiden, oder (2) Lek-

tionen mit ähnlichen beobachteten Lehr-Lernaktivitäten können daraufhin untersucht werden, ob sie sich auch im Qualitätsrating gleichen“ (S. 49).

Hugener et al. (2009) gingen im Rahmen der schweizerisch-deutschen Pythagoras-Videostudie der Frage nach, ob bestimmte Inszenierungsmuster des Unterrichts (hier der Oberflächenstruktur zugeordnet) mit der Einschätzung der kognitiven Aktivierung (Tiefenstruktur) zusammenhängen. Es zeigte sich, dass problemlösend-entdeckende Unterrichtsmuster im Vergleich zu darstellenden oder fragend-entwickelnden Vorgehensweisen im hoch inferenten Rating als kognitiv aktivierender eingeschätzt wurden. Weitere Analysen ergaben aber, dass zusätzliche Aspekte kognitiver Aktivierung (Anteil anspruchsvollen Übens, durchschnittliche Länge von Schülerbeiträgen im Klassengespräch und inhaltlich-strukturelle Klarheit von Theoriephasen) von den Inszenierungsmustern unabhängig sind (Pauli, Drollinger-Vetter, Hugener & Lipowsky, 2008).

Ein weiterer Versuch, Zusammenhänge zwischen Oberflächen- und Tiefenmerkmalen zu analysieren, wurde von Lotz (2015) unternommen. In dieser Studie wurde der Leseunterricht im ersten Schuljahr anhand niedrig inferenter Kodierungen in Phasen unterteilt und zusätzlich hoch inferent in seiner Qualität beurteilt. Anschließend wurden Korrelationen zwischen der Dauer und den prozentualen Anteilen verschiedener Sozialformen und der Ausprägung der Qualitätseinschätzungen des Unterrichts berechnet. Dabei zeigen sich einzelne Zusammenhänge: Beispielsweise korrelieren die klassenbezogenen Werte für die hoch inferent erfasste Skala „Anregung von Denkprozessen“ signifikant positiv mit der absoluten Dauer sowie dem prozentualen Anteil von Schülerarbeitsphasen im Unterricht. Auch die „Förderung einer eigenständigen Auseinandersetzung mit dem Lerngegenstand“ wird erwartungsgemäß signifikant besser eingeschätzt, wenn der Anteil an öffentlichem Unterricht geringer ist. Diese Ergebnisse können als erster Hinweis darauf gewertet werden, dass sich die Qualität des Unterrichts in Abhängigkeit von seiner Oberflächenstruktur unterscheiden kann. Da das hoch inferente Rating nicht für die verschiedenen Phasen getrennt durchgeführt wurde, können die berechneten Korrelationen zwischen der Ausprägung der Unterrichtsqualität und der Dauer und der Anzahl der Unterrichtsphasen lediglich als Hinweis, aber nicht als Beleg dafür betrachtet werden, dass Tiefenmerkmale mit Oberflächenmerkmalen systematisch kovariieren. Dazu sind phasenspezifische Analysen notwendig.

Daher werden im vorliegenden Beitrag Fragen von Lehrpersonen innerhalb des öffentlichen Unterrichts mit Fragen von Lehrpersonen während Schülerarbeitsphasen verglichen. Dieses Vorgehen ermöglicht den direkten Vergleich zwischen den Phasen.

2.3 Fragen von Lehrpersonen im Unterricht

Fragen von Lehrpersonen kommen sowohl im Unterrichtsgespräch als auch in der individuellen Lehrer-Schüler-Interaktion vor und werden vor allem anhand ihrer Funktion, eine Antwort zu verlangen, „der im allgemeinen ein bestimmter Denkproze(ss) vorausgeht“ (Spanhel, 1980, S. 89), definiert. Da durch Fragen kognitive Prozesse angeregt

werden können (z. B. Lipowsky, 2015), lassen sie sich den Tiefenmerkmalen des Unterrichts zuordnen (vgl. Decristan et al., in diesem Heft).

Geht es um die Qualität von Fragen in ihrer Funktion zur Förderung kognitiver Prozesse, erscheint die Unterscheidung ihres kognitiven Niveaus zentral (z. B. Levin, 2005). Die Grundlage für die meisten Klassifikationssysteme zum kognitiven Niveau von Fragen ist die Lernzieltaxonomie von Bloom, Engelhart, Furst, Hill und Krathwohl (1976), nach der Wissens-, Verständnis-, Anwendungs-, Analyse-, Synthese- und Bewertungsfragen unterschieden werden. Vereinfacht werden darauf aufbauend oftmals auch lediglich die beiden Kategorien Wissens- vs. Denkfragen bzw. Fragen mit höherem vs. niedrigerem kognitiven Niveau (higher vs. lower order questions) differenziert (z. B. Gayle et al., 2006; Lotz, 2015).

Wissensfragen zielen darauf ab, dass sich die Lernenden an bereits Bekanntes oder Erarbeitetes erinnern und dieses Wissen wiedergeben. Bei Denkfragen rückt hingegen eine neuartige Situation oder ein unbekanntes Problem in den Mittelpunkt. Vorerfahrungen müssen genutzt werden, um sie auf neue Situationen oder Probleme zu übertragen. Daneben können außerdem Reflexionsfragen, organisatorische und ablaufgerichtete Fragen im Unterricht vorkommen. Reflexionsfragen zielen darauf ab, die Lernenden zum Nachdenken über (ihre eigenen) Lernprozesse anzuregen. Organisatorische und ablaufgerichtete Fragen dienen hingegen eher der Unterrichtsorganisation (Lotz, 2015).[3]

3. Fragestellung

Der Beitrag geht der übergeordneten Frage nach, inwiefern ausgewählte Oberflächen- und Tiefenmerkmale des Unterrichts systematisch miteinander zusammenhängen. Dabei soll folgende Fragestellung fokussiert werden: *Unterscheidet sich das kognitive Niveau von Fragen im öffentlichen Unterricht und in Schülerarbeitsphasen?*

Sollten sich hier Unterschiede zeigen, wäre dies ein Hinweis dafür, dass verschiedene Sozialformen ein unterschiedlich hohes Potenzial zur Anregung kognitiver Prozesse bieten. Aus methodischer Sicht ist die Fragestellung des Beitrags für die (videobasierte) Unterrichtsforschung von Interesse, da sie Aufschluss darüber geben kann, ob bei der Beurteilung bestimmter Merkmale von Unterrichtsqualität die Rahmenbedingungen des Unterrichts (noch stringenter oder umfassender) berücksichtigt werden müssten, um systematische Verzerrungen oder Fehlinterpretationen der Ergebnisse zu vermeiden.

3 Als organisatorische Fragen wurden in der vorliegenden Studie Fragen zur Verwendung von Hilfsmitteln, zur Vorbereitung der Aufgabenbearbeitung, zum Aufrufen von Schülern zum Vorlesen oder Helfen sowie Aufforderungen zur Bearbeitung eines Auftrags sowie Ermahnungen in Form von Fragen kodiert. Ablaufgerichtete Fragen beziehen sich auf den Arbeitsstand, das Verständnis des Arbeitsauftrags, die Wahl der Aufgabe oder Textsorte sowie die Art der Aufgabenbearbeitung.

4. Datengrundlage: Die Videostudie im Fach Deutsch des PERLE-Projekts

Die den Analysen zugrundeliegenden Beobachtungsdaten stammen aus der Längsschnittstudie PERLE (Lipowsky, Faust & Kastens, 2013), in deren Rahmen die Persönlichkeits- und Lernentwicklung von Grundschulkindern sowie deren Determinanten vom ersten bis zum vierten Schuljahr untersucht wurden. Im März des ersten Schuljahres wurde im Fach Deutsch eine 90-minütige Unterrichtseinheit videografiert (Lotz & Corvacho del Toro, 2013). Zur Ermöglichung einer grundlegenden Vergleichbarkeit der videografierten Unterrichtseinheiten erhielten die Lehrpersonen inhaltliche Vorgaben zur Gestaltung der Stunde, in der u. a. eine Leseübung durchgeführt werden sollte. Über die Reihenfolge und alle weiteren didaktisch-methodischen Aspekte – wie beispielsweise die Sozialformen – konnten die Lehrpersonen selbst entscheiden.

Die Lehrpersonen waren bis auf eine Ausnahme weiblich und verfügten im Schnitt über 15 Jahre Berufserfahrung. Die Lernenden waren zum Zeitpunkt der Datenerhebung im Mittel sieben Jahre und einen Monat alt (Mädchenanteil: 52.4 %).

Für die Analyse der Fragestellungen werden nur Kodierungen der Leseübungsphase genutzt. Diese wurde vorab über eine niedrig inferente Kodierung identifiziert (vgl. Lotz, 2015), wobei nur Videos mit Leseübungen von mindestens fünf Minuten einbezogen wurden ($N = 47$; $Min = 6.17$ min; $Max = 55.50$ min; $M = 26.76$ min; $SD = 12.47$ min). Aufgrund der Varianz in der Unterrichtszeit wurden alle Analysen mit relativen Häufigkeiten oder prozentualen Anteilen durchgeführt.

5. Methodisches Vorgehen

5.1 Kodierung der Sozialformen des Unterrichts

Für die gesamte Phase der Leseübung wurden die Sozialformen niedrig inferent im Time-Sampling-Verfahren (10-Sekunden-Intervalle) kodiert (vgl. Lotz, 2013). Grundlegend werden durch das Kategoriensystem öffentliche Unterrichtsphasen von Schülerarbeitsphasen unterschieden. Zu den Schülerarbeitsphasen zählen die Kategorien „Einzelarbeit“, „Partnerarbeit“ und „Gruppenarbeit“. Der öffentliche Unterricht wurde mit den beiden Kategorien „Öffentlicher Unterricht im Sitzkreis“ und „Öffentlicher Unterricht ohne Sitzkreis“ erfasst. Für die folgenden Analysen werden nur die zusammengefassten Kategorien „Schülerarbeitsphasen“ („SAP“) und „Öffentlicher Unterricht“ („OEU“) differenziert. Zusätzlich wurde die „Restkategorie“ vergeben, wenn es während der Leseübung zu kurzen Unterbrechungen kam, in denen keine inhaltliche Arbeit stattfand.

Die Kodierungen wurden von vier trainierten Beobachterinnen durchgeführt (prozentuale Übereinstimmung $P\ddot{U} \geq 97.52\,\%$; Cohens Kappa $\kappa \geq .97$; vgl. Lotz, 2013; 2015). In den 47 hier analysierten Videos liegt der Anteil des öffentlichen Unterrichts durchschnittlich bei $M = 39.5\,\%$ ($SD = 25.5\,\%$), der Anteil an Schülerarbeitsphasen bei

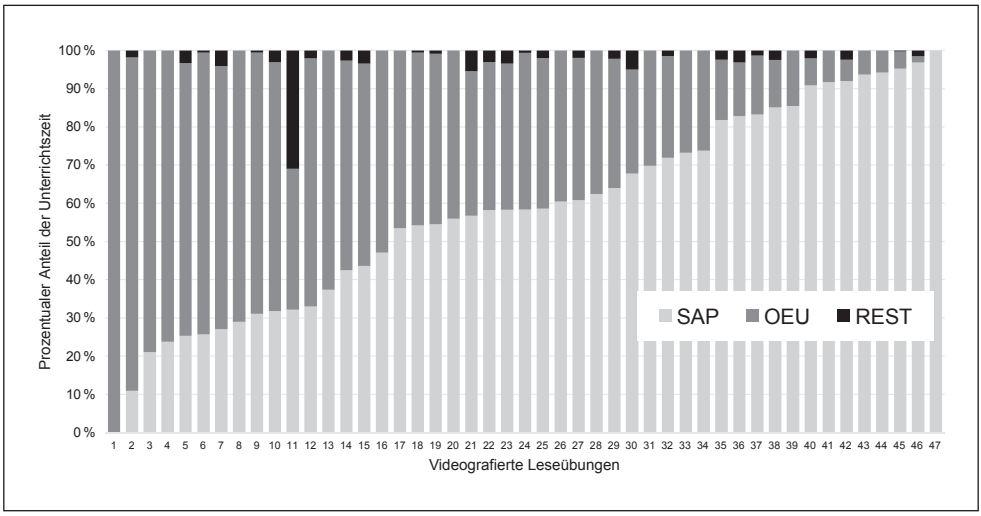

Abb. 1: Prozentuale Anteile von Schülerarbeitsphasen (SAP) und öffentlichem Unterricht (OEU) in den 47 videografierten Leseübungen

M = 58.4 % (SD = 25.8 %) und der Anteil der Restkategorie bei nur 2.0 % (SD = 4.6). Unter den 47 Videos gibt es sowohl ein Video, in dem die Leseübung zu 100 % in Schülerarbeitsphasen durchgeführt wurde als auch eine Leseübung, die ausschließlich im öffentlichen Unterricht stattfand.

Da sich – wie Abbildung 1 auch grafisch verdeutlicht – auf der Obenflächenebene des Unterrichts deutliche Unterschiede zwischen den videografierten Lerngruppen zeigen, bieten sich die Daten an, um Zusammenhänge zwischen der Oberflächen- und Tiefenstruktur des Unterrichts zu untersuchen.

5.2 Kodierung des kognitiven Niveaus der Fragen von Lehrpersonen

Als ein exemplarischer Aspekt auf der Tiefenebene des Unterrichts wurde jede während der Leseübung vorkommende Frage der Lehrperson mit einem niedrig inferenten Event-Sampling-Verfahren identifiziert. Als Fragen wurden dabei alle Äußerungen der Lehrperson definiert, die eine Schülerantwort oder -äußerung intendieren und inhaltliche Relevanz für die Leseübung besitzen. Damit beinhaltet die Kodierung auch Impulse. Für jede Frage wurde überdies festgehalten, in welcher Sozialform und auf welchem kognitiven Niveau sie gestellt wurde (vgl. Lotz, 2015). Das vollständige Kategoriensystem mit allen Kodierregeln findet sich bei Lotz (2015).[4] Dieses System wurde

4 Das Kategoriensystem mit den Kodierregeln ist als Anhang auch online verfügbar unter https://www.springer.com/de/book/9783658104351 (zuletzt geprüft: 17.01.2020).

basierend auf vorhandenen Klassifikationen verschiedener Fragearten deduktiv entworfen und induktiv anhand der eigenen Videos weiterentwickelt.

Bei der „Art der Frage" (kognitives Niveau) wurden insgesamt 26 einzelne Fragearten unterschieden, die zu den fünf übergeordneten Kategorien Wissensfragen, Denkfragen, Reflexionsfragen, organisatorische Fragen und ablaufgerichtete Fragen zusammengefasst wurden. Die Auswertung wurde auch hier von geschulten Kodiererinnen vorgenommen (PÜ ≥ 97.40 %; vgl. Lotz, 2015).

In den 47 videografierten Leseübungen wurden insgesamt 2 087 Fragen von Lehrpersonen an die Lernenden identifiziert. Im Schnitt entspricht dies 44.40 Fragen pro Video, wobei sich hier – wie bereits bei den Sozialformen – eine deutliche Varianz zeigt (*Min* = 7; *Max* = 118; *SD* = 30.61). Relativiert an der Dauer der jeweiligen Leseübungen wurden durchschnittlich 1.70 Fragen pro Minute gestellt (*Min* = 0.38; *Max* = 4.86; *SD* = 1.00).

5.3 Analysen der Zusammenhänge zwischen Oberflächen- und Tiefenmerkmalen

Zur Analyse der Zusammenhänge zwischen den Oberflächen- und Tiefenmerkmalen des Unterrichts werden zunächst deskriptive Analysen vorgenommen. Da die Fragen der Lehrpersonen im Event-Sampling-Verfahren einzeln kodiert wurden, kann für jede einzelne Frage analysiert werden, ob sie in einer Schülerarbeitsphase oder im öffentlichen Unterricht gestellt wurde. Auf dieser Basis können die Verteilungen und Häufigkeiten in Schülerarbeitsphasen und im öffentlichen Unterricht verglichen werden. Ob sich die Verteilungen systematisch unterscheiden, wird mit Chi-Quadrat-Tests überprüft.

6. Ergebnisse

Von den insgesamt 2 087 gestellten Fragen wurden 1 151 in Schülerarbeitsphasen gestellt, 931 im öffentlichen Unterricht und nur 5 Fragen während der Phase, die der Restkategorie zugeordnet wurde. Diese 5 Fragen werden im Folgenden nicht mehr berücksichtigt.

Im öffentlichen Unterricht wurden von den Lehrpersonen im Schnitt 2.04 Fragen pro Minute gestellt, während in Schülerarbeitsphasen die Anzahl der Fragen von Lehrpersonen mit 1.49 Fragen pro Minute etwas geringer ausfällt.[5] Insbesondere im öffentlichen Unterricht fällt aber eine starke Varianz zwischen den Klassen in der relativen

5 Dass auch in den Schülerarbeitsphasen häufig Fragen gestellt werden, lässt sich damit erklären, dass in den videografierten Leseübungen die Erstklasslehrpersonen auch in Schülerarbeitsphasen zu einem Großteil der Zeit mit einzelnen Lernenden interagieren (vgl. auch Lotz, 2015) und dabei auch Fragen stellen (z. B. *„Wie heißt dieses Wort?"; „Kommst du zu-*

Durchschnittliche Anzahl an Fragen pro Minute	*N*	*Min*	*Max*	*M*	*SD*
Gesamte Leseübung (OEU+SAP)	47	0.38	4.86	1.70	0.98
Öffentlicher Unterricht (OEU)	46	0.00	10.50	2.04	1.75
Schülerarbeitsphasen (SAP)	46	0.13	4.69	1.49	1.10

Anmerkung: *N* = 46 kommt zustande, da in jeweils einer Klasse die Leseübung ausschließlich in Schülerarbeitsphasen bzw. im öffentlichen Unterricht stattfand.

Tab. 1: Deskriptive Statistik der durchschnittlichen Anzahl von Fragen im öffentlichen Unterricht und in Schülerarbeitsphasen

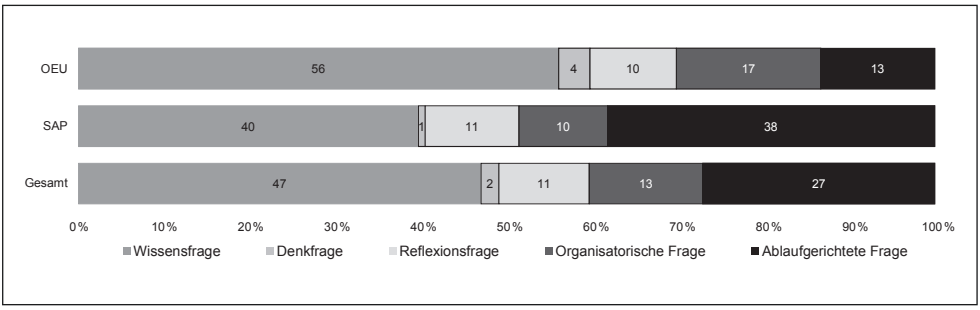

Abb. 2: Prozentuale Anteile der Arten von Fragen in Schülerarbeitsphasen (SAP) und im öffentlichen Unterricht (OEU)

Häufigkeit von Fragen von Lehrpersonen auf: Es gibt sowohl Klassen, in denen während des öffentlichen Unterrichts keine einzige Frage gestellt wird, als auch eine Klasse, in der durchschnittlich zwischen zehn und elf Fragen von Lehrpersonen pro Minute im öffentlichen Unterricht vorkommen (vgl. Tab. 1).

Abbildung 2 stellt die Verteilung der verschiedenen Arten von Fragen jeweils einzeln für den öffentlichen Unterricht, für die Schülerarbeitsphasen und für die gesamte Leseübung im Vergleich dar. Betrachtet man zunächst die Verteilung in der Leseübung insgesamt, so machen Wissensfragen mit 47.0 % den größten Anteil aus, wohingegen sich nur 2.1 % aller Fragen der übergeordneten Kategorie der Denkfragen zuordnen lassen. Auch ablaufgerichtete Fragen kommen mit 27.1 % relativ häufig vor, wohingegen Reflexionsfragen (10.6 %) und organisatorische Fragen (13.3 %) seltener gestellt werden.

Der Vergleich der Verteilungen zeigt, dass die relative Häufigkeit von Reflexions- und organisatorischen Fragen in Schülerarbeitsphasen und im öffentlichen Unterricht

recht?“; „Bist du schon fertig?“). Im Gegensatz zu den Fragen in öffentlichen Unterrichtsphasen richten sich die Fragen in Schülerarbeitsphasen meist an einzelne Schüler oder kleinere Schülergruppen.

relativ ähnlich ausfällt. Eine Verschiebung deutet sich aber zwischen Wissensfragen und ablaufgerichteten Fragen an. Während im öffentlichen Unterricht die Wissensfragen mit 56.1 % mehr als die Hälfte aller Fragen ausmachen und lediglich 13.3 % ablaufgerichtete Fragen gestellt werden, ist der Anteil von Wissensfragen in den Schülerarbeitsphasen geringer (39.7 %), wohingegen deutlich mehr ablaufgerichtete Fragen vorkommen (38.2 %). Der ohnehin sehr geringe Anteil von Denkfragen (insgesamt 2.1 %) ist in Schülerarbeitsphasen mit nur 0.8 % sogar noch einmal geringer als im öffentlichen Unterricht mit zumindest 3.7 %. Der geringe Anteil von Denkfragen drückt sich auch in ihrer geringen absoluten Zahl aus: Von den insgesamt über alle Videos hinweg lediglich 43 gestellten Denkfragen werden 34 (79.1 %) im öffentlichen Unterricht gestellt und nur 9 (20.9 %) in Schülerarbeitsphasen. Dabei muss berücksichtigt werden, dass der Anteil an Schülerarbeitsphasen mit 58.4 % der Unterrichtszeit sogar höher ist als der Anteil an öffentlichem Unterricht (39.5 %). Bei den ablaufgerichteten Fragen kehrt sich das Verhältnis um: 78.0 % aller ablaufgerichteten Fragen werden in Schülerarbeitsphasen gestellt und nur 22.0 % im öffentlichen Unterricht.

Der χ^2-Test ergibt einen Wert von 184.60 und liegt damit deutlich über dem kritischen Wert von $\chi^2 = 9.49$ (Kreuztabelle: 5×2; $df = 4$; $p \leq .05$; Effektstärke $d = 0.62$). Diese Abweichung der tatsächlichen Verteilung von der bei Gültigkeit der Nullhypothese erwartbaren Verteilung ist vorwiegend auf die Kategorien „Wissens-“, „Denk-“ und „ablaufgerichtete Fragen“ zurückzuführen.

Eine detailliertere Betrachtung zeigt, dass es sich bei 40.5 % aller im öffentlichen Unterricht gestellten Fragen um Fragen zum „Verständnis des Inhalts/Textes“ handelt, die zur übergeordneten Kategorie der Wissensfragen zählen (z. B. „Wie fängt die Geschichte an und wie hört sie auf?“ oder „Was wollte denn die Lucy [Hauptfigur des Buchs] machen?“). In Schülerarbeitsphasen hingegen machen solche Fragen nur 18.9 % der dort gestellten Fragen aus. In der individuellen Auseinandersetzung mit den Lernenden werden hingegen öfter Fragen gestellt, die den direkten Leseprozess und die Lesetechnik betreffen („Verständnis der Wörter/Laute“, z. B. „Was ist denn das für ein Buchstabe?“). Diese machen in Schülerarbeitsphasen 17.5 % der gestellten Fragen aus, im öffentlichen Unterricht nur 8.9 %. Die Dominanz ablaufgerichteter Fragen in Schülerarbeitsphasen lässt sich vor allem auf Fragen zum „Arbeitsstand“ zurückführen (z. B. „Wie viel habt ihr geschafft?“).

7. Zusammenfassung und Diskussion

Im Beitrag konnte gezeigt werden, dass sich die Arten von Fragen teilweise deutlich zwischen öffentlichem Unterricht und Schülerarbeitsphasen unterscheiden. Auffällig ist vor allem, dass Fragen im öffentlichen Unterricht stärker der Auseinandersetzung mit den Textinhalten zu dienen scheinen, wohingegen in Schülerarbeitsphasen mehr Fragen dazu genutzt werden, die Abläufe zu organisieren und aufrechtzuerhalten. Denkfragen kommen generell kaum vor, in Schülerarbeitsphasen allerdings noch seltener als im öffentlichen Unterricht. Dies deutet darauf hin, dass insbesondere in Schülerarbeitsphasen

Fragen weniger dazu genutzt werden, die Lernenden kognitiv herauszufordern, Wissen aufzubauen oder zu sichern.

Viele der gezeigten Unterschiede lassen sich gut anhand unterschiedlicher Funktionen und Erfordernisse in den beiden Phasen erklären (vgl. z. B. Krammer, 2009; Abschnitt 2.2). In Schülerarbeitsphasen dürfte es für viele Lehrpersonen wichtig sein, sich – beispielsweise anhand ablaufgerichteter Fragen – einen Überblick darüber zu verschaffen, wie die Lernenden zurechtkommen (z. B. „Hast du den Arbeitsauftrag verstanden?"). Schülerarbeitsphasen bieten außerdem die Möglichkeit, einzelne und insbesondere schwächere Lernende gezielt bei hierarchieniedrigeren Leseprozessen zu unterstützen (z. B. durch Fragen der Kategorie „Verständnis der Wörter/Laute"), wohingegen im Klassenunterricht der Fokus eher auf den Textinhalten liegt, was das häufigere Vorkommen von etwas hierarchiehöheren Fragen zum „Verständnis des Inhalts/Textes" erklären kann.

Anhand der Ergebnisse kann aber auch kritisch diskutiert werden, ob das Potenzial der Fragen von Lehrpersonen – insbesondere in Schülerarbeitsphasen – optimal zur kognitiven Aktivierung der Lernenden genutzt wird. Evtl. erfordert gerade im ersten Schuljahr das Gestalten einer Schülerarbeitsphase noch so viel Aufwand und Aufmerksamkeit, dass den Lehrpersonen kaum Zeit oder kognitive Ressourcen bleiben, um in der Interaktion mit einzelnen Lernenden vertieft auf die Inhalte einzugehen (vgl. Abschnitt 2.2). Auch empirische Arbeiten, die explizit (auch oder ausschließlich) Schülerarbeitsphasen in den Blick nehmen, deuten darauf hin, dass Lehrpersonen deren Potenzial oftmals nicht ausschöpfen (z. B. Krammer, 2009; Lotz, 2015; van de Pol, Volman & Beishuizen, 2010). Es dominiert meist einfaches gegenüber elaboriertem Feedback, Hilfestellungen werden selten im Sinne des Scaffolding erteilt und die kognitive Aktivierung wird insgesamt als eher gering eingeschätzt.

Für die Lehrerinnen- und Lehrerbildung dürfte es daher wichtig sein, Lehrpersonen insbesondere für die Gestaltung von Schülerarbeitsphasen Strategien an die Hand zu geben, damit auch diese Phasen gezielter zur kognitiven Aktivierung der Lernenden genutzt werden können. Dass dies möglich ist, unterstreicht die in Abschnitt 2.2 vorgestellte Studie von Galton et al. (2008), in deren Rahmen Lehrpersonen ein erfolgreiches Training zu ihrem Lehrverhalten während Gruppenarbeiten erhielten.

Wichtig ist, dass die hier aufgezeigten Unterschiede selbst noch keinen Aufschluss über die Ursachen der Differenzen liefern können. Hier könnten weitere Untersuchungen ansetzen, in denen analysiert wird, ob beispielsweise Drittvariablen identifizierbar sind, die sowohl die Tendenz der Lehrperson erklären können, bestimmte Fragen zu stellen als auch den Unterricht vorzugsweise eher schüler- oder aber lehrerzentriert zu gestalten (z. B. Überzeugungen von Lehrpersonen; Kompositionsmerkmale der Schulklassen). Anhand von Beobachtungsdaten der PERLE-Studie konnten Denn et al. (2017) beispielsweise zeigen, dass Lehrpersonen in Klassen mit hoher Leistungsstärke häufiger schülerzentrierte Sozialformen nutzen als Lehrpersonen in Klassen mit geringerer Leistungsstärke. Kompositionsmerkmale könnten sich aber ebenso auf die Tiefenebene des Unterrichts auswirken und somit teilweise erklären, wieso es zwischen Oberflächen- und Tiefenmerkmalen zu systematischen Abhängigkeiten und Zusammenhängen

kommt. Dabei könnten sich auch Analysen dazu anschließen, inwiefern ein unterschiedliches Frageverhalten von Lehrpersonen in verschiedenen Phasen auch differenzielle Effekte auf das Lernen hat.

Einschränkend muss beachtet werden, dass mit dem kognitiven Niveau von Fragen nur ein ausgewählter Aspekt der Tiefenebene des Unterrichts untersucht wurde. Selbstverständlich gibt es noch eine Reihe weiterer Tiefenmerkmale, wozu v. a. Merkmale kognitiver Aktivierung gezählt werden können, wie beispielsweise die Art der im Unterricht gestellten Aufgaben, die Anregung der Lernenden zum Einsatz von Strategien oder die gemeinsame Reflexion des Lernprozesses (zsf. Lotz, 2015). Auch sollte bedacht werden, dass die Art der Fragen und die Häufigkeit deren Vorkommens im Unterricht allein natürlich noch keine hohe Unterrichtsqualität garantieren. Daher müssten die hier dargestellten Analysen noch vertieft und um weitere Aspekte ergänzt werden, indem sowohl anhand niedrig als auch hoch inferenter Methoden gezielt unterschiedliche Phasen beobachtet und verglichen werden. Die hier analysierten Unterrichtsphasen sind darüber hinaus relativ kurz. Der Unterrichtsinhalt war standardisiert, was die generelle Art der Fragen beeinflusst haben dürfte. Da den Lehrpersonen die methodische Umsetzung aber freigestellt war und sich hier auf der Oberflächenebene auch deutliche Unterschiede zwischen den Stunden zeigten, eignen sich die Daten dennoch gut zur Analyse der Fragestellungen.

Im vorliegenden Beitrag wurde die Verteilung von Fragen im öffentlichen Unterricht und in Schülerarbeitsphasen klassenübergreifend miteinander verglichen. Vertiefend könnte zusätzlich analysiert werden, inwiefern bei einzelnen Lehrpersonen das Frageverhalten in unterschiedlichen Phasen differiert. Eventuell lassen sich dabei sowohl Lehrpersonen identifizieren, die unabhängig von der Unterrichtsphase eher kognitiv anregende Fragen stellen, als auch Lehrpersonen, bei denen die Qualität der Fragen in Abhängigkeit von der Unterrichtsphase stark divergiert.

Zur Frage, ob sich die Tiefenebene des Unterrichts abhängig von den Oberflächenmerkmalen unterscheidet, liefern die vorliegenden Ergebnisse erste Hinweise. Dies ist insbesondere vor dem Hintergrund relevant, dass häufig die Ergebnisse unterschiedlicher Beobachtungsstudien oder auch mehrerer Unterrichtsstunden einer Lehrperson miteinander verglichen werden, ohne dass dabei dezidiert auf die Oberflächenebene des Unterrichts eingegangen wird. Die Ergebnisse der vorliegenden Studie zeigen aber – wenn auch zunächst lediglich anhand eines ausgewählten Aspekts – dass die Oberflächenmerkmale des Unterrichts nicht vernachlässigbar ist. So können Unterschiede zwischen den deskriptiven Ergebnissen verschiedener Studien auch dadurch zustande kommen, dass in jeder Studie ein anderes Unterrichtssetting beobachtet wurde. Hier wären allerdings noch mehr Studien nötig, die diese Fragestellung systematisch und für verschiedene Aspekte auf Tiefen- und der Oberflächenebene analysieren. Denn es ist sehr wahrscheinlich, dass für einige Tiefenmerkmale eine stärkere Abhängigkeit von der Oberflächenebene verzeichnet werden kann als für andere. In diese Richtung weisen auch Analysen von Praetorius (2014), die zeigen, dass insbesondere zur Beurteilung kognitiver Aktivierung mehrere Unterrichtsstunden verwendet werden sollten, um Abhängigkeiten der Qualitätsbeurteilung von der konkreten Unterrichtsgestaltung zu

minimieren. Während für die Beurteilung von Klassenführung und Schülerorientierung eine Stunde ausreicht, um Reliabilitäten ≥ .70 zu erreichen, werden zur reliablen Beurteilung des instabileren Merkmals der kognitiven Aktivierung eigentlich neun Stunden benötigt.

Für die Planung und Durchführung von künftigen Beobachtungsstudien kann aus den Ergebnissen der vorliegenden Studie v. a. abgeleitet werden, dass die ohnehin bereits häufig geforderte Standardisierung (z. B. Pauli, 2008) nicht nur aus inhaltlichen Gründen, sondern auch vor dem Hintergrund der Bedeutsamkeit unterschiedlicher Unterrichtsgestaltung auf Oberflächenebene wichtig sein dürfte. Natürlich bleibt dabei eine relevante Frage, wie viel Eingreifen in die übliche Planung der Lehrperson für die Forschungsfragen unabdingbar ist oder aber, inwiefern dadurch der übliche Unterricht zu stark verzerrt wird. Entscheidet man sich, wofür es durchaus gute Gründe geben kann, aber für ein geringeres Maß an Standardisierung, so sollte bei der Interpretation der Ergebnisse – selbst, wenn es dabei primär um die Tiefenmerkmale geht – der potenzielle Einfluss unterschiedlicher Oberflächenmerkmale mitbedacht oder kontrolliert werden.

Literatur

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1976). *Taxonomie von Lernzielen im kognitiven Bereich.* Weinheim: Beltz.

Denn, A.-K., Hess, M., & Lipowsky, F. (2017). Hängen das Leistungsniveau und die Leistungsheterogenität von Grundschulklassen mit dem Anteil lehrerzentrierter Unterrichtsphasen im Deutsch- und Mathematikunterricht zusammen? Ergebnisse der PERLE-Studie. *Zeitschrift für Grundschulforschung, 10*(1), 162–176.

Galton, M., Hargreaves, L., & Pell, T. (2009). Group work and whole-class teaching with 11- to 14-year-olds compared. *Cambridge Journal of Education, 39*(1), 119–140.

Gayle, B. M., Preiss, R. W., & Allen, M. (2006). How effective are teacher-initiated classroom questions in enhancing student learning? In B. M. Gayle, R. W. Preiss, N. Burrell & M. Allen (Hrsg.), *Classroom communication and instructional processes: Advances through meta-analysis* (S. 279–293). Mahwah: Erlbaum.

Hugener, I. (2008). *Inszenierungsmuster im Unterricht und Lernqualität. Sichtstrukturen schweizerischen und deutschen Mathematikunterrichts in ihrer Beziehung zu Schülerwahrnehmung und Lernleistung – eine Videoanalyse.* Münster: Waxmann.

Hugener, I., Pauli, C., Reusser, K., Lipowsky, F., Rakoczy, K., & Klieme, E. (2009). Teaching patterns and learning quality in Swiss and German mathematics lessons. *Learning and Instruction, 19*(1), 66–78.

Hugener, I., Rakoczy, K., Pauli, C., & Reusser, K. (2006). Videobasierte Unterrichtsforschung: Integration verschiedener Methoden der Videoanalyse für eine differenzierte Sicht auf Lehr- und Lernprozesse. In S. Rahm, I. Mammes & M. Schratz (Hrsg.), *Schulpädagogische Forschung. Unterrichtsforschung, Perspektiven innovativer Ansätze* (S. 41–53). Innsbruck: Studien-Verlag.

Krammer, K. (2009). *Individuelle Lernunterstützung in Schülerarbeitsphasen. Eine videobasierte Analyse des Unterstützungsverhaltens von Lehrpersonen im Mathematikunterricht.* Münster: Waxmann.

Lepper, M. R., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Hrsg.), *Improving academic achievement: Impact of psychological factors on education* (S. 135–158). San Diego: Academic Press.

Levin, A. (2005). *Lernen durch Fragen. Wirkung von strukturierenden Hilfen auf das Generieren von Studierendenfragen als begleitende Lernstrategie.* Münster: Waxmann.

Lipowsky, F. (2015). Unterricht. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 69–105). Heidelberg: Springer.

Lipowsky, F., Faust, G., & Kastens, C. (2013). *Persönlichkeits- und Lernentwicklung an staatlichen und privaten Grundschulen. Ergebnisse der PERLE-Studie zu den ersten beiden Schuljahren.* Münster: Waxmann.

Lipowsky, F., & Lotz, M. (2015). Ist Individualisierung der Königsweg zum Lernen? Eine Auseinandersetzung mit Theorien, Konzepten und empirischen Befunden. In G. Mehlhorn, K. Schöppe & F. Schulz (Hrsg.), *Begabungen entwickeln & Kreativität fördern* (S. 155–219). München: kopaed.

Lotz, M. (2013). Die Kodierung der Sozialformen. In M. Lotz, F. Lipowsky & G. Faust (Hrsg.), *Technischer Bericht zu den PERLE-Videostudien* (S. 123–142). Frankfurt a. M.: Gesellschaft zur Förderung Pädagogischer Forschung (GFPF).

Lotz, M. (2015). *Kognitive Aktivierung im Leseunterricht der Grundschule. Eine Videostudie zur Gestaltung und Qualität von Leseübungen im ersten Schuljahr. Wiesbaden: VS Verlag für Sozialwissenschaften.* https://www.springer.com/de/book/9783658104351 [25. 09. 2019].

Lotz, M., & Corvacho del Toro, I. (2013). Die Videostudie im Fach Deutsch: „Lucy rettet Mama Kroko". In M. Lotz, F. Lipowsky & G. Faust (Hrsg.), *Technischer Bericht zu den PERLE-Videostudien* (S. 29–36). Frankfurt a. M.: Gesellschaft zur Förderung Pädagogischer Forschung (GFPF).

Pauli, C. (2008). Unterrichtsbeobachtung. In F. Hellmich (Hrsg.), *Lehr-Lernforschung und Grundschulpädagogik* (S. 143–155). Bad Heilbrunn: Klinkhardt.

Pauli, C., Drollinger-Vetter, B., Hugener, I., & Lipowsky, F. (2008). Kognitive Aktivierung im Mathematikunterricht. *Zeitschrift für Pädagogische Psychologie, 22*(2), 127–133.

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

Spanhel, D. (1980). Analyse der verbalen Kommunikation im Unterricht. In K. Boeckmann (Hrsg.), *Analyse von Unterricht in Beispielen* (S. 83–97). Stuttgart: Klett.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction 4*(4), 295–312.

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296.

Abstract: According to published literature about instructional quality it can be assumed that deep structure features of teaching are the crucial factors for the assessment of instructional quality. Surface structure features provide a framework for the deep structure features. To analyse this assumed relationship empirically, this paper, using video data of $N = 47$ reading lessons in first grade, examines whether surface features (here: whole class instruction vs. student work phases) have a connection to selected aspects of deep structure (here: cognitive level of teacher questions). Chi-square tests show that teacher questions differentiate considerably between whole class instruction and student work phases. Implications for video-based observation of instruction can be derived from this outcome (e. g. the significance of standardization).

Keywords: Surface Structure Features of Teaching, Deep Structure Features of Teaching, Instructional Quality, Teacher Questions, Elementary School

Anschrift der Autor_innen

Dr. Miriam Hess, Friedrich-Alexander-Universität Erlangen-Nürnberg,
Institut für Grundschulforschung,
Regensburger Str. 160, 90478 Nürnberg, Deutschland
E-Mail: miriam.hess@fau.de

Prof. Dr. Frank Lipowsky, Universität Kassel,
Fachgebiet für Empirische Schul- und Unterrichtsforschung,
Nora-Platiel-Straße 1, 34127 Kassel, Deutschland
E-Mail: lipowsky@uni-kassel.de

Christine Pauli

Kommentar zum Themenblock „Oberflächen- und Tiefenstruktur des Unterrichts“

Nutzen und Grenzen eines prominenten Begriffspaars für die Unterrichtsforschung – und das Unterrichten

Zusammenfassung: Ausgehend von einem Blick auf den Entstehungskontext der Unterscheidung von Oberflächen- und Tiefenstrukturen des Unterrichts geht der Kommentar anhand der Beiträge von Decristan, Hess, Holzberger und Praetorius (in diesem Heft) und von Hess und Lipowsky (in diesem Heft) der Frage nach dem Nutzen dieser Unterscheidung und nach der Beziehung zwischen Oberflächen- und Tiefenstrukturen nach. Die Autorin sieht die Unterscheidung als didaktisches Werkzeug für die Planung und Reflexion lernwirksamen Unterrichts und als heuristisches Instrument für die Unterrichtsqualitätsforschung, das u. a. auch dazu beiträgt, bisher eher vernachlässigte Forschungsfragen wie z. B. jene nach Zusammenhängen zwischen methodischen Gestaltungsmerkmalen und qualitätsvollen Lehr- und Lernprozessen, bezogen auf mehrdimensionale Bildungsziele, systematisch zu bearbeiten.

Schlagworte: Oberflächen- und Tiefenstrukturen von Unterricht, Unterrichtsqualität, Unterrichtsforschung, Didaktik, Lehr- und Lernprozesse

Was ist unter Oberflächen- und Tiefenstrukturen des Unterrichts zu verstehen, welche Zusammenhänge bestehen zwischen ihnen und welchen Nutzen bringt diese Unterscheidung? Diesen Fragen gehen die Texte von Decristan, Hess, Holzberger und Praetorius (in diesem Heft) sowie von Hess und Lipowsky (in diesem Heft) nach und leisten damit wichtige Beiträge zur Klärung eines Konzepts, das in den letzten Jahren Eingang in die (deutschsprachige) Unterrichtsforschung gefunden hat.

1. (Lehr- und) Lernprozesse im Zentrum – die Unterscheidung von Oberflächen- und Tiefenstrukturen des Unterrichts als Werkzeug für Didaktik und Unterrichtsforschung

Im Mittelpunkt des theoretischen Beitrags von Decristan und Kolleginnen (in diesem Heft) steht die Frage, inwieweit die Unterscheidung von Oberflächen- und Tiefenstrukturen des Unterrichts dazu beitragen kann, eine *systematische Brücke zwischen Unterricht und Lernen* zu schlagen, bzw. im Sinne einer Unterrichtstheorie Lehren und Lernen zu verknüpfen. Inwieweit das Begriffspaar diesen weit reichenden Anspruch erfüllen kann, wird im Text anhand einer gründlichen Analyse erörtert, beginnend mit

einem Blick auf die historischen Wurzeln und gefolgt von einem begrifflichen Klärungsversuch. Dabei zeigt sich, dass die Bestimmung von Oberflächen- und Tiefenstrukturen in der Literatur nicht einheitlich ist, was sich auch an den unterschiedlichen Bezeichnungen – oft in Form von Mehrfachumschreibungen – ablesen lässt: So ist z. B. die Rede von der „Oberflächen-, Sicht- oder Handlungsstrukturebene“ auf der einen, und von „Tiefenqualitäten oder Tiefenstruktur“ (Reusser, 2019, S. 144), „Basisstruktur“ oder „Basismodellen“ des Unterrichts (Oser & Patry, 1990) auf der anderen Seite.

Sowohl die Unterscheidung von Oberflächen- und Tiefenstrukturen (Reusser, 2009) als auch jene von Sichtstrukturen und Basismodellen (Oser & Patry, 1990) des Unterrichts wurden ursprünglich im Kontext der Didaktik entwickelt. Beide nehmen einen Grundgedanken psychologischer Didaktiken wie z. B. jener von Hans Aebli (1983) auf, nämlich Unterricht von den intendierten Lernprozessen der Schülerinnen und Schüler her zu denken (Messner, 2019; Messner & Reusser, 2006). Mit seinem „PADUA“-Modell vollständiger Lernprozesse (**P**roblemlösender **A**ufbau, **D**urcharbeiten, **U**eben/Wiederholen und **A**nwendung) hatte Aebli eine „moderne Version einer Formalstufentheorie“ (Reusser, 2009, S. 890) entwickelt, die im Prinzip auf alle kognitiven Aneignungsprozesse anwendbar sein sollte[1] – als „auf das Bildungsziel Verstehen gerichtete kognitive Tiefengrammatik des Lernens“ (Messner, 2019, S. 39). Die mit PADUA beschriebenen Formalstufen betreffen sowohl die Lern- als auch die darauf bezogenen Lehrfunktionen: Aufgabe der Lehrpersonen ist es, bei allen Schülerinnen und Schülern zielführende Lernaktivitäten zu initiieren, zu unterstützen und anzuleiten, um vollständige Lernprozesse zu ermöglichen. Dies erfolgt durch eine geeignete Inszenierung, d. h. Strukturierung der Unterrichtszeit als Abfolge von Sozialformen und Lehrmethoden. Daraus ergeben sich die beiden Betrachtungsebenen von Unterricht, für die sich u. a. die Bezeichnung Oberflächen- und Tiefenstrukturen eingebürgert hat: Einerseits können Merkmale der Inszenierung (z. B. die Lehrmethoden) fokussiert werden, andererseits ist es aber auch möglich, die Lernaktivitäten und -interaktionen der Schülerinnen und Schüler bzw. deren Qualität im Hinblick auf die Erreichung bestimmter Lernziele in den Blick zu nehmen. Diese Unterscheidung ist sowohl für die Didaktik als auch für die pädagogisch-psychologische Unterrichtsqualitätsforschung relevant: Aus didaktischer Perspektive stellt sie ein Werkzeug für das Unterrichten dar, indem sie Lehrpersonen dazu anhält, Unterricht mit Blick auf die angestrebten, auf mehrdimensionale Bildungsziele bezogenen Lernprozesse aller Schülerinnen und Schüler zu planen, durchzuführen und zu reflektieren, wobei Letzteres auch die Frage einschließt, wie gut sich die gewählten Formen der Unterrichtsinszenierung (d. h. die Oberflächenstrukturen) in Bezug auf die Qualität der Lernprozesse und des Lernerfolgs bewährt haben. Solche Fragen stehen auch im Mittelpunkt der pädagogisch-psychologischen Unterrichtsqualitätsforschung. Als heuristisches Instrument kann die Unterscheidung dazu beitragen, auf der Grundlage pädagogisch-psychologischer, aber auch (fach-)didaktischer Kriterien wirksamer

1 Demgegenüber postulierten Oser und Patry (1990) zwölf spezifische, auf bestimmte Lernzieltypen bezogene Basismodelle, die, wie auch Decristan et al. (in diesem Heft) darlegen, v. a. in fachdidaktischen Forschungskontexten zur Anwendung kommen.

Lehr- und Lernprozesse die Bedeutung bestimmter Unterrichtsqualitätsmerkmale (auch über die sog. *Basisdimensionen* hinaus) im Hinblick auf fachliche und überfachliche Bildungsziele sowie Zusammenhänge zwischen Merkmalen der Oberflächen- und Tiefenstrukturen zu untersuchen (vgl. auch den Beitrag von Hess und Lipowsky, in diesem Heft) und so eine Brücke zwischen Didaktik und Unterrichtsqualitätsforschung schlagen.

Eine offene Frage ist dabei, wie differenziert die theoretische Spezifizierung der Tiefenstrukturen erfolgen kann bzw. soll, jenseits relativ allgemeiner, aber empirisch überprüfter Erkenntnisse der Lern- und Gedächtnispsychologie und Lehr-Lernforschung. Mit Blick auf eine gewinnbringende Nutzung der Unterscheidung von Oberflächen- und Tiefenstrukturen sowohl als didaktisches Werkzeug für das Unterrichten als auch als heuristisches Instrument für die Unterrichtsforschung erscheint es sinnvoll, sie sparsam als zwei Betrachtungsebenen festzulegen und bei ihrer Anwendung im konkreten Fall, bezogen auf den Lerngegenstand und die Art der Lernziele, pädagogisch-psychologisch begründete und insbesondere auch fachlich bzw. fachdidaktisch ausdifferenzierte Indikatoren für die tiefenstrukturelle Qualität zu bestimmen, wobei nicht nur fachliche, sondern auch überfachliche und nicht-kognitive Lernziele zu berücksichtigen sind. Solche Indikatoren können sich sowohl auf das Verhalten von Schülerinnen und Schülern als auch auf die Interaktion zwischen Lehrkräften und Schülerinnen und Schülern beziehen, da es im Unterricht i. d. R. um sozial angeleitete Lernprozesse geht. So wurden beispielsweise in den letzten Jahren im Bereich der Lehr-Lernforschung und den Fachdidaktiken Erkenntnisse generiert, aus denen sich theoretisch und empirisch fundierte Hypothesen in Bezug auf die Gestaltung eines lernwirksamen Unterrichts ableiten lassen, beispielsweise in Bezug auf produktive Lernaktivitäten (Chi & Wylie, 2014) und Unterrichtsgespräche (vgl. z. B. Resnick, Asterhan, Clarke & Schantz, 2018), individuelle Lernunterstützung und Feedback (vgl. z. B. Schütze, Souvignier & Hasselhorn, 2018) oder die Stoffdarbietung (z. B. Drollinger-Vetter, 2011).

2. (Wie) hängen Oberflächen- und Tiefenstrukturen des Unterrichts zusammen?

Die Frage nach systematischen Zusammenhängen zwischen Oberflächen- und Tiefenstrukturen ist Gegenstand des empirischen Beitrags von Hess und Lipowsky (in diesem Heft). Ihre Analyse der Lehrpersonenfragen (als Merkmal der Tiefenstruktur) zeigt, dass sich deren Qualität je nach Sozialform unterscheidet und somit nicht unabhängig von der Gestaltung der Unterrichtsoberfläche ist. Dass zwischen Oberflächen- und Tiefenstrukturen systematische Zusammenhänge bestehen, indem z. B. Sozialformen und Lehrmethoden je spezifische Potenziale im Hinblick auf unterschiedliche Bildungsziele aufweisen, ist besonders im Zusammenhang mit der Förderung überfachlichen Kompetenzen, wie z. B. Kooperations- oder Lernkompetenzen, offensichtlich. Doch auch Prozesse der fachlichen Wissensaneignung (vgl. z. B. Glogger-Frey, Gaus & Renkl, 2017) und, wie die Ergebnisse von Hess und Lipowsky (in diesem Heft) zeigen, Merk-

male der Interaktion zwischen Lehrkräften und Schülerinnen und Schülern können von methodischen Gestaltungsmerkmalen beeinflusst werden.

Die Ergebnisse von Hess und Lipowsky (in diesem Heft) werfen sowohl inhaltlich wie methodisch gesehen interessante Fragen auf, wie z. B. die methodisch relevante Frage, inwieweit die Qualität von Interaktionen zwischen Lehrkräften und Schülerinnen und Schülern überhaupt unabhängig von ihrer Einbettung in den Verlauf einer Unterrichtsstunde erfasst werden kann. Denn es ist denkbar, dass bestimmten, beobachtbaren Lehr- und Lernaktivitäten je nach Kontext eine unterschiedliche Bedeutung für die Lernprozesse von Schülerinnen und Schülern zukommt, weil mit dem Wechsel der Sozialformen oder Lehrmethoden auch ein Wechsel der didaktischen Funktion verbunden sein könnte. Im Fall der Lehrpersonenfragen im öffentlichen Unterricht und in Phasen des selbstständigen Arbeitens der Schülerinnen und Schüler (im Folgenden *Arbeitsphasen* genannt) trifft dies vermutlich zu, dienen die Fragen doch einerseits der Leitung von Unterrichtsgesprächen im Hinblick auf einen gemeinsamen Wissensgenerierungsprozess und andererseits der individuellen Lernunterstützung, was, wie auch Hess und Lipowsky (in diesem Heft) feststellen, mindestens teilweise die nachgewiesenen Unterschiede erklären dürfte. Aus inhaltlicher Sicht lassen sich die Ergebnisse gemäß Hess und Lipowsky (in diesem Heft) auch als Hinweis auf ungenutztes Potenzial in Bezug auf die kognitive Aktivierung der Lernenden in Arbeitsphasen interpretieren. Angesichts der aktuell beobachtbaren Zunahme individualisierender Unterrichtsformen, die tendenziell mit einem höheren Anteil an Arbeitsphasen von Schülerinnen und Schülern einhergehen (Stebler, Pauli & Reusser, 2018), ist dies besonders bedeutsam. Damit leistet die Studie von Hess und Lipowsky (in diesem Heft) sowohl aufgrund der Ergebnisse als auch aufgrund der methodischen Folgerungen einen wichtigen Beitrag zur Fragestellung dieses Heftteils, nicht nur in Bezug auf die Frage nach Zusammenhängen zwischen Oberflächen- und Tiefenstrukturen sondern auch in Bezug auf den Nutzen dieser Unterscheidung.

Zusammengenommen scheint es auch aufgrund der beiden Beiträge sinnvoll, die Unterscheidung von Oberflächen- und Tiefenstrukturen trotz gewisser begrifflicher Unschärfen beizubehalten. Dies trifft insbesondere auch mit Blick auf die Aus- und Weiterbildung von Lehrpersonen zu, da sie für (angehende) Lehrpersonen ein hilfreiches didaktisches Werkzeug für die Unterrichtsplanung und -reflexion darstellt. Für die Unterrichtsforschung kann sie als nützliche Heuristik betrachtet werden, indem sie den Blick sowohl auf Kernmerkmale und -anforderungen wirksamer Lehr- und Lernprozesse als auch auf methodische Gestaltungsmerkmale des Unterrichts sowie auf mögliche Zusammenhänge zwischen Beidem lenkt und so beispielsweise dazu beitragen kann, Potenziale und Gelingensbedingungen methodischer Gestaltungsformen im Hinblick auf das Erreichen mehrdimensionaler Bildungsziele auf der Grundlage pädagogisch-psychologischer und (fach-)didaktischer Erkenntnisse systematisch zu untersuchen. Dies allerdings ohne den Anspruch, Lehren und Lernen im Sinn einer fach-, situations- und lernzielübergreifenden Unterrichtstheorie zu verknüpfen.

Literatur

Aebli, H. (1983). *Zwölf Grundformen des Lehrens.* Stuttgart: Klett-Cotta.

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243.

Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit. Fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht.* Münster: Waxmann.

Glogger-Frey, I., Gaus, K., & Renkl, A. (2017). Learning from direct instruction: Best prepared by several self-regulated or guided invention activities? *Learning and Instruction, 51*(Supplement C), 26–35.

Messner, R., & Reusser, K. (2006). Aeblis Didaktik auf psychologischer Grundlage im Kontext der zeitgenössischen Didaktik. 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. 52–73). Bern: h. e. p.

Messner, R. (2019). „Tiefen-Didaktik" – zur praktischen Wende der Lehr-Lernforschung. In U. Steffens & R. Messner (Hrsg.), *Unterrichtsqualität. Konzepte und Bilanzen gelingenden Lehrens und Lernens* (S. 29–56). Münster: Waxmann.

Oser, F., & Patry, J.-L. (1990). *Choreographien unterrichtlichen Lernens. Basismodelle des Unterrichts.* Freiburg: Pädagogisches Institut (Berichte zur Erziehungswissenschaft Nr. 89).

Resnick, L. B., Asterhan, C. S. C., Clarke, S. N., & Schantz, F. (2018). Next generation research in dialogic learning. In G. E. Hall, L. F. Quinn & D. M. Gollnick (Hrsg.), *Wiley handbook of teaching and learning* (S. 323–338). New York: Wiley.

Reusser, K. (2009). Unterricht. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 881–896). Weinheim: Beltz.

Reusser, K. (2019). Unterricht als Kulturwerkstatt in bildungswissenschaftlich-psychologischer Sicht. In U. Steffens & R. Messner (Hrsg.), *Unterrichtsqualität. Konzepte und Bilanzen gelingenden Lehrens und Lernens* (S. 105–128). Münster: Waxmann.

Schütze, B., Souvignier, E., & Hasselhorn, M. (2018). Stichwort – Formatives Assessment. *Zeitschrift für Erziehungswissenschaft, 21*(4), 697–715.

Stebler, R., Pauli, C., & Reusser, K. (2018). Personalisiertes Lernen – Zur Analyse eines Bildungsschlagwortes und erste Ergebnisse aus der perLen-Studie. *Zeitschrift für Pädagogik, 64*(2), 159–178.

Abstract: After a look at the context of the genesis of the distinction between surface and deep structures of teaching, the commentary turns to the contributions by Decristan, Hess, Holzberger und Praetorius (in this issue) and Hess und Lipowsky (in this issue). It adresses the question of the usefulness of this distinction and the relationship between surface and deep structures. The author regards the distinction as a pedagogical tool for the planning of and reflection on effective teaching and as a heuristic instrument for research into the quality of teaching. Besides serving other purposes, it provides a conceptual foundation for a systematic approach to deal with research questions that have so far been rather neglected, such as the relationship between teaching methods and the quality of teaching and learning in relation to multidimensional educational objectives.

Keywords: Surface and Deep Structure of Teaching, Teaching Quality, Teaching Research, Didactics, Teaching and Learning Processes

Anschrift der Autorin

Prof. Dr. Christine Pauli, Universität Freiburg,
Departement für Erziehungs- und Bildungswissenschaften,
Zentrum für Lehrerinnen- und Lehrerbildung,
Rue P.-A. de Faucigny 2, 1700 Fribourg, Switzerland
E-Mail: christine.pauli@unifr.ch

Themenblock IV: Zur Bedeutung unterschiedlicher Perspektiven bei der Erfassung von Unterrichtsqualität

Benjamin Fauth/Richard Göllner/Gerlinde Lenske/Anna-Katharina Praetorius/ Wolfgang Wagner

Who Sees What?

Conceptual Considerations on the Measurement of Teaching Quality from Different Perspectives

Abstract: One puzzling finding in education research is that teachers, students, and external observers agree only marginally on their ratings of teaching quality. In this theoretical contribution, we summarize and reappraise previous findings on agreement between different raters of teaching quality. We explain these findings by thoroughly examining the instruments that have been used to measure teaching quality. Building on this, we propose a reference perspective matrix, which should be useful in explaining perspective-specific rating mechanisms behind responses to certain survey or observation items. The reference perspective matrix could thus afford a theoretical foundation for future studies on the assessment of teaching quality.

Keywords: Teaching Quality, Measurement, Validity, Classroom Management, Agreement

1. Introduction

Teaching quality is one of the most prominent and powerful predictors of student learning in school (Hattie, 2009). However, properly measuring teaching quality is still a huge challenge. Researchers and practitioners often assess teaching quality by using ratings from students, teachers, or trained observers. Although researchers assume that these different approaches assess the same target constructs, empirical studies have repeatedly found low or even zero correlations between these data sources when they are applied to the same sample of classes (e. g., Clausen, 2002; Fauth, Decristan, Rieser, Klieme & Büttner, 2014b; Kunter & Baumert, 2006; Wagner, Göllner, Werth, Voss, Schmitz & Trautwein, 2016).

In the present contribution, we seek explanations for these results. Drawing on a variety of theoretical traditions, including personality and social psychology approaches, we present theoretical considerations that should help to explain previous findings and that may serve as a framework for future studies. Our approach is to examine the items

found in commonly used survey and observation instruments in order to investigate how the specific wording of items might shape responses by students, teachers, and observers. Focusing on the construct of classroom management, we show that items refer either to teacher behavior, to student behavior, or to a mixture of both, and that these different item references have consequences for assessments of classroom management. On the basis of this observation, we present a matrix of item references and rater perspectives that can help us understand how a certain item's wording may shape answers to this item from a certain perspective.

2. Teaching Quality: The Problem of Alignment Between Perspectives

The increased use of direct measures of teaching quality has been accompanied by a growing interest among researchers in the psychometric quality of these measures. An important indicator of psychometric quality is the degree to which different data sources produce the same results in evaluating the same instructor. In a seminal study, Clausen (2002) drew on high-inference video ratings of the German TIMSS 1995 video data and compared these ratings to survey responses collected from students, and teacher self-reports. The study found that only 13 out of 36 correlations between corresponding scales (as measured from the three perspectives) yielded values differing significantly from zero. The average correlation between the perspectives was .16, with a range between −.28 and .45 (Clausen, 2002, p. 129). By September 2019, this study had been cited 450 times, according to Google Scholar: this indicates that other researchers have indeed paid attention to this finding.

In the last 15 years, numerous studies have applied a similar approach and found relative agreements between rater perspectives roughly in the range reported by Clausen (2002; Camburn & Barnes, 2004; Chaplin, Gill, Thompkins & Miller, 2014; De Jong & Westerhof, 2001; Desimone, Smith & Frisvold, 2010; Fauth et al., 2014b; Gitomer et al., 2014; Kunter & Baumert, 2006; Wagner et al., 2016; Wettstein, Ramseier, Scherzinger & Gasser, 2016; Mayer, 1999; Kaufman, Stein & Junker, 2016). These studies have two consistent findings: First, overall, the correlations between measures obtained from student ratings, teacher self-evaluations, and observation protocols are low. Second, the highest correlations are among indicators of classroom management, rather than indicators of other constructs, such as cognitive activation or student support.

2.1 Perspective-Specific Validities?

The relatively low correlations between perspectives have led researchers to think about the relations between different data sources in terms of validity rather than reliability (Kunter & Baumert, 2006). Additionally, it has been put into question whether it makes sense from a methodological standpoint to think of teaching quality as a perspective-independent construct (Clausen, 2002). Accordingly, *perspective-specific validities* have

been hypothesized for the different data sources: "It is conceivable that students' and teachers' perceptions tap different aspects of the classroom environment, rather than the same underlying construct" (Kunter & Baumert, 2006, p. 234). Indeed, from an epistemological perspective, we have to acknowledge that humans' perceptions of their environment are perspective specific in nature (Graumann, 1960). Both philosophers and psychologists have argued convincingly that our knowledge of the world is and will always be an idiosyncratic construction that is fundamentally affected by our individual preconceptions and schemes of perception. This idiosyncratic way of perceiving our environment is rooted in previous experiences during the life course. As these experiences naturally differ between persons, their perceptions of the environment will also differ.

The literature nowadays commonly refers to perspective-specific validities to explain and/or justify low correlations between perspectives (e. g., Fauth, Decristan, Rieser, Klieme & Büttner, 2014a; Wettstein et al., 2016). In the present paper we argue, however, that this approach has at least two pitfalls.

First, the term teaching quality is currently not used in a perspective-specific way, either by teachers or by those conducting substantive research. In the contexts where these measures are usually applied, most people are interested in *teaching quality,* not in *teaching quality as perceived from a certain perspective.* When we think about classroom interactions that foster student learning, we usually do not think about 'teachers' perceived classroom interactions' or 'students' perceived classroom interactions.' Thus, from a scientific perspective, knowing that human perceptions are perspective specific in nature should not limit the search for the best instruments to measure teaching quality.

Second, the plausibility of the concept of perspective-specific validities may vary for different constructs. The student's perspective on the feeling of being emotionally supported by the teacher might have a special relevance. Some degree of nonagreement will be the standard for these support dimensions, even within one rater group's perspective (e. g., students in a class; Schweig, 2016). In contrast, classroom disruptions or teacher strategies to ensure smooth transitions could be perceived differently from different perspectives. But we would not expect different disruptions or different transitions to be in evidence, depending on who is doing the rating. The events rated are relatively distinct ones in the classroom that – in principle – everyone should be able to rate accurately. Accordingly, deviations between perspectives should be understood in light of reliability rather than validity.

Consequently, at least for classroom management, we assume that there is a 'true score' and that while deviations between perspectives are possible, they require explanation. Having acknowledged that agreement between perspectives can be expected, nonagreement has to be explained. The concept of perspective-specific validities is potentially attractive for researchers, as it offers a plausible explanation for nonagreement. But the risk that this concept entails is that deviations between perspectives may be unquestioningly accepted, instead of being properly investigated.

2.2 Approaches to Explaining Low Correlations Between Perspectives

A number of reasons for perspective-specific deviations have been advanced in the literature. For instance, researchers have considered that students might find it difficult to judge the didactic value of specific math assignments, or that teachers might find it difficult to judge the correct learning speed for students (Kunter & Baumert, 2006; Mayer, 1999). Some researchers have expressed doubts on whether *students* sufficiently understand the pedagogic principles underlying teaching (e. g., Fauth et al., 2014b). Thus, their agreement with teachers and observers would be lower for constructs requiring an understanding of pedagogy (Clausen, 2002). One example would be the 'Socratic-dialogue practice' (e. g., "In math class, our teacher lets us keep making the wrong assumption until we notice it ourselves.").

As *external observers* usually only get a short look at what is going on in the classroom during the school year (usually one to five lessons; Praetorius, Pauli, Reusser, Rakoczy & Klieme, 2014), a "sampling effect" (Clausen, 2002, p. 90) is assumed to limit the accuracy of observer ratings. Consequently, for constructs of low observability (e. g., scales that refer to seldom-occurring events such as the 'social orientation' of the teacher being demonstrated, as in the following, sample item: "Our math teacher cares about students' problems") or high variability across lessons (Kane et al., 2012; Praetorius et al., 2014), observer ratings would not correlate highly with student and teacher ratings. In the case of *teacher* ratings, self-serving biases can be a problem. Hence, one would expect lower correlations to student and observer ratings for scales with a strong evaluative component. Clausen (2002, p. 91) cites "discipline" as an example (sample item: "The students in this class mess around a lot" – see Section 3 for a discussion of this scale).

In an attempt to account for these challenges, Clausen (2002) named one characteristic of each rating perspective (didactic understanding, small sample of lessons, self-serving bias) that limits the accuracy of ratings from that perspective. Each major characteristic would lead to deviations from the other two perspectives. According to Clausen, such deviations should differ in extent, depending on the characteristics of the target constructs (demands regarding didactic understanding, observability, evaluativeness). Clausen (2002) rated these characteristics of the target constructs for each scale used in the TIMSS video data, to test the assumptions of this model empirically. However, the results showed little support for these assumptions.

On second glance, the categorization of the target constructs and their measurement shows that the observability or evaluativeness of an item is difficult to determine. We discussed these questions extensively and uncovered one possible explanation: That the inconsistent findings may be explained by the specific wording of survey and observation items. More precisely, we suspected that item characteristics such as observability or evaluativeness strongly depend on whom or what an item refers to (item reference): that is, whether the item refers to the teacher's or to students' classroom behavior. In order to explain this, we need to briefly discuss the theoretical foundations of teaching quality.

2.3 Two Sides of a Coin – the Significance of Teacher and Student Behavior for Teaching Quality

From a theoretical standpoint, teaching is interactive in nature, and teaching quality thus can only be understood as an interplay between teachers and students. Current research agrees with this line of reasoning, and accordingly has conceptualized teaching quality as a complex social process that takes place in the interactions between students and teachers (Doyle, 2013). This also implies that teaching quality is not completely determined by the teacher's behavior (Göllner, Fauth, Lenske, Praetorius & Wagner, in this issue; Fauth et al., in press). Instead, teaching must be understood as a "social practice that is co-constructed by students and teachers" (Praetorius, Klieme, Herbert & Pinger, 2018, p. 6). This theoretical view of teaching quality as a co-construction between teachers and students is also reflected in theoretical conceptualizations of classroom management. The ecological approach to classroom management established by Doyle (2013), and the early work of Kounin (1970) describe the characteristics of interactions between teachers and students rather than specific teacher behaviors (Klieme, 2006).

The quality of a certain teaching strategy will always be hard to evaluate without knowing the students' reactions to it (or the students' behavior preceding the teachers' action). For instance, the same classroom management strategy will have a completely different impact when it is applied in a class of well-behaved students compared to poorly behaved students. Accordingly, Praetorius et al. stated that "Classroom management is both a condition for students getting attentive (e. g., through teacher monitoring) and an indication of students being attentive (e. g., lack of interruptions)" (2018, p. 6).

This conceptualization has consequences. Given that most teaching quality assessments are designed to evaluate the teacher, the notion that teaching quality also depends on the students is not trivial.

The current policy press is to develop measures that allow for inferences about *teacher* effectiveness. Using particular measures, the goal is to be able to make some type of claim about the qualities of a teacher. Yet, to varying degrees, the measures we examine do not tell us only about the teacher. A broad range of contextual factors also contributes to the evidence of the *teaching* quality, which is more directly observable. (Gitomer & Bell, 2013, p. 416).

This understanding of teaching quality as a co-construction between teachers and students is seemingly shared by many researchers in the field of education research and educational psychology. For example, in a study in Germany (Clausen, Schnabel & Schröder, 2002), 22 researchers from these two fields were asked to rate student survey scales in regard to similarities and differences ('free pile sorting'). Using multidimensional scaling, the authors showed that participants used the degree to which the scales referred to student or teacher behavior, to sort items (Clausen et al., 2002). This result shows that – at least in researchers' understandings of these scales – item reference plays a central role: Items refer in varying degrees to student or to teacher behavior.

3. The Perspective Reference Matrix: A Taxonomy for Understanding Perspective-Specific Rating Processes

The notion that teaching quality is not limited to teacher behavior, but also depends on student behavior, is reflected in the items that researchers have used to measure teaching quality. We believe that this insight can play a particularly important role in understanding previous results on correlations between perspectives. In the present paper, we argue that how an item rates in terms of observability or evaluativeness will always depend on the kinds of questions asked, and on the person who has to answer these questions. It is therefore necessary to analyze perspective-specific judgments at the level of specific items.

Consider the above-mentioned sample item for *discipline* ("The students in this class mess around a lot"): This item refers not to teacher behavior but to student behavior. Thus, it seems at least questionable whether it is really highly evaluative *when answered from the teacher's perspective* (see Section 2.2). The claim that students in a class tend to mess around a lot might even be a good, self-serving explanation for teachers in chaotic classrooms. This would make the item nonevaluative for teachers. In contrast, imagine a student who has to respond to the item "In this class, I mess around a lot". As the student's (mis-)behavior is now being evaluated, this item becomes highly evaluative – but only *when it is answered from the student's individual perspective.* Finally, external observers do not need excuses for chaotic classrooms; nor will they feel evaluated when students mess around. This simple example shows that evaluativeness and observability (see Section 2.2) are not attributes of constructs or items, but rather of item-rater combinations: How evaluative an item is will always depend on who is answering it and to whom the item refers. These examples also suggest that these differences may relate to differences between self- and other-ratings: The same item can require a self-rating from one perspective but an other-rating from the other perspective. This thought can be formalized in a matrix (Fig. 1) of rater perspectives (who is rating?) and item references (what is rated?).

To understand the specific mechanisms that underlie responses to a certain item, it is crucial to be aware both of the perspective from which an item is answered (teacher, student, or external observer) and of what the item refers to (teacher actions, student actions, or a mixture of both). Additionally, the response to a certain item from a certain perspective will be driven by the quality and the quantity of information that is available to a rater, as well as the degree of ego-involvement that a specific item wording implies for a certain rater. We describe these psychological mechanisms in detail in Section 3.3.

In the following sections, we first of all evaluate whether the parameters of this taxonomy are the most relevant ones. To do so, we first review currently used instruments, to show that the item reference dimension is relevant to them. Afterwards, we examine the psychological processes that are assumed to be responsible for the differences in self- and other-ratings that occur when different item references are rated from different perspectives.

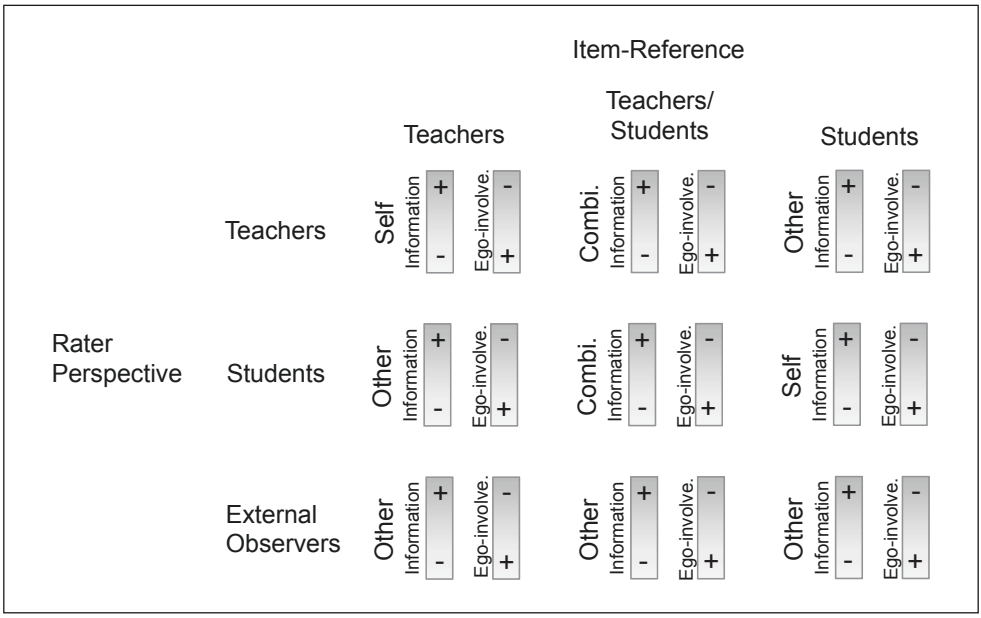

Fig. 1: Reference perspective matrix

3.1 Different Item References in Assessments of Classroom Management

In the following section, we review whether and how the different item references (columns in the taxonomy; see Fig. 1) relate to the measures that are commonly used to assess teaching quality from different perspectives (rows in the taxonomy). Here, we concentrate on the field of classroom management. The instruments discussed below were selected according to two criteria: First, we selected instruments and items for measuring classroom management from each of the three perspectives of external observations, student ratings, and teacher self-ratings. Second, within each of the perspectives we concentrated on those instruments that are either most popular in the educational system (e. g., the Tripod student survey in the US) or that are most frequently used in empirical education research. In our review, we did not make any a priori assumption that one of the perspectives (e. g., external observations) would be superior to the others, in the sense that one group of instruments would – in general – provide more accurate ratings than the others. Additionally, the following sections should not be read as an evaluation of specific instruments. Rather, the instruments reviewed below serve as examples of the way teaching quality is assessed from the different perspectives. The primary focus of this review is: To what extent do these instruments refer to teacher and/or student actions in the classroom?

External observations. The CLASS framework (Classroom Assessment Scoring System; Pianta & Hamre, 2009), which is one of the most frequently used classroom obser-

vation systems, explicitly considers the role of student behavior in successful classroom management. "In contrast to traditional observation protocols that focus on teacher actions, CLASS-S is representative of more recent evaluation protocols that focus on the actions and interactions of both teachers and students" (Gitomer et al., 2014, p. 9). In the CLASS manuals (e. g., Pianta, La Paro & Hamre, 2008), we find items both on teachers' classroom management behavior (e. g., "The teacher is consistently proactive and monitors the classroom effectively"; Pianta et al., 2008, p. 45) and on students' behavior (e. g., "There are few, if any, instances of student misbehavior in the classroom"; Pianta et al., 2008, p. 45). The authors point out in a footnote to the *behavior management* scale that in certain classrooms the teacher strategies described in the protocol might not be observable – "because behavior is so well managed. If there is no evidence of student misbehavior, it is assumed that effective behavioral strategies are in place and a classroom may score in the high range" (Pianta et al., 2008, p. 45). The fact that an assumption about effective behavioral strategies replaces observed teacher behavior in such cases shows that there is indeed a dependency of ratings on student behavior.

Other rating instruments also explicitly address students' behavior when evaluating classroom management. In the Framework for Teaching (FFT; Danielson, 2007), the absence of student misbehavior serves as an indicator of *managing student behavior.* In the video rating instruments that capture the three basic dimensions of teaching quality (Klieme, Pauli & Reusser, 2009), student disruptions are regularly assessed as an indicator of a teacher's classroom management (see Lipowsky et al., 2009; Praetorius et al., 2018).

Student ratings. The Tripod Classroom Environment Survey (Ferguson, 2010) is one of the most frequently applied student surveys in the United States, and has also been used in the Measures of Effective Teaching (MET) study (Kane et al., 2012). Interestingly, in the field of classroom management, the questionnaire asks only about student behavior, not about teacher behavior. These are the items of the *control* scale, which represents one of the Tripod's 7 Cs (sample item: "Student behavior in this class is a problem"; Wallace, Kelcey & Ruzek, 2016, p. 1857; the "Cs" refer to care, captivate, challenge, confer, clarify, consolidate, and control).

Göllner, Wagner, Rose, Fauth, and Nagengast (2018) reviewed student survey items from five large-scale studies conducted in Germany in recent years, and pooled the data collected in these studies into one integrated data set. This final data set included data from a total of 95,328 students. A scale was only included in this data set when it was applied in at least two different studies to 5 % of the whole sample (5,766 students). This approach justifies the authors' assumption that the scales they included constitute a representative sample of what is usually used to capture student ratings of teaching quality in German large-scale assessments. In the field of classroom management, each of the five large-scale studies included items on students' discipline or disruptions in the classroom (student reference). In three out of five studies, items on teachers' monitoring behavior were used (teacher reference). Additionally, two studies asked about the "inefficient use of time" in class (e. g., "In math, a lot of time is wasted"), where the item's wording leaves it open as to whom it refers (combined item reference).

Teacher self-ratings. In the teacher version of the Classroom Assessment Scoring System (CLASS-T; Hamre, 2008), teachers were asked to judge their "areas of strength and growth" (ranging from 1 = area of much growth to 5 = area of great strength). The item "Using time productively" is described as follows: "Productive classrooms are like 'well-oiled machines' – students in these classrooms know what they should be doing and always have something to do" (Gitomer et al., 2014, p. 30). Thus, this item refers to both teachers and students.

Studies that use teacher self-evaluations often use items similar to those used with students (where the item refers to teacher behavior, the third-person perspective is changed to a first-person formulation; e. g., Clausen, 2002; Kunter & Baumert, 2006; Wagner et al., 2016; Wettstein et al., 2016). Accordingly, we find a similar variety of item references (teachers, students, and a mixture of both) in teacher survey items and student survey items.

Summary. Summing up, we can conclude that all of the items used in the aforementioned studies can be categorized into three groups: (a) items that refer to student actions (e. g., student behavior, control, discipline, and disruption), (b) Items referring to teacher actions (e. g., monitoring, teacher awareness of student conduct), and (c) items that are open-ended as to whose actions exactly they are referring to (e. g., inefficient use of time; using time productively). We can draw a relatively well-founded distinction between items that clearly refer to teacher actions and items that clearly refer to student actions. In addition to these easily classifiable cases, there are items that do not spell out a specific referent but that allow the responder to easily infer to what or whom the items refer (e. g., "In math, the lesson is often disrupted," Göllner, Wagner, Rose et al., 2018, where responders will probably think of student misbehavior); there are also items without a clear referent that could be interpreted as referring to teachers, students, or interactions between both (e. g., "In math, a lot of time in class is wasted").

We found studies that use items referring only to student actions to operationalize classroom management (Kunter et al., 2013; Fauth et al., 2014b; Wagner et al., 2016; Wallace et al., 2016) and studies that use items referring only to teacher actions (de Jong & Westerhof, 2001; Wagner, Göllner, Helmke, Trautwein & Lüdtke, 2013; Mayr, Eder, Fartacek, Lenske & Pflanzl, 2013). Finally, there are studies that use both kinds of items, either combined in a single scale (Fauth et al., 2014a; Hochweber, Hosenfeld & Klieme, 2014) or separated in different scales (Clausen, 2002; Wettstein et al., 2016).

Interestingly, the studies mentioned above do not make a clear distinction between scales referring to the teacher's behavior and scales referring to the students' behavior. All of these different items are subsumed under the term classroom management. Additionally, studies tend to attribute well-managed classrooms to the teacher (e. g., Hochweber et al., 2014, p. 289). For example, the Tripod's *control* dimension (items referring only to student behavior) is meant to evaluate whether teachers "are able to manage the class in a way that teaching and learning occur efficiently, without being derailed by misbehavior or distractions" (Ferguson & Danielson, 2015, p. 106). As with the Tripod student survey, many studies use instruments that focus primarily on student behavior to measure classroom management. Oftentimes, students' discipline is the only

indicator of a teacher's classroom management (Fauth et al., 2014; Kunter et al., 2013; Wagner et al., 2016).

In the taxonomy presented here, we assume that the different item references described above would play a role in the relative agreement between different perspectives on teaching quality. In the following section, we explain our hypothesis that differences between self- and other-ratings play an important role in these rating mechanisms, in more detail.

3.2 Self- and Other-Ratings

The reference perspective matrix makes the assumption that the various item references described in the previous section can have an impact on item responses. Depending on who is responding to a certain item (students, teachers, or external observers), the same item can be an invitation to judge one's own behavior (e. g., a teacher's judgment of his/her own monitoring behavior) or another person's behavior (a student's judgment of the teacher's monitoring). The item reference, in combination with the identity of the person answering the item, determines whether an item represents a self- or an other-rating. We assume that this is a crucial issue in the assessment of teaching quality. In the following section, we further outline how the distinction between self- and other-ratings may be highly relevant for understanding perspective-specific ratings of teaching quality.

Think about an item like "Our teacher immediately notices when students start doing something else" (Baumert et al., 2009, p. 211). A student who has to judge this item will have to rely on indirect behavioral indicators that a teacher has noticed something (e. g., the teacher steps nearer to a student who is seemingly not paying attention, or the teacher starts staring at the student while continuing to speak), which the student would have to interpret correctly (see the realistic accuracy model of personality judgement: Funder, 1995). When the same item is answered from the teacher's perspective, we instantly notice significant differences. First, the teacher has privileged access to his/her own thoughts and thus is directly privy to his/her own noticing of something (although the teacher may find it difficult to judge whether he/she *immediately* notices when the students start doing something else). Second, teachers will be much less prone to error than students or external observers, who have to interpret a certain teacher behavior as an indicator of noticing students' attention behaviors. The students, in contrast, are limited to drawing inferences from the teacher's behavior and to interpreting overt behavior, to determine whether this indicates noticing, on the part of the teacher. External observers are in a similar position to students. However, they will have very specific indicators for a teacher noticing something in their rating manual. This will make their ratings more reliable. However, it is very unlikely that these indicators would be the same ones that students use. Teachers' thoughts are hard to read, even for trained observers.

3.3 Information and Motivation as Central Dimensions of Differences Between Self- and Other-Ratings

Such differences between self- and other-ratings also form the foundation of a theoretical model that has been developed in the field of personality research: the SOKA model (self-other knowledge asymmetry) of Vazire (2010). This model has been very influential in personality and social psychology, and has proven to be a powerful framework when it comes to explaining differences between self- and other-ratings in personality research. In the following discussion, we apply this model to our research on different perspectives on teaching quality.

The SOKA model assumes two major asymmetries between self- and other-ratings: (1) the quality and the quantity of information that is available and salient for a rater ("informational difference in perspective"), and (2) the degree of ego involvement that goes along with a rating ("motivational significance"; Vazire, 2010, p. 283). Regarding ego involvement, Vazire (2010) states that "judges have a lot more at stake when they are also the target than when they are judging someone else" (p. 284).

By considering the two aspects of information and motivation, this approach takes into account that "human perceivers act as both intuitive scientists and intuitive politicians – their judgments are influenced by both 'cold' information-processing goals (i. e., understanding and predicting the actor's behavior) and by 'hot' motivational goals (i. e., protecting or enhancing their own self-worth)" (Vazire, 2010, p. 283).

In our example of a teacher noticing whether students are paying attention, we have discussed the informational asymmetries between the teacher's and students' perspectives. What about the second asymmetry, of difference in motivation? Teachers certainly have a professional ethos that highlights that noticing what students do is good and necessary for teachers. So responding to this item will somehow activate the "intuitive politician" (Vazire, 2010) who is motivated to protect his or her self-worth. Additionally, teachers will differ in the importance they place on noticing student actions, and thus they will differ in how strong their intuitive politician is. Students will probably not be as interested in protecting their teachers' self-worth – they have less at stake in this evaluation. However, as described above, if they want to answer this item honestly they face another severe problem: They lack relevant *information.* In fact, the double asymmetry of *motivation* and *information* makes deviations between the three perspectives more than plausible.

Let us have a look at an item already discussed above: "The students in this class mess around a lot" (item from PISA 2003 assessment; see Kunter & Baumert, 2006, p. 245). The asymmetries in information and motivation will probably be less important in responses to this item from different perspectives. The *information* available will not be very different for the different perspectives. Students messing around do not refer to someone's thoughts, but to openly displayed behavior. Concerning *ego involvement,* teachers have much less at stake when student behavior is under scrutiny (in fact, it might even be self-protective for a teacher to agree with this item – see our discussion of this item example above). Students – who are rating their own behavior in this case –

are probably not explicitly motivated to answer the item in a certain way either. That is because this item is not a pure self-rating but a combination of self- and other-ratings. The item leaves open the degree to which an individual student is responsible for the messing around. This phenomenon will be very common in items referring to students. Many surveys contain items that refer to individual students ("I-form") as well as to the whole class ("We-form," according to Sirotnik, 1980; see also Den Brok, Brekelmans & Wubbels, 2006; Wagner et al., 2013). To assess items that focus on student behavior and that are rated by students, one should therefore always take into account whether the item asks respondents to assess their individual student behavior or to assess class behavior as a whole.

Thus, we have identified differences in information and ego involvement as two central factors that operate in a perspective-specific way rating teaching quality. Regarding the reference perspective matrix, we now have a better idea of why students, teachers, and external observers might disagree in their ratings of teaching quality. Different processes are at work depending on whether an item represents a self- or an other-rating. These different processes can be explained by differences in the information available to raters and how the information is used to respond to an item. To make predictions about the accuracy of a rater's response to an item, it is necessary to begin by identifying the cell where the item is located from a certain perspective, and then to evaluate the extent of information available to answer this item, and the degree of ego involvement that could motivate the rater to answer in a certain way.

4. Applying the Reference Perspective Matrix to Previous Findings

In this section, we revisit previous studies and reinterpret their results in light of the reference perspective matrix. The main questions in this section are: Does the item reference really make a difference? And second: Can these differences between perspectives be explained by informational and motivational differences in self- and other ratings?

4.1 Factorial Structure in Student Ratings

In Fauth et al. (2014a) the factorial structure of primary school students' teaching quality ratings were examined. The expected three factors could be distinguished relatively well. However, the model fit might have been artificially inflated, due to the fact that one factor consisted only of items referring to student behavior (classroom management), whereas all the items of the other two factors referred to the teacher. The comparably low correlations between the classroom management factor and the two other factors are in line with this interpretation (Fauth et al., 2014b, p. 6).

In a recent study on the Tripod student survey by Wallace et al. (2016), the authors presented a bi-factor model of teaching quality ratings, with one general factor representing all items and one specific factor in classroom management. Their interpretation

suggested that there was a general teaching competence and an additional specific classroom management competence. However, taking a closer look at the items, it turns out that the items on classroom management (the control dimension of Tripod) are also the ones that refer to students' actions (e. g., "Student behavior in this class is a problem"), whereas almost all of the other items[1] refer to the teacher (e. g., "My teacher explains difficult things clearly"). Thus, at this point we do not know whether the distinctive aspect of the specific classroom management factor is the substantive focus on classroom management or the reference to students' behavior rather than teacher behavior. A similar factor structure emerged in the analyses of Schweig (2014), who considered schools, as level-2 units, rather than classes.

4.2 Correlations Between Perspectives

These studies indicate that item references make a difference within the perspective of student ratings. When we take into account the reference perspective matrix and the considerations of informational and motivational differences, we would additionally predict higher between-perspective correlations for those scales that refer to student behavior (see the extensive discussion above on the "students messing around" items).

In Kunter and Baumert (2006), there was one factor (classroom management) that emerged in both the students' and in the teachers' responses (while all other items revealed a different factor structure in both perspectives). Again, the items on classroom management were the ones that referred to the students, while all other items focused on teacher behavior. In accordance with our assumptions, it was also this factor that showed the highest correlations among the few significant relationships between perspectives in this study (latent correlation of 0.64, see Kunter & Baumert, 2006, p. 240). This pattern was confirmed in Wagner et al. (2016) and Fauth et al. (2014a), where the highest correlations between teacher and student ratings were again found for classroom management measured with items referring to student behavior.

However, in the four studies mentioned above, the relationship between item reference and substantive focus on classroom management was confounded. That is, classroom management was measured using items referring to the students, whereas all other factors were measured with items referring to the teacher. In Wettstein et al. (2016), the authors included three scales with items referring to students (e. g., "Some students don't really listen to the teacher") and one scale referring to the teacher (e. g., "The teacher notices when students are not on task"). Correlations between teacher and student ratings were only found in the three scales that referred to students. No correlation was found in the scale referring to the teacher. Again, these results are in line with the assumption of different informational and motivational processes in different cells of the reference perspective matrix. Wettstein et al. (2016) also examined the correlations between student ratings of two different teachers teaching the same class. The results revealed a simi-

1 Four out of the remaining 29 items had a student referent.

lar pattern: No correlation between ratings of the two teachers on the scale referring to teacher behavior, but high correlations in scales referring to student behavior.

4.3 Different Item References for Different Perspectives

Clausen (2002) is another example of a study that considered items with both referents – teacher behavior (in the form of teachers' *monitoring*) and student behavior (in the form of students' *discipline*). Consistently with the results of Wettstein et al. (2016), the author found no significant correlations between ratings of teachers' monitoring behavior. In the case of discipline, only student and observer ratings were correlated. A closer look at the items used for the teacher scale of discipline reveals, however, that these items actually do not refer to student behavior but rather to the teacher (e. g., "Right in the beginning of a new course I explain the rules that students have to stick to in round terms"; Clausen, 2002, p. 219). Thus, regarding the perspective reference matrix, we have a comparably high correlation between those scales with the same reference (same column in the matrix) and no significant correlations between scales with different item references (different columns in the matrix; see Fig. 1).

In summary, the results of the studies reviewed above provide strong evidence that item reference really matters. Additionally, we have found indications that differences between self- and other-ratings have important implications for responses to measures of teaching quality. Certainly, the results presented above indicate that we should pay more attention to these issues in future research on the assessment of teaching quality.

5. Limitations and Future Research

In the present contribution, we started with the question of how to explain the low agreement between different rater perspectives in teaching quality assessments. With the reference perspective matrix developed here, we now offer more general considerations of how teaching quality can be best assessed. This contribution might thus be helpful when it comes to developing high-quality survey and observation instruments in the future.

The presented matrix provides a taxonomy that can serve as a heuristic for examining survey items and that may also be helpful in explaining results from previous studies. However, the explanations given above are as yet post hoc hypotheses. To properly validate the assumptions made above, we would need strong, preferably experimental research designs. One possibility would be to systematically manipulate assessment items regarding their item reference (teacher/student behavior) but also in regard to the differences in information and motivation that a certain item from a certain perspective is likely to trigger. These factors would not be easy to manipulate in an isolated way, but we believe that the effort invested in this endeavor would be worth it.

Another possibility would be to make more use of recent approaches to video observation data. For instance, the 'advocatory' approach proposed by Oser, Curcio &

Düggeli (2007) explicitly takes into account multiple perspectives on teaching as well as a combination of self- and other-ratings. In this approach, the competence of a teacher is not being evaluated with very broad items trying to capture what is going in the classroom *in general.* Instead, teachers are invited to judge very specific situations in videos of another teacher's instruction. The quality of these judgements can then serve as indicators of the observing teacher's competence. Although this approach targets teacher competence rather more than actual teaching quality, we believe that such innovations could also be helpful in addressing some of the issues raised in this paper.

The present contribution has limited application of the reference perspective matrix to the field of classroom management. Whether the matrix will also be applicable to other constructs of teaching quality such as cognitive activation and individual support, is as yet unclear. Idiosyncratic perceptions play a more important role in such constructs (Göllner, Wagner, Eccles & Trautwein, 2018). We have reason to believe that the reference perspective matrix will also prove helpful to understanding perspective-specific ratings in these areas, where perspective-specific differences in information and ego-involvement will also likely be crucial factors. This assumption will have to be examined in future contributions.

References

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., Krauss, S., Kunter, M., Löwen, K., Neubrand, M., & Tsai, Y.-M. (2009). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente* (Materialien aus der Bildungsforschnung Nr. 83). Berlin: Max-Planck-Institut für Bildungsforschung.

Camburn, E., & Barnes, C.A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal, 105,* 49–73.

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh public schools.* Regional Educational Laboratory Mid-Atlantic.

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

Clausen, M., Schnabel, K., & Schröder, S. (2002). Konstrukte der Unterrichtsqualitat im Expertenurteil. *Unterrichtswissenschaft, 30,* 246–260.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: ASCD.

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

Den Brok, P., Brekelmans, M., & Wubbels, T. (2006). Multilevel issues in research using students' perceptions of learning environments. *Learning Environments Research, 9,* 199–213.

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.

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., Wagner, W., Bertram, C., Göllner, R., Roloff-Bruchmann, J., Lüdtke, O., Polikoff, M., Klusmann, U., & Trautwein, U. (in press). Don't blame the teacher? The need to account for classrooms characteristics in evaluations of teaching quality. *Journal of Educational Psychology.*

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 Instruction, 29,* 1–9.

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

Ferguson, R. (2010). *Student perceptions of teaching effectiveness.* Boston, MA: Harvard University.

Ferguson, R.F., & Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T.J. Kane, K.A. Kerr, & R.C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 98–143). San Francisco, CA: John Wiley & Sons.

Funder, D.C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652–670.

Gitomer, D.H., & Bell, C.A. (2013). Evaluating teaching and teachers. In K.F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology* (pp. 415–444). Washington, DC: American Psychological Association.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B.K., & Pianta, R.C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record, 116*(6), 1–32.

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.

Göllner, R., Wagner, W., Rose, N., Fauth, B., & Nagengast, B. (2018). Students' perceptions of teaching quality in mathematics: An integrated data analysis of five large-scale assessments. Manuscript submitted for publication.

Graumann, C.F. (1960). *Grundlagen einer Phänomenologie und Psychologie der Perspektivität.* Berlin: de Gruyter.

Hamre, B.K. (2008). *My areas of strength and growth.* Unpublished manuscript, University of Virginia, Charlottesville, VA.

Hattie, J. (2009). *Visible learning.* New York: Routledge.

Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology, 106,* 289–300.

Kane, T.J., Staiger, D.O., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., Kerr, K., Kawakita, T., & Parker, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Kaufman, J.H., Stein, M.K., & Junker, B. (2016). Factors associated with alignment between teacher survey reports and classroom observation ratings of mathematics instruction. *The Elementary School Journal, 116,* 339–364.

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

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

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

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

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

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.

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

Mayr, J., Eder, F., Fartacek, W., Lenske, G., & Pflanzl, B. (2013). *Linzer Diagnosebogen zur Klassenführung. Version LDK-P-WP.* Alpen-Adria-Universität Klagenfurt.

Oser, F., Curcio, G. P., & Düggeli, A. (2007). Kompetenzmessung in der Lehrerbildung als Notwendigkeit – Fragen und Zugänge. *Beiträge zur Lehrerinnen- und Lehrerbildung, 25,* 14–26.

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS).* Baltimore: Paul H. Brookes.

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

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.

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.

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.

Schweig, J. D. (2016). Moving beyond means: Revealing features of the learning environment by investigating the consensus among student ratings. *Learning Environments Research, 19,* 441–462.

Sirotnik, K. A. (1980). Psychometric implications of the unit-of-analysis problem. *Journal of Educational Measurement, 17*(4), 245–282.

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.

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.

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.

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.

Wettstein, A., Ramseier, E., Scherzinger, M., & Gasser, L. (2016). Unterrichtsstörungen aus Lehrer- und Schülersicht. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 48,* 171–183.

Zusammenfassung: Die Urteile von Schüler*innen, Lehrkräften und externen Beobachter*innen zur Unterrichtsqualität stimmen häufig nur in geringem Maße überein. In dem vorliegenden konzeptuellen Beitrag geben wir einen Überblick über bisherige empirische Befunde zu Perspektivenvergleichen und versuchen diese Befunde durch eine Analyse der eingesetzten Erhebungsinstrumente zu erklären. Darauf aufbauend skizzieren wir eine Referenten-Perspektiven-Matrix zur Klassifikation existierender Fragebogen- und Beobachtungsitems. Diese kann zur Erklärung der perspektivenspezifischen Mechanismen beitragen, welche der Beantwortung von Items zugrunde liegen. Die vorgestellte Matrix bietet damit auch eine Grundlage für künftige Arbeiten zur Erfassung von Unterrichtsqualität.

Schlagworte: Unterrichtsqualität, Messung, Validität, Klassenführung, Perspektivenvergleich

Contact

Prof. Dr. Benjamin Fauth, Institut für Bildungsanalysen Baden-Württemberg (IBBW),
Heilbronner Str. 172, 70191 Stuttgart, Germany
E-Mail: benjamin.fauth@ibbw.kv.bwl.de

Prof. Dr. Richard Göllner, University of Tübingen,
Hector Research Institute of Education Sciences and Psychology,
Europastraße 6, 72072 Tübingen, Germany
E-Mail: richard.goellner@uni-tuebingen.de

Prof. Dr. Gerlinde Lenske, University of Koblenz-Landau,
Institute of Primary Education,
Fortstraße 7, 76829 Landau, Germany
E-Mail: lenske@uni-landau.de

Prof. Dr. Anna-Katharina Praetorius, University of Zurich,
Institute of Education,
Freiestrasse 36, 8032 Zurich, Switzerland
E-Mail: anna.praetorius@ife.uzh.ch

Dr. Wolfgang Wagner, University of Tübingen,
Hector Research Institute of Education Sciences and Psychology,
Europastraße 6, 72072 Tübingen, Germany
E-Mail: wolfgang.wagner@uni-tuebingen.de

Richard Göllner/Benjamin Fauth/Gerlinde Lenske/Anna-Katharina Praetorius/ Wolfgang Wagner

Do Student Ratings of Classroom Management Tell us More About Teachers or About Classroom Composition?

Abstract: The present study investigated whether the varying referents (i. e., teacher or student referent) of student ratings of established classroom management measures differ in their associations with compositional classroom characteristics and students' math achievement. Re-analysis of a large-scale dataset (PISA 2003) showed that classrooms with a higher proportion of male students, as well as those with lower math performance, exhibited lower scores on classroom management factors referring more to students than the teacher. These were in turn related to lower pre-adjusted math achievement of students. There were no associations with a measure referring to the teacher. Our results indicate that varying referents tap into different aspects of the classroom management process.

Keywords: Classroom Management, Student Ratings, Indicator References, Classroom Composition, Math Achievement

1. Introduction

In the educational literature focused on effective teaching, no other aspect of teaching quality receives as much attention as classroom management. Classroom management is a central element in most theories of good teaching, and has been shown to be a consistent predictor of students' learning and development (e. g., Creemers, Kyrikides, & Antoniou, 2013; Hamre & Pianta, 2010; Hattie, 2009).

Oftentimes, student ratings are used to assess classroom management, with students asked about their teachers' ability to provide clear and consistent behavioral expectations, monitor the classroom for potential problems, and spend a minimal amount of time on behavior management issues (e. g., Aldrup, Klusmann, Lüdtke, Göllner, & Trautwein, 2018; Kunter & Baumert, 2006; Wagner et al., 2016). Other indicators of high-quality classroom management that are used in empirical studies are the number of disruptions by students, and the extent to which classroom time is used for learning-related purposes.

All of these constructs of the classroom management domain address quite different aspects of classroom management. Some of the constructs, for instance, address teachers' managing actions (e. g., monitoring, or the establishment of rules and classroom procedures), whereas other constructs seem to tap more into the consequences of teachers' classroom management (e. g., number of disruptions). At the same time, frequently

used constructs differ as to which referent of actions they address. Whereas some constructs of the classroom management domain focus directly on teacher behavior (and as such may be assessed solely by items with a teacher referent), others are more related to students in classrooms, or to the interplay between the teacher and students (see also Fauth, Göllner, Lenske, Praetorius & Wagner, in this issue).

In the present article, we examine whether such differences in referent (i. e., teacher vs. students) have an impact on the assessment of classroom management. We argue that shifting the reference from the teacher to the students makes classroom management measures (more strongly) dependent on classroom student composition. Thus, we measure not only the quality of a certain teacher's classroom management, but also the characteristics of classroom students being taught by this teacher.

We begin by summarizing past research on teachers' classroom management, as well as findings using student ratings of classroom management. We then present the results of an empirical study investigating three sub-dimensions of classroom management (i. e., monitoring, the absence of disturbances, and effective time use), examining to what extent measures with varying referents are associated with student characteristics at the classroom level. On the basis of the assumption that there is a higher reliance on measures referring to students in compositional classroom characteristics, finally we investigate the associations between measures of classroom management and students' math achievement.

2. Conceptualizing Classroom Management

Several aspects of teaching quality are currently seen as important for students' learning. Existing conceptions contain numerous factors that are relevant in describing the complex nature of teaching quality (Creemers et al., 2013; Hamre & Pianta, 2010; Helmke, 2010; Klieme, Pauli & Reusser, 2009; Kunter & Baumert, 2006). Classroom management is seen as a central element of good teaching, and has an important place in many conceptualizations of teaching quality. Related measures, such as a lack of student misbehavior and effective management of time and classroom routines, have been found to be consistently related to students' learning outcomes (e. g., Aldrup et al., 2018; Kunter et al., 2013; Wagner et al., 2016).

Various conceptualizations and measurement instruments cover a variety of aspects of the construct of classroom management. That is, classroom management is in modern conceptualizations seen as a hierarchical structure, consisting of a broad quality domain encompassing various key aspects, such as the absence of disturbances, effective time use, the existence of classroom rules, and classroom monitoring (e. g., Hamre & Pianta, 2010). A majority of studies have used student ratings to assess teachers' classroom management (e. g., Aldrup et al., 2018; Kunter & Baumert, 2006; Wagner et al., 2016). A quick glance at existing measures, however, shows that student ratings differ not only with respect to the specific quality aspect (i. e., monitoring, time use, disturbance prevention) but also with respect to the reference object. In some instances, operationaliza-

tions clearly refer to the teacher (e. g., "Our math teacher always knows exactly what is happening in class"; Baumert, Gruehn, Heyn, Köller & Schnabel, 1997, p. 85), whereas other measures are vaguer ("Our teacher has to wait a long time before it gets quiet"; Bos, Gröhlich, Dudas, Guill & Scharenberg, 2010, p. 163) or refer more to students and the interplay between the teacher and students ("In math class, the lesson is often disrupted"; Ramm et al., 2006, p. 191).

The use of different referents usually goes along with changes in the construct to be assessed. Indeed, current assessments of teaching quality, and interpretations of study findings, typically take a teacher-oriented perspective. The main question they address is whether teachers are equipped with the abilities and skills they need to manage the classroom in a productive and effective way. However, the organization and management of students' behavior, time, and attention in the classroom can also be seen from an ecological perspective. According to Doyle (2013), classrooms are quite complex systems of individuals. Thus, the assumption that students' behavior in the classroom is almost completely dependent on their teacher's classroom management ability is an oversimplification. Rather, teachers and students both contribute to classroom interactions, and thus, teachers and students are jointly enacted in the classroom management process (Doyle, 2013). It can be argued that this process finds expression the more that indicators refer to students rather than teachers. In addition, one further reason for varying referents in the classroom management indicators used may lie in the distinction between teacher's classroom management actions and their successful realization during a regular class. Whereas indicators with an explicit reference to the teacher focus more on the teacher's management actions aimed to achieve effective classroom management (e. g., monitoring, structure, or rule setting), indicators referring to the students (e. g., disturbances or time use) provide more information as to whether this objective is actually achieved in a classroom. Existing indicators may reflect operationalizations from different theoretical perspectives in a spectrum ranging from the strong focus on behavioral operations on the part of teachers (e. g., Landrum & Kauffman, 2006) to the achievement of an effective classroom management in a class comprising students with specific needs and learning requirements (e. g., Gettinger & Kohler, 2006).

3. Classroom Management and Classroom Student Composition

The idea that classroom management measures vary in the extent to which they refer more to the teacher or to the students in a classroom, raises the question of what classroom characteristics, in terms of student composition, are included in these measures. In general, classroom composition characteristics such as students' SES or academic performance are prominent candidates, and have been shown to be relevant predictors of students' learning over and above their individual learning backgrounds. Consequently, a more favorable classroom composition with higher student SES and higher performance should lead to higher learning achievement because effective teaching is easier to implement and students are equipped with academically-oriented social networks that

facilitate the spread of academic norms and higher academic aspirations (e. g., Harker & Tymms, 2004).

Guided by this perspective, and given the fact that classroom management is revealed to be a consistent predictor of student learning, measures of classroom management that refer to students might represent an important means of understanding the effects of class composition. In other words, using classroom management measures that refer to students makes it possible to test that the repeatedly reported associations between classroom management and students' learning outcomes are at least partly due to classroom composition. In fact, previous research has indicated that ratings of classroom management referring to students make measures of classroom management more dependent on class composition. For instance, a study by Praetorius, Vieluf, Saß, Bernholt, and Klieme (2016) using a measure that mainly refers to discipline problems, showed that student ratings were highly consistent across two subjects (German, and English as a foreign language). This was also the case if subjects were taught by two different teachers. Furthermore, research on observational data has shown that school and classroom composition with respect to achievement, gender, and migration background, is systematically related to teaching quality ratings (Campbell & Ronfeldt, 2018). However, it remains open whether these findings might have arisen from the use of measurement indicators that refer more to students rather than the teacher.

4. The Present Study

In this study, we used data on student ratings of classroom management, including measures with different referents, and examined whether these exhibited different associations with classroom composition and students' math achievement. We addressed two research questions: First, we investigated whether student ratings of multiple aspects of classroom management (monitoring, the absence of disturbances, and effective time use) were associated with class composition in terms of math performance, gender, and socioeconomic background. We expected that the associations with compositional characteristics would be more pronounced for ratings of classroom management referring to students or the interplay between teacher and students, than for ratings that clearly refer to the teacher. Second, we tested whether the measures differed in their prediction of students' math achievement. Assuming that classroom management measures that refer more to the students are associated with class compositional characteristics, which in turn could be shown to be related to students' learning, we hypothesized that the measures referring to students would be more predictive of students' pre-adjusted math achievement.

5. Method

5.1. Sample

We used data from the German extension of the 2003 cycle of the Programme for International Student Assessment (PISA; Organisation for Economic Co-operation and Development 2004).[1] In this national extension, a subsample of 15-year-old PISA students in Grade 9 and their teachers, took part in an additional longitudinal study, with reassessment in Grade 10. Student assessment took place in the second half of the year in Grade 9, and one year later in Grade 10. Participation in the study was voluntary. Students in PISA classes were administered achievement tests as well as questionnaires concerning background data and aspects of teaching quality in math lessons. We used the sample of N = 4,645 students from intermediate and academic track schools (K = 259 classes).[2] On average, 12 students per class provided data on their math teachers' teaching quality. Due to time constraints, PISA used a matrix design, where half of the students in each class were chosen to complete one set of items while the remaining students answered a different set of items. In the present study, teaching quality perception measures from N = 2,508 students were able to be used.

5.2 Instruments

Classroom management. We used three well-known measures of classroom management: teachers' monitoring activity, the absence of disturbances, and efficient time use. All measures were drawn from the Grade 10 measurement point. The indicators for monitoring refer to the teacher, while the indicators for disturbances and ineffective time use refer more to the students (see Table 1 for a complete set of items). All responses were given on a 4-point Likert scale ranging from 1 (completely disagree) to 4 (completely agree). *Monitoring* assesses the extent to which the teacher keeps an eye on students' actions and is alert to any behavioral problems or learning difficulties. The construct was operationalized using four items (e. g., "The teacher always knows exactly what is going on in class"). Scale reliability was α = .71. *The absence of disturbances* was assessed with two items referring to difficulties in maintaining discipline in the classroom (e. g., "In math, the lesson is often disturbed"). The scale was re-

---

1 We thank the German PISA consortium (Prenzel et al., 2007; Prenzel et al., 2013) and the Research Data Centre (FDZ) at the IQB in Berlin for their approval and support in conducting the secondary analysis.

2 A "tripartite" system of lower track schools (Hauptschule), intermediate track schools (Realschule), and academic track schools (Gymnasium) is the most common system in German states; some states offer multitrack schools that serve lower and intermediate track students in joint classes. The present study sample consisted only of students from intermediate track schools and academic track schools, because lower track students finish school at the end of Grade 9.

verse-coded so that higher values indicated higher classroom management. Finally, *effective time use* captured the amount of time lost through disciplinary problems in class via two items (e. g., "In math, it is long after the lesson starts that students become quiet and start working"). This scale was also reverse-coded, so that higher values would indicate higher classroom management. Descriptive statistics of the classroom management indicators are given in Table 1. For all indicators, a substantial degree of variance could be attributed to the classroom level (ICC1) and assure a reliable assessment of classroom management at the classroom level (ICC2).

Students' mathematics achievement. To measure students' mathematics achievement, we used the standardized math achievement test scores that were applied in prior teaching quality research with the PISA dataset (see Kunter et al., 2013 for a detailed description). As we focused on the effect of classroom management on students' learning gains, we also included students' prior math performance at Grade 9, which was measured as part of the international PISA 2003 assessment. Students' mathematics achievement at the end of Grade 10 was assessed with a test covering standard content from the federal states' curricula for Grade 10 mathematics. All tests were scaled using Rasch analysis. Test items had a closed response format, and subsets of test items were administered us-

Quality dimension		*M*	*SD*	*ICC(1)*	*ICC(2)*
Monitoring (Cronbach's α = .73)					
Mo1	My teacher always knows exactly what is going on in the class	2.52	0.97	.26	.81
Mo2	My teacher always checks our homework very accurately	2.23	0.98	.29	.83
Mo3	My teacher makes sure that we pay attention	2.83	0.92	.23	.78
Mo4	My teacher immediately notices when students start doing something else	2.61	0.96	.20	.75
Absence of disturbances (Cronbach's α = .81)					
AD1	In math class, the lesson is often disturbed (recoded)	2.43	0.99	.32	.85
AD2	In math, a lot of nonsense is going on all the time (recoded)	2.63	1.02	.31	.84
Effective time use (Cronbach's α = .74)					
ET1	In math, it is long after the beginning of the hour by the time the students get quiet and start working (recoded)	2.62	1.00	.29	.83
ET2	In math, a lot of time in class is wasted (recoded)	2.66	1.01	.26	.80

Note. The scale for each item was 1 to 4.

Tab. 1: Summary of item indicators for students' ratings of classroom management

ing a multi-matrix design. Item and person parameters for students' math achievement were estimated, and the weighted likelihood estimates were used as person parameters for individuals' math achievement in Grades 9 and 10. Again, students' math achievement scores revealed a substantial amount of variance at the classroom level, both for the Grade 9 scores (ICC1 = .39, ICC2 = .89) and for the Grade 10 scores (ICC1 = .39, ICC2 = .89).

Students' background. In addition to student ratings of classroom management and math achievement in Grade 9, we used students' gender (0 = female, 1 = male) and socioeconomic background (SES) as further measures to assess their learning background. The social status of the students' families was operationalized by the International Socio-Economic Index, which was developed by Ganzeboom and Treiman (2003) on the basis of the International Labour Office's occupation classification system. The score from the parent with the highest index ranking was used in the analyses. At the classroom level, we also included the school track in which students were enrolled (non-academic or academic track). Students from the non-academic track served as reference.

5.3 Statistical Analysis

Preliminary analysis. Based on a two-level first-order factor model that contained correlated first-order factors corresponding to the dimensions of monitoring, effective time use, and absence of disturbances at both levels (within and between classes), we began by checking whether student ratings reflected multiple dimensions of classroom management. Even though a model with three factors exhibited a good model fit, $\chi^2(34) = 195.62$, $p < .001$; CFI = .97; TLI = .95; RMSEA = .04; $SRMR_{within} = .04$; $SRMR_{between} = .04$, scaling correction factor = 1.11, a nearly perfect correlation was revealed between absence of disturbances and effective time use, both at the within- ($r = .94$) and the between-classroom levels ($r = .97$). In contrast, correlations between the absence of disturbances and effective time use with monitoring were substantially smaller (within level: $.31 \leq r \leq .32$; between level: $.81 \leq r \leq .85$). For this reason, we tested a second model, combining the absence of disturbances and effective time use, resulting in a two-factor model. This model's fit was similarly good to the three-factor model, $\chi^2(38) = 212.45$, $p < .001$; CFI = .97; TLI = .95; RMSEA = .04; $SRMR_{within} = .04$; $SRMR_{between} = .04$, scaling correction factor = 1.13. Thus, in the interest of parsimony we decided to combine the two factors in all subsequent analyses (see Table 2).

Nested factor model. In order to address our research questions, we then proceeded to examine a nested factor model. A nested factor model (or bifactor model) assumes the existence of a general factor that directly influences all observed measurement indicators and one or more additional components that account for different specific subsets of indicators (Gustafsson & Åberg-Bengtsson, 2010). In order to deal efficiently with the different referents in the measurement indicators, we used all of the different indicators

	Student level				Classroom level			
	3 factors		2 factors		3 factors		2 factors	
Quality dimension	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***
Monitoring								
Mo1	0.59	0.02	0.59	0.02	0.86	0.03	0.86	0.03
Mo2	0.43	0.02	0.43	0.02	0.49	0.06	0.49	0.06
Mo3	0.67	0.02	0.67	0.02	0.91	0.03	0.91	0.03
Mo4	0.68	0.02	0.68	0.02	0.96	0.02	0.96	0.02
Absence of disturbances								
AD1	0.74	0.02	0.73	0.02	0.99	0.01	0.99	0.01
AD2	0.76	0.02	0.75	0.02	1.00	0.01	1.00	0.01
Effective time use								
ET1	0.73	0.02	0.71	0.02	0.99	0.01	0.96	0.01
ET2	0.64	0.02	0.63	0.02	0.91	0.02	0.89	0.02

Note. [a] Factor loadings are shown in their standardized form.

Tab. 2: Factor loadings of first-order factor multilevel models with two and three factors

for monitoring to identify the general factor, thus controlling for differences in monitoring between teachers/classes and for specific references to the teacher in all of the remaining classroom indicators (see the measurement model in Fig. 1).

Here, the specific factors represent classroom management aspects that vary between students within classes (student level) and between classes (classroom level) with identical levels of perceived monitoring.

Classroom compositional effects on ratings of classroom management. As we were interested in whether classroom management measures were associated with classroom student composition, we further conducted multilevel regression analyses. The central question in the analysis of composition effects is whether an aggregated classroom characteristic is associated with an outcome measure after controlling for individual differences among students on that characteristic (Kreft, de Leeuw & Aiken, 1995). Based on the nested factor model, we included classroom composition characteristics as predictors of the general and the specific classroom management factors (see the covariate model in Figure 1), using an approach proposed by Koch, Holtmann, Bohn and Eid (2018). Specifically, we used their residual approach, in which either the general or the specific factors are partialed out from the covariates and are then used as an independent variable in an additional model to predict the general or specific factors. This procedure ensures that the implied model assumption of zero correlations between the general and

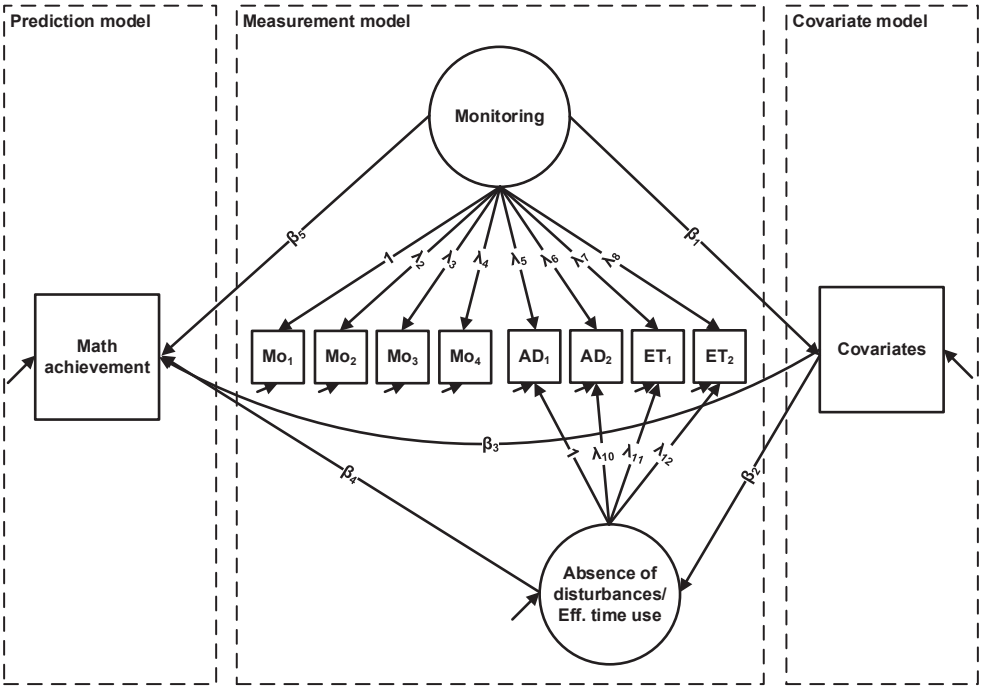

Fig. 1: Nested factor model representing a general (monitoring, Mo) and a specific (absence of disturbances, AD; effective time use, ET) classroom management factor. Covariates of the general and specific factors were included using an approach proposed by Koch et al. (2018). Models were estimated simultaneously at the student and classroom levels. Covariates (math performance, SES, and gender) at the classroom level were modeled by manifest aggregation. School track was used only at the classroom level.

specific factors is compatible with testing these same predictors on the general and the specific factors (see Koch et al., 2018 for more detail).

Effects of classroom management on student achievement. Finally, we extended the analytical model by including student math achievement at Grade 10 in the multilevel analysis (see the prediction model in Figure 1) and tested whether student ratings of classroom management were associated with students' math achievement at Grade 10 after controlling for students' prior math achievement at Grade 9.

One important prerequisite for testing such effects with latent measures is the cross-level invariance of measures. Otherwise, the comparison of level-specific associations may not be meaningful, as the factor variances of the same set of indicators at the two levels – and thus, also the corresponding regression coefficients – are not comparable. Consequently, we constrained the measurement models (i. e., factor loadings) to be equal across levels (see Wagner, Göllner, Helmke, Trautwein & Lüdtke, 2013). For nested factor models, this procedure is somewhat more complicated as the differ-

ences between factor intercorrelations at the different levels in a simple first-order factor model (i. e. the structural model) are represented by different loadings on the general factor (i. e. the measurement model) for indicators that also load on specific factors. In other words, some parts of the measurement model (i. e., loadings on the general factor from indicators that also loaded on specific factors) in the nested factor model reflected differences in the structural model. For this reason, we tested only for partial measurement invariance, by applying equality constraints to the monitoring indicators' loadings on the general factor, and effective time use and absence of disturbance loadings on the specific factors.

We conducted all analyses using the Mplus 7.3 software (Muthén & Muthén, 1998–2012). Robust maximum likelihood estimation (MLR) for continuous data was used to obtain reliable standard errors and fit tests for non-normally distributed data. In addition, we utilized full information maximum likelihood (FIML) estimation, where model variables are used to predict missing data. All continuous variables were z-standardized ($M = 0$, $SD = 1$) before analysis.

6. Results

6.1. Nested Factor Model

To address our research questions, we applied a nested factor model with one general classroom management factor and one specific factor, with factor loadings on absence of disturbances/effective time use (see Figure 1). All descriptive fit indices indicated a good model fit, $\chi^2(33) = 140.80$, $p < .001$; CFI = .98; TLI = .97; RMSEA = .03; $SRMR_{within} = .02$; $SRMR_{between} = .04$, scaling correction factor = 1.07. Constraining the factor loadings for monitoring (general factor) and for absence of disturbances/effective time use (specific factor) led to a highly comparable model fit, $\chi^2(39) = 144.23$, $p < .001$; CFI = .98; TLI = .97; RMSEA = .03; $SRMR_{within} = .02$; $SRMR_{between} = .04$; scaling correction factor = 1.11, indicating measurement equivalence across levels.

6.2 Compositional Classroom Effects

We then examined the association between class composition and classroom management factors (research question 1). Specifically, we included students' gender, math performance at Grade 9, and SES into the multilevel regression model and examined the associations both at the within- and the between- classroom levels, between students' background and their ratings of classroom management. In addition, we controlled for potential school track differences at the classroom level. The results are shown in Table 3 and can be summarized as follows: First, the results at the within level revealed no statistically significant associations between students' background and classroom management. Neither students' math performance, SES, nor gender was associated with class-

	Monitoring				Absence of disturbances/effective time use[c]			
	Student level		Classroom level		Student level		Classroom level	
Variables	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***
Math performance	−0.02	0.03	−0.06	0.13	0.02	0.02	0.15	0.06*
SES	−0.03	0.04	−0.03	0.09	0.01	0.02	−0.06	0.08
Gender[a]	−0.04	0.02	−0.05	0.03	0.02	0.03	−0.36	0.18*
School track[b]			−0.13	0.09			0.01	0.08

Note. [a] Gender (male); [b] School track (Gymnasium). [c] Based on the modeling approach proposed by Koch et al. (2018), predictors and outcomes were residualized by monitoring. Regression coefficients are shown in their unstandardized form.
* $p < .05$

Tab. 3: Multilevel regression models predicting monitoring and disturbances with compositional classroom characteristics

room management factors. At the classroom level, two findings are noteworthy. First, all of the classroom student characteristics were unrelated to monitoring ($p > .05$; see Table 3). Second, the results for the specific factor of absence of disturbances/effective time use revealed two statistically significant findings. Classrooms with higher math performance at Grade 9 ($b = 0.15$, SE = 0.06, $p = .015$) and a higher proportion of female students ($b = −0.36$, $SE = 0.18$, $p = .047$) reported greater absence of disturbances and more effective time use. School track was unrelated to the two classroom management factors (monitoring: $b = −0.13$, $SE = 0.09$, $p = .155$; absence of disturbances/effective time use: $b = 0.01$, $SE = 0.08$, $p = .891$). In order to test whether these associations also held after controlling for individual differences among students (i. e., compositional effects), we compared the level-specific coefficients. The results showed that the association between absence of disturbances/effective time use (math performance: $b = 0.13$, $SE = 0.06$, $p = .047$; gender: $b = −0.38$, $SE = 0.18$, $p = .037$) remained statistically significant, even after accounting for individual differences among students.

6.3 Associations with Math Achievement

Finally, for our second research question we examined the association between classroom management factors and student achievement at Grade 10. To do this, we additionally regressed students' math achievement at Grade 10 on the classroom factors and controlled for students' gender, school track differences, SES, and math performance at Grade 9. The complete model including the classroom management factors, predictor variables and achievement outcomes elicited a good model fit, $\chi^2(95) = 265.69$, $p < .001$; CFI = .98; TLI = .97; RMSEA = .02; $SRMR_{within} = .02$; $SRMR_{between} = .05$, scaling cor-

rection factor = 1.01. The associations found between student background variables and classroom factors remained unchanged at both the student and classroom levels. In addition, monitoring and absence of disturbances/effective time use were not associated with students' achievement at the within-classroom level (monitoring: $b = 0.02$, $SE = 0.03$, $p = .436$; absence of disturbances/effective time use: $b = -0.01$, $SE = 0.03$, $p = .821$).

However, the prediction results at the classroom level revealed a significant association between absence of disturbances/effective time use and achievement at Grade 10 ($b = 0.20$, $SE = 0.08$, $p = .008$). Monitoring, on the contrary, was not associated with math achievement ($b = 0.02$, $SE = 0.04$, $p = .643$). The results for absence of disturbances/effective time use also held after controlling for individual differences among students ($b = 0.21$, $SE = 0.08$, $p = .014$).

In sum, these findings demonstrate that classroom management was associated with students' achievement, although only those measures referring to the students showed statistically significant results.

7. Discussion

In the present study, we examined student ratings of classroom management and potential differences between different measures. We tested whether classroom management measures, according to the extent to which they referred more to the teacher or to the students in a classroom, differed in their associations with classroom student composition and students' achievement in mathematics.

Our results show that the way students are specifically asked about classroom management is of direct relevance to which aspect of the classroom management process is being addressed. Classroom management indicators that do not directly address the teacher as the referent of actions, yield information about students as an equally important part of the classroom management process. Specifically, the results revealed associations between classroom student composition and students' ratings of classroom management, but only for measures that referred to the students or to the interplay between teacher and students (absence of disturbances and effective time use). The same pattern of results was also found for the prediction of students' pretest-adjusted math achievement at Grade 10. Whereas the absence of disturbances and effective time use were associated with students' pretest-adjusted math achievement, teachers' monitoring exhibited a non-significant effect.

In sum, our results add to recent literature showing that classroom management indicators referring more to students in the class, correspondingly capture information about the composition of the class, and thus, tap into a conceptually different aspect of the classroom management process (Fauth et al., see in this special issue).

7.1. Classroom Management and Classroom Student Composition

Classroom management is known to be one of the most consistent predictors of students' learning, and is frequently assessed via student ratings. In this study, we compared three measures of classroom management that differed not only in the aspect of classroom management addressed, but also in the extent to which they referred more to the teacher or to the students in a given class. Consistently with prior research, the results of factor analysis showed that students are able to distinguish different aspects of classroom management. However, students' perceptions were much less differentiated when measures referred to students. In fact, the association between the absence of disturbances and effective time use was nearly perfect, indicating that the concordance of measures assessing the same underlying domain also depends on the referent to which the measures refer. This may sound trivial but this point is largely ignored in teaching quality research, where different aspects of classroom management such as the absence of disturbances and monitoring are often used more in the sense of interchangeable domain indicators. In addition, our results showed that these two sets of measures were differently related to classroom student composition. Teachers who taught lower-performing students and a higher proportion of male students received lower ratings on classroom management measures referring to students, but not on a measure referring to the teacher.

These differences are of high practical and theoretical importance, as they show that compositional effects cannot be explained purely by a general bias in students' perceptions, as compositional effects depend on which measures of classroom management are used. Rather, we believe that measures referring more to students than to the teacher provide information on classroom management from a perspective that combines both teachers' abilities and class characteristics. That is, lower or higher ratings on indicators referring to the students cannot simply be equated with the teacher's ability, but need to be considered from an interactionist perspective that takes into account both teachers and the students taught.

7.2 Classroom Management and Student Achievement

Furthermore, our results showed that the examined measures of classroom management were differently related to students' later achievement. Whereas factors referring to students were related to students' pretest-adjusted class achievement, teachers' monitoring did not exhibit a statistically significant effect. Thus, the findings of the present study provide important insights into the frequently observed association between classroom student composition and students' learning. Over the last two decades, empirical studies have shown that a favorable classroom composition provides important benefits above and beyond students' individual learning backgrounds (e. g., Harker & Tymms, 2004). This research suggests that being part of higher-achieving classrooms leads to greater learning, even after controlling for students' personal characteristics (e. g., learning ca-

pabilities and parents' educational background). The present study makes a strong contribution to this research. At the same time, however, these findings also raise the question to what extent the much-reported associations between classroom management and student learning are merely due to classroom composition effects.

7.3 Limitations and Future Research

In sum, our results raise questions about the differences between frequently-used classroom management measures assessed via student ratings. However, the study has important limitations that should be addressed in future research. First, the main aim of our study was to compare classroom management measures with varying referents. Thus, we used classroom management measures that differed in respect of whether or not they referred to the teacher. Even though the study was based on frequently used and well-known aspects of classroom management, the measures used do not allow us to separate the referent from the content of the classroom management constructs examined. For instance, the reason that classroom student composition was related to the absence of disturbances/effective time use but not to monitoring, may be due to content-related differences between measures, instead of the varying referents between measures.

A second limitation refers to the conceptual differences underpinning classroom management measures with varying referents. The present findings indicate that existing indicators tap into theoretically distinct aspects of the classroom management process, ranging from the assessment of concrete behavioral operations and diagnostic aspects on the side of the teacher, to the effectiveness of these operations, given the specific characteristics of the class being taught. Future research should systematically examine to what extent the referent of measures reflects conceptually different aspects of the classroom management process, including teachers' classroom management-related diagnostic abilities, their behaviors, and students' responses in a specific class. Particularly productive would be a longitudinal study using measures with varying referents and applying an interactional perspective, to explain how well teachers' management behavior meets the specific requirements of the class.

Third, considering a larger number of classroom management measures is also important for another reason. Although monitoring reflects a prototypical measure of a teacher's classroom management, other teacher-directed measures of classroom management were not able to be examined in the present study. For this reason, the inclusion of additional measures with a teacher referent (e. g., rule setting) would further clarify potentially confounding effects of teacher-directed measures on measures that are not explicitly teacher-referenced.

Finally, our results relied solely on student ratings of classroom management. The findings showed that students' views on classroom management reflected theoretically-assumed differences, in terms of the underlying management processes, but we were not able to include alternative methods of assessment. Thus, it remains an open question as to whether similar findings would also result from observation ratings or from teacher

self-reports (see Campbell & Ronfeldt, 2018). In addition, it has to be borne in mind that the sole reliance on student ratings has potential to impede the accessibility of teachers' monitoring, leading to non-statistically significant results. As pointed out by Fauth and colleagues (submitted), judgments about teacher-directed behavior place higher cognitive demands on the judgment process of students than do judgments about their own behavior. In the same vein, the use of alternative methods of assessment would also help to clarify whether the high correlations between measures referring to students are a result of using the same referent or rather, of students' lower ability to differentiate between theoretically distinct dimensions of classroom management.

In conclusion, the present study has shown that the way students are asked about classroom management can make a difference to the extent to which such measures provide information about teachers and students in the classroom. The results point to the complex nature of the classroom management process, which involves both teachers and students within a particular class.

References

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*(8), 1066–1083.

Baumert, J., Gruehn, S., Heyn, S. Köller, O., & Schnabel, K. U. (1997). *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU) – Dokumentation (Band 1) [Educational trajectories and psychosocial development during adolescents (BIJU) – Documentation (Issue 1)]*. Berlin: Max-Planck-Institut für Bildungsforschung.

Bos, W., Gröhlich, C., Dudas, D.-F., Guill, K., & Scharenberg, K. (2010). *KESS 8 – Skalenhandbuch zur Dokumentation der Erhebungsinstrumente [KESS 8 – Codebook to documenting study measures]*. Münster: Waxmann.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*(6), 1233–1267.

Creemers, B. P. M., Kyriakides, L., & Antoniou, P. (2013). Establishing theoretical frameworks to describe teacher effectiveness. In B. P. M. Creemers, L. Kyriakides & P. Antoniou (Eds.), *Teacher professional development for improving quality of teaching* (pp. 101–135). Dordrecht, Netherlands: Springer.

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

Ganzeboom, H. B. G., & Treiman, D. J. (2003). Three internationally standardized measures for comparative research on occupational status. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in cross-national comparison: A European working book for demographic and socio-economic variables* (pp. 159–193). New York: Kluwer Academic/Plenum.

Gettinger, M., & Kohler, K. (2006). Process-outcome approaches to classroom management and effective teaching. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 73–96). Mahwah, NJ: Erlbaum.

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97–121). Washington, DC: American Psychological Association.

Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement, 15,* 177–199.

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

Helmke, A. (2010). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts [Instructional quality and teacher professionalism. Diagnosis, evaluation, and improvement of instruction].* Seelze, Germany: Klett-Kallmeyer.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–21.

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

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study: Investigating the 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). Münster: Waxmann.

Koch, T., Holtmann, J., Bohn, J., & Eid, M. (2018). Explaining general and specific factors in longitudinal, multimethod, and bifactor models: Some caveats and recommendations. *Psychological Methods, 23,* 505–523.

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

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

Landrum, T. J., & Kauffman, J. M. (2006). Behavioural approaches to classroom management. In C. M. Everton & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 47–72). Mahwah, NJ: Erlbaum.

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

Organisation for Economic Co-operation and Development. (2004). *Learning for tomorrow's world: First results from PISA 2003.* Paris: OECD.

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

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rost, J., & Schiefele, U. (2007). *PISA 2003 [Version 1, Code book and dataset].* Berlin: Institute for Educational Quality Improvement.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rost, J., & Schiefele, U. (2013). *PISA-I-Plus 2003, 2004 [Version 1, Code book and dataset].* Berlin: Institute for Educational Quality Improvement.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U. (2006). *PISA 2003. Dokumentation der Erhebungsinstrumente [PISA 2003. Documentation of the study instruments].* Münster: Waxmann.

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.

Zusammenfassung: In der vorliegenden Studie wurden die Konsequenzen variierender Referentenbezüge (Lehrkraft vs. Schülerinnen und Schüler) bei etablierten Skalen zur Erfassung der Klassenführung aus Schülersicht untersucht. Die Ergebnisse einer Reanalyse der PISA 2003-Daten zeigten, dass Klassen mit höherem Jungenanteil und niedrigerem mittleren Leistungsniveau niedrigere Werte auf Skalen mit stärkerem Schülerbezug aufwiesen. Diese waren wiederum mit einer geringeren Leistungsentwicklung von Schülerinnen und Schülern im Fach Mathematik assoziiert. Für eine Skala mit Lehrkraftbezug fanden sich hingegen keine Zusammenhänge. Die Ergebnisse legen nahe, dass der Referentenbezug von Skalen für die erfassten Aspekte der Klassenführung entscheidend ist.

Schlagworte: Klassenführung, Schülerurteile, Itemreferenten, Klassenkomposition, Mathematikleistung

Contact

Prof. Dr. Richard Göllner, University of Tübingen,
Hector Research Institute of Education Sciences and Psychology,
Europastraße 6, 72072 Tübingen, Germany
E-Mail: richard.goellner@uni-tuebingen.de

Prof. Dr. Benjamin Fauth, Institut für Bildungsanalysen Baden-Württemberg (IBBW),
Heilbronner Str. 172, 70191 Stuttgart, Germany
E-Mail: benjamin.fauth@ibbw.kv.bwl.de

Prof. Dr. Gerlinde Lenske,
Universitätsallee 1, 21335 Lüneburg, Germany
E-Mail: gerlinde.lenske@leuphana.de

Prof. Dr. Anna-Katharina Praetorius, University of Zurich,
Institute of Education,
Freiestrasse 36, 8032 Zurich, Switzerland
E-Mail: anna.praetorius@ife.uzh.ch

Dr. Wolfgang Wagner, University of Tübingen,
Hector Research Institute of Education Sciences and Psychology,
Europastraße 6, 72072 Tübingen, Germany
E-Mail: wolfgang.wagner@uni-tuebingen.de

Marten Clausen

Commentary Regarding the Section "The Role of Different Perspectives on the Measurement of Teaching Quality"

Abstract: In this commentary, the paper by Fauth, Göllner, Lenske, Praetorius, and Wagner as well as the paper by Göllner, Fauth, Lenske, Praetorius, and Wagner published in this special issue of the Zeitschrift für Pädagogik are discussed. In the context of specificity and perspective-agreement in perceptions of teaching quality, merits and possible limitations of theoretical conceptions and empirical analyses are reviewed for both papers. In a final section, the complexity inherent to teaching research is addressed on a more general level focussing on differences between psychometric and edumetric approaches.

Keywords: Teaching Quality, Perspective Agreement, Classroom Management, Item Wording, Ratings of Teaching

1. Introduction

In this discussion, I will comment on the papers by Fauth, Göllner, Lenske, Praetorius, and Wagner (in this issue) and Göllner, Fauth, Lenske, Praetorius, and Wagner (in this issue) presented in this issue of the *Zeitschrift für Pädagogik.* From a broader perspective, both papers deal with the relation between classroom management aspects of teaching quality and their operationalization and measurement in items and scales. I will first address some more specific aspects of both papers and then will discuss more general aspects that are relevant to both papers.

2. Who Sees What? Conceptual Considerations on the Measurement of Teaching Quality from Different Perspectives

In a comparison between the different perspectives of the teacher, his students, and external observers, Clausen (2002) found evidence for a low level of agreement of teaching quality ratings. In the following years, a growing number of studies could replicate these results. Similarly to Clausen (2002), several researchers (e. g. Kunter & Baumert, 2006) have moved away from the idea of a common true score and concluded that for perceptions of instruction, perspective-specific validity should be assumed. Clausen's theoretical approach to predicting agreement as an interaction of perspectives and teaching construct characteristics (observability, demands regarding didactic understanding, evaluativeness) has had limited success in respect of empirical validation, with observability being the only dimension for which some support has been found.

Fauth et al. (in this issue) propose a similar but different approach. Instead of classifying characteristics of teaching constructs to predict agreement, they look at the item level to determine the *referent* of an item (teacher, students, both/unclear) – the perspective that determines whose behavior is focused on by an item. If a perspective is the referent of an item, Fauth et al. (in this issue) conclude this entails a higher ego involvement and more and better information. They integrate their considerations in the *perspective reference matrix* which, as a theoretical framework, allows the prediction of agreement between various combinations of item-referent and perspective. By applying this framework to reinterpret the empirical results of other studies, they find some confirmation for referent-dependent variance in the factor structures of rating instruments, and in the correlations between perspectives in agreement studies.

The paper by Fauth et al. (in this issue) offers a very valuable critical approach to the measurement of classroom management as a central aspect of the quality of teaching and to the instruments recently and currently used in empirical teaching research. Their precise focus on item level is similar to the classic facet theoretic approach to item analysis (Guttman & Greenbaum, 1998), in which the relevant item aspects are analyzed on a theoretical level to build a basic item structure (mapping sentence). The elements of the mapping sentence are then tested empirically for their effects on item inter-correlations (e. g. Alt, 2018). Lenske (2016, p. 95) provides a list of facets that may generate specific variances in ratings. While the overall approach is very convincing to me, there are two smaller points specific to the paper that caught my attention.

(1) From my view, there is no clear information advantage for the teacher in regard to the example item "Our teacher immediately notices when students start doing something else." Teachers can only remember those times when they noticed the behavior. Both the students and external observers have at least the possibility of seeing both cases: that is, including those cases where the teacher does not notice the off-task behavior.

(2) I am not convinced that for all student behavior items there is a higher ego-involvement for students compared to the ego-involvement of the teacher. Even if a construct or item focuses only on the behavior of the student, a quality of teaching approach would regard the teacher as implicitly responsible or at least to some degree in control. There is a difference between "The class does not behave in an orderly manner." and "In mathematics lessons of Mr. Smith, the class does not behave in an orderly manner". Furthermore, I assume that ego-involvement and the resulting self-serving bias have to be regarded in the context of the data collection. If a teacher and his students fill out ratings questionnaires for TIMSS or PISA, there is little chance that the answers will have direct consequences for themselves. If these assessments take place with feedback in the context of teacher evaluation or professional development, there may be a lot more weight on the answers because they are communicated and discussed; so that disparities between the ratings of the students and the view of their teacher may become obvious.

I do agree with the authors that the deviations between perspectives should be investigated. On the other hand, I am very much convinced that a common true score does not represent the social reality of classroom interaction. Idiosyncratic perspectives may

result from different logics of action. They are real, in the sense that they influence attitudes and behavior, as noted in the Thomas Theorem (Thomas & Thomas, 1928). External observers such as school inspectors transport, communicate and establish outside norms and standards into the school – they are not supposed to take the teacher's view, nor should they try to adopt a student perspective. Students should be free to state their perceptions, which may be mainly focused on being treated respectfully, being treated fairly, being entertained, and being motivated. Teachers have to develop a realistic self, and should use feedback from other perspectives as a means of professional development. To some degree they should learn to look at their teaching from a student perspective (Wettstein, Ramseier, Scherzinger & Gasser, 2016). However, this does not mean they should make the student perspective their own. To me, there is still potential to further explore social perspectives in teaching research by applying general concepts of social perspectivity (Strack, 2004).

3. Do Student Ratings of Classroom Management Tell us More About Teachers or About Classroom Composition?

The paper by Göllner et al. (in this issue) focuses on whether the difference in referent of an aspect of classroom management does generate variation that depends more on the composition of the class than on the actual teaching. In an elaborate statistical model using student ratings from PISA 2003, the authors test the assumption that aspects of classroom composition influence ratings of those aspects of classroom management that mainly refer to the students' behavior. Results indicate that teachers of classes with more low-achieving students and more male students are rated lower on aspects of classroom management that refer to the students, in contrast to measures referring to the teacher.

The paper by Göllner et al. (in this issue) has a sound theoretical background closely related to the theoretical approach of Fauth et al. (in this issue). The authors apply a methodologically sophisticated multi-level model to work out the importance of the specific referent-dependent variance of the "student referent items" for student ratings and their relations to academic achievement and classroom composition. From my point of view, the referent-dependent variance results from characteristics of general class behavior and its corresponding psychological dimension, classroom climate; both of which influence the specific behavior towards the teacher who is being rated. The classroom climate should stabilize this behavior at the class level. These results are of special importance when ratings are used in the context of teacher evaluation, because they might impair the fairness of the evaluation (Campbell & Ronfeldt, 2018).

4. General Comments

Aspects of teaching and classroom interaction are still being treated like psychological constructs, with unidimensionality and normal distribution being common assumptions. Looking at some possible "states" of classroom management along a very basic abstract continuum of "order vs. chaos", it is obvious however that the continuum is not normally distributed. (1) A class that is very difficult – perhaps because of its composition – will make teaching hard for almost any teacher. (2) A teacher with great authority will control even the most difficult class. (3) With a class of committed and intrinsically motivated students, there is almost no need for classroom management. (4) A stable balanced state in the middle of the continuum has a relatively low probability. Disturbances and disciplinary problems that are not dealt with effectively have a tendency to escalate and spread by contagion. All of the different examples described here have different referents or combinations of referents.

As Fauth et al. (in this issue) and Göllner et al. (in this issue) point out, for typical aspects of classroom management the item referent varies on the item level, and the item referent influences both the internal structure and the correlations to external criteria. On the other hand, this is also true on a construct level. In their analyses of expert sortings of teaching constructs, Clausen, Schnabel & Schröder (2002, p. 254) generated a differentiation regarding the "subject" and "object" of a construct (in brackets):

1) students → (students): e. g., student-student relations
2) students → (teacher): e. g., student engagement
3) teacher → (students): e. g., monitoring and
4) teacher → (teacher): e. g., clarity and structure.

This differentiation is not a question of item formulation, but a question of the construct. At the core of teaching research, the constructs of quality of teaching represent a dynamic social interaction of teachers and students which is also stressed in the German "Angebot-Nutzungs-Modell" by Helmke (2003; roughly translated as "offer-use model" or "utilization-of-learning-opportunities" model). From this point of view, the most relevant constructs are the interactive dimensions that focus the students' behavior towards the teacher (category 2) and the teacher's behavior towards the students (category 3). For these constructs the reference perspective matrix does not make clear predictions. Reducing constructs to a "referent differentiated" level would avoid the referent-dependent variance problem, but would also reduce the behavioral universe to the less-relevant pure teacher variables and pure student variables. To put all referents into every single item, would make the resulting items more complex and more difficult to agree to. Balancing the different referents relevant to a construct across different items of a scale would not solve the referent problem – and yet in a way, this is how the instruments discussed by Fauth et al. (in this issue) work.

Edumetrics is not psychometrics – some of the complexity is inherent to the domain of teaching research, and thus cannot be solved easily. Classroom management is too

broad and too complex a field for one to be able to expect it to be homogeneous in terms of internal structure, as well as in terms of correlations to other constructs. Both papers focus mainly on behavior and classroom management, while the more cognitive, motivational, and emotional basic dimensions of teaching play marginal roles. If the authors were able to extend the theoretical scope of the reference perspective matrix, it would gain more relevance.

Updating the theoretical and empirical approaches to teaching research and perspective agreement, and connecting them closer to the Anglo-American research, is a major merit of both papers. The new approaches still need to be further elaborated and validated. The authors of both papers are well aware of this, and many of the points discussed in this comment paper are addressed at least to some degree in these papers. The quantitative dimensional approach to describing the quality of teaching has its limitations, and the more elaborate the methodology and the design, the more obvious are these limitations. Many of the items and scales discussed in Fauth et al. (in this issue) and Göllner at al. (in this issue) have their roots in instruments from the German classroom climate research of the 1980s and 1990s (e. g. Saldern & Littig, 1987). They were created with different theoretical and empirical frameworks in mind. This new generation of researchers has developed promising ideas and they are capable of testing them in elaborate designs with state-of-the-art methodology. I am very curious to see them apply these ideas to more recent data with modern multi-perspective instruments.

References

Alt, D. (2018). Students' perceived constructivist learning environment. *European Journal of Psychological Assessment, 34*(6), 432–443.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*(6), 1233–1267.

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

Clausen, M., Schnabel, K., & Schröder, S. (2002). Konstrukte der Unterrichtsqualität im Expertenurteil. *Unterrichtswissenschaft, 30*(3), 246–260.

Guttman, R., & Greenbaum, C. W. (1998). Facet theory. *European Psychologist, 3*(1), 13–36.

Helmke, A. (2003). *Unterrichtsqualität erfassen, bewerten, verbessern.* Schulisches Qualitätsmanagement. Seelze: Kallmeyer.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*(3), 231–251.

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

Saldern, M. v., & Littig, K.-E. (1987). *Landauer Skalen zum Sozialklima: 4.–13. Klassen. LASSO 4–13.* Weinheim: Beltz.

Strack, M. (2004). *Sozialperspektivität – Theoretische Bezüge, Forschungsmethodik und wirtschaftliche Praktikabilität eines beziehungsdiagnostischen Konstrukts.* Göttingen: Universitätsverlag.

Thomas, W. I., & Thomas, D. S. (1928). *The child in America.* New York: Knopf.

Wettstein, A., Ramseier, E., Scherzinger, M., & Gasser, L. (2016). Unterrichtsstörungen aus Lehrer- und Schülersicht. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 48*(4), 171–183.

Zusammenfassung: In diesem Kommentar werden die in diesem Themenheft der Zeitschrift für Pädagogik erschienenen Beiträge von Fauth, Göllner, Lenske, Praetorius und Wagner sowie Göllner, Fauth, Lenske, Praetorius und Wagner diskutiert. Vor dem Hintergrund der Frage von Spezifität und Perspektivenübereinstimmung von Unterrichtswahrnehmungen werden Verdienste und mögliche Einschränkungen der theoretischen Ansätze und empirischen Analysen beider Beiträge reflektiert. Im abschließenden Abschnitt wird die inherente Komplexität der Unterrichtsforschung auf allgemeinerer Ebene thematisiert mit Blick auf Unterschiede zwischen psychometrischen und edumetrischen Ansätzen.

Schlagworte: Unterrichtsqualität, Perspektivenübereinstimmung, Klassenführung, Itemformulierung, Unterrichtsbeurteilungen

Contact

Prof. Dr. Marten Clausen, Universität Duisburg-Essen,
AE Unterrichtsforschung,
Universitätsstr. 2, 45117 Essen, Deutschland
E-Mail: marten.clausen@uni-due.de

Themenblock V: Modellierung der Wirkungen von Unterrichtsqualität

Alexander Naumann/Susanne Kuger/Carmen Köhler/Jan Hochweber

Conceptual and Methodological Challenges in Detecting the Effectiveness of Learning and Teaching

Abstract: One major goal of research on educational effectiveness is to detect the effects of teaching and learning. Reliably detecting the effects of teaching and learning requires the identification and adequate measurement of (a) the relevant classroom processes and (b) outcomes on the student and the classroom level and also (c) modeling the link between both. The present paper aims to identify and discuss current conceptual and methodological challenges in regard to making inferences on the effectiveness of teaching and learning. We give a brief overview of current practices, discuss key quality criteria with respect to these three aspects, and identify areas in need of further development.

Keywords: Educational Effectiveness, Measurement, Student Outcomes, Multilevel Modeling, Validity

1. Introduction

Key to research on educational effectiveness is detecting the effects of teaching and learning. However, learning in schools and classes is a complex interaction of students and teachers (Helmke, 2015). Accordingly, the detection of factors fostering students' learning in schools and classes is a demanding task that requires both sound theory and elaborate research methodology. Thus, our paper focuses on conceptual and methodological challenges that one faces in detecting the effectiveness of teaching and learning. We address three different yet interrelated methodological aspects of effectiveness research: (a) the identification and measurement of relevant processes, (b) the measurement of outcome variables, and finally (c) modeling the link between these processes and outcomes of interest. Our focus is on multilevel modeling, as multilevel models have become standard in educational effectiveness research (e. g., Marsh et al., 2012). Multilevel models account for nested data structures and allow for the partitioning of variances at the different levels. In the following sections, we first give a brief overview of current practice, before discussing quality criteria and pointing to recent developments that may have the potential to improve future effectiveness research.

2. Identification and Measurement of the Relevant Processes and Outcomes

In general, detecting the effectiveness of teaching and learning requires the identification and measurement of the relevant variables. Essentially, this initial step involves two central questions: (1) *what are the key variables* and (2) *what are adequate ways of measuring them?*

Engaging the first question requires researchers to define their study's key variables. That is, researchers need to explicitly state and theoretically substantiate which variables they expect to be associated with or to contribute to their evaluation of effectiveness. These key variables comprise the independent variables and at least one or more dependent variables or outcomes, respectively. Prominent frameworks highlighting key dependent and independent variables in educational research are, for example, the dynamic model of educational effectiveness (Creemers & Kyriakides, 2008) or the utilization of opportunity to learn models (e. g., Helmke, 2015; Seidel, 2014). Such utilization of opportunity to learn models distinguish characteristics related to (a) the provision of learning opportunities, (b) the use of the learning opportunities, and (c) the learning outcomes. Further, they visualize potential empirical relationships and interactions.

Moreover, these models demonstrate that key variables are located at different levels. Following Seidel (2014), the provision of learning opportunities comprises characteristics related to the context (e. g., context of the educational system, the school, the classroom), the teacher (e. g., professional experience, competence), and the teaching processes (e. g., quality of the materials). That is, different layers of the model cover variables at the system, school, classroom or teacher levels. In contrast, characteristics related to the use of the learning opportunities (e. g., learning prerequisites) or the learning outcomes (e. g., achievement) are predominantly located at the level of individual students. Accordingly, in many cases, modeling educational effectiveness means modeling the cross-level effects of group-level processes on individual-level outcomes (Creemers, Kyriakides & Sammons, 2010; Marsh et al., 2012). In addition, the various levels have implications for the second question relating to the measurement of the group-level processes and the individual-level outcomes themselves.

2.1 Measurement of the Processes

Suppose the aim of a study is to evaluate whether teaching quality (e. g., Klieme, 2018) is effective in fostering students' learning. The relevant processes related to teaching quality dimensions are located at the classroom level. Such classroom level processes are typically measured in three ways: via (a) classroom observations, (b) teacher ratings or (c) student ratings (e. g., Fauth, Decristan, Rieser, Klieme & Büttner, 2014). Observers and teachers provide information on the same level as the process is located on, while students provide individual level data that are aggregated to provide information on the classroom level (De Jong & Westerhof, 2001; Lüdtke, Robitzsch, Trautwein &

Kunter, 2009). Recent research has provided a methodological foundation for dealing with measurement error and sampling error in such situations.

Manifest or observed scale scores carry measurement error, due to the sampling of the items serving as indicators for the latent variables (Skrondal & Rabe-Hesketh, 2004). Thus, latent variable models like item response theory (IRT; e. g., Embretson & Reise, 2000) or confirmatory factor analysis (CFA; e. g., Bollen, 1989) models are commonly-applied tools for dealing with such measurement error. Still, in situations where individual level indicators are aggregated to form group-level constructs, additional sampling error arises, due to the sampling of students within classrooms.

More specifically, Marsh and colleagues (2009, 2012) distinguish two types of group-level constructs: namely, constructs based on (a) true group-level measures (e. g., classroom observations, teacher responses) and those based on (b) aggregates of individual level responses (e. g., student ratings of teaching quality, gender ratio). While variables in both categories are subject to measurement error, variables in the latter category additionally contain sampling error. The latter category comprises so-called contextual variables and climate variables.[1] Contextual and climate variables differ insofar as the former are reflective aggregations, while the latter are formative aggregations of the individual level measures to the group level (Lüdtke et al., 2008). Reflective and formative aggregation have different referents; this has implications for sampling error.

For classroom climate variables, the referent is a student's classroom or teacher. That is, each student rates some aspect of his or her classroom or teacher. Conceptually, the classroom level construct is a latent variable based on multiple indicators (i. e., individual students' ratings). The underlying assumption is that each student rates the same classroom level construct. Hence, the expectation is to achieve high agreement among students within the same classroom. In line with this reasoning, the suggestion has been made to treat the idiosyncratic proportion of the students' ratings as sampling error: that is, students' ratings are treated as exchangeable (Lüdtke et al., 2009, 2011; Marsh et al., 2009). To handle both measurement error and sampling error in climate variables assessed by individual student ratings, Lüdtke, Marsh and colleagues recommend latent measurement and latent aggregation using multilevel CFA or structural equation (SEM) models, given that the sample size is sufficiently large (Lüdtke et al., 2011; Marsh et al., 2009, 2012).

Nevertheless, within-classroom variation in ratings does not necessarily represent merely undifferentiated 'noise', but will often be of substantial interest and importance (Cole, Bedeian, Hirschfeld & Vogel, 2011; Schweig, 2016). Among other possibilities, such variation may point to actual or perceived differences in treatment by teachers (e. g., based on differential teacher expectations) between students or subgroups of students, or may indicate interpersonal friction in a classroom. Student characteristics (e. g., demographic variables) can be systematically related to student ratings, potentially leading to bias in the group-level measures if classrooms differ considerably in their student

1 Other terms in the literature that have been used synonymously to 'climate' and 'contextual' are 'shared' and 'configural' (see Stapleton, Yang, & Hancock, 2016).

composition (Wagner, Göllner, Helmke, Trautwein & Lüdtke, 2013). Also, systematic error may be introduced in measures of group-level constructs by certain specifics of the measurement situation or item content (Lüdtke et al., 2011). Hence, to facilitate the reliable and valid measurement of group-level constructs, the factors introducing variation in students' ratings between and within classrooms should be further investigated.

For contextual variables, the referent is the individual student; the group-level construct is an aggregation of the individual level characteristics. That is, in contrast to climate variables, the students are not interchangeable, as the expectation is that students within the same group differ with respect to their individual characteristics. Consequently, the measurement precision of contextual variables depends on the proportion of students assessed per class: For example, if all students within a classroom are assessed, a classroom's gender ratio can be determined perfectly. On the other hand, measurement error increases as the number of students decreases (Lüdtke et al., 2008; Marsh et al., 2012). In the latter case, latent aggregation of formative variables is still appropriate to account for sampling error, while in the former case, it is reasonable to assume that the contextual variable is free of sampling error (Lüdtke et al., 2008).

Nevertheless, the boundaries between contextual and climate variables are fluid. Hence, a group-level construct may simultaneously be both a contextual and a climate variable (Stapleton, Yang & Hancock 2016). In such cases, group-level constructs comprise both aforementioned sources of variation – that is, variation due to individual level differences (i. e., the contextual part) – and sampling error due to the idiosyncratic proportions of the individual ratings (i. e., the climate part).

2.2 Measurement of the Outcomes

Education has manifold outcomes, comprising student achievement, cognitive outcomes, and motivational-affective outcomes (e. g., Seidel & Shavelson, 2007). In recent years, non-cognitive outcomes such as well-being or interest have received increasing attention as prerequisites for successful student learning (e. g., Cappella, Aber & Kim, 2016). Nevertheless, the most commonly applied criterion for judging educational effectiveness is student achievement (Klieme, 2018).

Student achievement may be assessed in multiple forms, such as educational degrees or grades; in educational research, student achievement is usually conceived of as a latent variable measured via multiple indicators in standardized tests. Standardized tests may be administered at one time point or on multiple measurement occasions, in order to assess growth. A first step in constructing the outcome measure is scaling.

In recent years, IRT has become the method of choice for scaling achievement test data. Typical IRT models, such as the Rasch model (Rasch, 1961) relate students' observed item responses to underlying latent variables – that is, parameters describing a student's ability and parameters related to item characteristics. While the Rasch model is unidimensional, as it assumes one latent ability dimension, student achievement is oftentimes more complex, with tasks requiring multiple abilities and skills (Klieme,

Hartig & Rauch, 2008). Multidimensional IRT models (MIRT; Reckase, 2009) allow for the incorporation of multiple person characteristics that describe the abilities and skills needed to solve the items, thus increasing the assessment's informative value. Within MIRT models, each test item may be related to either one dimension only or to multiple ability dimensions at the same time (Adams, Wilson & Wang, 1997; Hartig & Höhler, 2008). Also, IRT models may be extended to account for (a) multilevel structures, with students nested in classes, courses, schools or measurement occasions (e. g., Hartig & Kühnbach, 2006; Kamata, 2001) or (b) predictors, to explain variation in ability or item parameters (De Boeck & Wilson, 2004), thus providing a flexible framework for scaling and analyzing cross-sectional and longitudinal data. Alternatively, researchers may resort to, for example, mixture distribution Rasch models (Rost, 2004) or cognitive diagnosis models (e. g., Leighton & Gierl, 2007).

Still, not all of the aforementioned features of IRT are used in educational research practice. While many studies rely on IRT for scaling (e. g., Decristan et al., 2015), very few use IRT or – more generally – latent variable models when analyzing process and outcome variables, thereby accounting for the different sources of error.

2.3 Modeling the Link between Processes and Outcomes

The previous sections have addressed the measurement of the relevant processes and outcome variables located at different levels within the educational system. Accordingly, linking processes to outcomes also requires multilevel modeling approaches. In the following, we address the issue of modeling the link beween processes and outcomes from a conceptual perspective, while an accompanying empirical paper by Köhler, Kuger, Naumann, and Hartig in this issue presents different modeling examples.

Essentially, researchers need to specify (a) the level of analysis, (b) the model, and (c) control variables. Specifying the level of analysis relates to the question of where we expect effectiveness to become visible (Morin, Marsh, Nagengast & Scalas, 2014). That is, researchers are required to clarify the level(s) that may exhibit the effects of processes on outcomes. These levels affect the ways data are analyzed and interpreted.

In recent years, multilevel regression (MLR) models (e. g., Raudenbush & Bryk, 2002) have become the standard method for determining the effectiveness of learning and teaching (e. g., Creemers et al., 2010; Marsh et al., 2012). MLR models account for nested data structures with students in classes, schools or time points, respectively. Even when all variables are on the same level, multilevel analysis is usually advised, since neglecting nested data structures may lead to biased estimates (Gelman & Hill, 2006). Moreover, MLR models offer great flexibility when relating individual- or group- level independent variables to individual-level dependent variables – for example, by allowing the inclusion of both manifest and latent dependent and independent variables in the model (e. g., Lüdtke et al., 2008).

Lüdtke, Marsh, Robitzsch and Trautwein (2011) have provided a framework to distinguish approaches using either manifest or latent variables or a mixture of both. Four

types of such manifest or latent covariate models are distinguished: (1) doubly manifest (no dealing with measurement or sampling error), (2) manifest-latent (no dealing with measurement error, but accounting for sampling error in contextual and climate variables), (3) latent-manifest (accounting for measurement error, but manifest aggregation), and (4) doubly latent models (accounting for both measurement and sampling error in covariates). While the doubly latent models are conceptually preferable, estimates of group-level effects may be unstable in practice – for example, due to small group-level sample sizes (Lüdtke et al., 2008, 2009). Hence, applying either manifest-latent or latent-manifest models is recommended if the doubly latent approach is not feasible.

In addition, linking processes and outcomes requires specifying linear or nonlinear relationships in the model. While there is theoretical support for nonlinear relationships (e. g., Creemers, 2006), the empirical evidence is ambiguous. For example, Polikoff (2016) recently investigated linear and nonlinear relationships in various measures of teaching quality and student achievement. His teaching quality measures comprised student ratings as well as observations, including, amongst others, the CLASS observation system (Pianta & Hamre, 2009). Polikoff found some indication supporting linear relationships, but no evidence supporting nonlinear relationships. In contrast, both Caro, Lenkeit, and Kyriakides (2016) and Teig, Scherer, and Nilsen (2018) recently found indications supporting curvilinear relationships of student achievement and teaching practices in 62 countries' PISA 2012 and Norwegian TIMSS 2015 data. While the latter findings are in line with positions arguing that curvilinear relationships require extensive data to prevent variance restrictions (Creemers, 2006), results stemming from stronger experimental studies would be desirable.

Finally, modeling the link of processes and outcomes entails choosing a reasonable set of control variables. Control variables may be conceived of as study design factors that, if neglected, are detrimental to the drawing of valid inferences. In practice, researchers often control for specific variables because (a) it is common practice in their field, (b) treatment groups differ significantly on these variables, or (c) due to theoretical considerations. In educational research, controlling for students' background or prior achievement has become standard practice (Sammons, 2012). In practice, many more control variables are used. Consequently, a more systematic approach to covariate selection would appears beneficial. We elaborate on this issue in a following section, describing causal models.

3. Future Methodological Developments in Educational Effectiveness Research

In the previous sections, we briefly described current practices associated with detecting the effectiveness of teaching and learning. In this concluding section, we address recent methodological developments that we expect to have the potential to open new possibilities for future educational effectiveness research. Specifically, we point to three selected areas where recent developments may foster the connection of theory and re-

search practice: the use of causal models in effectiveness research, Bayesian inference, and recent trends in validity.

3.1 Causal Models

Although many effectiveness studies aim to establish causal relationships, strict causal claims are oftentimes precluded, due to the use of cross-sectional or correlational designs (Creemers et al., 2010). In recent years, educational research has put increasing emphasis on quasi-experimental and longitudinal designs, as well as analytical methods that allow for a more unequivocal attribution of outcomes to classroom- or school-level processes (Rowan & Raudenbush, 2016; Sammons, 2012). In particular, for non-randomized designs, matching methods (e. g., propensity score matching) have become increasingly common alternatives to linear regression with adjustment for covariates (e. g., Becker, Lüdtke, Trautwein, Köller & Baumert, 2012). Still, thinking and expressing causal relationships in educational settings in a more formal way may be helpful in fostering the development of more rigorous research designs backed by sound theory. For example, Hedges (2007) suggested distinguishing between an inference model that is used to specify the relationship between a hypothesized causal factor and its predicted effect, and the statistical procedures that are used to determine the strength of this relationship. One way to articulate such inference models is in directed acyclic graphs (DAGs; e. g., Pearl, Glymour & Jewell, 2016).

DAGs are formal visual representations of researchers' (expert) knowledge and beliefs about the working mechanisms within a domain (Elwert, 2013). The two basic elements of DAGs are nodes (i. e., variables) and arrows, which express relationships between the nodes. Missing arrows denote the lack of a relationship. Connections of two nodes via one or more arrows are called paths. "Acyclic" means that DAGs may not contain paths that can be traced along the direction of the arrows so as to arrive back at the starting point. Given at least three nodes A, B, and C, there are three types of paths:

- Causal paths with A influencing C through B (A → B → C).
- Non-causal paths with A and C being influenced by B (A ← B → C). Then, B is a so-called confounder.
- Non-causal paths with A and C influencing B (A → B ← C). Then, B is a so-called collider.

Using this notation, DAGs make explicit the assumptions on central interactions of variables, in a way that is very similar to path diagrams. However, DAGs are not statistical but rather are hypothesized causal models (cf. Hedges, 2007). Hence, if specified correctly, a DAG captures the hypothesized (causal) structure of the relevant elements of a process.

In addition to making theoretical assumptions on relations explicit, DAGs provide one potentially beneficial way of supporting the choice of control variables. Consider

again three nodes A, B, and C, and suppose we would like to investigate the influence of A on C using linear regression. There are three causal models for these variables which have implications for statistical control: (1) If the 'true' causal model (i. e., the DAG) is a mediation model with A influencing C directly and also indirectly through B, controlling for B in a linear regression model would only reveal the direct effect of A on C, while not controlling for B would reveal the total effect of A on C. (2) If B is not a mediator but a confounder in the causal model, controlling for B is necessary in the linear regression model. Otherwise, associations of A and C might be overestimated up to the point where an artificial relationship is found between A and C. (3) Finally, if B is a collider, controlling for B in the linear regression model leads to biased estimates of the relationship of A and C. In summary, whether or not to control for variable B depends on its status in the causal model, which should be substantiated by theory.

As an illustration, we draw on a study from medical education, which investigated the relationship between medical educators' teaching performance and the extent to which they were perceived by students as a role model as (a) teacher-supervisor, (b) physician, and (c) person (Boerebach, Lombarts, Scherpbier & Arah, 2013). DAGs were used to depict alternative conceptions of the causal associations between these variables. For the sake of brevity, we focus on the association between teaching performance, teacher-supervisor and physician role models. Several control variables considered in Boerebach et al. (2013) are not included here. Figure 1 shows three of the DAGs that were considered as theoretically plausible in Boerebach et al. in a simplified form.

In Figure 1 A, teaching performance (TP) affects both teacher-supervisor (RM-TS) and physician (RM-phy) role models. Hence, teaching performance is a confounder and has to be controlled for when estimating the association between the role model variables. On the other hand, given no directed paths between the role model variables, the paths from teaching performance to each of the role model variables can be estimated without considering the other role model variable, respectively. In Figure 1 B, teaching performance and teacher-supervisor role model are linked causally by a direct path, and additionally by an indirect path via physician role model. That is, at least part of the causal effect of teaching performance on role model as teacher-supervisor is mediated by the educator being perceived as a role model as physician. To estimate the paths of role model as teacher-supervisor, teaching performance and physician role models have to be controlled for. In contrast, the teacher-supervisor role model should not be controlled for when estimating the path from teaching performance to role model as physician.

Finally, in Figure 1 C, physician role model acts as a collider, having directed paths pointing toward both teaching performance and teacher-supervisor role models. Thus, role model as physician *must not* be controlled for when estimating the directed path from teaching performance to teacher-supervisor role models. In contrast, to estimate the paths for role model as physician, teaching performance and teacher-supervisor role models have to be controlled for. As noted, the example adapted for this illustration has been simplified considerably. A more comprehensive treatment, including empirical results, can be found in Boerebach et al. (2013). Nevertheless, although we have only

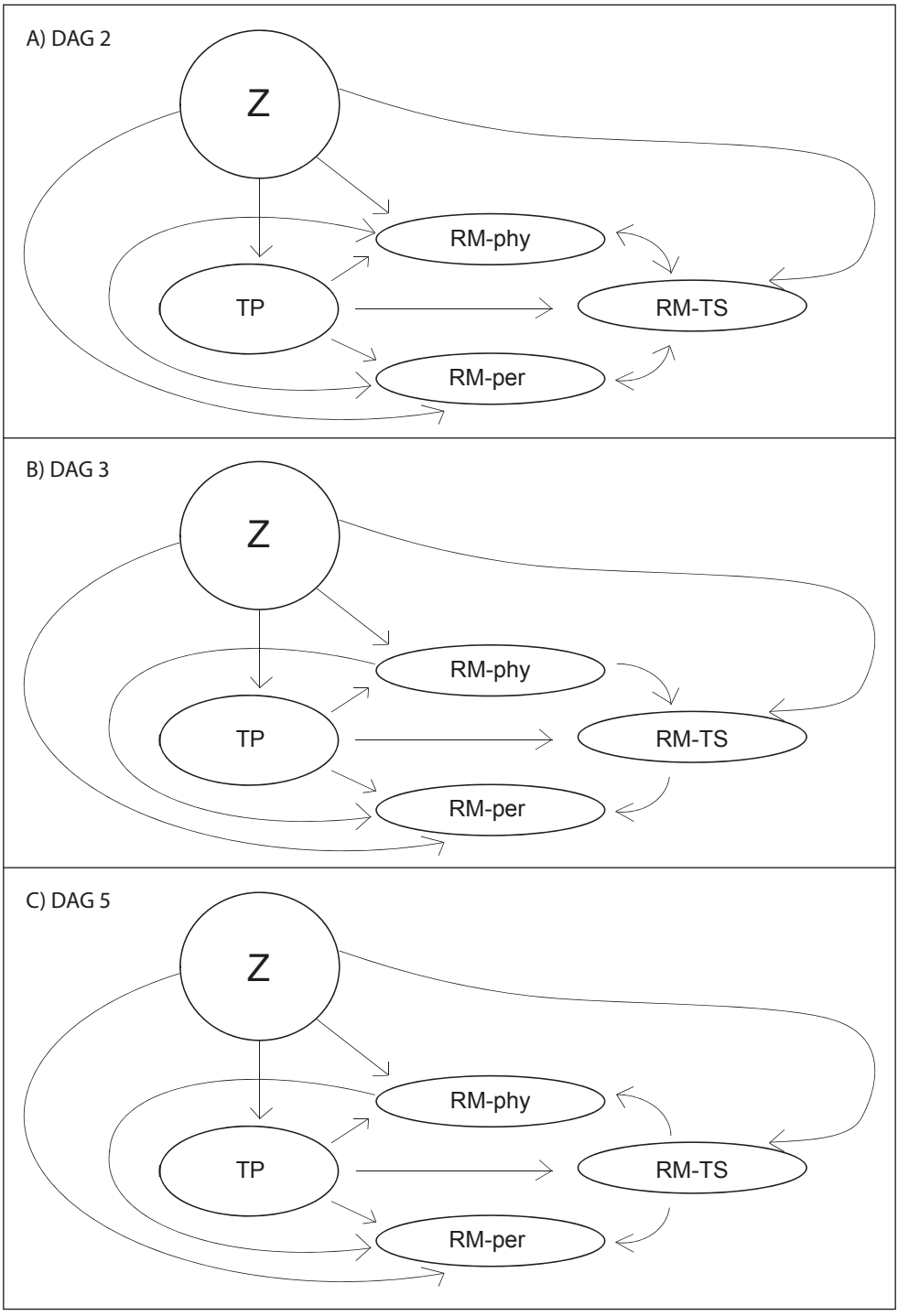

Fig. 1: Examples for DAGS (adapted from Boerebach et al., 2013)

considered three variables, the underlying principles are also applicable to more complex settings. Accordingly, we are confident that such a way of thinking and expressing the hypothesized interconnections of variables and their theoretical role in our studies, might strengthen the ties between theory and research practice, and contribute to a more thoughtful and theory-based selection of control variables beyond simple statistical significance.

3.2 Bayesian Inference

Bayesian inference plays an increasingly important role in the social and psychological sciences (Kaplan, 2014). While Bayesian inference in the past has been exclusive to a rather small community, due to the high computational demands, great progress has been made in making Bayesian inference accessible to a wider scientific public. Today, Bayesian estimation is readily available in R (R Core Team, 2017) through, for example, JAGS (Plummer, 2003) or Stan (Stan Development Team, 2017) interfaces, and it has also been implemented in the widely-used Mplus software (Muthén & Muthén, 1998–2017). The enhanced availability of software has led to an increasing application of Bayesian inference to IRT (e. g., Fox, 2010), SEM (e. g., Kaplan, 2014) and multilevel regression (e. g., Gelman & Hill, 2006).

Bayesian inference relies on Bayes' theorem to make probability statements about hypotheses or model parameter values (e. g. Gelman et al., 2013). Probability statements are expressed as probability distributions. That is, parameters are treated not as fixed but as random quantities. Three components are key to Bayesian inference: (a) the likelihood, describing the relationships within the data, (b) the prior distribution, which expresses the researcher's prior knowledge or belief with respect to the parameter values, and (c) the parameter's posterior distribution, which is the product of the prior distribution and the likelihood, and thus the foundation of Bayesian inference: The more uncertainty there is with respect to a parameter's values, the wider is its corresponding posterior distribution, and thus the wider is the range of values the parameter might probably take on. Contrariwise, if there is high certainty in a parameter's value, its posterior distribution becomes comparably narrow, with the highest probability mass or density in areas of the most probable values that the parameter might take on. That is, the posterior distribution provides information on the given probability of values a parameter might take on. Usually, posterior distributions are summarized using point estimates (i. e. mean, median or mode) and interval-based measures (e. g., Bayesian credible intervals). For example, if a regression coefficient's posterior mean is 0.5 and the 95 % Bayesian credible interval ranges from 0.3 to 0.7, one may infer that there is at least 95 % certainty that the regression coefficient is unequal to zero, indicating a statistically meaningful association of the predictor and the dependent variable.

With respect to educational effectiveness research, Bayesian inference offers two potential benefits. From a conceptual perspective, Bayesian inference allows for "learning" about parameters by updating prior knowledge with new data, resulting in a poste-

rior distribution that may in turn serve as a prior distribution in future analyses (Gelman et al., 2013). The concept of the prior distribution is thus key to Bayesian inference. Prior distributions may be either non-informative – that is, they carry no information on whether specific values are more likely than others – or informative: that is, specific values are a priori more likely than others. Whether researchers choose informative or non-informative prior distributions should depend on how much information is available, and how accurate researchers believe this information to be. On the one hand, Bayesian inference has regularly been criticized for its incorporation of such so-called subjective beliefs (e. g., Gelman, 2008). On the other hand, it has been argued that previous findings play a major role in designing empirical studies, and therefore the incorporation of substantiated knowledge into statistical models is consistent with common research practice. Still, this idea of Bayesian learning has rarely been implemented in educational research so far. Hence, there is little systematic knowledge available on the consequences of the purposeful inclusion of prior information obtained from previous effectiveness studies.

One of the few studies comparing informative and non-informative approaches has been conducted by Kuger, Klucziok, Kaplan and Roßbach (2016). They specified highly informative priors from previous research on relations between structural conditions and the quality of interactions in a classroom, and compared the results to those of a model with non-informative priors. In this case, due to major changes in classroom composition and educational standards, the ten-years-old information included in the prior was less informative than what the authors had hoped for, and model fit turned out to be better with uninformative priors.

From a more practical perspective, a second benefit of Bayesian estimation is that it conveniently allows for estimating parameters in very complex models – for example, longitudinal multilevel growth curves, IRT or SEM models with multiple (latent) variables, or cross-classified multilevel structures (van den Noortgate, De Boeck & Meulders, 2003). Bayesian estimation does not rely heavily on large sample asymptotic assumptions (Fox, 2010). Thus, Bayesian statistics allows for complex modeling even in situations with comparatively small sample sizes. Small sample sizes usually are detrimental to parameter estimation using maximum likelihood (ML) estimation (e. g., Maas & Hox, 2005). In Bayesian estimation, the data will still dominate the posterior distribution if the data contain a sufficient amount of information. For example, Zitzmann, Lüdtke and Robitzsch (2015) recently demonstrated for the aforementioned multilevel latent covariate model that Bayesian estimation, in comparison to ML provides more stable estimates of group-level effects in settings with a small number of groups ($n < 50$ groups). However, researchers still need to be aware that if the data contain little information, estimates will be sensitive to the specification of the prior distribution.

In summary, Bayesian inference bears the potential to foster educational effectiveness research (a) on a conceptual level by integrating prior knowledge into statistical modeling and (b) on a practical level by allowing the application of sound models (e. g., dealing with measurement and sampling error) that fit the demands of theory in the complex field of educational effectiveness. As there is increasing literature on the

application of Bayesian estimation of latent variable models, including very complex multilevel, structural equation or IRT models (e. g., Levy & Mislevy, 2016), adaptation of Bayesian estimation in applied educational research should in the future become straightforward.

3.3 Validity

Stimulated by the release of the latest *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), there has been a paradigm shift in the process of validation. The focus of validity has moved from the validity of instruments to the validity of the diverse uses and interpretations of educational assessments (Kane, 2001, 2013; Messick, 1995). Following this argumentative approach to validity, essentially, researchers are no longer required to provide all kinds of information on internal, content, criterion-related (and so on) validity, but rather are expected to provide such validity evidence fitting and supporting their intended use and interpretation of assessments.

In educational assessments, evidence may relate either to cognitive, instructional or inferential components of validity (Pellegrino, DiBello & Goldman, 2016). A cognitive validity component addresses domain knowledge and skills tapped by an assessment, while instructional components target the assessment's alignment with the curriculum and teaching. Finally, the inferential component relates to the degree an assessment provides information on student achievement.

The living debate on how to assess student achievement and competencies is a prominent example of this emphasis on the need for adequate validity evidence related to the cognitive component. For example, Blömeke, Gustafsson, and Shavelson (2015) argue that achievement or competency measures serve different purposes (e. g., testing whether a person will be able to accomplish a job or whether a student as successfully mastered the content of teaching) that impact the sampling of tasks (items), how the tasks are implemented (e. g., assessment center vs. paper-pencil tests), and scaling procedures. Essentially, the authors suggest that test developers and users should be required to provide corresponding evidence suitable to the purpose of their study.

More generally, the question arises how individual level measures are conceived of when used to make inferences at the group level, especially on student achievement. While educational research has put much effort into fostering valid measurements of group-level constructs using individual level data, comparatively less effort so far has been put into the meaning of achievement measures at the group level. For example, relevant classroom, school or teacher characteristics are oftentimes assessed via student reports (e. g., Fauth et al., 2014). In such scenarios, the general strategy is to aggregate the student level variables to form group-level constructs that serve as predictors of student achievement (Marsh et al., 2012). Lüdtke and colleagues (2011) argue that when applying this strategy, it is important to evaluate the psychometric properties of the aggregated student ratings and to determine whether it even makes sense to form aggregate variables in the first place.

Consequently, researchers have investigated the multilevel factor structure of many group-level constructs, especially when it comes to students' perceptions or evaluations of teaching. For example, Fauth and colleagues (2014) analyzed the multilevel factor structure of primary school students' ratings on the three basic dimensions of teaching quality (Klieme, 2018). They found a three-dimensional structure both on the individual and on the group levels. Similarly, Kuger and colleagues (2017) investigated the dimensionality of student ratings on teaching quality obtained from PISA participants. Other constructs under consideration include motivation (e. g., Martin, Malmberg & Liem, 2010) and subject-related interest (e. g., Drechsel, Carstensen & Prenzel, 2011). Yet, while the dimensional structure of student ratings at the different levels is investigated regularly nowadays, the dimensional structure of student achievement at the group level is not. Consequently, whether the same dimensional structure of achievement holds at the different levels or – with respect to growth or change measures – at different points in time, is far less frequently investigated for achievement than for questionnaire measures. In particular, there is little knowledge on the dimensional structure, with respect to repeated measurements of student ability at the level of schools or classrooms.

Moreover, validity evidence is required not only for substantiating the measurement of achievement, but also for analyzing its relation to other variables. Educational effectiveness research regularly uses student test scores as dependent variables in statistical models (e. g., Klieme, 2018; Marsh et al., 2012). However, the assumption that the individual-level student outcome measures are indeed sensitive to classroom-level teaching is scarcely substantiated in effectiveness research, although researchers have regularly asserted the need for evidence of instructional sensitivity (e. g., Naumann, Musow, Aichele, Hochweber & Hartig, 2019; Popham, 2007). Recent studies have found the association of teaching measures and student achievement to vary across different tests (e. g., Grossman, Cohen, Ronfeldt & Brown, 2014; Polikoff, 2016). That is, the capacity to capture the effects of teaching may vary across tests, resulting in inconsistent conclusions on teaching effectiveness. Consequently, if the degree of instructional sensitivity of a test has not been clarified prior to effectiveness analyses, it may remain unclear whether the teaching was ineffective or whether the test was insensitive (Naumann, Hartig & Hochweber, 2017). Accordingly, substantiating whether, or to what degree, tests are actually capable of capturing the effects of teaching is vital to establishing the validity of inferences based on the scores.

In summary, the instructional sensitivity and dimensionality of achievement measures are two exemplary areas addressing the validity of inferences and uses of assessments in educational effectiveness research. In general, we recommend that researchers adapt the principles of the argumentative approach to validity, as promoted by the latest *Standards for Educational and Psychological Testing* (AERA et al., 2014), as it may strengthen the persuasive power of their studies when they provide validity evidence fitting to their claims on the effectiveness of teaching and learning. While we are aware that the Standards themselves do not provide hands-on guidelines on how to implement the argumentative approach in research practice, there are frameworks avail-

able that offer at least some practical guidelines on the incorporation of validity evidence: for example, evidence-centered design (e.g, Levy & Mislevy, 2016; Mislevy & Haertel, 2006).

4. Concluding Comments

In this paper, we have discussed methodological and conceptual challenges associated with current practices, and future directions in educational effectiveness research. Our focus was on multilevel modeling. The following paper by Köhler and colleagues (this issue) addresses these and related issues from a more applied perspective and provides a demonstration of how elaborate multilevel modeling can be implemented in educational effectiveness research.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

AERA, APA, & NCME (2014). *Standards for educational and psychological testing.* Washington, DC.

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology, 104*(3), 682–699.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Approaches to competence measurement in higher education. *Zeitschrift für Psychologie, 223*(1), 3–13.

Bollen, K. A. (1989). *Structural equations with latent variables.* Oxford: Wiley.

Boerebach, B. C. M., Lombards, K., Scherpbier, A., & Arah, O. (2013). The teacher, the physician and the person: Exploring causal connections between teaching performance and role model types using directed acyclic graphs. *PLoS ONE 8*(7): e69449.

Cappella, E., Aber, J. L., & Kim, H. Y. (2016). Teaching beyond achievement tests: Perspectives from developmental and education science. In D. H. Gitomer & C. A. Bell (Eds.). *Handbook of research on teaching* (pp. 249–347). Washington, D. C.: AERA.

Caro, D. H., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. *Studies in Educational Evaluation, 49,* 30–41.

Cole, M. S., Bedeian, A. G., Hirschfeld, R. R., & Vogel, B. (2011). Dispersion-composition models in multilevel research: A data-analytic framework. *Organizational Research Methods, 14*(4), 718–734.

Creemers, B. P. M. (2006). The importance and perspectives of international studies in educational effectiveness. *Educational Research and Evaluation, 12*(6), 499–511.

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools. Context of learning.* London/New York: Routledge.

Creemers, B. P. M., Kyriakides, L., & Sammons, P. (2010). *Methodological advances in educational effectiveness research.* London/New York: Routledge.

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer.

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

Decristan, J., Hondrich, A. L., Büttner, G., Hertel, S., Klieme, E., Kunter, M., Lühken, A., Adl-Amini, K., Djakovic, S.-K., Mannel, S., & Naumann, A., & Hardy, I. (2015). Impact of additional guidance in science education on primary students' conceptual understanding. *The Journal of Educational Research, 108*(5), 358–370.

Drechsel, B., Carstensen, C., & Prenzel, M. (2011). The role of content and context in PISA interest scales: A study of the embedded interest items in the PISA 2006 science assessment. *International Journal of Science Education, 33*(1), 73–95.

Elwert, F. (2013). Graphical causal models. In S. Morgan (Ed.), *Handbook of causal analysis for social research*. Springer.

Embretson, S., & Reise, S. (2000): *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Gelman, A., & Hill, J. (2006). *Data analysis sing regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis, 3*(3), 445–450.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis*. CRC/Chapman & Hall.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293–303.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional irt models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology, 216*(2), 89–101.

Hartig, J., & Kühnbach, O. (2006). Schätzung von Veränderung mit Plausible Values in mehrdimensionalen Rasch-Modellen. In A. Ittel & H. Merkens (Eds.), *Veränderungsmessung und Längsschnittstudien in der Erziehungswissenschaft* (S. 27–44). Wiesbaden: Verlag für Sozialwissenschaften.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341–70.

Helmke, A. (2015). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts*. (6. überarb. Auflage) Seelze: Klett-Kallmeyer.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79–93.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York/London: The Guildford Press.

Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In D. Leutner, E. Klieme & J. Hartig (Eds.), *Assessment of competencies in educational contexts. State of the art and future prospects* (S. 3–22). Göttingen: Hogrefe Publishing.

Klieme, E. (2018). Unterrichtsqualität. In M. Gläser-Zikuda, M. Harring & C. Rohlfs (Hrsg.). *Handbuch Schulpädagogik*. Münster: Waxmann.

Kuger, S., Kluczniok, K., Kaplan, D., & Roßbach, H. (2016). Stability and patterns of classroom quality in German early childhood education and care. *School Effectiveness and School Improvement, 27*(3), 418–440. doi: 10.1080/09243453.2015.1112815.

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

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Boca Raton, FL: Chapman & Hall/CRC.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group level effects in contextual studies. *Psychological Methods, 13*(3), 203–229.

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

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias tradeoffs in full and partial error-correction models. *Psychological Methods, 16,* 444–467. doi 10.1037/a0024376.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86–92.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44,* 764–802.

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 effects. *Educational Psychologist, 47,* 106–124.

Martin, A. J., Malmberg, L.-E., & Liem, G. A. D. (2010). Multilevel motivation and engagement: Assessing construct validity across students and schools. *Educational and Psychological Measurement, 70*(6), 973–989.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25,* 6–20.

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Double latent multilevel analyses of classroom climate: An illustration. *Journal of Experimental Education, 82*(2), 143–167.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide.* Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Naumann, A., Hartig, J., & Hochweber, J. (2017). Absolute and relative measures of instructional sensitivity. *Journal of Educational and Behavioral Statistics, 42*(6), 678–705.

Naumann, A., Musow, S., Aichele, C., Hochweber, J., & Hartig, J. (2019). Instruktionssensitivität von Tests und Items. *Zeitschrift für Erziehungswissenschaft, 22*(1), 181–202.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer.* John Wiley & Sons.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81.

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.

Plummer, M. (2003). JAGS. A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.* Vienna.

Polikoff, M. S. (2016). Evaluating the instructional sensitivity of four states' student achievement tests. *Educational Assessment, 21*(2), 102–119.

Popham, W. J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan, 89*(2), 146–155.

R Core Team (2017). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from http://www.R-project.org [08. 10. 2019].

Rasch, G. (1961). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research, Copenhagen.

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

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer New-York.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion.* Bern: Huber.

Rowan, B. & Raudenbush, S. W., (2016) Teacher evaluation in American schools. In D. H. Gitomer & C. A. Bell (Eds.). *Handbook of research on teaching* (pp. 1159–1216). Washington, DC: American Education Research Association.

Sammons, P. (2012). Methodological issues and new trends in educational effectiveness research, In C. Chapman, P. Armstrong, A. Harris, D. Muijs, D. Reynolds & P. Sammons (Eds.). *School effectiveness and improvement research, policy and practice: Challenging the orthodoxy?* (pp. 9–26). London: Routledge.

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.

Seidel, T. (2014). Angebots-Nutzungs-Modelle in der Unterrichtspsychologie. Integration von Struktur- und Prozessparadigma. *Zeitschrift für Pädagogik, 60*(6), 850–866.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton: Chapman & Hall/CRC.

Stan Development Team (2017). *RStan: the R interface to Stan. R package version 2.16.2.* http://mc-stan.org [08. 10. 2019].

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*(5), 481–520.

Schweig, J. D. (2016). Moving beyond means: Revealing features of the learning environment by investigating the consensus among student ratings. *Learning Environments Research, 19*(3), 441–462.

Teig, N., Scherer, R., & Nilsen, T. (2018). More isn't always better: The curvilinear relationship between inquiry-based teaching and student achievement in science. *Learning and Instruction, 56,* 20–29.

van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28,* 369–386.

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 domain-generalizability of domain-independent assessments. *Learning and Instruction, 104,* 148–163.

Zitzmann, S., Lüdtke, O., & Robitzsch, A. (2015). A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivariate Behavioral Research, 50,* 688–705.

Zusammenfassung: Ein zentrales Ziel der Schul- und Unterrichtseffektivitätsforschung ist die Erfassung der Effektivität von Lernen und Unterricht. Die zuverlässige Erfassung der Wirkungen erfordert die Identifizierung und angemessene Messung von (a) den relevanten Unterrichtsprozessen und (b) den Ergebnissen auf Schüler- und Klassenebene sowie (c) die Modellierung der Verbindung zwischen eben diesen. Unser Beitrag zielt darauf ab, aktuelle konzeptuelle und methodische Herausforderungen zu identifizieren und zu diskutieren, wenn es um Rückschlüsse auf die Effektivität von Lernen und Unterricht geht. Wir geben einen kurzen Überblick über die aktuelle Praxis, erörtern wichtige Qualitätskriterien in Bezug auf die drei genannten Aspekte und benennen Bereiche, die weiterentwickelt werden müssen.

Schlagworte: Schul- und Unterrichtseffektivität, Schulische Lernergebnisse, Measurement, Mehrebenenmodellierungen, Validität

Contact

Dr. Alexander Naumann, DIPF | Leibniz-Institut für Bildungsforschung
und Bildungsinformation,
Rostocker Straße 6, 60323 Frankfurt a. M., Deutschland
E-Mail: Naumanna@dipf.de

Dr. Susanne Kuger, Deutsches Jugendinstitut (DJI),
Nockherstr. 2, 81541 Munich, Germany
E-Mail: kuger@dji.de

Dr. Carmen Köhler, DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation,
Rostocker Straße 6, 60323 Frankfurt a. M., Deutschland
E-Mail: carmen.koehler@dipf.de

Prof. Dr. Jan Hochweber, University of Teacher Education St. Gallen (PHSG),
Notkerstrasse 27, CH-9000 St. Gallen, Switzerland
E-Mail: Jan.Hochweber@phsg.ch

Carmen Köhler/Susanne Kuger/Alexander Naumann/Johannes Hartig

Multilevel Models for Evaluating the Effectiveness of Teaching

Conceptual and Methodological Considerations

Abstract: In research on teaching, the primary focus lies in identifying teacher behavior that positively influences relevant student outcomes. To adequately design the study, statistically model and interpret the results poses challenges for researchers. For example, the inherent multilevel structure in studies on teaching requires the application of multilevel models. This research used one exemplary data set, to which varying multilevel models were applied, thus illustrating how these models variously affect the substantial interpretation of the research question. The research question in all settings concerned the effects of teacher behavior on student outcomes. The overall purpose of this paper is to give an overview of modeling and interpreting results regarding the effectiveness of teaching appropriately.

Keywords: Multilevel Models, Repeated Measurement, Effectiveness of Teaching, Shared Construct, Configural Construct

1. Introduction

In research on the effectiveness of teaching, a primary focus has been on identifying teacher behavior that positively influences relevant student outcomes such as achievement, self-concept, or motivation. Typical research questions thus relate to the classroom level: How do teachers with different levels of, for example, supportiveness, affect the mean learning outcome in their class? An important aspect of the data in this research area is that the responses of students from the same class are typically not independent of each other, since their context is more similar, compared to students from other classes. Thus, the nested data structure is an explicit part of research on teaching; this requires the use of multilevel models. Multilevel models take the clustered structure of the data into account, allowing for inferences at the classroom level (L2), even if the information were obtained at the student level (L1). However, expanding regression models to multiple levels entails a number of methodological considerations (Byrne, 2012).

Also, research questions in the area of teaching typically revolve around making justifiable assumptions about causal inferences. Through sophisticated study planning and controlling for other potential causes, longitudinal analyses allow the drawing of conclusions about change processes and possible causes. Since research involving repeated measurements at different time points is amenable to stronger causal inferences, it is considered superior to studies that investigate relationships at only one point in time.

Such studies make analytical models more complex, however, and raise further methodological questions and challenges.

The main aim of this paper is to outline the relevant aspects that researchers working in the area of teaching need to consider when using multilevel latent variable models to answer their research questions. Our focus lies on latent variable approaches, in which several manifest indicators inform about the disposition on a latent construct, since such models control for measurement error and can be considered state of the art. We firstly consider important aspects of the research design and the data structure: (a) the number of students per class, and the number of classes; (b) the type of L2 construct; (c) the number of measurement occasions. These aspects play a vital role in regard to the precise framing of the research question and the types of models that can be applied to it. Secondly, we point out the necessary steps prior to the main data analysis, including (a) standardization of variables, (b) testing of reliability, and (c) testing of invariance assumptions.

In a third step, we introduce three multilevel models: The first involves data from only one measurement occasion, whereas the last two are examples of models that deal differently with data from two measurement occasions. In each case, the same independent and dependent variables are used, but each model is apt to answer a different question on the effectiveness of teaching. The first example, with only one measurement occasion, covers a data setup that is often found in freely available (international) large-scale data sets such as PISA (Programme for International Student Assessment) or TIMSS (Third International Mathematics and Science Study). Research questions that might be worth pursuing under such study designs can, at best, concern relationships between features of teaching and student features at a specific point in time. For example, "Do students in classes with a supportive class climate show higher competence levels compared to students in classes with a less-supportive class climate?"

However, in order to make substantial arguments for the effectiveness of teaching, analysis of relationships at one measurement occasion hardly suffices. To infer that teaching has an effect requires observations that classes develop differently under different forms or levels of teaching. To draw such conclusions, it is necessary to observe at least the dependent variable at two time points. This allows investigating whether the differences in learning growth can be attributed to teaching. Examples 2 and 3 therefore take on a longitudinal data perspective and utilize a repeated measurement design with two measurement occasions. In Example 2 we illustrate a latent regressor variable approach and investigate the relationship between the teacher variable and the outcome variable, where we condition the outcome variable on the outcome variable at a previous time point; in Example 3 we demonstrate a latent change score approach in investigating the question whether the teacher variable is related to changes in the outcome variable.

2. Aspects of the Design and the Data Structure

Prior to conducting analyses of research on teaching, several aspects of the data set should be considered:

1) The number of classes and the number of students in each class: the number of classes needs to be sufficiently large in order to obtain reliable and unbiased parameter estimates (Lüdtke, Marsh, Robitzsch & Trautwein, 2011), whereas the number of students in each class affects the reliability of the variable modeled at L2 but measured at L1 (i. e., the L2 teacher variable): A higher number of student evaluations of the teacher lead to more reliable teacher variables (Marsh et al., 2012).
2) In research on teaching, the unit of interest is typically the class. Item responses at the individual level often inform about constructs at L2, which are separated into two types: shared constructs and configural constructs (Stapleton, Yang & Hancock, 2016).[1] Shared constructs are based on items that inquire directly about the construct of interest, such as the shared classroom environment. Configural constructs, on the other hand, refer to constructs that exist at L1 and are aggregated to inform about the average within the cluster: for example, the mean motivational level in a class. In general, theoretical and empirical arguments should guide the decisions as to whether a variable is treated as a configural or a shared construct.
3) Another relevant criterion for the study design is the number of measurement occasions. In general, the research questions determine whether a study with a repeated measurement design is necessary. Multiple time points allow for questions on growth (e. g., improvement of cognitive or social skills), whereas cross-sectional data can only reveal relationships between variables at one point in time.

3. Preliminary Steps of Analysis

Before conducting the main analyses, preliminary steps should be taken. These include the standardization of variables, testing the reliability of the measures, and testing for model assumptions.

3.1 Standardizing Variables

Manifest predictors or covariates at L1 can be centered either at the cluster mean or at the grand mean. In the former case, the measure of the individual is expressed in relation to the cluster the individual belongs to; in the latter, it represents the difference to the overall mean. Since all individuals belonging to the same cluster have identical

1 Other terms in the literature that have been used synonymously to *shared* and *configural* are *climate* and *contextual* (see Marsh et al., 2012).

scores at L2 constructs (i. e., the average class level), centering is possible for L2 constructs only at the grand mean (Enders & Tofighi, 2007). The choice of centering is vital for the interpretation of model effects, and should be based on the research question (Marsh et al., 2012).

3.2 Testing Reliability

For a configural construct that is measured at L1 and is of interest at both L1 and L2, we would not necessarily expect the individuals within a cluster to respond similarly (Stapleton et al., 2016). For shared constructs, on the other hand, the measures should correlate to a high degree between individuals providing information on the same construct, demonstrating agreement amongst students. Unconditional multilevel models without predictors at either L1 or L2 can be used to measure the degree of item variance that exists at the cluster level and thus inform about how reliably the construct is measured at L2. Bliese (2000), as well as Raudenbush and Bryk (2002), proposed two kinds of intraclass correlation coefficients (ICCs) that measure the proportion of variance that is due to the clustering (ICC1) and reliability of the cluster-level components (ICC2). Low ICC1 values indicate that hardly any variance in item responses is due to the clustering of students, and that any two students within the same cluster give more similar responses than two students from different clusters. ICC2 values express the reliability of the cluster components, and should exceed .5 (Klein et al., 2000)

3.3 Testing Model Assumptions

Imposing equal factor loadings across levels implies that constructs have the same meaning at both levels (Stapleton et al., 2016; Zyphur, Kaplan & Christian, 2008). The fixing of factor loadings is also typically done across measurement occasions in longitudinal studies, thus presuming that the measured construct has the same meaning over time (Morin, Marsh, Nagengast & Scalas, 2014). These invariance assumptions can be tested by comparing multilevel confirmatory factor analysis (CFA) models that make various invariance assumptions. If the models with invariance assumptions have a similar fit as the models without invariance assumptions, imposing equal factor loadings is justified.

4. Multilevel Models

The data we used to introduce three different multilevel models came from the German DESI (Deutsch Englisch Schülerleistungen International) study, which was conducted to assess different competence areas in German and English as a foreign language, of ninth graders in Germany (Beck & Klieme, 2007). The students were tested at the begin-

ning and at the end of the school year 2003/2004. The sample size was $N = 10{,}985$; the number of classes was 427 (minimum of 9 students, maximum of 36 students in a class).

Keeping the analytical model as simple as possible, the exemplary research question throughout this article concerns the effect of teacher supportiveness on learning outcomes in English. Teacher supportiveness measures a student's perceived individualized help, the teacher's interest in his or her progress, and general experiences of teacher support for learning success (Praetorius, Klieme, Herbert & Pinger, 2018). The construct was assessed with four items, rated on a four-point Likert scale (1 = *Untrue,* 2 = *Somewhat untrue,* 3 = *Somewhat true,* 4 = *True*), inquiring about the teacher's supportiveness (TS) towards the student (e. g., "My English teacher gives me advice on how to improve"). The instrument used to assess an English learning outcome was the C-test (Harsch & Schröder, 2007), which measures text reconstruction (TR), and consists of short English texts in which half of every third word is missing and has to be completed. The test contained 12 texts with 25 incomplete words each. Some texts were only presented in specific school tracks. In our analyses, we based the latent variable TR on the four texts that were presented in all school tracks. We calculated the mean number of correctly completed words per text, using them as manifest indicators.

Note that teacher supportiveness is a configural construct, existing at both L1 and L2. In this article, we also briefly discuss the models when a shared construct is of interest. We therefore redid the analyses, using student orientation (SO) as the independent variable. Student orientation was measured on the same Likert scale as teacher supportiveness; the four items revolved around teaching practices with a particular student-centered focus (e. g., "My English teacher takes our suggestions into account"). Student orientation describes the teachers' tendency to incorporate students' interests in the class, and to use methods that focus on high student engagement.

Note that all involved manifest variables were observed at L1. Information on average text reconstruction ability in the class, average perceived teacher supportiveness and average perceived student orientation of the teacher were obtained by modeling those variables at L2 also. The advantage of the resulting doubly-latent models is that they control for measurement error at L1 and L2 as well as for sampling error with respect to the aggregation of L1 scores to form L2 constructs (Marsh et al., 2009). For the shared construct, we simply let the manifest variables correlate at the within level, without imposing a latent factor structure. The underlying assumption for the shared construct was that differences in how students perceived the student orientation of their teacher resulted from random error, and thus that an individual student rating was unrelated to the individual skill level in English text reconstruction. All analyses were conducted using the software M*plus* 7.4 (Muthén & Muthén, 1998–2015). The full-information-maximum-likelihood (FIML) approach in M*plus* was used to deal with the missing data.[2]

2 Before the main analyses, we conducted all preliminary analyses described in the previous sections. Results are not presented here due to limited space, but will be provided by the author on request.

5. Example 1: One Measurement Occasion

The first example illustrates a scenario in which data is obtained on only one measurement occasion. To estimate the relationship between text reconstruction and teacher supportiveness, we analyzed a doubly-latent multilevel model, as depicted in Figure 1(a). The significant positive slope coefficient at L2 indicates that classes with higher teacher supportiveness have, on average, higher scores on the text reconstruction test. Note that the standardized regression coefficient of .292 does not represent the effect of text reconstruction on teacher supportiveness, controlling for this effect at L1. Instead, due to the implicit group mean centering, the effect at L1 is controlled for the L2-effect, but the L2 effect is confounded with the L1-effect (Enders & Tofighi, 2007; Kreft, de Leeuw & Aiken, 1995). In order to evaluate whether the cluster has any explanatory power additional to L1, a difference parameter between the L2 and L1 slope coefficients can be estimated, representing the actual contextual effect (Marsh et al., 2009, 2012). We calculated the standardized contextual effect parameter by putting the contextual effect of teacher support in relation to the overall variance of the dependent variable:

$$\beta = (b_B - b_W) \frac{\sqrt{\sigma^2_{TS_B}}}{\sqrt{\sigma^2_{TR_B} + \sigma^2_{TR_W}}}, \tag{1}$$

where b_B and b_W are the unstandardized regression coefficients at the between level (L2) and the within level (L1), respectively, $\sigma^2_{TS_B}$ is the variance of teacher support at L2, and $\sigma^2_{TR_B}$ and $\sigma^2_{TR_W}$ are the variances of text reconstruction at L2 and L1, respectively. In our example, the contextual effect was 0.206.

For the climate variable, we calculated the effect of text reconstruction on student orientation at L2 only (see Fig. 1, b). There was no regression coefficient at L1 because the items on student orientation were intended to measure a shared construct only. Results showed, however, that there was considerable variation and covariation of the responses regarding student orientation at L1. This, alongside the low ICC1 values, raises doubts as to whether student orientation is truly a shared construct that represents a characteristic of the classroom only (Stapleton et al., 2016). In general, inspection of the variation and covariation of the responses can give insight into whether a construct also exists at L1 and should be taken into account at that level.

How can the results from the contextual and climate analyses at one measurement occasion be interpreted? The positive standardized coefficients simply inform us that classes with higher average English text reconstruction skills also report more supportive teachers, and teachers with a higher student orientation. This relationship, however, might simply reflect an existing state and not a result in the sense that more supportive teachers and teachers with a high student orientation foster text reconstruction skills. The results thus do not allow conclusions regarding the effectiveness of teaching.

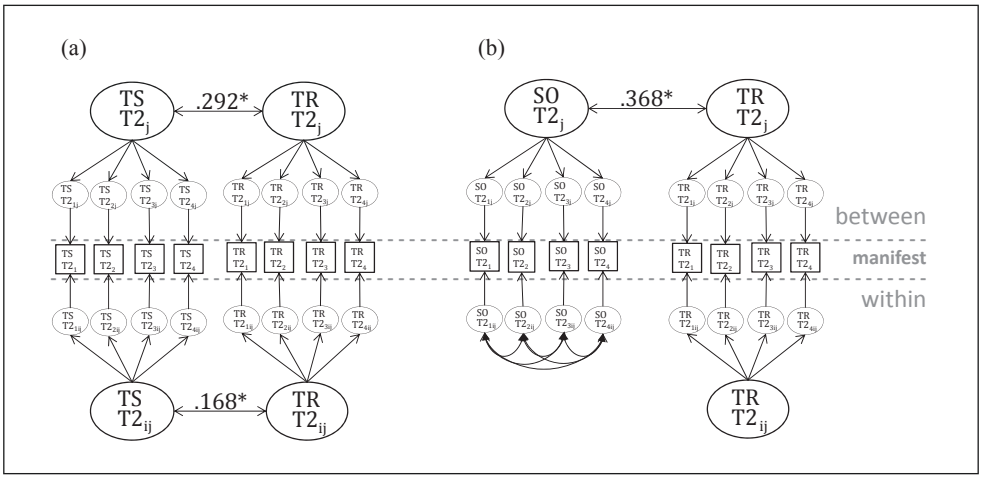

Fig. 1: Doubly-latent multilevel model estimating (a) the relationship between text reconstruction (TR) and the configural construct teacher supportiveness (TS) and (b) the relationship between text reconstruction (TR) and the shared construct student orientation (SO) at the second measurement occasion.

Researchers conducting similar analyses should be aware that data from only one measurement occasion allow no assumptions regarding causality or any underlying process. For example, it would be inappropriate to conclude that a highly supportive class climate *leads to* higher class competence levels, or that a teacher who encourages students to read at home *results in* better reading skills. Alternatively, teachers might act more supportively in classes with highly skilled students, or highly skilled students might only perceive their teachers as being more supportive.

6. Example 2: Two Measurement Occasions, Regressor Variable Approach

Literature on causality recommends conducting experiments in such a way that participants are randomly assigned to either the treatment or the control group, and that the treatment should take place between the pre-test and the post-test (Allison, 1990; Steyer, 2005). In educational assessments, however, the prototypical panel design consists of one or several measurement occasions where the independent variables are assessed simultaneously with the outcome variables, and not necessarily on all measurement occasions. Further, no true intervention takes place, in the sense that some classes are assigned a supportive teacher whereas others are not. It is also debatable as to which time frame students consider when they are asked about a teacher behavior or class characteristic. Specifying the time frame in the item wording (e. g., "Within the last 3 months, did your teacher …") makes the assumption more plausible that the responses

to items measuring the independent variable relate to a time frame that precedes the measure of the outcome variable. It would thus strengthen the argument that the teacher characteristics were causally prior to the outcome. Also, information on the independent variable prior to T1 would be useful to identify classes in which a change occurred (e. g., classes that switched from a less-supportive teacher to a supportive teacher).

When two or more measurement occasions are involved, the researcher needs to decide how to model change over time. Two prominent models for dealing with longitudinal data are (1) regressor variable approaches and (2) change score approaches (Allison, 1990).[3] The regressor variable approaches are basically covariance analytical approaches in which the variable of the previous measurement occasion is included as a form of control variable in the regression model, thus predicting the outcome variable at a later time point from the measure at an earlier time point. In change score approaches, the difference of the outcome variable between the two time points is calculated and this change score is used as the dependent variable. Both approaches have been thoroughly discussed in the literature in terms of their advantages and disadvantages (Allison, 1990; Cronbach & Furby, 1970; McArdle, 2009). We apply the approaches to our data in Examples 2 and 3, discussing how they differ, and which interpretations they each allow. As they answer distinct research questions, we do not expect matching results.

The DESI study tested the students at the beginning (T1) and at the end of ninth grade (T2) in regard to their skills; the student questionnaire was only administered at T2. Unfortunately, when the students were asked about their teacher or their school, no time reference was included in the item wording. Therefore, when evaluating questions such as "My English teacher takes our suggestions into account", it is somewhat unclear as to what time frame the students had in mind when responding. Nevertheless, we argue that the situations that came to mind must have occurred sometime prior to the second test situation, although we cannot rule out that the student had already had this specific teacher prior to the first test situation. In order to control for text reconstruction skills at T1, we included text reconstruction at T1 at both levels in the model (see Fig. 2). We thus accounted for individual levels of previous achievement and class average levels of previous achievement (Morin et al., 2014). Note that the modeling of T1 at L2 is vital, because otherwise the previous average class level would not be controlled for. The context variable teacher supportiveness was also regressed on text reconstruction at T1, since student performance might be related to the behavior of the teacher towards the students at both L1 and L2. As recommended by Jöreskog (1979), as well as Marsh and Hau (1996), we included correlated uniquenesses between each item pair that was assessed at both T1 and T2.[4]

3 Certainly other models, such as SEM growth models, are used for answering the types of research questions we consider here. Due to space restrictions, we limit our study to the models at hand.

4 Note that for reasons of simplicity, the correlated uniquenesses are not presented in the figures.

Results illustrate that, at both levels, text reconstruction at T1 was highly predictive of text reconstruction at T2. Text reconstruction at T1 also significantly predicted teacher support at L1 and L2. This could be due to some form of selectivity, adaptiveness of the teacher, or different student perceptions of the same teacher behavior. The result regarding our main research question – the influence of teacher support on text reconstruction at T2 – was also significant at both levels. Note that, as in Example 1, all standardized regression coefficients at L2 represent effects without controlling for that same relationship at L1. As in the previous example, we calculated the standardized contextual effect parameter of text reconstruction at T2 on teacher supportiveness using Equation 1, which was .016. This means that, when controlling for previous skill levels, and also accounting for the relationship between text reconstruction and teacher supportiveness at L1, the relationship between teacher supportiveness within a class and class average English text reconstruction performance was not strong

The climate variable was only introduced at L2 (see Fig. 2, b). The L2 standardized coefficient of text reconstruction at T2 on student orientation was .035, and can be interpreted to mean that, controlling for previous average English text reconstruction skill levels, the effect of the average perceived student orientation within a class on class average skill level, was rather small.

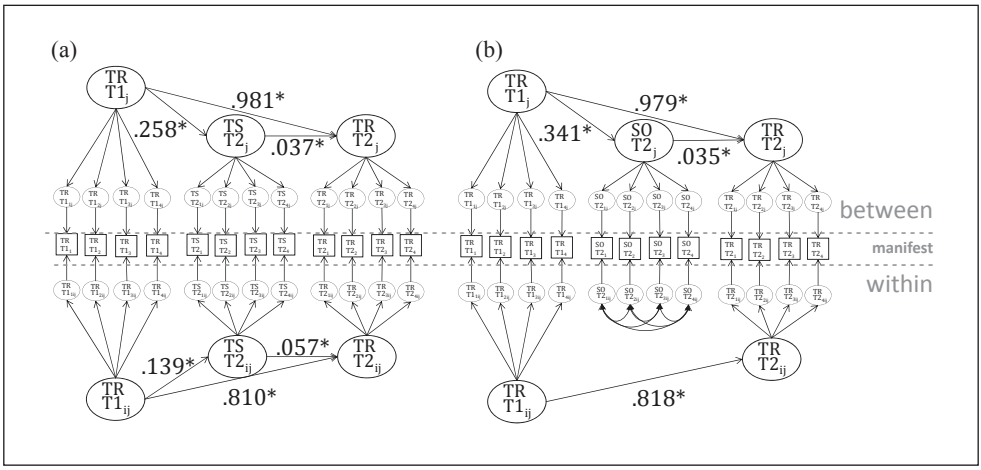

Fig. 2: Doubly-latent multilevel model estimating (a) the relationship between text reconstruction (TR) and the configural construct teacher supportiveness (TS) and (b) the relationship between text reconstruction (TR) and the shared construct student orientation (SO) at the second measurement occasion, after controlling for text reconstruction at the first measurement occasion.

7. Example 3: Two Measurement Occasions, Change Score Approach

For the change score approach, we adapted a latent structural equation model for measuring change at the individual level (McArdle, 2009) in order to apply it to the multilevel case. At both levels, an additional latent change-score variable ΔTR was introduced (see Fig. 3). At L1, the change score represented the difference between a student's skill at T2 and at T1; at L2, it represents the difference between the average classroom skill level at T1 and the average classroom skill level at T2. The advantage of this change score variable is that the variance and the mean of this variable, as well as covariances with other variables, are directly estimable model parameters (McArdle, 2009). Thus, we could directly regress the latent variable representing the change in skill level on the independent variable.

The standardized effect of the latent change variable on the configural construct variable teacher supportiveness was .211. This means that 4.5 % ($.211^2$*100) of variance of the change score was explained by L2 teacher supportiveness. Note that this approach cannot be directly compared to the regressor variable approach, since the dependent variable in the regressor variable approach is not the change score, but the class average text reconstruction at T2. The regression coefficient is also not comparable to Example 1, in which a cross-sectional relationship between the average classroom level of TR at T2 and teacher support was estimated.

For the latent change score approach using the climate variable student orientation as the independent variable, the standardized regression coefficient was .51. This means that the climate variable explained 26.2 % of variance of the change score variable.

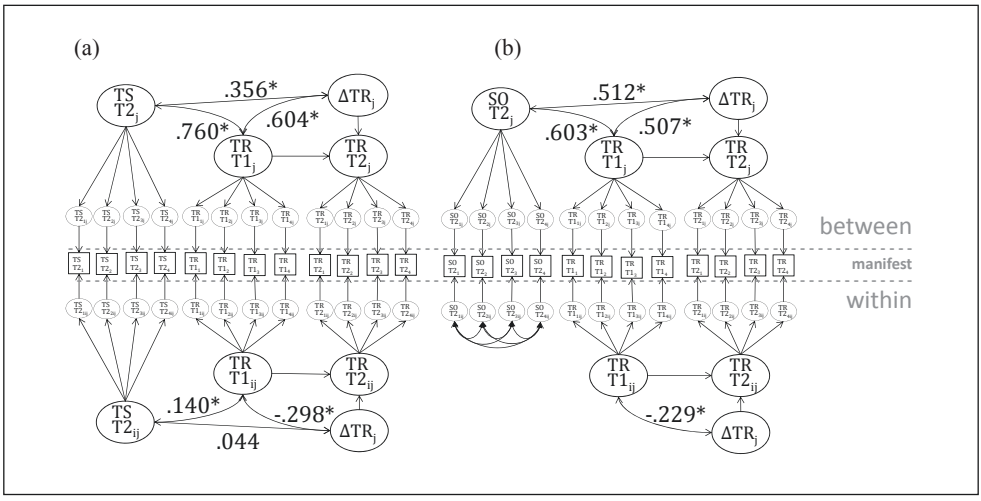

Fig. 3: Doubly-latent multilevel model estimating (a) the relationship between the text reconstruction change score (ΔTR) and the configural construct teacher supportiveness (TS) and (b) the text reconstruction change-score (ΔTR) and the shared construct student orientation (SO).

8. Discussion

The present paper has pointed out methodological issues relevant to examining the effectiveness of teaching. It highlights essential considerations, including sample and cluster size, construct measurement levels, the number of measurement occasions, variable standardization, testing the reliability of the included variables, and testing model assumptions. The article also points to the close link between data structure, measurement level and analytical models. Essentially, these form the basis for the specific research question. Using one exemplary data set and the same constructs of interest, we analyzed three different latent multilevel models to demonstrate which specific research question each answers. As expected, the results showed varying effects across the models. These differences might, in part, explain the diverse findings on the effectiveness of teaching (Praetorius et al., 2018). Although all three examples essentially deal with the effectiveness of teaching, they differ in respect of the specific research question and the modeling approach. This raises the question as to which method should be considered the standard method, in order to allow comparisons of findings across different studies. A decision to use either the regressor variable or the latent-change approach should be based on content aspects: for example, the type of outcome. To test proficiency growth, the goal is not to measure change in previous knowledge but rather, to explain which class – after controlling for the initial level – learned more, and why. Instead of choosing one particular method, another option could be to use various methods and to base conclusions on the conglomerate of those findings (Allison, 1990).

From our examples using a configural and a shared construct, respectively, we would infer that both variables relate to the dependent variable of text reconstruction to a considerable degree, but that they fail to explain additional variance in competence at the second measurement occasion after controlling for competence at the first measurement occasion. Note further that at the classroom level, hardly any competence change actually occurred, and hence there was little explanatory potential. In this regard, it is vital to discuss the quality of the measurement instrument. In order to explain changes, changes first need to actually occur. Second, they need to be detected by the measurement instrument. This means that the instrument should be sensitive enough to identify competence acquisition in educational settings (Naumann, Hartig & Hochweber, 2017).

In order to comprehensively answer research questions on teaching, and to draw more general conclusions, it is necessary to investigate multiple scenarios. These include various time points and various time intervals, several – and preferably sensitive – measurement instruments, and diverse study designs (Marsh et al., 2012). True experiments would assist in bringing forth more reliable statements on causality. Further discussion on this topic can be found in the theoretical article on this contribution (see Naumann, Kuger, Köhler & Hochweber, in this issue).

References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20,* 93–114.

Beck, B., & Klieme, E. (Eds.) (2007). *Sprachliche Kompetenzen. Konzepte und Messungen. DESI-Studie (Deutsch Englisch Schülerleistungen International).* Weinheim: Beltz.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming.* New York, NY: Taylor and Francis Group.

Cronbach, L. J., & Furby, L. (1970). How we should measure change – or should we? *Psychological Bulletin, 74,* 68–80.

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

Harsch, C., & Schröder, K. (2007). Textrekonstruktion: C-Test. In B. Beck & E. Klieme (Eds.) *Sprachliche Kompetenzen. Konzepte und Messungen. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (pp. 212–225). Weinheim: Beltz.

Jöreskog, K. G. (1979). Statistical models and methods for the analysis of longitudinal data. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation Models.* Cambridge, MA: Abt Books.

Klein, K. J., Bliese, P. D., Kozlowski, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A., Hofmann, D. A., James, L. R., Yammarino, F. J., & Bligh, M. C. (2000). Multilevel analytical techniques: Commonalities, differences, and continuing questions. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 512–553). San Francisco, CA: Jossey-Bass.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–21.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias tradeoffs in full and partial error-correction models. *Psychological Methods, 16,* 444–467. doi: 10.1037/a0024376.

Marsh, H. W., & Hau, K-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64,* 364–390.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthen, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44,* 764–802.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Adbuljabbar, A., & Köller, O. (2012). Classroom climate and contextual effects. Methodological issues in the evaluation of group-level effects. *Educational Psychologist,* 47, 106–124.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60,* 577–605. doi: 10.1146/annurev.psych.60.110707.163612.

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *Journal of Experimental Education,* 82, 143–167. doi: 10.1080/00220973.2013.769412.

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide.* Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Naumann, A., Hartig, J., & Hochweber, J. (2017). Absolute and relative measures of instructional sensitivity. *Journal of Educational and Behavioral Statistics.* Journal of Educational and Behavioral Statistics, 42(6), 678–705. doi: 10.3102//1076998617703649.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of the three basic dimensions. *ZDM, 50*(3), 407–426.
Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.
Stapleton, L.M., Yang, J.S., & Hancock, G.R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*(5), 481–520.
Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology, 1,* 39–54.
Zyphur, M.J., Kaplan, S.A., & Christian, M.S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice, 12,* 127–140.

Zusammenfassung: In der Unterrichtsforschung liegt ein Schwerpunkt auf der Identifizierung von Lehrpersonalverhalten, welches Lernende positiv beeinflusst. Ein angemessenes Studiendesign sowie die statistische Modellierung und die Ergebnisinterpretation bergen einige Herausforderungen. Beispielsweise erfordert die dem Forschungsbereich inhärente Mehrebenenstruktur mehrstufige Analysemodelle. Im folgenden Artikel wurde ein exemplarischer Datensatz verwendet, auf den verschiedene mehrstufige Modelle angewendet wurden, um zu veranschaulichen, wie diese Modelle die substantielle Interpretation der Forschungsfrage beeinflussen. Die Forschungsfrage in allen Settings bezog sich auf die Auswirkungen des Lehrpersonalverhaltens auf die Ergebnisse der Lernenden.

Schlagworte: Multilevel-Modelle, Messwiederholung, Wirkung von Unterricht, Geteilte Konstrukte, Konfigurale Konstrukte

Contact

Dr. Carmen Köhler, DIPF | Leibniz Institute for Research and Information in Education,
Rostocker Str. 6, 60323 Frankfurt a.M., Germany
E-Mail: carmen.koehler@dipf.de

Dr. Susanne Kuger, Deutsches Jugendinstitut (DJI),
Nockherstr. 2, 81541 Munich, Germany
E-Mail: kuger@dji.de

Dr. Alexander Naumann, DIPF | Leibniz Institute for Research and Information in Education,
Rostocker Str. 6, 60323 Frankfurt a.M., Germany
E-Mail: Naumanna@dipf.de

Prof. Dr. Johannes Hartig, DIPF | Leibniz Institute for Research and Information in Education,
Rostocker Str. 6, 60323 Frankfurt a.M., Germany
E-Mail: hartig@dipf.de

Oliver Lüdtke/Alexander Robitzsch

Commentary Regarding the Section "Modelling the Effectiveness of Teaching Quality"

Methodological Challenges in Assessing the Causal Effects of Teaching

Abstract: In this comment paper, we focus on three particular challenges in specifying appropriate models that can be used to estimate the causal effects of teaching in nonrandomized designs. First, we clarify that from a causal perspective the ANCOVA and change score approaches address the same research question (i. e., estimating the causal effects of teaching) but rely on different assumptions to identify the causal effects. Second, we argue that the cumulative effects of teaching (over several years) are often underestimated with two-occasion data. Thereby, we also point out the great potential of marginal structural models for analyzing the effects of time-varying treatments. Finally, we briefly discuss the role of measurement error and compositional effects, which we believe deserve further attention in future methodological research.

Keywords: Causal Effects, ANCOVA, Change Scores, Compositional Effects, Measurement Error

1. Introduction

Assessments of the effects of teaching tend to suffer from several methodological challenges. The articles in this special issue by Naumann, Kuger, Köhler, and Hochweber (in this issue) and Köhler, Kuger, Naumann, and Hartig (in this issue) provide an excellent overview of the many important statistical and methodological developments that have been achieved in the last two decades. In this comment paper, we focus on the particular challenges in specifying appropriate models that can be used to estimate the causal effects of teaching in nonrandomized designs. In line with Naumann et al. (in this issue), we want to show the potential of directed acyclic graphs (DAGs; Pearl, Glymour & Jewell, 2016) for clarifying the – often not articulated – causal assumptions of different modeling choices. More specifically, we use a structural modeling perspective that relies on DAGs to discuss three analytical issues that we believe are particularly relevant in targeting the causal effects of teaching. First, we clarify that the ANCOVA and change score approaches to analyzing two-occasion data discussed by Köhler et al. (in this issue) address the same research question (i. e., estimating the causal effects of teaching) but rely on different assumptions to identify causal effects. Second, we argue that the cumulative effects of teaching (over several years) are often underestimated with two-occasion data (Raudenbush, 2008). Thereby, we introduce a structural model

for three-occasion data and show how the causal effect of a sequence of teaching regimes (e. g., the cumulative effect of teaching across 2 school years) can be estimated. We also point out the great potential of marginal structural models for analyzing the effects of time-varying treatments (Robins, Hernán & Brumback, 2000). Finally, we briefly discuss the role of measurement error and compositional effects, which we believe deserve further attention in future methodological research.

2. ANCOVA versus Change Scores: A Structural Model Perspective

In the following discussion we introduce a structural model for two-occasion data that represents the causal relationships between the variables and allows us to clearly state the causal assumptions that are made by the different analytical approaches (see also Allison, 1990; Kenny, 1975; Kim & Steiner, 2019). More specifically, we assume that a student outcome (e. g., mathematics achievement) is measured at two measurement occasions (e. g., Grades 7 and 8), denoted as Y_1 and Y_2 respectively. We are interested in the effect of a treatment A_2 (e. g., quality of math teaching in Grade 8) on the outcome Y_2. Furthermore, we assume that a confounding variable U (e. g., socioeconomic background, gender) that affects both the student outcomes (Y_1 and Y_2) and the treatment is present. In the interests of simplicity and transparency, we assume that all effects are linear and that the variables are standardized.

To estimate the causal effect of A_2, at least three different approaches can be distinguished (see Köhler et al., in this issue). First, a naive estimator that ignores the pretest measure Y_1 is given by

$$Y_2 = \tau_{\text{naive}} A_2 + \varepsilon \tag{1}$$

Note that the naive estimator is a simple regression of Y_2 on the treatment variable A_2.

Second, an ANCOVA estimator that is conditioned on the pretest measure and has been used in many studies can be represented as

$$Y_2 = \tau_{\text{ANCOVA}} A_2 + \beta_{21} Y_1 + \varepsilon \tag{2}$$

The ANCOVA approach can be considered a special case of a more general class of conditioning methods (e. g., matching methods) in which the causal effect is obtained by conditioning on the pretest (and other observed covariates; see Morgan & Winship, 2015).

Third, a change score approach has been recommended to estimate treatment effects with two-occasion data (e. g., Allison, 1990). In this approach, the difference between the Time 2 and Time 1 scores is regressed on the treatment variable

$$Y_2 - Y_1 = \tau_{\text{change}} A_2 + \varepsilon \tag{3}$$

There has been a longstanding debate among methodologists about whether the ANCOVA approach or the change score approach is more appropriate for analyzing two-occasion data (Lord, 1967). From a descriptive perspective, it can be argued that the two approaches address different questions. In the change score approach, one would be interested in whether differences in the quality of teaching are associated with changes in student achievement. By contrast, the ANCOVA approach estimates whether differences in the quality of teaching predict achievement at Time 2 after controlling for the initial level. However, from a causal perspective, the two approaches address the same question (i. e., estimating the causal effect of the treatment) but rely on different assumptions about potential unobserved confounders. These assumptions can be clarified using the structural model in Figure 1.

It can be shown (see Appendix) that the naive estimator provides an unbiased estimate only if the treatment is unrelated to the pretest and to the unobserved confounder – a condition that is rarely met in nonrandomized designs. The ANCOVA approach produces an unbiased estimate of the treatment effect with two-occasion data if, conditional on Y_1, the unobserved confounder U does not affect the treatment (i. e., $\gamma_A = 0$) or the outcome Y_2 (i. e., $\gamma_2 = 0$). Another view of the ANCOVA approach is that it uses the past outcome and other observed covariates as a proxy for the unobserved confounder (Kim & Steiner, 2019). The change score approach is based on a more subtle set of causal assumptions. First, it is assumed that the treatment is not affected by the past outcome Y_1 (i. e., $\delta = 0$) – an assumption that does not seem very plausible in studies on the effects of teaching. Second, the effect of the unobserved confounder U needs to fulfill a very specific constraint (i. e., $\gamma_2 + \beta\gamma_1 = \gamma_1$) which essentially means that the effects of the (time-invariant) variable U are stable across time (also known as the common trend assumption; Allison, 1990).

Table 1 further illustrates the performances of the ANCOVA and change score approaches under different scenarios. We assumed that the true treatment effect would be

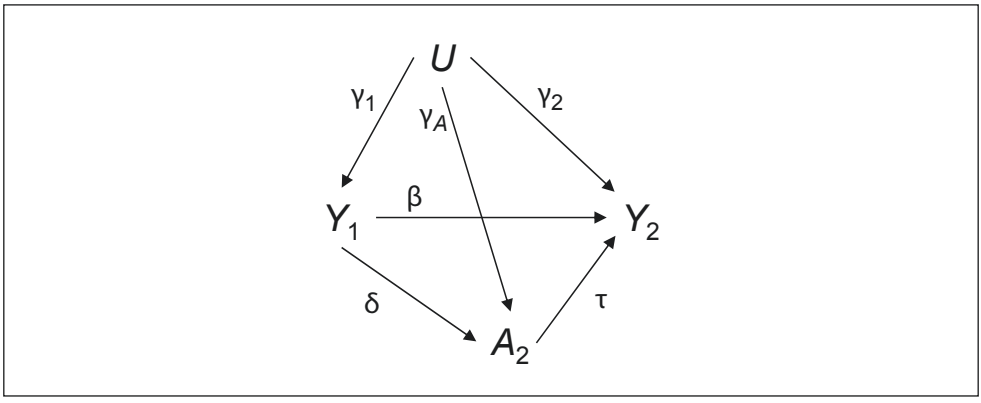

Fig. 1: Structural model for two-occasion data: Effect of treatment A_2 (e. g., quality of teaching) on outcome Y_2 with effects of the past outcome Y_1 and confounder U

γ_2	δ	γ_A	τ_{naive}	τ_{ANCOVA}	τ_{change}
0	0	0	**.20**	**.20**	**.20**
0	0	.3	.26	**.20**	.14
0	.3	0	.35	**.20**	.05
0	.3	.3	.41	**.20**	−.01
.1	0	0	**.20**	**.20**	**.20**
.1	0	.3	.29	.23	.17
.1	.3	0	.36	**.20**	.06
.1	.3	.3	.45	.23	.03
.2	0	0	**.20**	**.20**	**.20**
.2	0	.3	.32	.25	**.20**
.2	.3	0	.37	**.20**	.07
.2	.3	.3	.49	.26	.07

Note. It is assumed that the true treatment effect is $\tau = .20$, the stability of Y is moderate ($\beta = .50$), and the effect of the confounder U on Y_1 is $\gamma_1 = .40$. Unbiased estimates are printed in bold.

Tab. 1: Illustration of bias in the naive, ANCOVA, and change score estimators in two-wave design (see Fig. 1): Size of the estimated treatment effect as a function of γ_2, δ, and γ_A

modest in size (i. e., $\tau = .20$) and that the outcome Y would be moderately stable (i. e., $\beta = .50$), but we manipulated the effect of U on Y_2 (i. e., γ_2) and the treatment A_2 (i. e., γ_A). We also varied whether the past outcome Y_1 had an effect on the treatment (i. e., δ). As expected, the naive estimator in general overestimated the size of the treatment effect, and was only unbiased if the confounder U and the pretest Y_1 were not related to the treatment. The ANCOVA estimator tends to overestimate the true treatment effect and is unbiased under conditions in which U does not have an effect on either Y_2 (i. e., $\gamma_2 = 0$) or the treatment (i. e., $\gamma_A = 0$). By contrast, the change score estimator is only unbiased if $\delta = 0$ and either U does not affect the treatment (i. e., $\gamma_A = 0$) or the common trend assumption is met. Interestingly, the ANCOVA and change score estimators have a useful bracketing property (Angrist & Pischke, 2009; see also Ding & Li, 2019). Under reasonable conditions, the ANCOVA estimator provides an upper bound and the change score provides a lower bound for the true treatment effect.[1]

1 It can be shown that this bracketing property holds under mild assumptions about the data-generating model. More specifically, it needs to be assumed that the (cumulative) effect of U on the outcome is smaller for the posttest than for the pretest – that is, $(1 - \beta)\gamma_1 - \gamma_2 > 0$ – and that $\beta < 1$ (Angrist & Pischke, 2009).

Overall, we tried to clarify that from a causal perspective, the ANCOVA and change score approaches rely on different assumptions for identifying causal effects. Unfortunately, the identifying assumptions of these methods cannot be tested, and in practice it is possible that neither of these assumptions will reflect the true data-generating model. We tend to prefer the ANCOVA approach because it provides a clear rationale for including observed covariates in the analysis (VanderWeele, 2019). However, the change score approach offers the option of controlling for the effect of unobserved confounders. This comes at the price of a very restrictive assumption about the effects of the unobserved confounder (i. e., common trend assumption), an assumption that often does not seem plausible in practice (see Imai & Kim, 2019; Sobel, 2012). In addition, it can be shown that even if the confounder U is observed and included in Equation 3, the change score approach will in general produce biased estimates of the causal effect as long as the past outcome affects the current treatment (see the Appendix).

3. Assessing the Effects of a Sequence of Teaching Experiences

As aptly pointed out by Raudenbush (2008), "whether children can read or reason mathematically is the cumulative result of sequences of teaching experiences over several years" (p. 221). However, the two-occasion design is usually limited to assessing the teaching effects that occur during a single year. To better understand the limitations of two-occasion data for estimating the cumulative effects of teaching, we extended our structural model to include three-occasion data (Fig. 2) in which Y_0 now denotes baseline achievement, and A_1 and A_2 denote a sequence of two treatments (e. g., the quality of teaching over 2 years).

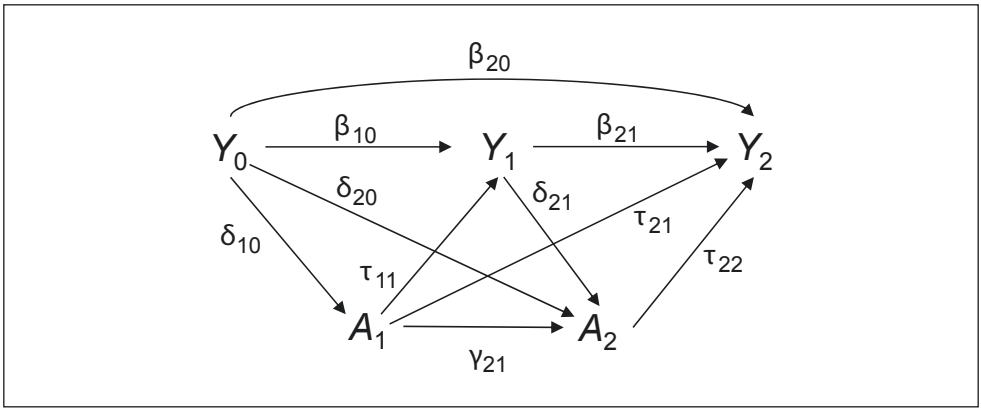

Fig. 2: Structural model for three-occasion data: Effect of the sequence of treatments A_1 and A_2 (e. g., quality of teaching across 2 school years) on outcome Y_2 with the effects of past outcomes Y_0 and Y_1

Estimation of the causal effect of the sequence of treatments A_1 and A_2 (also called an instructional regime; Raudenbush, 2008) on the final student outcome Y_2 thus becomes complicated by the fact that Y_1 is a time-varying confounder that is affected by prior treatment status (i.e., A_1). On the one hand, Y_1 is a confounder of the relationship between A_2 and Y_2. Thus, it is necessary to condition on Y_1 to obtain an unbiased estimate of the effect of A_2. On the other hand, Y_1 is on the causal pathway between past treatment A_1 and Y_2. Thus, controlling for Y_1 would block some of the effect of A_1 on Y_2.

Marginal structural models (MSMs) have been developed as powerful tools that can be used to address issues of time-varying confounders under different potential time-varying treatment regimes (see also Daniel, Cousens, de Stavola, Kenward & Sterne, 2013; Robins et al., 2000). The basic idea of MSMs is to specify a model for the treatment regime in which the effects of confounding variables have been removed. In most cases, MSMs are estimated by weighting methods (Daniel et al., 2013). However, under the assumption of a data-generating model with only linear relations, the coefficients of an MSM can be computed from the coefficients of a path model using tracing rules (Moerkerke, Loeys & Vansteelandt, 2015). In our case, the joint and direct effects of A_1 and A_2 would be given as follows

$$E(Y_2(a_1, a_2)) = (\tau_{21} + \beta_{21}\tau_{11})a_1 + \tau_{22}a_2 \tag{4}$$

where $E(Y_2(a_1, a_2))$ denotes the expected potential outcome that would have resulted if the treatment status for A_1 and A_2 had been set to levels a_1 and a_2, respectively. Thus, the joint effect of increasing the quality of teaching by 1 unit in both time intervals would be $(\tau_{21} + \beta_{21}\tau_{11}) + \tau_{22}$, whereas the direct effects of A_1 and A_2 would be $\tau_{21} + \beta_{21}\tau_{11}$ and τ_{22}, respectively.

In the following, we use the structural model for the three-occasion data (see Figure 2) to illustrate how the ANCOVA or change score approaches that rely on only two-occasion data (i.e., only considering Y_1, Y_2, and A_2) will result in biased estimates of the cumulative effects of a sequence of teaching experiences (see Appendix). We assumed that the outcome Y would show moderate stability across time and that past outcomes (Y_0 and Y_1) would have a small effect on the current treatment (A_1 and A_2). In Table 2, we varied the effects of A_1 (i.e., τ_{11} and τ_{21}) and A_2 (i.e., τ_{22}) and the stability of the treatment (i.e., γ_{21}). The MSM estimates (see Equation 4) provide the joint effect of the treatment sequence. For example, in the penultimate row, the joint effect of increasing both treatments by 1 unit is given by $(.1 + .55 \cdot .2) + .2 = .41$, which is the sum of the direct effect of A_1 and the direct effect of A_2. However, both the ANCOVA and change score approaches underestimate the joint effect of the treatment. Note that the ANCOVA approach is particularly biased when the treatment shows only moderate stability (i.e., $\gamma_{21} \leq .4$). This is a reasonable scenario when classes change their teacher after a year. In practice, effects of teaching (or observed covariates) are expected to deviate from linearity (see Naumann et al., in this issue). In this case, the MSMs would also be nonlinear, and weighting or Monte Carlo-based approaches would be recommended (Daniel

τ_{11}	τ_{21}	τ_{22}	MSM			$\gamma_{21} = 0$		$\gamma_{21} = .4$		$\gamma_{21} = .8$	
			Joint	A_1	A_2	τ_{ANCOVA}	τ_{change}	τ_{ANCOVA}	τ_{change}	τ_{ANCOVA}	τ_{change}
0	0	0	.00	.00	.00	.02	−.05	.05	−.05	.08	−.04
0	.1	0	.10	.10	.00	.02	−.04	.09	.00	.17	.04
0	.2	0	.20	.20	.00	.02	−.04	.14	.05	.27	.13
.1	0	.1	.16	.06	.10	.12	.05	.14	.04	.16	.02
.1	.1	.1	.26	.16	.10	.12	.06	.18	.08	.26	.11
.1	.2	.1	.36	.26	.10	.12	.07	.23	.13	.36	.20
.2	0	.2	.31	.11	.20	.22	.15	.23	.12	.24	.09
.2	.1	.2	.41	.21	.20	.22	.16	.27	.17	.34	.18
.2	.2	.2	.51	.31	.20	.22	.17	.31	.22	.45	.27

Note. MSM = Marginal structural model. It is assumed that the outcome *Y* shows moderate stability across time ($\beta_{10} = .60$, $\beta_{20} = .30$, and $\beta_{21} = .55$) and that past outcomes affect the current treatment ($\delta_{10} = .30$, $\delta_{20} = .10$, and $\delta_{21} = .20$). MSM estimates are based on Equation 4.

Tab. 2: Illustration of bias in the ANCOVA and change score estimators with the three-occasion data (see Fig. 2) as a function of the effects of A_1 and A_2 and the stability of the treatment (γ_{21})

et al., 2013). We believe that MSMs have great potential and deserve more attention in research on the effectiveness of learning and teaching (see Vandecandelaere, Vansteelandt, De Fraine & Van Damme, 2016).

4. Further Challenges in Estimation of the Causal Effects of Teaching

In our discussion of different approaches for estimating the causal effects of teaching, we have made several simplifying assumptions. First, we did not mention the multilevel structure of educational data. Usually, multilevel models are applied to take into account a nested data structure, and to estimate the effects of variables that are located at different levels. It should be emphasized that our remarks about the performance of the ANCOVA or change score estimators would also apply to specification of the structural model at the class level in multilevel structural equation models (MSEMs). As pointed out by Naumann et al. (in this issue), measures of teaching are affected by different kinds of error (e. g., sampling error, measurement error; Kane & Brennan, 1977). MSEMs provide a powerful tool that can be used to take these errors into account when

estimating the effects of teaching, but could provide unstable estimates in certain data constellations (e. g., a small number of classes, many items, low intraclass correlations). Bayesian methods have been shown to provide improved parameter estimates even under such challenging conditions (Zitzmann, Lüdtke, Robitzsch & Marsh, 2016). Alternatively, estimation of the measurement model (i. e., model for items) could be separated from estimation of the structural model, which could also result in more stable and robust estimates (see Anderson & Gerbing, 1982).[2]

Second, the correct way to treat compositional effects can be debated. More specifically, when conditioning on the pretest measure (e. g., Y_1 in Fig. 1 and 2), it is a crucial question as to whether the class mean (or school mean) should also be included in the regression. MSEMs decompose Level 1 predictors into a within-part and a between-part, and the group means of the Level 1 predictors are introduced into the model by default (Rabe-Hesketh, Skrondal & Zheng, 2012). This strategy of including the group means and controlling for compositional effects was also recommended by Köhler et al. (in this issue), who noted that "the modeling of T1 at L2 is vital, because otherwise the previous average class level would not be controlled for" (p. 204). However, it has been argued that controlling for compositional effects can bias the potential effects of teaching quality (Castellano, Rabe-Hesketh & Skrondal, 2014). Imagine that in the transition from elementary to secondary school, students with more favorable background characteristics are more likely to be sent to better schools (e. g., schools with greater resources, more motivated staff, better expected performance). Furthermore, it could be possible that better teachers (i. e., higher teaching quality) are attracted by better schools. As can be seen in the structural model in Figure 3, this would result in positive associations of student achievement Y_1 (more exactly, its between-part Y_{B1}) as well as the treatment A_2 with the random school effect U on the posttest (i. e., $\rho_{Y_{B1}U}\sigma_U > 0$, and $\rho_{A_2U}\sigma_U > 0$). Note that the pretest is determined before U and affects the grouping of students into different schools, resulting in an artificially increased composition effect (Castellano, Rabe-Hesketh & Skrondal, 2014; see also Cronbach, 1976). Hence, the positive covariance $\rho_{Y_{B1}U}\sigma_U$ will positively bias the estimate of β_B (i. e., "overcontrolling" for compositional effects; see the Appendix), which in turn could negatively bias the estimate of the treatment effect τ. However, the ANCOVA estimator could also be positively biased if higher teaching quality is associated with better schools. In practice, it is likely that both bias contributions are present and the ANCOVA estimator would be unbiased in the special case that they cancel each other out (i. e., $\rho_{A_2U} - \rho_{Y_{B1}U}\rho_{Y_{B1}A_2} = 0$). Interestingly, the change score estimator has the potential to control for the artificial grouping effect of students (unlike the ANCOVA), but will still be biased if the treatment is affected by the pretest, and if the correlation between the pretest and posttest differs sub-

2 For example, in generalizability theory, less parameterized measurement models are used to decompose the different error components (Brennan, 2001). Further integration of these measurement models would be a promising way to obtain more stable estimates in multilevel models.

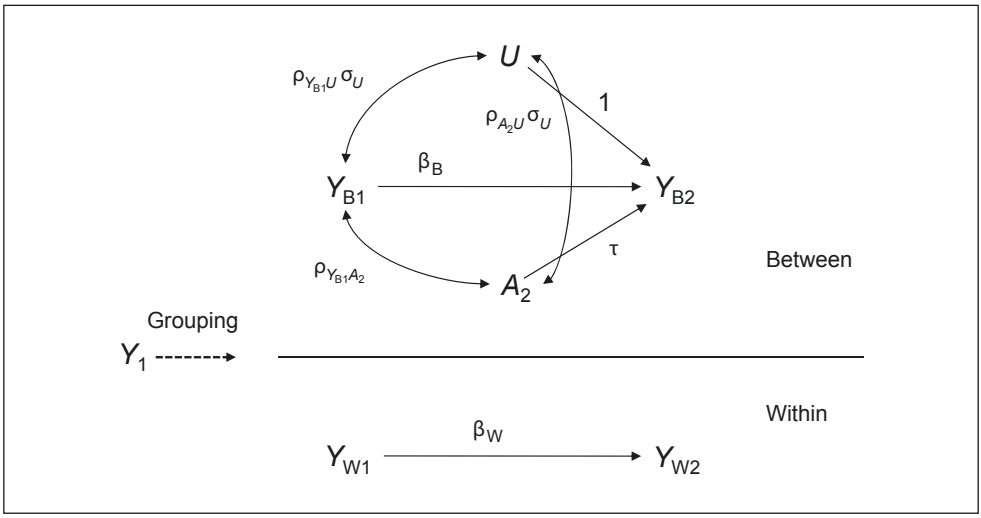

Fig. 3: Structural model for the role of composition effects with two-occasion data. The assignment to different clusters depends on the pretest Y_1 (e.g., children with high pretest scores are more likely to be sent to better schools by their parents), resulting in a covariance between Y_{B1} and U ($\rho_{Y_{B1}U}\sigma_U$).

stantially from one. Again, this illustrates how a structural model perspective can help to clarify the assumptions behind different modeling approaches.

References

Allison, P. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20,* 93–114.

Anderson, J. W., & Gerbing, D. W. (1982). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics, 39,* 333–367.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulations of questions, design and analysis.* Stanford, CA: Stanford Evaluation Consortium.

Daniel, R. M., Cousens, S. N., De Stavola, B. L., Kenward, M. G., & Sterne, J. A. C. (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine, 32,* 1584–1618.

Ding, P., & Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis.* https://doi.org/10.1017/pan.2019.25.

Imai, K., & Kim, I. S. (2019). When should we use fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science, 63,* 467–490.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 47,* 267–292.
Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the non-equivalent control group design. *Psychological Bulletin, 82,* 345–362.
Kim, Y., & Steiner, P. M. (2019). Gain scores revisited: A graphical modeling perspective. *Sociological Methods and Research.* https://doi.org/10.1177/0049124119826155.
Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305.
Moerkerke, B., Loeys, T., & Vansteelandt, S. (2015). Structural equation modeling versus marginal structural modeling for assessing mediation in the presence of posttreatment confounding. *Psychological Methods, 20,* 204–220.
Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge: University Press.
Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer.* John Wiley & Sons.
Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2012). Multilevel structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 512–531). New York, NY: Guilford.
Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal, 45,* 206–230.
Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11,* 550–560.
Sobel, M. E. (2012). Does marriage boost men's wages? Identification of treatment effects in fixed effects regression models for panel data. *Journal of the American Statistical Association, 107,* 521–529.
Vandecandelaere, M., Vansteelandt, S., De Fraine, B., & Van Damme, J. (2016). Time-varying treatments in observational studies: Marginal structural models of the effects of early grade retention on math achievement. *Multivariate Behavioral Research, 51,* 843–864.
VanderWeele, T. (2019). Principles of confounder selection. *European Journal of Epidemiology, 34,* 211–219.
Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling, 23,* 661–679.

Appendix: Derivations of Bias for the ANCOVA and Change Score Approaches

In this Appendix, we sketch the bias derivations for the ANCOVA and change score approaches. All variables are assumed to be standardized.

Two-Occasion Data

Given the structural model in Figure 1, the covariances between the observed variables Y_1, A_2, and Y_2 are derived as follows: $\text{Cov}(A_2, Y_1) = \delta + \gamma_1\gamma_A$, $\text{Cov}(Y_2, Y_1) = \beta + \delta\tau + \gamma_1\gamma_2 + \gamma_1\gamma_A\tau$, $\text{Cov}(Y_2, A_2) = \tau + \beta\delta + \gamma_2\gamma_A + \beta\gamma_1\gamma_A + \delta\gamma_1\gamma_2$. Note that the naive estimator τ_{naive} is given by $\text{Cov}(Y_2, A_2)$. The ANCOVA estimator (without controlling for the unobserved confounder U) is given as follows:

$$\tau_{ANCOVA} = \frac{Cov(Y_2, A_2) - Cov(Y_2, Y_1)\, Cov(A_2, Y_1)}{1 - Cov(A_2, Y_1)^2} = \tau + \frac{\gamma_2 \gamma_A (1 - \gamma_1^2)}{1 - 2\delta\gamma_1\gamma_A - \delta^2 - \gamma_1^2\gamma_A^2} \quad (A.1)$$

The ANCOVA estimator is unbiased if $\gamma_A = 0$ or if $\gamma_2 = 0$. Furthermore, the change score estimator is given by

$$\tau_{change} = Cov(Y_2, A_2) - Cov(Y_1, A_2) = \tau + \gamma_A(\gamma_2 + \beta\gamma_1 - \gamma_1) + \delta(\beta + \gamma_1\gamma_2 - 1) \quad (A.2)$$

The change score estimator is unbiased if $\delta = 0$ and if $\gamma_2 + \beta\gamma_1 = \gamma_1$ (i. e., the effect of U is stable with respect to Y_1 and Y_2). In addition, it is evident that the ANCOVA estimator is unbiased if the confounder U is included in the regression in Equation 2. However, it can be shown that the change score estimator is not unbiased even if the confounder U is included in the regression in Equation 3

$$\tau_{change,U} = \frac{Cov(Y_2 - Y_1, A_2) - Cov(Y_2 - Y_1, U)\, Cov(A_2, U)}{1 - Cov(A_2, U)^2} = \tau + \frac{\delta(1 - \beta)\,(1 - \gamma_1^2)}{1 - 2\delta\gamma_1\gamma_A - \gamma_A^2 - \delta^2\gamma_1^2}$$

In this case, the change score estimator would be unbiased if $\delta = 0$ (i. e., past outcome Y_1 does not affect the treatment).

Three-Occasion Data

The structural model for the three-occasion data consists of five observed variables (i. e., Y_0, Y_1, Y_2, A_1, and A_2). The implied covariances between Y_1, Y_2, and A_2 are given as follows:

$$Cov(A_2, Y_1) = \delta_{21} + \beta_{10}\,\delta_{20} + \gamma_{21}\,\tau_{11} + \beta_{10}\,\delta_{10}\,\gamma_{21} + \delta_{10}\,\delta_{20}\,\tau_{11}$$

$$\begin{aligned} Cov(Y_2, Y_1) = {} & \beta_{21} + \beta_{10}\,\beta_{20} + \delta_{21}\,\tau_{22} + \tau_{11}\,\tau_{21} + \beta_{10}\,\delta_{10}\,\tau_{21} + \beta_{10}\,\delta_{20}\,\tau_{22} + \beta_{21} + \beta_{10}\,\beta_{20} \\ & + \delta_{21}\,\tau_{22} + \tau_{11}\,\tau_{21} + \beta_{10}\,\delta_{10}\,\tau_{21} + \beta_{10}\,\delta_{20}\,\tau_{22} \end{aligned}$$

$$\begin{aligned} Cov(Y_2, A_2) = {} & \tau_{22} + \beta_{20}\delta_{20} + \beta_{21}\,\delta_{21} + \gamma_{21}\,\tau_{21} + \beta_{10}\,\beta_{20}\,\delta_{21} + \beta_{10}\,\beta_{21}\,\delta_{20} + \beta_{20}\,\delta_{10}\,\gamma_{21} + \\ & \beta_{21}\,\gamma_{21}\,\tau_{11} + \delta_{10}\,\delta_{20}\,\tau_{21} + \delta_{21}\,\tau_{11}\,\tau_{21} + \beta_{10}\,\beta_{21}\,\delta_{10}\,\gamma_{21} + \beta_{10}\,\delta_{10}\,\delta_{21}\,\tau_{21} + \\ & \beta_{20}\,\delta_{10}\,\delta_{21}\,\tau_{11} + \beta_{21}\,\delta_{10}\,\delta_{20}\,\tau_{11} \end{aligned}$$

The ANCOVA and change score estimators that ignore the baseline measure Y_0 and the previous treatment A_1 can now be derived by inserting the covariances into Equations A.1 and A.2. The calculations, however, are cumbersome and do not provide any further insights.

Role of Compositional Effects

The covariances between the observed variables at the between level are given as follows (see Fig. 3): $\text{Cov}(A_2, Y_{\text{B1}}) = \rho_{Y_{\text{B1}}A_2}$; $\text{Cov}(Y_{\text{B2}}, Y_{\text{B1}}) = \beta_{\text{B}} + \tau\rho_{Y_{\text{B1}}A_2} + \rho_{Y_{\text{B1}}U}\sigma_U$; $\text{Cov}(Y_{\text{B2}}, A_2) = \tau + \beta_{\text{B}}\rho_{Y_{\text{B1}}A_2} + \rho_{A_2U}\sigma_U$. The between group coefficient of the pretest Y_{B1} is given as

$$\hat{\beta}_{\text{B}} = \beta_{\text{B}} + \frac{(\rho_{Y_{\text{B1}}U} - \rho_{A_2U}\rho_{Y_{\text{B1}}A_2})\sigma_U}{1 - \rho^2_{Y_{\text{B1}}A_2}}$$

which is positively biased if $\rho_{Y_{\text{B1}}U} - \rho_{Y_{\text{B1}}A_2}\rho_{A_2U} > 0$. Typically, the ANCOVA estimator of the treatment effect at the between level is biased

$$\tau_{\text{ANCOVA}} = \tau + \frac{(\rho_{A_2U} - \rho_{Y_{\text{B1}}U}\rho_{Y_{\text{B1}}A_2})\sigma_U}{1 - \rho^2_{Y_{\text{B1}}A_2}}$$

It overadjusts for the compositional effect if $\rho_{A_2U} - \rho_{Y_{\text{B1}}U}\rho_{Y_{\text{B1}}A_2} < 0$ which, in turn, results in a negatively biased treatment effect estimate. The change score estimator can be calculated as

$$\tau_{\text{change}} = \tau + \rho_{A_2U}\sigma_U - (1 - \beta_{\text{B}})\rho_{Y_{\text{B1}}A_2}$$

If $1 - \rho^2_{Y_{\text{B1}}A_2} \approx 1$, it can be seen that the change score estimator has a lower potential for negative bias if $1 - \beta_{\text{B}} < \rho_{Y_{\text{B1}}U}\sigma_U$, which is fulfilled if pretest Y_{B1} and posttest Y_{B2} are highly correlated at the between level.

Zusammenfassung: Der vorliegende Kommentar konzentriert sich auf drei Herausforderungen, die bei der Spezifikation des Analysemodells auftreten. Erstens wird gezeigt, welche Annahmen über die Wirkung konfundierender Variablen sowohl mit dem ANCOVA- als auch mit dem Differenzwert-Ansatz getroffen werden müssen. Zweitens wird argumentiert, dass die kumulativen Effekte des Unterrichts (über mehrere Jahre) mit Zwei-Wellen-Daten häufig unterschätzt werden. Dabei wird das große analytische Potential von Marginal Structural Models betont, die sich besonders zur Schätzung zeitlich variierender kausaler Effekte eignen. Abschließend werden mit der Rolle des Messfehlers und der Behandlung von Kompositionseffekten zwei Themen diskutiert, die aus unserer Sicht in zukünftiger Forschung noch mehr beachtet werden sollten.

Schlagworte: kausale Effekte, ANCOVA, Differenzwerte, Kompositionseffekte, Messfehler

Contact

Prof. Dr. Oliver Lüdtke, IPN – Leibniz Institute for Science and Mathematics Education,
Department of Educational Measurement,
Olshausenstr. 62, 24118 Kiel, Germany
E-Mail: oluedtke@leibniz-ipn.de

Dr. Alexander Robitzsch, IPN – Leibniz Institute for Science and Mathematics Education,
Department of Educational Measurement,
Olshausenstr. 62, 24118 Kiel, Germany
E-Mail: robitzsch@leibniz-ipn.de

Kommentare

Ewald Terhart

Unterrichtsqualität zwischen Theorie und Empirie

Ein Kommentar zur Theoriediskussion in der empirisch-quantitativen Unterrichtsforschung

Zusammenfassung: In diesem Kommentar zum vorliegenden Beiheft wird der Anspruch der Herausgeberschaft positiv aufgenommen. Nach einer kurzen Skizze der Theorie- und Methodendiskussion in der empirischen Unterrichtsforschung werden aus dem gesamten Spektrum des Beiheftes zentrale Themen selektiv herausgegriffen und kommentiert: das Modell der drei Basisdimensionen, das Verständnis von „Theorie" beim Reden über Theorien des Unterrichts sowie schließlich das Angebot-Nutzungs-Modell in seinen Hintergründen, Varianten und Grenzen. Nach einer Erörterung der Internationalität der deutschsprachigen Unterrichtsforschung wird abschließend nur knapp auf zwei weiterführende Aspekte hingewiesen, die teilweise über das Spektrum des Beiheftes hinausgehen.

Schlagworte: Unterrichtsqualität, Unterrichtstheorie, Unterrichtsforschung, Angebots-Nutzungs-Modell, Internationalisierung

1. Einleitung

Unterricht als gesellschaftliche Realität in Schulen ist auf unterschiedliche Weise mit dem Wissenschaftssystem verbunden: Erstens sollten die *fachlichen Inhalte,* die im Unterricht vermittelt werden, so weit wie möglich durch Wissenschaft fundiert sein, mit anderen Worten: Es sollte nichts Falsches unterrichtet werden. Zweitens sind die *Lehrkräfte anhand ihrer Aus- und Weiterbildung* mit dem Wissenschaftssystem verbunden: Sie studieren in Universitäten und befassen sich dort mit den wissenschaftlich begründeten Inhalten ihrer Fächer sowie mit Ergebnissen der wissenschaftlichen Forschung zu Bildung und Erziehung, Schule und Unterricht, Lehren und Lernen in fachlichen und überfachlichen Kontexten. Die (Aus)Bildung durch, mit und an Wissenschaft(en), so die Annahme, führt zu einer höheren Qualität des Unterrichtens im Vergleich zu einer Fundierung der Lehrarbeit allein durch Alltagswissen, Berufstraditionen und Ähnliches. Und drittens: *Unterricht selbst wird im Wissenschaftssystem erforscht,* und zwar von

mehreren Disziplinen, mit unterschiedlichen Erkenntnisinteressen sowie anhand verschiedener theoretischer Konzepte und Forschungsmethodiken. Die gewonnenen Erkenntnisse sollen sowohl in die Ausbildung zukünftiger als auch in den Berufsalltag im Dienst befindlicher Lehrkräfte einfließen.

Damit bin ich beim Thema: Der wissenschaftlichen Befassung mit Unterricht mit dem doppelten Ziel sowohl der Erkenntnisbildung über diesen Gegenstandsbereich als auch der Einflussnahme auf die Arbeit der Lehrkräfte, um deren Arbeit in Unterricht und Schule positiv zu beeinflussen. Einer derart allgemein formulierten Aufgabenbeschreibung könnten vermutlich noch alle Personen zustimmen, die sich wissenschaftlich mit Unterricht befassen. Jede darüber hinaus gehende, konkretere Bestimmung führt jedoch innerhalb und außerhalb der *scientific community* zu Meinungsunterschieden, Differenzbehauptungen und Abgrenzungen.

Die gegenwärtige Unterrichtsforschung weist national wie international eine ungeheure theoretische und methodische Breite auf; sie wird von sehr unterschiedlichen sozial- und humanwissenschaftlichen Groß-Theorien und Global-Methodologien angetrieben: Man trifft auf sozialphilosophische, kommmunikations- und sozialisationstheoretische, bildungstheoretische, existenzialistische, praxistheoretische, strukturationstheoretische, austauschtheoretischer, identitätstheoretische, konstruktionistische, dekolonialistische, antirassistische, poststrukturalistische etc. Ansätze. Je nach disziplinärem Kontext (von Philosophie bis Ökonomie bis hin zu den vielen forschenden Fachdidaktiken), von dem aus man auf Unterricht schaut, fließen unterschiedliche normative Annahmen ein, werden unterschiedliche theoretische Denkmodelle zugrunde gelegt und ebenso unterschiedliche Methodiken eingesetzt. Weiterhin prägend und orientierend ist die Unterscheidung zwischen einer empirisch-analytischen, im Wesentlichen auf quantifizierende Methodik basierenden Forschungstradition einerseits und einer im weitesten Sinne qualitativ-hermeneutischen Forschungstradition andererseits.[1] Entgegen manchen Prognosen hat sich die Unterschiedlichkeit dieser beiden Traditionen – nicht zuletzt bedingt durch die erwähnte Binnendifferenzierung und -spezialisierung – nach meiner Wahrnehmung eher noch erhöht. Man könnte von zwei getrennten Theorie- und Methodenwelten sprechen, die sich kaum noch berühren. Die Binnendifferenzierung innerhalb dieser beiden Traditionen ist bereits jetzt sehr groß und wird vermutlich noch steigen.

Angesichts dieser vielfachen Differenziertheit und einer gewissen Tendenz, sich jeweils nur innerhalb des einmal gewählten ‚Paradigmas' zu bewegen, wird die Idee einer umfassenden oder integrativen, gar irgendwie vereinheitlichten Sichtweise auf Unterricht zunehmend unplausibel, wenn nicht gar illusionär. Zugleich wird das Gespräch über die Grenzen des je eigenen Ansatzes schwierig und m. E. auch gar nicht mehr gesucht; findet es doch statt, sind nicht selten Unverständnis und Fremdheitserfah-

1 Als aktuelle Übersicht über die qualitative Unterrichtsforschung vgl. Proske & Rabenstein (2018). Über die quantitative und z. T. auch qualitative Unterrichtsforschung informieren die Handbücher von Gitomer & Bell (2016) und Hall, Quinn & Gollnick (2018), stärker auf Theorielinien und Entwicklungen bezogen Steffens & Messner (2019).

rung, schlimmstenfalls wechselseitige Vorhaltungen die Folgen. Solche Abschließungstendenzen sind Resultat einer zunehmenden Spezialisierung innerhalb eines Paradigmas – und Spezialisierung treibt immer weitergehende, immer feinere Spezialisierung an. Nicht nur hier, sondern ganz generell prämiert das moderne Wissenschaftssystem epistemische Spezialisierung. Insofern muss man diesen Prozess wohl als Normalität ansehen.

Für ‚praktizierende' Lehrerinnen und Lehrer ist Unterricht demgegenüber eine drängende und häufig unübersichtliche Handlungsaufgabe, die jeden Tag, Schulstunde für Schulstunde bewältigt werden muss. Die zunehmende Spezialisierung und gelegentlich ausufernde Kleinteiligkeit beim Umgang mit dem *Forschungsproblem Unterricht* hier – die komplexe, ganzheitlich und zugleich vereinfacht wahrgenommene *Handlungsaufgabe Unterricht* dort: diese Differenz der Perspektiven und Aufgaben ist nicht zu beklagen, sondern zu konstatieren. In modernen Gesellschaften ist sie typisch für das Verhältnis von Wissenschaften und (immer mehr) akademischen Berufen; die entsprechenden Wissenschaft-Beruf-Komplexe haben unterschiedliche Strategien des Umgangs mit diesem Problem gefunden.

2. Der Anspruch des Beiheftes

Für die *Erziehungswissenschaft* allgemein sowie für die *Allgemeine Didaktik* insbesondere sind diese Entwicklungen innerhalb des breiten interdisziplinären Feldes der Forschung und Theoriebildung zu Unterricht von sehr großer Bedeutung. Weiterhin findet der allergrößte Teil der wissenschaftlichen Thematisierung von Unterricht inklusive Unterrichtsforschung im disziplinären Feld der Erziehungswissenschaft statt. Aber schon immer waren die Einflüsse (bezüglich Theorien, Modellen, Methoden, Erkenntnissen, Empfehlungen etc.) vor allem aus der Psychologie, den Fachdidaktiken, der Soziologie, den Sprachwissenschaften etc. sehr stark. Zugleich ist die Komplexität und Pluralität des Forschens und Theoretisierens zunehmend schwerer zu überschauen und zu bündeln. Gleichwohl drängt das durch alle Formen von Unterrichtsforschung wachsende wissenschaftliche Wissen über Unterricht weiterhin nach Systematisierung und Ordnung. Dies gilt nicht nur auf wissenschaftlicher Ebene, sondern insbesondere für die Frage nach der Bedeutung der gewonnenen Erkenntnisse für die Arbeit und Ausbildung von Lehrkräften.

In dieser Situation bietet das vorliegende Beiheft der Zeitschrift für Pädagogik eine sehr gute Orientierungsmöglichkeit über einen wichtigen Ausschnitt der gegenwärtigen Unterrichtsforschung: der empirisch-quantitativen Forschung über Unterricht, genauer: über diejenigen Faktoren des Unterrichts, die seine Qualität ausmachen.[2] Herausgeberschaft wie auch Autorinnen und Autoren sind sich darüber im Klaren, dass mit der Verwendung des Begriffs „Qualität" Fragen nach dem Kriterium, mit anderen Worten,

2 Die Beiträge sind im Kontext des seit 2017 von der Leibniz-Gemeinschaft geförderten „Leibniz-Netzwerks Unterrichtsforschung" entstanden und stellen eine Zwischenbilanz dar.

dass mit der Verwendung dieses Begriffs normative Fragen aufgeworfen werden. Dieser Problemkomplex bildet jedoch nicht den Gegenstand des vorliegenden Beiheftes. Im Mittelpunkt stehen *dezidiert grundlegende Theorie- und Methodenprobleme* dieses im Wesentlichen von der empirisch-quantitativ arbeitenden Lern- und Unterrichtspsychologie inspirierten Segments der Unterrichtsforschung. Dies geschieht einerseits zur Selbstverständigung innerhalb des eigenen Paradigmas, aber womöglich auch als Reaktion auf die von gänzlich anderen theoretischen und methodischen Richtungen kommende und bekannte Kritik, empirisch-quantitative Unterrichtsforschung verfahre theorielos und reflektiere nicht die Eigenarten und Grenzen ihrer Methodik. Dieser Kritik arbeiten die Autorinnen und Autoren und die Herausgeberschaft dieses Beiheftes entschlossen und erfolgreich entgegen. Die Struktur des Beiheftes befördert dieses Ziel: Es werden fünf Theorie- und Forschungsprobleme der aktuellen empirisch-quantitativen Unterrichtsforschung identifiziert. Zu jedem dieser fünf Probleme werden ein theoriebezogener Übersichtsbeitrag und ein empirischer Forschungsbeitrag präsentiert; beide Beiträge werden dann durch einen Experten bzw. eine Expertin kommentiert. Auf diese Weise wird eine anregende Übersicht über den Stand der Forschung und Diskussion in den fünf Problemfeldern vermittelt.

Was kann eine abschließende Gesamt-Kommentierung dann noch leisten? Im Folgenden soll versucht werden, den Anspruch und die Art der Theorie-Diskussion zu kommentieren sowie auf einige weiterführende Aspekte näher einzugehen, und zwar aus der Sicht der Erziehungswissenschaft und der Allgemeinen Didaktik. Dies geschieht in einem begrenzten Rahmen und sehr selektiv.

3. Modelle und Theorien

Es ist völlig angemessen, dass in dem eröffnenden Beitrag von Praetorius, Klieme, Kleickmann, Brunner, Lindmeier, Taut und Charalambous (in diesem Heft) u. a. die Frage nach einer Theorie des Unterrichts aufgeworfen wird *(theory of teaching)*. Unter der Überschrift „Theorien des Lehrens" gab es vor Jahrzehnten hierzu eine Diskussion, die durch die Übersetzung einschlägiger englischsprachiger Fachbeiträge zum Thema ausgelöst wurde (vgl. den Sammelband von Loser & Terhart, 1977), aber damals keinen größeren Nachhall gefunden hat. Theorien des Lehrens oder, mit einer gewissen Ausweitung des Perspektive: Theorien des Unterrichts waren und sind schon immer Thema der Didaktik, aber auch der Unterrichtspsychologie etc. gewesen. Praetorius et al. (in diesem Heft) geht es um den Status von Theorien innerhalb der quantitativen Unterrichtsforschung, wobei die Annahme zugrunde liegt, dass es einen *„lack of theorizing, systematic revision and verification of theories in quantitative research on teaching"* gibt.

Die Frage nach Theorie wird folgendermaßen gestellt: Gibt es im Kontext der empirisch-quantitativen Unterrichtsforschung eine begrenzte Zahl von empirisch begründeten und messbaren Dimensionen, die ein theoretisch konsistentes Ganzes bilden und zugleich weitgehend die Wirkung von Unterricht auf das Lernen der Schüler erklären?

Diese Dimensionen werden als die zentralen Dimensionen der Qualität von Unterricht verstanden. Aus der vom Max-Planck-Institut gestifteten Tradition der Unterrichtsforschung (Ursprung: TIMSS-Video 1995) wird als ein solches, empirisch gestütztes und durch statistische Faktorenanalyse fundiertes Modell die Identifikation von *drei Basisdimensionen des Unterrichts* betrachtet und erörtert, inwiefern sich hieraus eine Theorie entwickeln lässt. Dem Modell zufolge gibt es drei basale Kennzeichen jedweden lernwirksamen Unterrichts: gute Klassenführung, konstruktive Unterstützung, kognitive Aktivierung der Schülerinnen und Schüler. Diese drei basalen Dimensionen lernwirksamen Unterrichts konnten in weiteren Studien repliziert werden; Praetorius et al. (in diesem Heft) bezeichnen das Modell kurz als *three basic dimensions* bzw. *TBD*.

Dieses Modell der Basisdimensionen, verbunden mit der Vorstellung, dass diese Basisdimensionen gewissermaßen die elementare Tiefenstruktur von Unterricht ausmachen, wird seit einigen Jahren innerhalb der einschlägigen Fachdebatte erörtert und ausdifferenziert; in einigen Beiträgen des Beiheftes wird die Genealogie differenziert dargestellt. Entscheidender Zugewinn ist nun, dass Praetorius et al. (in diesem Heft) nunmehr prüfen, ob diese Erkenntnisse den Charakter einer Theorie des Unterrichts bzw. der Unterrichtsqualität *(theory of teaching* rsp. *theory of teaching quality)* haben. Noch einmal: Eine solche Theorie soll auf empirisch gestützter Basis Unterricht selbst sowie auch seine Wirkung auf das Lernen der Schülerinnen und Schüler erklären. Als Bezugsbasis für die Prüfung der Theoriequalität von TBD wählt die Gruppe der Autorinnen und Autoren die von (Kane & Marsh, 1980) aufgestellten Kriterien für eine allgemeine Unterrichtstheorie. Im Wesentlichen stützen sich diese beiden Autorinnen und Autoren auf das klassische Verständnis von „Theorie" im empirisch-analytischen Wissenschaftsansatz: Theorien sind in sich logische, widerspruchsfreie, hierarchisch geordnete Begriffs- und Satzsysteme, deren Aussagen untereinander in klar definierter Beziehungen stehen. Sie sind empirisch getestet und können (vorläufig) verifiziert oder falsifiziert werden; ebenso benennen sie ihre Grenzen. Theorien erlauben zu einem gewissen Grad auch Prognosen über zu erwartende Ereignisse und Abläufe in ihrem Gegenstandsbereich. Und schließlich sollten Theorien präskriptive Handlungsanweisungen dazu enthalten, was zu tun ist, wenn man bestimmte Ziele erreichen will.

Im Lichte dieses klassischen und – mit Verlaub – recht betagten *statement-view of theory*[3] erweist sich, so das Urteil von Praetorius et al. (in diesem Heft), die Theorie der drei Basisdimensionen von Unterricht als unvollständig bzw. unzureichend. Dem kann man nur zustimmen – und sich zugleich fragen, warum diese sehr traditionelle metatheoretische Bewertungsgrundlage gewählt wurde, in deren Licht letztlich *alle* damals wie heute real-existierenden Theorien unvollständig, inkonsistent und unzureichend erscheinen. Darüber hinaus ist in früheren Analysen breit dokumentiert worden (z. B. Drerup, 1979), dass metatheoretische Leitkonzepte nie hinreichend instruktiv sind für

3 Diesem klassischen Theorieverständnis des *statement-view* wird seit den 1970er Jahren der „non-statement-view" gegenübergestellt, der (nach Kuhn, Feyerabend und Toulmin) stärker Theoriedynamik und den Wandel von Theorien in der Zeit bzw. in der Geschichte betont (vgl. einführend als Übersicht Chalmers, 2007, Kap. 8–10).

eine genaue Anleitung einzelwissenschaftlicher Theoriearbeit und Forschungspraxis, da sie sich in (außergeschichtlichen, die konkrete Wissenschaftspraxis und -dynamik souverän ignorierenden) epistemologischen Kunstwelten bewegen. Für die Schärfung der theoretischen Debatte um den Status von modellhaften Annahmen über Unterricht und seine Struktur wäre z. B. ein Eingehen auf die Allgemeine Modelltheorie selbst (Stachowiak, 1973; Saam, 2009) und deren Erörterung in der Erziehungswissenschaft und Didaktik weiterführend gewesen; in Büchern von MacMillan und Garrison (1988) bis Biesta (2017) werden gänzlich andere Grundlagen für eine Theorie des Lehrens entwickelt.

Vieles ist also noch im Fluss und Alternativen müssen geprüft werden. Im nächsten, empirischen Beitrag wird z. B. dargelegt, ‚dass es sehr gute Gründe gibt, nicht nur drei, sondern *vier* Basisdimensionen anzunehmen: Kleickmann, Steffensky und Praetorius (in diesem Heft) belegen empirisch schlüssig, dass *kognitive Unterstützung* eine gesonderte, von kognitiver Aktivierung und auch von emotionaler Unterstützung zu unterscheidende *vierte* Basisdimension darstellt – auch dann, wenn man das Sparsamkeitskriterium für Theorien berücksichtigt. Es drängt sich die Frage auf, ob demnächst mit weiteren Basisdimensionen zu rechnen ist. Dahinter steht die Frage nach der Existenz eines logischen und/oder eines empirischen und/oder auf Konvention basierenden Kriteriums für die Abgeschlossenheit einer Theorie.

4. Angebote und Nutzungen

Praetorius et al. (in diesem Heft) weisen in ihrem Beitrag zu recht und mehrfach auf einen uneinheitlichen Gebrauch zentraler Begriffe innerhalb dieses Spektrums der empirisch-quantitativen Unterrichtsforschung hin. Dieser Eindruck verfestigt sich, wenn man den zweiten thematischen Komplex zur Kenntnis nimmt, der in diesem Beiheft erörtert wird: Die Diskussion um das vielzitierte und sehr bekannte *Angebots-Nutzungs-Modell* von Unterricht (ANM). Der sehr differenzierte und kenntnisreiche Beitrag von Vieluf, Praetorius, Rakoczy, Kleinknecht und Pietsch (in diesem Heft) macht unter Rückgriff auf die Genese dieses Modells[4] und mit Blick auf seine aktuellen Varianten

4 M. E. hätte als Ursprung deutlicher auf den Ansatz bzw. das Konzept der „opportunity to learn“ hingewiesen werden können, das inhaltlich zutreffend, aber doch etwas ungelenk in der deutschsprachigen Fachliteratur im Allgemeinen mit „Lerngelegenheit“ übersetzt wird (vgl. zur Geschichte und zum Wandel dieses Konzepts McDonnell, 1995). Eine übergreifende, eventuell paradigmenübergreifende Heuristik könnte das von Cappella, Aber & Kim, 2016, S. 250) entwickelte „Modell von Unterricht jenseits von Leistungstests“ darstellen. – Das ANM des Unterrichts ist auch auf die Lehrerbildung übertragen worden, so dass ein *doppeltes* Angebot-Nutzungs-Modell entstand: Von der Lernsituation Lehrerbildung (ANM 1) und ihren Folgen, die dann in die Lernsituation Unterricht einfließen (ANM 2); auf diese Weise wird ein mehrschrittiger Angebot-Nutzungs-Prozess modelliert. Ebenso werden Varianten von ANMs bei der Modellierung des Prozesses der Übernahme von Innovationen durch Praktiker verwendet.

deutlich, dass es *das* ANM gar nie gab, nicht gibt und wohl auch nie geben wird. Das Modell hat unterschiedliche Quellen und Hintergründe, wurde von den Protagonisten unterschiedlich verstanden, in unterschiedlicher Weise weiterentwickelt und hat heute eine Vielfalt der Rezeptions- und Verwendungsformen gefunden, die beeindruckend ist.

Vieluf et al. (in diesem Heft) präsentieren als Ergebnis ihrer Rekonstruktion eine neue, die bisherigen Modellvarianten klar verändernde Fassung des ANM, die die Kreisförmigkeit aller Abläufe bzw. die wechselseitige, also in beide Richtungen gehenden Beeinflussung aller am Unterricht beteiligten Faktoren (Rahmenbedingungen, Voraussetzungen und Folgen auf Seite der Lehrkräfte, Voraussetzungen und Folgen auf Seite der Lernenden, Angebote und Nutzungen auf Seite der Lehrkräfte, Angebote und Nutzungen auf Seite der Lernenden etc.) berücksichtigt. Am Ende entsteht ein äußerst komplexes Bild, welches deutlich macht, dass letztlich alle durch Forschung identifizierbaren, benannten und gemessenen Faktoren oder Dimensionen des Unterrichts miteinander interagieren und sich in Regelkreisen wechselseitig in alle Richtungen beeinflussen (vgl. das transaktionale Modell der Lehrer-Schüler-Beziehung bei Nickel, 1976, S. 165). Die empirische Forschung bzw. die Techniken der statistischen Analyse sind mittlerweile auch so weit vorangeschritten, dass sie diesen sehr komplexen, auf mehreren Ebenen stattfindenden, miteinander verbundenen Prozesse und Beeinflussungsformen in Teilen, aber nie insgesamt abzubilden in der Lage sind (vgl. die Beiträge zum abschließenden Themenkomplex des Beiheftes, in dem es um die Frage der statistischen Modellierung der Zusammenhänge zwischen Lehren und Lernen geht, z. B. die *directed acyclic graphs* in Abb. 1 im Beitrag von Naumann, Kuger, Köhler & Hochweber, in diesem Heft). In der Tat hängt im Unterricht sowohl aus Forschungs- wie auch aus Gestaltungsperspektive letztlich alles mit allem zusammen. In Theorie und Forschung führt dies zu stetiger und nie abzuschließender Komplexitätssteigerung – wohin führt diese Entwicklung in pragmatischer Hinsicht?

Sowohl die Diskussion um die drei Basisdimensionen lernwirksamen Unterrichts als auch die Inspektion und Neuformulierung des Angebots-Nutzungs-Modells von Unterricht machen deutlich, und die Autoren dieses Beiheftes sprechen es mehrfach aus, dass die Theorie- und Methodenlage der empirisch-quantitativen Unterrichtsforschung äußerst unübersichtlich, ungeordnet und sprachlich uneinheitlich ist – fast so wie die (gerne ausufernden) Theorie- und Methodendebatten in der philosophischen, praxis- oder kulturtheoretischen, struktur-rekonstruktiven, qualitativen, differenzsensiblen, postkolonialen etc. Unterrichtsforschung! Zweitens wird deutlich, dass die Lehre von den drei Basisdimensionen, von der der Differenz zwischen Oberflächen- und Tiefenstrukturen sowie schließlich das Angebot-Nutzungs-Modell eigentlich zu früh ihren Weg in die Lehrbücher der Unterrichtspsychologie und der Allgemeinen Didaktik gefunden haben (vgl. Kunter & Trautwein, 2013; Lipowsky, 2015; Terhart, 2019a). Die theoretischen Reflexionen und empirischen Befunde in diesem Beiheft haben diese Inhalte gewissermaßen ent-kanonisiert. Denn bei Licht besehen stellen die Vertreter bzw. Konstrukteure dieser Ansätze selbst fest, dass es sich eher um instabile, im Fluss befindliche semantische Felder als um stabile und präzise abgrenzbare Modelle oder Theorien handelt.

Diese Einschätzung ist nicht als ein von außen vorgehaltener, irgendwie ‚kritischer' Befund zu lesen, sondern macht nur deutlich, dass es in der aktuellen empirisch-quantitativen Unterrichtsforschung genauso zugeht wie in allen anderen Forschungsfeldern der Human-, Sozial- und Bildungswissenschaften: Die Erosion klassischer wissenschaftstheoretischer Großmodelle in Verbindung mit den verschiedenen Varianten des konstruktivistischen Denkens hat teils zu einer Komplexitätssteigerung, teils zu einer Entspannung der Methodenproblematik und zu einer positiven Bewertung des pluralen und fluiden Charakters von Theorien und Theorielinien geführt. (Inter-)Disziplinäre Abgrenzungsübungen und Platzanweisungen sind unangemessen und führen nicht weiter, da immer weniger klar ist, was Disziplinen eigentlich noch sind. Generell wirken methodologische Rechthaberei und Scharfmacherei – von welcher Seite auch immer – wie aus der Zeit gefallen.

Wie erwähnt, gehört es zu den klassischen Kriterien von voll entwickelten Theorien, dass sie die Grundlage für zielorientiertes Handeln in demjenigen Gegenstandsfeld bieten, auf das eine Theorie sich bezieht. Das Verhältnis von Theorien und darauf basierenden oder vorsichtiger: dadurch inspirierten Handlungen, Entscheidungen, praktischen Konsequenzen, ‚Technologien' etc. war und ist eines der Dauerthemen des klassischen wissenschaftstheoretischen Diskurses. In den Sozial- und Bildungswissenschaften, hier: bei der Frage nach der praktischen und praxisgestaltenden Bedeutung von durch empirische Unterrichtsforschung entwickelten Unterrichtstheorien wurden mittlerweile alle naiven Ableitungs- und Anwendungsversprechungen zurückgelassen. Die Unterscheidung von Wissensformen (Beschreibungswissen, Erklärungswissen, Gestaltungswissen) weist darauf hin, dass diese Wissensformen zwar in Verbindung stehen, sich aber nicht auseinander ergeben, sich nicht wechselseitig ersetzen können, mit anderen Worten jede dieser Wissensarten eine eigene Qualität und Dignität aufweist. In dem einleitenden Beitrag weisen Praetorius et al. (in diesem Heft) explizit darauf hin, dass beim Theoriekriterium der Praxisrelevanz eigentlich alle Fragen offen sind. Die verschiedenen Modelle zu bzw. über Unterricht werden vor allem als *Heuristiken für Forschung* verstanden, und *nur vereinzelt als Grundlage für didaktische* Gestaltung (didactical design).

Die klassischen und neueren Theorien (oder ‚Ansätze') der Allgemeinen Didaktik (vgl. dazu das Jahrbuch Allgemeine Didaktik von Bohl, Hanke, Koch-Priewe & Zierer, 2013) waren und sind dort offensiver – ob mit guten Gründen, sei an dieser Stelle dahingestellt. Hier paust sich die unterschiedliche wissenschaftsgeschichtliche und -institutionelle Herkunft und Platzierung von Unterrichtspsychologie (Lehr-Lern-Forschung) einerseits und Allgemeiner Didaktik andererseits durch: Letztere stand immer schon im Kontext der Bildung von Lehrkräften; allgemeindidaktische Theorien und Entwürfe versuchten und versuchen in aller Regel, zu einem Set von Gestaltungsempfehlungen für Unterricht zu kommen, wobei als oft gewählte Konkretionsstufe Berufsneulingen ein Muster für das Erlernen des Unterrichtens, konkret: etwa eine Anleitung zur Vorbereitung von Unterricht an die Hand gegeben wurde. Diese dreifache Verpflichtung auf Theorie, Forschung und Anleitung belastet bekanntermaßen den wissenschaftlichen Status der Allgemeinen Didaktik bis heute (vgl. auch Rothland, 2018). Für die em-

pirisch-quantitative Unterrichtsforschung resultiert aus dem weiterhin ungelösten Problem des Praxisbezugs bzw. der Transferfähigkeit ihrer Erkenntnisse: *further research is needed!*

5. Internationalität

Die empirisch-quantitative Unterrichtsforschung in Deutschland wird zunehmend internationaler, wenn man die stark steigende Zahl von englischsprachigen Veröffentlichungen in den entsprechenden Fachjournalen (in Europa und weltweit) als Indikator nimmt.[5] Auch in deutschen erziehungswissenschaftlichen Zeitschriften werden immer häufiger englischsprachige Beiträge gedruckt bzw. ganze Thementeile, Sonderhefte etc. in englischer Sprache veröffentlicht – so auch größere Teile in diesem Beiheft. Ich halte das für ebenso bemerkenswert wie erfreulich. Eine zunehmende Tendenz zu englischsprachigen Veröffentlichungen ist auch in den verschiedenen Zweigen der theoretisch und methodisch anders ausgerichteten, qualitativen Unterrichtsforschung bzw. in der Allgemeinen Didaktik festzustellen, allerdings noch keineswegs in diesem Ausmaß.

Die Autorinnen und Autoren dieses Beiheft tragen durch ihre Veröffentlichungspraxis (nicht nur durch dieses Beiheft!) selbst dazu bei, dass das Modell der drei Basisdimensionen inklusive der Unterscheidung zwischen Oberflächen- und Tiefenstrukturen und das Angebot-Nutzungs-Modell etc. dem internationalen Fachpublikum in englischsprachigen Beiträgen bekannt gemacht werden.[6] Während innerhalb der deutschsprachigen Debatte die Auseinandersetzung mit diesen und ähnlichen Konzepten sehr intensiv ist, ist der Bekanntheitsgrad – Ausnahme: das Modell der drei Basisdimensionen – dieser deutschen Theorie- und Forschungslinie im internationalen Feld nicht sehr hoch; Tina Seidel weist in ihrem Kommentar nach meiner Ansicht zu Recht darauf hin. Ob es der internationalen Verbreitung förderlich ist, wenn man den deutschen Ursprungsort dieser Forschungen programmatisch (z. B. in den Titeln der Beiträge) betont, ist m. E. durchaus fraglich: Generell herrscht in großen Teilen der englischsprachigen Forschungsliteratur ein starker Internationalismus. D. h. ob ein Modell, ein Ansatz aus diesem oder jenem Land kommt, ist nicht von zentraler Bedeutung bzw. könnte gar als eine problematische Betonung nationaler Denkgewohnheiten angesehen werden.[7] Dabei ist mir durchaus bewusst, dass der erwähnte *Internationalismus* durchaus und de facto als ein starker *Ame-*

5 Hinsichtlich der eingesetzten Theorien und Methoden war der Einfluss der US-amerikanischen Pädagogischen Psychologie bzw. des „research on teaching" auf die deutschsprachige allgemeine und fachdidaktische Unterrichtsforschung und Didaktik traditionell sehr groß (vgl. zur Geschichte dieses Einflusses: Terhart, 2016).

6 Wobei es hilfreich wäre, wenn sich die Autorinnen und Autoren auf eine einheitliche englischsprachige Bezeichnung des ANM und seiner Elemente einigen würden!

7 Ein früherer Versuch seitens des IPN Kiel zur Vermittlung zwischen der US-amerikanischen Curriculum- und der bundesdeutschen bildungstheoretischen Didaktik-Tradition führte zwar zu einer Reihe von Publikationen (Hopmann & Riquarts, 1995; Westbury, Hopmann & Richquarts, 2000), ist aber (Ausnahme: Skandinavien) m. E. ohne allzu großen Nachhall (Zita-

rikanismus daherkommt, dass also die US-amerikanische Tradition der empirisch-quantitativen Unterrichtsforschung für diesen Ansatz als internationaler Standard auftritt und gilt. Zugleich muss man auch sehen, dass die Arbeit in den im weitesten Sinne (sozial-) philosophisch-qualitativen und schulkritischen Formen von Unterrichtstheorie und -forschung ebenfalls eine starke Rezeption US-amerikanischer Positionen zu beobachten ist (vgl. Postkoloniale Studien, Anti-Rassismus-Studien zu Schule und Unterricht). Man sieht: Quer über die verschiedenen Paradigmen hinweg ergeben sich Perspektiven und Potentiale für interessante international-vergleichende Studien.

6. Weiterführende Aspekte

Abschließend zwei kurze Hinweise auf ein ganz altes und ein recht neues, sich aber ausweitendes Thema für Unterrichtstheorie und -forschung, die über den unmittelbaren Kontextes dieses Beiheftes hinausgehen.

Wo bleiben die Inhalte? Eine schon traditionsreiche, aus der Erziehungswissenschaft und Didaktik heraus formulierte Kritik besagt, dass die von psychologischen Lern- und Interaktionsmodellen inspirierte Unterrichtsforschung ‚inhaltsleer' sei, mit anderen Worten, dass sie Unterricht als Lehr-, Lern- und Interaktionssituation weitgehend ohne Beachtung der konstitutiven Bedeutung der Gegenstände, die im Unterricht verhandelten werden, um deren Vermittlung und Aneignung Unterricht und Schule überhaupt institutionalisiert worden seien. Dem Hinweis, dass Inhaltsfragen in der Curriculumforschung bearbeitet werden, wurde und wird von erziehungswissenschaftlicher und didaktischer Seite entgegengehalten, dass die formale und materiale Seite des Unterrichtsgeschehens eben unauflöslich miteinander verbunden seien; die Aufteilung in Curriculumforschung hier und Unterrichtsforschung dort sei eben problematisch. Durch das sehr starke Wachstum der *fachdidaktisch* ausgerichteten Unterrichtsforschung (quantitativer wie qualitativer Art) ist dieses Argument mit Erfolg relativiert worden. Lindmeier & Heinze (in diesem Heft) machen auf diese positive Entwicklung, aber auch auf die damit verbundenen Probleme sehr schön aufmerksam (vgl. zum Thema auch Martens et al., 2018). Bei den fachunabhängigen Bemühungen um eine allgemeine (generische) Theorie des Lehrens bzw. des Unterrichts treten naturgemäß die Inhalte wieder zurück – das grundlegende Dilemma jeder Allgemeinen Didaktik holt mithin auch die neuere theoretische und methodische Debatte in der empirischen Unterrichtsforschung ein.

Die grundlegende Herausforderung ist darin zu sehen, dass mit der Frage nach den Inhalten des Unterrichts unmittelbar *normative* Fragen aufgeworfen werden, die hier nur angedeutet werden können: Was soll warum auf welchem Niveau welchen Schülergruppen im Unterricht bzw. allgemeiner: im Laufe ihres schulischen Bildungsganges angeboten und vermittelt werden – und was wird damit implizit wem vorenthalten? An

---

tionsquoten etc.) in der Breite der einschlägigen englischsprachigen internationalen Fachdiskussion geblieben.

welchen normativen und inhaltlichen Grundprinzipien soll sich die Komposition der Lehrpläne für schulische Bildungswege und -abschlüsse orientieren, wenn als Aufgabe nicht nur die Tradierung von und Initiation in Kultur, sondern auch die Entfaltung individueller Potentiale sowie die Fortschreibung erreichter kulturellen Niveaus definiert wird – und das nicht nur für die Gegenwart, sondern vor allem mit Blick auf die Zukunft der Schülerinnen und Schüler (ausführlicher dazu Terhart, 2019b)?

Learning Analytics, Big Data und Didaktik: Stellt man den Trend zur zunehmenden Digitalisierung aller menschlichen Lebensbereiche und -vollzüge in Rechnung, so muss man sich der Tatsache stellen, dass nicht nur das Lehren und Lernen in der beruflichen Schulung und Weiterbildung bzw. generell im informellen lebensweltlichen Lernen im Alltag zunehmend digitalisiert wird, sondern zunehmend auch der schulische Unterricht. Die aktuelle „DigitalPakt Schule" von Bund und Ländern mag da nur ein äußeres, beinahe rührendes Indiz sein. Blickt man in die Zukunft, so wird alles in allem ein immer größerer Teil des Lehrens und Lernens in der Schule auf digitalisierter Basis verlaufen. Dadurch werden begleitend und ‚automatisch' riesige Datenmengen erzeugt, die einen bisher nicht möglichen Einblick in schulisches Lehren, Lernen und Interagieren und deren Zusammenhang ermöglichen. Schon jetzt ist absehbar, dass das formative und summative Erfassen und Bewerten der Lernfortschritte und -leistungen der Schüler maschinell erfolgen kann, zumindest eine neue Basis für das weiterhin bestehende Lehrerurteil liefern kann. Und weiter: Wie in vielen anderen Berufsfeldern mittlerweile üblich, wird man in dem Zusammenhang auch die berufliche Kompetenz der Lehrkräfte und ihre Folgen kontinuierlich datenbasiert erfassen können. Diese Datenmengen sind derart groß, dass vermutlich künstliche Intelligenz eingesetzten werden wird, um solche Analysen durchzuführen. Das gesamte Bildungsmonitoring – von der einzelnen Schulstunde bis hin zur Gesamtebene eines Bildungssystems – würde damit auf ein neues Fundament gestellt. Die technischen, rechtlichen, pädagogischen, kulturellen etc. Probleme und Folgen sind gegenwärtig nur zu ahnen.[8]

8 In China wird mit smarten Schuluniformen der Schülerinnen und Schüler und der mittels Kameras erfolgenden Erfassung von Unterricht (Schülerinnen und Schüler und Lehrkräfte) auf breiter Front experimentiert; dies ist Teil des staatlichen, auf jede einzelne Bürgerin bzw. jeden einzelnen Bürger gerichteten allgemeinen *social credit system* (vgl „Gläserne Schüler", in: FAS, 28.4.2019; „Chinesische Schulen überwachen Schüler per Uniform", in: ZEIT ONLINE, 21.12.2018; „Chinas intelligenter Schule entgeht nichts", in: Deutschlandfunk, 21.2.2019).

Literatur

Biesta, G. J. J. (2017). *The rediscovery of teaching.* New York: Routledge.

Bohl, T., Hanke, U., Koch-Priewe, B., & Zierer, K. (Hrsg.) (2013). *Neuere Ansätze in der Allgemeinen Didaktik* (Jahrbuch Allgemeine Didaktik). Baltmannsweiler: Schneider Verlag.

Cappella, E., Aber, J. L., & Kim, H. Y. (2016). Teaching beyond achievement tests. Perspectives from developmental and educational science. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching.* (Fifth Edition, S. 249–347). Washington: American Educational Research association.

Chalmers, A. F. (2007). *Wege der Wissenschaft. Einführung in die Wissenschaftstheorie.* Berlin: Springer.

Drerup, H. (1979). *Wissenschaftstheorie und Wissenschaftspraxis. Probleme der Vermittlung zwischen metawissenschaftlichen Forschungsprogrammen und einer Praxis der Sozial- und Erziehungswissenschaft.* Bonn: Bouvier.

Gitomer, D. H., & Bell, C. A. (Hrsg.) (2016). *Handbook of research on teaching* (Fifth Edition). Washington: American Educational Research Association.

Hall, G. E., Quinn, L. F., & Gollnick, D. M. (Hrsg.) (2018). *The wiley handbook of teaching and learning.* New York: Wiley.

Hopmann, S., & Riquarts, K. (Hrsg.) (1995). *Didaktik and/or curriculum.* Kiel: IPN.

Kane, R., & Marsh, C. J. (1980). Progress towards a general theory of instruction? *Educational Leadership, 38*(3), 253–255.

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

Lipowsky, F. (2015). Unterricht. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 69–105). Berlin: Springer.

Loser, F., & Terhart, E. (Hrsg.) (1977). *Theorien des Lehrens.* Stuttgart: Klett.

Martens, M., Rabenstein, K., Bräu, K., Fetzer, M., Gresch, H., Hardy, I., & Schelle, C. (Hrsg.) (2018). *Konstruktionen von Fachlichkeit. Ansätze, Erträge und Diskussionen in der empirischen Unterrichtsforschung.* Bad Heilbrunn: Klinkhardt.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305–322.

MacMillan, C. J. B., & Garrison, J. W. (1988). *Logical theory of teaching. Erotetics and intentionality.* Dordrecht: Kluwer Academic Publishers.

Nickel, H. (1976). Die Lehrer-Schüler-Beziehung aus der Sicht neuerer Forschungsergebnisse. *Psychologie in Erziehung und Unterricht, 23*(2), 153–172.

Proske, M., & Rabenstein, K. (Hrsg.) (2018). *Kompendium Qualitative Unterrichtsforschung. Unterricht beobachten – beschreiben – rekonstruieren.* Bad Heilbrunn: Klinkhardt.

Rothland, M. (2018). Allgemeine Didaktik und empirische Unterrichtsforschung als Teilgebiete der Schulpädagogik. *Die Deutsche Schule, 110*(4), 369–382

Saam, N. J. (2009). Modellbildung. In S. Kühl, P. Strodtholz & Taffertshofer, A. (Hrsg.), *Handbuch Methoden der Organisationsforschung* (S. 517–513). Wiesbaden: VS Verlag für Sozialwissenschaften.

Stachowiak, H. (1973). *Allgemeine Modelltheorie.* Wien: Springer.

Steffens, U., & Messner, R. (Hrsg.) (2019). *Unterrichtsqualität. Konzepte und Bilanzen gelingenden Lehrens und Lernens.* Münster: Waxmann.

Terhart, E. (2016). „Research on teaching" in the USA and „Didaktik" in (West-)Germany. Influences since 1945. In J. Overhoff & A. Overbeck (Hrsg.), *German-American educational history: Topics, trends, fields of research* (*Studien zur Deutsch-Amerikanischen Bildungsgeschichte, Band 1,* S. 159–174). Bad Heilbrunn: Klinkhardt.

Terhart, E. (2019a). *Didaktik. Eine Einführung* (Neuausgabe). Stuttgart: Reclam.

Terhart, E. (2019b). Die Frage nach den Inhalten schulischen Lehrens und Lernens. Alte und neue Antworten. In L. Haag & K. Zierer (Hrsg.), *Unterrichten wir das „Richtige"? – Die Frage nach zeitgemäßen Bildungsinhalten der Schule.* (Jahrbuch für Allgemeine Didaktik 2018, S. 169–185). Bad Heilbrunn: Klinkhardt.

Westbury, I., Hopmann, S., & Riquarts, K. (Hrsg.) (2000). *Teaching as a reflective practice: The German Didaktik tradition.* Mahwah: Erlbaum.

Abstract: In this commentary on the present supplement, the claim of the editorship is received positively. After a short sketch of the current debate about theory and research methods in empirical research on teaching, a few topics are selected and commented on. These topics are: the model of the three basic dimensions of teaching, the understanding of 'theory' when talking about theories of teaching, and finally the so-called offer-use-model of classroom teaching with its background, variations and limitations. After a discussion of the international character of German research on teaching and teaching quality, two additional aspects are briefly pointed out that – in part – lie beyond the scope of the supplement.

Keywords: Quality of Teaching, Theories of Teaching, Models of Teaching, Research on Teaching, Internationalization

Anschrift des Autors

Prof. i. R. Dr. Ewald Terhart, Westfälische Wilhelms-Universität Münster,
Institut für Erziehungswissenschaft,
Georgskommende 26, 48143 Münster, Germany
E-Mail: ewald.terhart@uni-muenster.de

Kurt Reusser

Unterrichtsqualität zwischen empirisch-analytischer Forschung und pädagogisch-didaktischer Theorie

Ein Kommentar

Zusammenfassung: Nachfolgend werden die Beiträge des Beihefts aus einer psychologisch-didaktischen Perspektive kommentiert. Nach einleitenden Bemerkungen zur Stellung des Themas der Unterrichtsqualität in der bildungswissenschaftlichen Diskussion folgen zuerst allgemeine Bemerkungen zu den bearbeiteten fünf Theoriesträngen und deren Verbindung zueinander. Anschließend folgen selektive Kommentare insbesondere zu den Themen Basisdimensionen (als Beispiele für die Tiefenqualität) des Unterrichts, zum Angebots-Nutzungsmodell und zur Perspektivenabhängigkeit der Wahrnehmung von Unterrichtsqualität. Der Beitrag schließt mit Bemerkungen zu Desideraten der empirisch-quantitativen Unterrichtsforschung, insbesondere vor dem Hintergrund eines beobachtbaren didaktischen Gestaltwandels der Schule.

Schlagworte: Unterrichtsqualität, Didaktik, Unterrichtsforschung, Unterrichtstheorie, Lehren

1. Einleitung

Die Frage nach der Qualität von Unterricht beschäftigt die pädagogische Reflexion seit ihren Anfängen. Erst mit der Entwicklung der empirischen Bildungsforschung wurde jedoch damit begonnen, Unterrichtsqualität nicht mehr allein an in der pädagogischen Tradition ausgeformte – über pädagogisch-didaktische Prinzipien transportierte – Normen und Ergebniserwartungen, sondern an messbare Wirkungen wie das Erreichen fachlicher und überfachlicher Kompetenzen zu koppeln. Im Sinne von Berliner (2005): an ein Qualitätsverständnis von Unterricht, das sich weniger an pädagogisch-gesellschaftlichen Normen orientiert (good teaching), sondern daran, ob Bildungsziele auch tatsächlich erreicht werden (successful teaching; Klieme, 2019; Reusser, 2008). Im Zuge eines von theoretischen und methodischen Innovationen begleiteten Aufschwungs einer transdisziplinär und multimethodisch arbeitenden empirischen Bildungsforschung hat sich im deutschsprachigen Raum die Dynamik und das Kräftespiel um die Deutungshoheit in Bildungsfragen und insbesondere in der Lehrkräftebildung deutlich verändert. Gleichzeitig ist die Zahl der Partner und Disziplinen, die unser Wissen über qualitätsvollen Unterricht repräsentieren und die in der Bildung von Lehrpersonen zusammenarbeiten müssen, größer geworden. Während die (forschungsferne) Allgemeine Didaktik als jahrzehntelange Hüterin der Tradition und Referenzdisziplin zum Thema *guter*

Unterricht in den vergangenen Jahren massiv unter Druck geraten ist, hat die pädagogisch-psychologisch-sozialwissenschaftliche Bildungsforschung und in deren Schlepptau die Fachdidaktik an Reputation zugelegt – nicht nur bei der mehr-methodischen wissenschaftlichen Erforschung des Unterrichts, sondern auch in der Lehrkräftebildung. Erkenntnisse einer in Bezug auf theoretische Ansätze, Forschungsmethoden sowie Disziplinen vielfältigen, internationalen Unterrichtsforschung finden über Metastudien (prominent: die Hattie-Studie) sowie über neue Lehrmittel zunehmend Eingang in die Ausbildung und in den Berufsalltag von Lehrkräften. Sodann haben nebst TIMSS, PISA und Co vor allem eine Reihe von (inter)nationalen videobasierten Unterrichtsstudien dazu beigetragen, dass eine allgemein und fachdidaktisch ausgerichtete, quantitative (empirisch-analytische) *und* qualitative (phänomenologisch-hermeneutische) Unterrichtsforschung im deutschsprachigen Raum Fuß gefasst und zu einem sehr aktiven Teil einer zunehmend transdisziplinär arbeitenden Bildungsforschung geworden hat. Auch wenn es der Forschungslandschaft noch deutlich an Kohärenz fehlt, gemeinsame Verständnishorizonte Mangelware sind und theoretische sowie methodologische Abgrenzungen zwischen wie innerhalb von Paradigmen die akademische Diskussion prägen, ist das Thema der Unterrichtsqualität – im Sinne der Identifikation von inhalts- und prozessbezogenen Merkmalen und Praktiken, von denen vermutet wird, dass sie mit dem Erreichen von (wünschbaren) Bildungszielen in bedeutsamen Zusammenhängen stehen – in der Mitte der Diskussion in den Bildungswissenschaften, auch in der Erziehungswissenschaft, angekommen. Das vorliegende Beiheft legt davon Zeugnis ab.

Dass die prominente Stellung, die das Thema der Unterrichtsqualität in der aktuellen bildungswissenschaftlichen Diskussion einnimmt, eine durchaus neue Situation darstellt, zeigt der Blick zurück in eine (nicht ferne) Zeit, als die klassischen Modelle der Allgemeinen Didaktik das Meinungsfeld (bezüglich eines normativen Verständnisses von gutem Unterricht) bestimmten. So galt in der geisteswissenschaftlich-bildungstheoretischen Didaktik lange Zeit das auf Klafki (1964) zurückgehende Diktum des *Primats der Didaktik vor der Methodik.* Das heißt, im Zentrum des didaktischen Denkens stand die *WAS-Frage* des Unterrichts: *Was* soll mit welchem *Bildungsanspruch,* in welcher *sachlichen Ordnung* und mit welchen *Bildungs- und Sozialisationszielen in den Fächern* gelehrt werden, und welche Stoffe und Aufgaben ermöglichen in exemplarischer Weise Zugänge zu Wissen und Verstehen. Vor dieser bildungstheoretischen, das Gravitationszentrum der Didaktik prägenden Frage trat die *WIE-Frage* der methodischen Gestaltung der Unterrichtsprozesse in den Hintergrund, bzw. wurde als nachrangige Frage von niedrigerer Wertigkeit betrachtet. Auch wenn praktizierende Lehrpersonen sich seit jeher täglich fragen mussten, *auf welche Weise* (in welcher Konfiguration von Methoden, Sozial- und Unterrichtsformen), *in welchem funktionalen Nacheinander* (Artikulation, Lernzyklus) und assoziiert mit *welchen angestrebten Wissensniveaus und Prozessqualitäten* (elementares Wissen aufbauen, Konzepte verstehen, das Gelernte anwenden, dabei Arbeitsmethoden und Lernstrategien erwerben) gelehrt und gelernt werden soll, gehörten diese Fragen nicht zum geisteswissenschaftlich- bildungstheoretischen Kern didaktischen Denkens, sondern wurden der individuellen Lehrkunst bzw. der Praxis zugeordnet. Vor diesem Hintergrund hatten psychologisch-didaktische Ansätze und ganz

allgemein die (bedingungsanalytisch arbeitende) empirisch-psychologische Lehr-Lern-Forschung in der geisteswissenschaftlich ausgerichteten, deutschsprachigen Erziehungswissenschaft über Jahrzehnte einen schweren Stand. So wurde die von Piagets epistemologischem Konstruktivismus inspirierte, am Lernen des Kindes orientierte *psychologische Didaktik* von Hans Aebli (1951, 1961) von der (akademischen) deutschen Didaktik jahrzehntelang ignoriert und als *technologisch,* als Methodik ohne bildungstheoretischen Tiefgang beargwöhnt (vgl. dazu Kiper, 2006; Messner & Reusser, 2006). Dies ungeachtet des Umstandes, dass es sich um einen in der Tradition Herbarts und Deweys stehenden, pädagogisch motivierten Ansatz handelte, der den problemorientierten *Wissensaufbau* der *Schüler/innen* auf der Tiefenebene des fachlichen Verstehens und der Denkfähigkeit zu seinem Ankerpunkt machte, und ungeachtet des Sachverhalts, dass Aeblis *Grundformen des Lehrens* seit 1961 in Deutschland sehr rege gelesen wurden und die aus der Psychologie Piagets hervorgehende, denkpsychologisch orientierte *operative Didaktik* sehr früh auch in die Mathematikdidaktik Eingang gefunden hat (vgl. Wittmann, 1974).[1]

2. Übergreifende Bemerkungen zum Theorieanspruch und zu den Schwerpunkten des Beiheftes

Die im Beiheft versammelten, von einem interdisziplinären Kreis von Unterrichtsforschenden im Rahmen eines von der Leibniz-Gemeinschaft geförderten Netzwerks verfassten Beiträge lassen sich dem quantitativ-empirischen Paradigma der bedingungsanalytisch arbeitenden Unterrichtsforschung zuordnen. Den Ankerpunkt des Heftes bildet der im ersten Heftbeitrag von Praetorius et al. (in diesem Heft) theoretisch-systematisch erörterte und in aktuelle, erziehungs- und sozialwissenschaftliche Diskussionszusammenhänge eingeordnete Begriff der *Unterrichtsqualität* in seiner Bedeutung von *effective* bzw. *successful teaching* (im Sinne von Berliner, 2005). Die Beitragenden gehen – mit guten Gründen – davon aus, dass die dem Prozessgeschehen des Lehrens und Lernens zugewandte, empirisch-analytische Unterrichtsforschung als eines der aktivsten Segmente der deutschsprachigen Bildungsforschung in den vergangenen Jahren erhebliche Fortschritte gemacht hat und es angezeigt ist, den Theorie- und Forschungsstand zum Konzept lernwirksamen Unterrichts zusammenhängend darzustellen und gleichzeitig weiter voranzubringen. Insgesamt möchte das Heft zur Schärfung und weiteren Ausdifferenzierung von empirisch erhärteten, auf der Basis erziehungs- und sozialwissenschaftlicher Theorien gewonnener Aussagen über zentrale Merkmale von

1 Zwar hat sich die *lehrtheoretische (Berliner,* später*: Hamburger) Didaktik* um Heimann, Otto und Schulz (1965) ebenfalls mit dem Prozess des Unterrichtens beschäftigt. Heimann schlug vor, den empirisch schwer zugänglichen Bildungsbegriff durch die beobachtungsnähere Kategorie des *Lernens* zu ersetzen. Allerdings bestand der Mangel des Ansatzes darin, dass vor allem in abstrakten *Strukturmodellen* gedacht und empirisch-psychologische Prozessqualitäten des Unterrichts und des Lernens der Schüler/innen kaum thematisiert wurden.

Unterricht und das damit verbundene Bedingungs- und Wirkungsgefüge beitragen und damit dem international beobachtbaren „lack of theorizing and systematic revision of theories in quantitative research on teaching“ (Praetorius et al., S. 17, in diesem Heft) entgegenwirken.

Der Herausgeberschaft ist bewusst, dass mit der Beschränkung auf empirisch-quantitativ ausgerichtete Beiträge, die sich überwiegend mit Mathematik als Unterrichtsgegenstand beschäftigen, der Prozess des Unterrichts nicht in der inhaltlichen Breite seiner Bedeutung abgedeckt wird. Gleichwohl ist die Anlage des Bandes von einer erstaunlichen Vielfalt, indem nicht weniger als fünf gewichtige Theoriestränge einbezogen werden, die in den vergangenen Jahren – verteilt auf zahlreiche Forschungsgruppen und Netzwerke – viel beachtete Themen der deutschsprachigen Unterrichtsforschung darstellten: Neben (i) dem seit Jahren intensiv beforschten *Modell der Basisdimensionen* handelt es sich um (ii) das breit diskutierte *Angebots-Nutzungs-Modell,* (iii) die aus immer mehr Studien als Destillat sich ergebende *Unterscheidung zwischen Oberflächen- und Tiefenstrukturen,* (iv) die vor allem im deutschsprachigen (kaum je im US-amerikanischen) Raum als Theorieproblem erkannte *Perspektivenabhängigkeit der Einschätzung von Unterrichtsqualität* sowie (v) um Überlegungen zur Frage, mit welchen *Instrumenten und Analysestrategien die Wirkungen von qualitätsvollem Unterricht überprüft werden können.* Das Modell der (drei) Basisdimensionen bildet dabei so etwas wie den Ankerpunkt des Heftes, die daran anschließenden Folgeteile beschäftigen sich mit zur Erklärung der Wirkung von Unterricht auf Schülerinnen und Schüler herangezogenen (auf die Basisdimensionen Bezug nehmenden) theoretischen Konstrukten. Alle fünf Theorie- und Forschungsstränge werden im Heft durch je zwei Beiträge abgedeckt.

Das Beiheft ist in mehrfacher Hinsicht sehr anregend. Auch wenn die Diskussion über Unterrichtsqualität auch in Zukunft zwischen einem normativen und einem empirischen Diskurs oszillieren wird, lohnt sich die Lektüre von Beiträgen, die sich der im Heft klar verfolgten Strategie der empirisch-analytischen Forschung zuordnen und das Paradigma in allen seinen Stärken und Schwächen dokumentieren. Aus meiner Sicht besteht die augenfälligste Stärke des Bandes darin, dass es sich nicht um einen herkömmlichen Sammelband handelt, dessen Beiträge bloß rhetorisch zusammengefügt sind. Der Band stellt das (Zwischen)Ergebnis eines Netzwerks dar, in dem nach einem Verständnis von Unterrichtsqualität gesucht wird, das über die seit Brophy (2000)[2], Meyer (2004) und Helmke (2006) (u. a.) kursierenden Listen von empiristisch (mit geringer theoretischer Rahmung) gewonnenen Qualitätsmerkmalen hinausgeht. Die das Heft ausrichtenden fünf Theorie- und Forschungsthemen lassen sich als Pfeiler einer Gesamtsicht verstehen, die anders als bloß variablenzentrierte Ansätze die Messlatte für eine empirische Unterrichtsforschung deutlich höher ansetzt und eine konzeptuelle

2 Jere Brophy (2000) hat in einer für die International Academy of Education der UNESCO verfassten Broschüre mit dem Thema *teaching* zwölf Merkmale identifiziert, welche als generisch verstandene Qualitätsmerkmale des Lehrens seither vielfach aufgegriffen worden sind.

Rahmung mit einem hohen Theorieanspruch anstrebt. Die Herausgeberschaft und auf weite Strecken auch die übrige Autorschaft stellen sich dieser spannenden Herausforderung zudem auf sehr selbstkritische Weise.

Dass im Heft eine mehrere Forschungsstränge verbindende – Theoriebildung angestrebt, vorgedacht und in Teilaspekten auch realisiert wird, erkennt man vor allem an dem in den Heftbeiträgen omnipräsenten Modell der von zahlreichen Forschungsgruppen aufgegriffenen Basisdimensionen von Unterrichtsqualität, dessen Statusklärung und Weiterentwicklung ein erklärtes Ziel des Bandes ist, und zahlreichen damit verbundenen Fragen. Dazu gehört beispielsweise die Erörterung, wie viele und welche Basisdimensionen es sein sollen (Praetorius et al. sowie Kleickmann, Steffensky & Praetorius, in diesem Heft), wie generalisierbar sie sich empirisch verankern und über einheitliche Indikatoren abbilden lassen. Des Weiteren, wie sich die Basisdimensionen zur Unterscheidung von Oberflächen- und Tiefenstrukturen oder zum Angebots-Nutzungs-Modell verhalten. Auch wenn das Heft keine die fünf Stränge integrierende Theorie anbietet, ist das Bemühen nach Verbindung und Kohärenz gegenwärtig. Das äußert sich etwa im Bestreben, nicht bloß selbstreferenziell im deutschsprachigen Theoriekontext zu argumentieren, sondern die eigene Forschung näher an den teils anders tickenden anglo-amerikanischen Forschungskontext heranzuführen (vgl. dazu auch weitere Diskussionsbeiträge in diesem Heft).

Wie steinig der Weg ist, um zu einem integrativen und empirisch tragfähigen Modell zu kommen, zeigt nicht nur der selbstkritische Blick der Autorschaft des Eingangskapitels (Praetorius et al., in diesem Heft) auf die Theoriequalität des Modells der Basisdimensionen, sondern ebenfalls der Blick auf die Herkunft der mit damit zu verknüpfenden weiteren Theoriestränge. Auch wenn die in der deutschsprachigen Unterrichtsforschung und Didaktik geformten thematischen Stränge ein denkanregendes Konglomerat von Begriffen und Theorietraditionen bilden, entstammen sie sehr unterschiedlichen Kontexten und sind auch unabhängig voneinander entstanden. Während das *Konzept der generischen Basisdimensionen* faktorenanalytisch aus der durch eine längsschnittliche Mehr-Ebenen-Studie ergänzten TIMSS-Videostudie gewonnen wurde, stammt das – wesentlich allgemeinere (!) – *Angebots-Nutzungs-Modell* ursprünglich von Fend (1980) – bevor es von Helmke und andern formalisiert und verbreitet wurde. Ganz anders verhält es sich mit der in der Aebli-Tradition (Aebli, 1961) der Didaktik in der Schweiz entstandenen, durch die Ergebnisse der Forschungssynthese von Hattie und weitere Forschungsstudien (z. B. Pythagoras) zunehmend empirisch aufgeladene *Unterscheidung von Oberflächen- und Tiefenstrukturen.* Bei aller Plausibilität und Akzeptanz, welche die Unterscheidung mittlerweile geniesst, handelt es sich ebenfalls um kein wirklich ausgeschärftes Konstrukt. Die beiden Ebenen lassen zwar im ersten begrifflichen Zugriff gut begründen, jedoch fällt eine trennscharfe theoretische und empirische Unterscheidung nicht leicht, was die beiden Kapitel des Beiheftes deutlich machen. Nochmals ganz anders verhält es sich mit der *Perspektivenabhängigkeit* der Erfassung von Unterrichtsqualität und damit der Frage, inwieweit überhaupt *wahre Werte* von Qualität angenommen werden können, wenn nicht sicher ist, dass alle Beobachtenden einer identischen Unterrichtssequenz dasselbe sehen bzw. dieselbe Sprache sprechen –

was nicht nur über Beobachtende/Akteure (Lehrpersonen, Schülerinnen und Schüler, externe Expertinnen und Experten), sondern auch über kulturelle Kontexte hinweg alles andere als selbstverständlich ist.

Das heißt, um die Theoriestränge zu etwas Gemeinsamem zusammenzufügen, bräuchte es weitere Rahmenüberlegungen und Ausarbeitungen der Teilbereiche. Sodann müsste man sich verständigen, wie eine integrative Theorie überhaupt aussehen könnte, und was diese für welche Ziele und für wen leisten sollte. Geht es vor allem um bedingungsanalytisches Erklären (explanation) in einem naturgesetzlichen, Hempel-Oppenheim-Sinn, oder geht es auch um Brückenbau zwischen Pädagogischer Psychologie und Schulpädagogik/Didaktik – oder um beides? Unrealistisch erschiene mir die Orientierung an einem Theorieverständnis, das sich exklusiv an ein mit universalistischen Generalisierungsansprüchen verbundenes, nomothetisches Verständnis von Forschung anlehnt. Bereits bei den im Lead-Kapitel des Beiheftes (Praetorius et al., in diesem Heft) ausbuchstabierten Metakriterien von Kane und Marsh (1980), von denen die Autorinnen und Autoren, wie sie sagen, *pragmatisch Gebrauch machen,* ergeben sich mir Zweifel, ob diese breit genug sind und inwieweit sie sich auf ein komplexes Modell von Unterricht und als dem (von Normen und Zielen mit gesteuerten) intentionalen Handeln von Lehrpersonen anwenden lassen. Sodann stellt sich die grundsätzliche Frage, ob das den Kriterien zugrundeliegende wissenschaftstheoretische Verständnis einer im klassischen Sinne explanativen und zu Voraussagen führenden Theorie (a) noch unangefochten zeitgemäß ist (vgl. Terhart, in diesem Heft) und (b) sich für das Feld einer auch didaktisch tragfähigen Unterrichtstheorie eignet (vgl. auch Bell, in diesem Heft). Da es sich beim Unterricht (i. S. von *instruction*) und dem mit diesem assoziierten Lehrhandeln *(teaching)* um einen genuin kulturell geprägten, eine hohe personen- und kontextabhängige Situativität aufweisenden Forschungsgegenstand handelt, dürfte die Reichweite universalistischer Erklärungsweisen begrenzt sein. Dies zeigt sich auch an den Herausforderungen, pädagogische Unterrichtskulturen innerhalb eines Schulsystems sowie komparativ über Systeme und (nationale) Kulturen hinweg zu rekonstruieren. Gegenüber einem rigorosen (naturwissenschaftlichen) Theorieanspruch skeptisch zu sein, bedeutete dabei keineswegs, auf den quantitativen Apparat der Forschung zu verzichten (wozu ja auch PISA gehört), diesen aber nicht in einem strengen Sinn nur dafür einzusetzen, um naturgesetzlich *wahre Werte und Strukturen* zu finden, sondern um pädagogisch-kulturelle (normativ aufgeladene), konfigurative Handlungs- und Wahrnehmungsmuster von Unterricht herauszuarbeiten. Man darf in diesem Zusammenhang gespannt sein, welche Ergebnisse die das Leibniz-Netzwerk umklammernde, einen ganzheitlichen Blick auf den Mathematikunterricht anstrebende TALIS-Video-Studie, die ein grosses multimethodisches Potenzial mit Bezug auf die Rekonstruktion pädagogisch-kultureller Kontexte innerhalb Deutschlands sowie der an der Studie beteiligten Länder aufweist, dazu bringen wird.

Die das Beiheft gliedernden Theoriestränge *in Kombination* zu denken ist nicht nur von einem psychologisch-prozesstheoretischen Verständnis von Unterricht her interessant. Auch aus einer didaktischen Optik erscheint sie ertragreich. Obgleich in mehreren Beiträgen handlungsbezogene Dimensionen angesprochen werden, wird das Verhält-

nis zwischen einer empirisch fundierten und gleichzeig didaktisch ertragreichen Unterrichtsforschung (vgl. Reusser, 2008, 2019) im Heft nirgends angesprochen. Auch aus einer didaktischen Perspektive bedeutet Unterricht ein durch Angebote und seine Nutzungen geprägtes, durch Choreographien gesteuertes Interaktionsgeschehen zwischen Lehrenden und Lernenden auf der Oberflächen- und der Tiefenebene. Das didaktische Dreieck als Bild für die Grundstruktur des schulpädagogischen Fragezusammenhangs (vgl. Reusser, 2008, 2019) würde sich dazu als Rahmung anbieten. Der *heuristische* Wert des didaktischen Dreiecks besteht darin, dass die *fachübergreifenden Handlungsaufgaben des Unterrichts, die sich auch* in den klassischen Modellen der Allgemeinen Didaktik spiegeln, in den Blick genommen werden: Die Herstellung (didaktische Konstruktion) einer *bildungsinhaltlichen Ziel- und Stoffkultur, einer prozessbezogenen Lern- und Verstehenskultur sowie einer personalen Unterstützungs- und Interaktionskultur.* Die drei pädagogischen Teilkulturen stehen in ihrer Interdependenz für die transfachlichen Grundqualitäten von Unterricht (vgl. Reusser, 2008, 2019). Eine didaktische Rahmung hätte den Vorteil, dass die im Beiheft angesprochenen Qualitätsdimensionen und die mit ihnen verbundenen Theorieprobleme und Forschungsthemen nicht nur aus einer empirisch-psychologischen Prozessoptik, sondern mitsamt ihren normativen, fachdidaktisch und bildungstheoretisch konnotierten Implikationen (als Reflexion darüber, welche Inhalte und human-kulturellen Werten die Gesellschaft will, dass sie von der Schule als Ziele verfolgt und in der Prozessgestaltung von Unterricht umgesetzt werden) in den Blick kämen.

3. Basisdimensionen – wie viele, welche, und in welcher inneren Ordnung?

Bei dem im Beiheft als Theoriekonzept behandelten Nukleus der Basisdimensionen stellt sich Frage warum (außer aus Gründen seiner faktorenanalytischen Herleitung aus der TIMSS-Videostudie), die Zahl DREI (oft in Verbindung mit dem definiten Artikel) in der Bezeichnung des Konstrukts stets mitgeführt wird. Sollte die Frage nach ihrer Anzahl, inneren Ordnung und nach ihrem Theoriestatus[3] gerade nach der in den Beiträgen deutlich werdenden, durchaus kritischen Befunden zur Stabilität und zur Generalisierbarkeit (über Stufen, Fächer, Schultypen, Kulturen) nicht offener gelassen, und könnte die Befunde nicht umfassender reflektiert werden? Dies bei aller eindrücklichen Evidenz bezüglich der durch sie ermöglichten, in zahlreichen Untersuchungen belegten (differenziellen) Abbildbarkeit von Unterrichtsbeobachtungen und -wahrnehmungen. Wie ist der Status eines Konstrukts zu beurteilen, wenn es keine anerkannte Art der Operationalisierung dazu gibt, dessen generische Qualität (auch durch die Außerachtlassung

3 Die insbesondere auch wissenschaftstheoretisch vertiefungsfähige Frage nach dem Theoriestatus (nicht nur der TBD, sondern allgemein einer Theorie des Unterrichts!) wird im Beitrag von Praetorius et al. (in diesem Heft) in Fußnote 3 nur kurz angesprochen (vgl. für eigene weiterführende Überlegungen auch Reusser, 1983).

fachdidaktischer Variabilität) genuin unscharf bleiben muss, und dessen Merkmale vor allem im Kontext der Mathematik (also in spezifischer Weise kontextabhängig) erhoben worden sind? Auch wenn bei der kognitiven Aktivierung und bei der konstruktiven Unterstützung eine inhaltliche Auseinandersetzung mit Inhalten stets mitgemeint ist, fehlt insbesondere diesen beiden Dimensionen im Prinzip die Inhaltlichkeit (was Lindemeier & Heinze, in diesem Heft, als Paradoxie bezeichnen). Eine weitere Frage bezieht sich auf die konstruktive Unterstützung: Handelt es sich bei der kognitiv-gegenstandsbezogene sowie affektive Merkmale beinhaltenden konstruktiven Unterstützung wirklich um *eine* Dimension (vgl. Kleickmann et al., in diesem Heft)? Anders gefragt: wie würde das empirische Ergebnisbild aussehen, wenn man diese der Perspektiven-Referenz-Matrix, wie sie im Beitrag von Fauth, Göllner, Lenske, Praetorius und Wagner (in diesem Heft) am Beispiel der Klassenführung durchdekliniert wird, unterziehen würde? Und stehen die drei Basisdimensionen als induktiv gewonnene Kondensate der Unterrichtsqualität nebeneinander, oder sind auch anders geordnete, z. B. hierarchische Beziehungen denkbar?[4] Aus meiner Sicht würde es Sinn machen, die Klassenführung als Grundschicht der Unterrichtsqualität zu betrachten, deren Qualitätsniveau (wenn Unterricht noch funktionieren soll) einen gewissen Schwellenwert nicht unterschreiten darf, während die beiden anderen Dimensionen linear (im Sinne eines *je-mehr-desto-besser*) modelliert werden können.

Stellt man des Weiteren den Befund der mangelnden Stabilität der Dimension der kognitiven Aktivierung auf den Prüfstand der Diskussion, so ergibt sich eine Frage, die uns bereits in der erweiterten schweizerischen TIMSS-Video und danach in der Pythagoras-Videostudie beschäftigt hat. In beiden Studien haben wir die Beobachtung gemacht, dass die Unterrichtsgestaltung zwischen Lektionen in Abhängigkeit von ihrer Stellung im Lernprozess (Einführung vs. Vertiefung) variiert (Hugener, 2008; Pauli & Reusser, 2011). Könnte die basisdimensionale Qualität des Unterrichts nicht auch davon abhängen, ob es sich bei der beobachteten Lektion um eine Einführungs- oder eine Vertiefungs- oder Übungslektion handelt? Da in deutschsprachigen Videostudien zumeist Einführungslektionen analysiert werden, sollte man (bevor man z. B. weitere Dimensionen, etwa *Üben* als Basisdimensionen einführt) die als *analytische Verdichtungen von Unterrichtsmerkmalen* (Klieme, 2019) etablierten Basisdimensionen durch die funktionalen Stufen eines Lernzyklus durchdeklinieren: Was heißt kognitive Aktivierung oder konstruktive Unterstützung (sowohl theoretisch-didaktisch als auch messtechnisch) in der Einführungs-, Durcharbeitungs- oder Übungsphase einer Unterrichtseinheit. Bedeuten sie überall dasselbe? Das heißt, ein wichtiger Grund für die bei geringer Beobachtungsbreite mangelnde Stabilität des Konstrukts der kognitiven Aktivierung könnte in deren Konfundierung mit der Artikulation des Unterrichtsprozesses begründet sein (vgl. auch Praetorius, Pauli, Reusser, Rakoczy & Klieme, 2014).[5]

4 Einen diesbezüglich interessanten Versuch hat Klieme (2011) in einer Reanalyse der Pythagoras-Daten auf der Basis von Arbeiten von Marcus Pietsch gemacht.

5 Analoge Überlegungen ließen sich auch zu fachlich-fachdidaktischen Aspekten des Unterrichts, auf die aus Raumgründen hier nicht eingegangen wird, anstellen.

Den Gedanken einer Konfundierung von Basisdimensionen und Artikulationsstufen fortführend wäre überdies zu erwägen, auch im Sinne einer weiteren Komplettierung des im Beiheft angelegten, multiple Theoriestränge umfassenden Blicks auf Unterricht, die seit Johann Friedrich Herbart[6] als Artikulation des Unterrichts bekannte psychologische Dimension des *Lernzyklus* (meist unter dem missverständlichen Begriff der *Formalstufen* gehandelt) zu den fünf Theoriesträngen hinzuzunehmen. In Bezug auf die Modellbildungen im Beiheft würde damit dem Problem begegnet, dass Unterricht nicht nur hinsichtlich seiner oberflächenstrukturellen Gestaltung sehr variabel ist, sondern auch bezüglich seiner Ziele und Funktionen alles andere als uniform erscheint: *Vollständige* Unterrichts- und Lernzyklen bestehen nicht nur aus Einführungslektionen, auch Übungs-, Durcharbeitungs- und Anwendungslektionen gehören dazu. Aebli hat dazu in seinen Grundformen das PADUA-Modell (mit den psychologischen Funktionsmomenten *Problemorientierter Aufbau, Durcharbeiten, Üben und Anwenden*) entwickelt (Aebli, 1983).

Eine letzte hier zu erwähnende Frage zu den Basisdimensionen bezieht sich auf die Beziehung zwischen den semantischen Feldern *Basisdimensionen* und *Tiefenstrukturen* einerseits und konkreten *Handlungsstrategien und Praktiken des Lehrens und Unterrichtens*[7] andererseits. Die Frage, wie sich wissenschaftliches Wissen (über Basisdimensionen, Tiefenstrukturen bildungswirksamen Lehrens und Lernens) als Kondensat der analytischen Forschung auf das praktische Unterrichtshandeln beziehen lassen, ist deshalb von Bedeutung, weil aus einer bedingungsanalytisch gerahmten (explanativ ausgerichteten) Prozesstheorie des Unterrichts schließlich kompetentes und kompetenzförderndes professionelles Handeln hervorgehen muss. Auch diese im Prinzip schon sehr alte Frage der *Lernübertragung,* des *Transfers* oder der *Anwendung* ist nicht ohne epistemologische Tücken, tangiert sie doch das u. a. vom finnischen Wissenschaftstheoretiker Henrik von Wright (1974) untersuchte Problem von zwei Denkkulturen bzw. die Grundfrage, wie sich wissenschaftliches *Erklären* auf *Verstehen* sowie auf praktisches Handeln beziehen lässt. Auf die Unterrichtsforschung bezogen geht es um das insbesondere die Lehrerinnen- und Lehrerbildung beschäftigende, trotz einer langen Problemgeschichte immer noch unzureichend verstandene Problem, wie das professionelle pädagogische Sehen, Denken und Handeln von Lehrpersonen von der Erklärung der diesem Handeln zugrundeliegenden psychologischen Prozesse profitieren kann. Wie können

---

6 J. F. Herbart (1831) beschreibt die Grundfigur des Lernens im Unterricht als Abfolge von vier fundamentalen erkenntnispsychologischen Stufen im Dienste des geordneten *Aufbaus von Gedankenkreisen:* (1) *Klarheit* über das Vorwissen schaffen; (2) *Assoziation* als Aufbau neuer Wissenselemente; (3) Einbau des Neuen in das *System* des vorhandenen Wissens; (4) durch Einüben das neue Wissen als *Methode* anwendbar machen.

7 Was bildungswirksame Praktiken des Unterrichtens sind, wurde immer wieder in der Literatur thematisiert und zu systematisieren versucht, z. B. als Grundformen des Lehrens von Aebli (1961, 1983) oder – neueren Datums – als *high leverage* bzw. *core practices* von Ball & Forzani (2009), Grossmann & Pupik Dean (2019), Fraefel & Scheidig (2018), vgl. auch Seidel, in diesem Heft. Kennedy beschreibt den Anspruch des Ansatzes als „shifting from a focus on bodies of knowledge to a focus on depictions of practice“ (2006, S. 6).

Theorien ÜBER Unterricht zu Theorien (und Praktiken) FÜR Lehrende und Lernende werden? (vgl. Messner & Reusser, 2000). Der Frage nach tauglichen Brückengliedern zwischen wissenschaftlicher Erklärung und Handlungsfähigkeit, Beschreibungs- und Handlungswissen sollte sich auch die empirisch-analytische Unterrichtsforschung vermehrt – auch als Forschungsproblem (!) – zuwenden. Dies auch deshalb, weil die Frage mit der Kleinteiligkeit und ausufernden Spezialisierung zu tun hat (vgl. Terhart, in diesem Heft), in der Lehr-Lernprozesse (als Interaktion zwischen spezifischen Inhalten, (Gruppen von) individuellen Lernenden und unter bestimmten Kontextbedingungen arbeitenden Lehrpersonen) mit immer ausgefeilteren Methoden und Modellen untersucht und abgebildet werden, und wo sich die Frage stellen lässt, wie *vermittelbar* und an die Praxis anschlussfähig die gewonnen Erkenntnisse z. B. in der Lehrerinnen- und Lehrerbildung sind.

4. Unterricht als koproduktives Handeln – oder das Verschwimmen von Ursachen und Wirkungen in Angebots-Nutzungmodellen

Während Angebots-Nutzungsmodelle im deutschsprachigen Raum weiträumig akzeptiert sind, sind sie als Theoriekonzept im anglo-amerikanischen Raum kaum bekannt. Wohl ist seit langem von situiertem Lernen (Greeno, 1998; Lave & Wenger, 1991) und von der Verschränkung von Lernen und Lehren die Rede. Jedoch wird die Erkenntnis, dass in einem konstruktivistischen Verständnis von Lernen und Lehren – das eigentlich ein ko-konstruktivistisches ist (Reusser & Pauli, 2015) – Unterrichtsangebote von Lernenden aktiv genutzt werden müssen, und deren Nutzungshandeln wiederum auf die nachfolgenden Angebote zurückwirkt, außerhalb des deutschsprachigen Raums durch kein gleichermaßen prägnantes und eingängiges Modell thematisiert.

Zu den Herausforderungen und Implikationen des reziproken Wirkmodells gehört, dass es einfache, von Angeboten zu Nutzungen linear verlaufende, kausale Denkweisen in Frage stellt (Klieme, 2019), und dass dies, wie Vieluf, Praetorius, Rakoczy, Kleinknecht und Pietsch (in diesem Heft) am Ende ihrer vergleichenden Darstellung von mehrdimensional sich unterscheidenden Varianten von Angebots-Nutzungsmodellen herausarbeiten, nicht ohne Folgen für Strategien und Studiendesigns sein kann. Was eine derart gerahmte Forschung anspruchsvoll macht ist, dass Unterricht nach diesem Verständnis kein von der Lehrperson in seiner Prozessqualität deterministisch gestalteter (gemachter), sondern ein ebenfalls von den Lernenden einer Klasse (koproduktiv) ermöglichter Unterricht ist. In der jüngeren Unterrichtsforschung finden sich nicht allzu viele Belege, die sich explizit auf Angebots-Nutzungsmodelle beziehen. Wie anspruchsvoll es ist, Interaktionen im Rahmen von Angebots-Nutzungsmodellen zu untersuchen, zeigt der Heftbeitrag von Meissner, Merk, Fauth, Kleinknecht und Bohl (in diesem Heft). Auch der im Theoriestrang *Perspektivenabhängigkeit* der Unterrichtswahrnehmung verortete Heftbeitrag von Göllner, Fauth, Lenske, Praetorius und Wagner (in diesem Heft) zur Klassenführung kann als Beleg für die Sichtweise reziproker Wechselwirkungen herangezogen werden. Komplexitätssteigernd kommt hinzu, dass

Wechselwirkungen im Rahmen des Angebots-Nutzungs-Denkens ganz allgemein über die Wahrnehmung und Interpretation der Geschehnisse durch unterschiedliche Akteure (Lehrperson, Schüler/innen) vermittelt werden. Damit werden auf kognitive und nichtkognitive Merkmale gerichtete Mediationsprozesse (insbesondere auf der Nutzungsseite) besonders wichtig. Aus einer didaktischen Perspektive, zu der sich im Kapitel von Vieluf et al. (in diesem Heft) ebenfalls einige Bemerkungen finden, besteht die wohl größte Krux des Modells darin, dass Lehrpersonen und Schülerinnen und Schüler zwar vordergründig die Verantwortung dafür, dass gelernt wird, teilen – die Schülerinnen und Schüler sollen ja Mitverantwortung für Ihr Lernen übernehmen. Hintergründig fällt den Lehrpersonen dadurch jedoch eine doppelte Verantwortung zu, da sie nicht mehr allein nur für die methodische und lehrstoffbezogene Angebotsgestaltung verantwortlich sind, sondern auch dafür, dass durch die Unterrichtsgestaltung stets auch die Nutzungsfähigkeiten, vor allem von schwächeren Lernenden gestärkt werden. Ein die Komplexität nochmals erhöhender Umstand besteht schließlich darin, dass Angebots-Nutzungsmodelle sich an einem multikriterialen Wirkungsverständnis orientieren. Angesichts der durch Angebots-Nutzungsmodelle erzeugten Komplexität ist der Bemerkung von Klieme zuzustimmen, dass in der Unterrichtsforschung nicht mehr von Wirkungen, die von einem Input zu einem Output verlaufen, sondern von „Zusammenhängen", gesprochen werden sollte, „weil eine kausale Darstellung von Unterrichtsmerkmalen als ‚Ursache' mehr oder weniger positiver Entwicklungen angesichts der Verschränkung von Angebot und Nutzung kaum möglich ist" (2019, S. 396). Unter dem Aspekt der Integration der im Beiheft behandelten Theoriestränge bedeutet dies, dass sich auch das Wirkungsmodell der Basisdimensionen nicht ohne Weiteres mit dem reziproken Angebots-Nutzungs-Modell verbinden lässt. Legt man das Angebots-Nutzungs-Modell, in dem Ursachen und Wirkungen tendenziell verschwimmen, der Analyse von Wechselwirkungen zwischen Lehrenden, Lerngruppen und einzelnen Lernenden mit ihren je individuellen Lernvoraussetzungen zugrunde, so stellt sich zudem die Frage nach der Ergänzung des quantitativ-analytischen Forschungsinstrumentariums durch phänomenologisch-rekonstruktive, fallanalytische und hermeneutisch-mikroanalytische Verfahren. Der Beitrag von Vieluf et al. (in diesem Heft), in dem die einschlägigen Modelle verglichen werden, schließt mit einem integrierten, angesichts der vorausgehenden Differenzierungen jedoch zwangsläufig abstrakten eigenen Modell, das eindrücklich ist in seiner mehrfachen Rahmung und in den darin sichtbar werdenden Wechselwirkungen. Als heuristisches Modell erscheint es geeignet, nicht nur Denkpfade für Selektionsentscheidungen bei der Planung konkreter Forschungsprojekte zu unterstützen, sondern auch zur Ausschärfung von mit dem Angebots-Nutzungs-Denken assoziierten Theorie- und Forschungsthemen beizutragen.

5. Oberflächen- und Tiefenstrukturen: Von der Suche nach der besten Methode zur Suche perspektiven(un)abhängiger Tiefenqualitäten von Lehren und Lernen

Ein von Hans Aebli angestoßenes, in den vergangenen Jahren u. a. durch die Forschungssynthese von Hattie in seiner Bedeutung bestätigtes Konzept ist die Unterscheidung zwischen Oberflächen- und Tiefenstrukturen als den beiden Qualitätsebenen des Unterrichts. Aebli hat in seinem Werk *Grundformen des Lehrens* (1961) bei jeder Grundform zwischen einem psychologischen und einem methodisch-didaktischen Teil unterschieden. Während sich der psychologische Teil mit der kognitionspädagogischen Tiefenschicht der Lehrformen beschäftigt, thematisiert der didaktische Teil deren wirkungsvolle praktische Umsetzung. Mit der in den Grundformen getroffenen Unterscheidung und der damit verbundenen psychologischen Fundierung des Unterrichtsgeschehens wandte sich Aebli schon sehr früh gegen eine über Jahrhunderte tradierte Auffassung, dass es im Unterricht vor allem auf als didaktische Selbstläufer verstandene Methoden ankomme. Heute ist man sich einig, *dass psychologisch-(fach)didaktische Qualitätsmerkmale methodischen und lernorganisatorischen Merkmalen der Inszenierung überlegen sind.* Insbesondere fachlich konnotierte Unterrichtsmerkmale haben sich als sehr erklärungsstark erweisen (Drollinger-Vetter, 2011; Lipowsky, Drollinger-Vetter, Klieme, Pauli & Reusser, 2018; Reusser & Pauli, 2013). Nicht die Wahl *bestimmter* Methoden und Inszenierungsformen, sondern die *Qualität der Umsetzung eines variablen Sets von Gesamtorientierungen des didaktischen Handelns unter Beachtung von Tiefenqualitäten* (wie kognitive Aktivierung, Verständnisorientierung, konstruktive Lernunterstützung, Beziehungs- und Lernklima) ist für die Qualität des Lernens von Schülerinnen und Schüler verantwortlich, d. h. entscheidet darüber, ob strukturklare Begriffe, verständnistiefes, beweglich-nutzbares Wissen aufgebaut wird und dabei auch überfachliche Lernprozesse bei den Schülerinnen und Schülern kultiviert werden (Reusser, 2019). Das bedeutet, dass Konfliktdebatten um Methoden, insbesondere über offene Unterrichtsformen wenig ertragreich sind, genauso wenig wie es sinnvoll ist, die Kritik an einer konstruktivistischen Unterrichtsgestaltung an der Oberflächenstruktur des Unterrichts festzumachen. Das heißt jedoch nicht, dass man die Oberflächenstruktur des Unterrichts vernachlässigen sollte. Inszenierungsformen konstituieren, insbesondere in Verbindung mit unterschiedlichen Lernzielen (z. B. dem Ziel der Förderung von Lernkompetenzen; vgl. Pauli, in diesem Heft), oder der Stellung einer Unterrichtsphase in einem Lernzyklus (z. B. gemeinsamer Frontalunterricht versus selbständige Arbeit von Schülerinnen und Schülern) stets mehr oder weniger begünstigende, oftmals auch differenzielle Opportunitäten für Tiefenlernprozesse. Der Heftbeitrag von Hess und Lipowsky (in diesem Heft) zur lernphasen-abhängigen Qualität von Fragen (Fragenniveaus) liefert dazu ein instruktives Beispiel. Für *alle* Unterrichtsformen gilt jedoch, dass ihre Qualität immer nur so gut ist wie ihre pädagogisch-psychologische Umsetzung (vgl. auch Raudenbush, 2008).

Ähnlich wie bei der Unterscheidung von Angebot und Nutzung ist es auch bei der Erforschung von Tiefenstruktur-Merkmalen schwierig, sich auf einheitlich abgrenzbare

und standardisierbare Konstrukte zu einigen und ihnen allgemein anerkannte Indikatoren zuzuordnen – zumal sich Merkmale häufig sowohl der Angebots- als auch der Nutzungsseite zuordnen lassen. Decristan, Hess, Holzberger und Praetorius (in diesem Heft) ist zuzustimmen, dass es (auch) hier nicht um eine methodologische, sondern eine viele Fragen aufwerfende theoretische Unterscheidung geht – auch wenn sich oberflächenstrukturelle Merkmale reliabler als Tiefenmerkmale messen lassen. Als Schweizer Projektpartner der 7-Länder TIMSS-1999 Videostudie (Reusser & Pauli, 1999) haben wir versucht, das von Aebli (1961) im Anschluss an Piaget inspirierte tiefenstrukturelle Konzept des *operativen Durcharbeitens* in der internationalen Kodierung zu verankern – und sind damit gescheitert. Da es sich als unmöglich erwies, im landeskulturell gemischten (internationalen) Kodierteam zu einem einheitlichen Verständnis der Unterscheidung zwischen *klassischem* und *operativem* Üben zu kommen, war auch deren reliable Verankerung in den an der Kodierung beteiligten Landeskulturen nicht erfolgreich. Vielleicht hätten wir heute (d. h. 20 Jahre später) mehr Erfolg, steht im anglo-amerikanischen Raum doch heute das Konzept der *deliberate practice* (Ericsson, Krampe & Tesch-Römer, 1993) zur Verfügung, das nicht nur dem Konzept des Durcharbeitens ähnlich ist, sondern sich auch auf Unterrichtsprozesse beziehen lässt (Lehtinen, Hannula-Sormunen, McMullen & Gruber, 2017).

Ein bereits mehrfach angesprochenes, in den Heftbeiträgen wiederholt zutage tretendes Grundproblem, welches jedoch nicht den Autorinnen und Autoren anzulasten ist, sondern seinen Ursprung im Konzept der Unterrichtsqualität selbst hat, ist, dass die im Heft behandelten fünf Theoriefelder (zu denen ich als sechstes Feld auch die Artikulation dazu nehmen würde) nicht unabhängig voneinander sind, sondern sich *mehrfach überkreuzen*. Auch wenn der Sachverhalt sich in mehreren Beiträgen zeigt, werden die Implikationen für die im Heft verfolgte, empirisch-quantitative Forschungsstrategie nur punktuell angesprochenen. Welche Theorieprobleme sich ergeben, wird vor allem in den Heftbeiträgen zur Perspektivenspezifität in der Einschätzung von Unterrichtsqualität deutlich, einem seit Clausen (2002) in der deutschsprachigen Unterrichtsforschung viel beachteten Sachverhalt (vgl. auch Pauli, 2012). Interessant ist, dass es sich beim Perspektivenproblem um ein doppeltes Problem handelt. Einerseits sind es *Akteursperspektiven,* die dazu führen, dass je nach Blickwinkel (Lehrperson, Schülerinnen und Schüler, externe Beobachtende) bei der Wahrnehmung von Unterricht unterschiedliche Realitäten in den Blick treten. Die Beiträge von Fauth et al. (in diesem Heft) sowie von Göllner et al. (in diesem Heft) demonstrieren dies eindrücklich an der Dekonstruktion des Begriffs der Klassenführung. Andererseits ist davon auszugehen, dass es auch kulturelle – *zwischen* Landes-, Schul-, Theorie- und Forschungskulturen und nicht bloß innerhalb von Akteursgruppen einer Schule/eines Systems auftretende – Differenzen gibt, welche unterschiedliche Beobachtende oder Forschende Unterschiedliches sehen lassen. So wird, wer Urteile von Lehrpersonen oder Schülerinnen und Schülern zur Qualität von Unterricht grundsätzlich misstraut, eine oftmals beobachtete Nicht-Übereinstimmung von Wahrnehmungen von Lehrpersonen und Schülerinnen und Schülern anders interpretieren (als Verzerrung, Verfälschung oder Messfehler) als jemand, der Perspektivendifferenzen für so etwas wie einen Normal-

fall[8] hält und Wahrnehmungsunterschiede darauf zurück führt, dass Gruppen keine gemeinsame Sprache haben und deshalb nicht dasselbe sehen.[9] Daraus ergibt sich, dass man sich bei der Beurteilung von Unterrichtsqualität von perspektiven*un*abhängigen, wahren Werten wohl zum Teil verabschieden muss (Clausen, in diesem Heft) bzw. prüfen muss, inwiefern es sich bei Unterschieden in der Unterrichtsqualitätswahrnehmung um theoretisch plausibilisierbare Perspektivendifferenzen handelt, wie wir sie mehrfach in der Pythagorasstudie und in der Folgestudie DIDKOM gefunden haben (Hugener, 2008; Pauli, 2012; Pauli, Reusser & Grob, 2007), und wie sie in weiteren Untersuchungen dokumentiert sind (z. B. Fauth, Decristan, Rieser, Klieme & Büttner, 2014). Daran schließen sich weitere interessante Fragen nach der Entstehung kultureller Sprachmuster an, in denen wir durch pädagogische Kulturen geprägte Unterschiede benennen. Vielleicht bräuchten wir eine Video-Enzyklopädie, mit deren Hilfe sich die (international) vielfältigen kulturellen Muster von Unterricht zeigen und diskutieren ließen.[10] Sodann sollten wir unsere in Videostudien überaus zahlreich gefilmten Unterrichtslektionen nicht nur dafür nutzen, dass wir sie immer wieder neuen Qualitäts-Ratings unterwerfen, sondern vermehrt versuchen, sie als kulturelle Muster u. a. mit phänomenologischen und hermeneutischen Methoden zu rekonstruieren.

6. Schlussbemerkungen

Auch nach über zwei Jahrzehnten Forschung zu einem Nukleus von Basisdimensionen sowie zu Oberflächen- und Tiefenmerkmalen des Lehrens und Lernens bleibt die Frage nach einem psychologischen Prozessmodell, das diese und weitere Dimensionen (Angebot und Nutzung, Artikulationsstufen, Diskursqualität des Unterrichts sowie fachdidaktische Merkmale) zu integrieren und mit multikriterialen Wirkungsebenen zu ver-

8 Der Psychologe Carl Friedrich Graumann (1960) bezeichnet die Perspektivität als Prinzip schlechthin unseres Erkennens; Perspektiven gehören konstitutiv zu unserem *Weltinnewerden*. – George Herbert Mead (1969) sah sie ebenfalls weder als „Verzerrungen von irgendwelchen vollkommenen Strukturen noch Selektionen des Bewusstseins aus einer Gegenstandsmenge, deren Realität in einer Welt der Dinge an sich (noumenal world) zu suchen ist. Sie sind in ihrer wechselseitigen Bezogenheit aufeinander die Natur, die die Wissenschaft kennt." (Mead, 1969, S. 215).

9 Dieses Problem hat auch uns in der TIMSS-Videostudie beschäftigt: so gehört die oben berichtete Schwierigkeit, operatives Üben anhand von Videodaten dingfest zu machen, zu dieser Kategorie kultureller Perspektivendifferenzen.

10 Wir haben selber dazu bereits 2005 einen Anfang gemacht, indem wir zahlreiche schweizerische (jedoch auch internationale) Unterrichtsvideos aus mehreren Studien interessierten Forschenden sowie in der Lehrerinnen- und Lehrerbildung tätigen Personen in einem (geschützten) Videoportal niederschwellig zugänglich gemacht haben. Immer noch kommen neue Videos zum Portal dazu (http://www.unterrichtsvideos.ch/). Wichtig ist, dass wir uns dabei *nicht* an einem Verständnis von *best practice* orientieren, sondern alltäglichen Unterricht einem an multiperspektivischen und multimethodischen Analysen interessierten Kreis von Kolleginnen und Kollegen zugänglich machen.

binden vermag, weit offen. Etwas anderes zu erwarten wäre auch naiv und wird auch in Zukunft eine Illusion bleiben, die nur jene haben können, die verkennen, dass es beim Unterricht und seiner Qualität nicht allein um ein naturgesetzliches, sondern um ein hochgradig kulturell und normativ aufgeladenes, durch vielfältige Wechselwirkungen zwischen intentional handelnden Akteuren, Ebenen und Wirkungsdimensionen geprägtes, kontextuell situiertes Handlungs- und Prozessgeschehen geht. Dennoch hat die empirisch-quantitative Unterrichtsforschung dazu beigetragen (und dieses Beiheft leistet dazu einen wichtigen Beitrag), dass die Konturen einer psychologisch-didaktisch fundierten, *prozessbezogenen,* sich von normativen Prinzipien emanzipierenden Unterrichtstheorie zunehmend klarer hervortreten und deren Forschungsergebnisse geeignet sind, das Unterrichtshandeln von Lehrpersonen – auch evidenzbasiert – zu reflektieren. Dennoch bleibt die Brücke von gutem zu effektivem zu qualitätsvollem Lehren im Sinne Berliners (2005) nach wie vor bruchstückhaft – so dass man dem Leibniz-Netzwerk auf jeden Fall ein langes Leben wünschen sollte!

Unterricht im Gestaltwandel. Bei aller eindrücklichen Steigerung der Qualität und Perspektivenvielfalt der im vorliegenden Beiheft dokumentierten, empirisch-quantitativen Unterrichtsforschung hinsichtlich ihres Beitrags zur Klärung des Zusammenspiels von Basisdimensionen und Tiefenmerkmalen der Unterrichtsqualität in einem durch Angebote und Nutzungen geprägten Unterrichtshandeln gibt es auch Desiderata, wovon ein mir wichtig erscheinendes zum Schluss kurz angesprochen werden soll. Was als Desiderat immer dringlicher wird ist, dass sich die quantitative Unterrichtsforschung vermehrt auf eine in zahlreichen, hoch entwickelten Ländern beobachtbare *Veränderung der grammar of schooling* (Tyack & Tobin, 1994) einstellt. Zahlreiche von der quantitativ-empirischen Unterrichtsforschung entwickelte und standardmäßig verwendete Erhebungs- und Analyseinstrumente und Auswertungsstrategien passen bereits heute nur mehr bedingt zu den Choreographien einer sich teils dramatisch verändernden Unterrichtsarchitektur, wo nicht mehr Frontal- und Ganzklassenunterricht vorherrschen, sondern Wochenplanunterricht, Projektarbeit und selbständiges Arbeit in Lernateliers das Bild individualisierter Praktiken des Unterrichts immer stärker dominieren. Wie erforscht man Unterrichtskulturen, in der die Kinder über mehrere Räume verteilt sind, parallele Unterrichts- und Lernprozesse in gleichzeig mehreren Fächern stattfinden, Klassenverbände aufgelöst, die Lerngruppen altersgemischt und sehr heterogen sind und zum Teil von fachfremden Lehrpersonen, die nicht ihre Stammlehrpersonen sind, unterstützt werden – und dies unter der Leitidee *personalisierten Lernens*?

Will die quantitativ arbeitende Unterrichtsforschung ihre Bedeutung nicht langfristig verlieren (und auch gegenüber der qualitativen Forschung an Terrain einbüssen), sollte sie sich bezüglich Studiendesigns, Erhebungsmethoden, Instrumenten und Forschungskooperationen auf diese Veränderungen einstellen. Dass dies keine triviale Aufgabe ist, haben wir in unserer PerLen-Studie (*Per*sonalisiertes *L*ern*en* in heterogenen Lernumgebungen) erfahren (Stebler, Pauli & Reusser, 2018), deren Stichprobe zahlreiche bezüglich ihrer Lerndesigns sehr innovative Schulen umfasst, deren schulorganisatorische und didaktische Architektur sich von einem Mehrebenen-System mit fixen Lerngruppen und stabilen Zuordnungen von Lehrpersonen zu Schülerinnen und Schüler

weit entfernt hat (Reusser, Pauli, Stebler & Grob, 2015). Auch wenn der Begriff eines guten und qualitätsvollen Unterrichts als Umsetzung eines variablen Sets von Methoden, Interaktions- und Diskursformen, als *Orchestrierung von Inhalten und Methoden unter Beachtung der generischen Qualitätsdimensionen* (Klieme, 2019) damit nicht obsolet geworden ist, muss der *Gestaltwandel des Unterrichts,* insbesondere von sich neu herausbildenden Praktiken ernst genommen werden, was bedeutet, dass viele der für den traditionellen Ganzklassenunterricht entwickelten Modelle, Erhebungsmethoden, -instrumente und Analysestrategien an stärker individualisierte und fragmentierte Settings angepasst werden müssen.

Literatur

Aebli, H. (1951). *Psychologische Didaktik.* Stuttgart: Ernst Klett.

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

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

Ball, D. L., & Forzani, F. M. (2009). The work on teaching and the challenge for teacher education. *Journal of Teacher Education, 60(5),* 497–511.

Berliner, D. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56(3),* 205–213.

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

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

Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit. Fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht.* Münster: Waxmann.

Ericsson, K. A., Krampe, R., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406.

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

Fend, H. (1980). *Theorie der Schule.* München: Urban & Schwarzenberg.

Fraefel, U., & Scheidig, F.(2018). Mit Pragmatik zu professioneller Praxis? Der „Core Practices"-Ansatz in der Lehrpersonenbildung. *Beiträge zur Lehrerinnen- und Lehrerbildung, 36(3),* 344–364.

Graumann, C. F. (1960) *Psychologie der Perspektivität.* Berlin: Walter de Gruyter.

Greeno, J. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53(1),* 5–26.

Grossman, P., & Pupik Dean, C. G. (2019). Negotiating a common language and shared understanding about core practices: The case of discussion. *Teaching and Teacher Education, 80,* 157–166.

Heimann, P., Otto, G., & Schulz, W. (1965). Unterricht: Analyse und Planung. Hannover: Schroedel.

Helmke, A. (2006). Was wissen wir über guten Unterricht? *Pädagogik, 2,* 42–45.

Herbart, J. F. (1831/1964). Von der Erziehungskunst. In W. Asmus (Hrsg.), *Pädagogische Schriften. Erster Band: Kleinere Pädagogische Schriften.* Düsseldorf: Küpper vormals Bondi.

Hugener, I. (2008). *Inszenierungsmuster im Unterricht und Lernqualität. Sichtstrukturen schweizerischen und deutschen Mathematikunterrichts in ihrer Beziehung zu Schülerwahrnehmung und Lernleistung – eine Videoanalyse.* Münster: Waxmann.

Kane, R., & Marsh, C. J. (1980). Progress toward a general theory of instruction? *Educational Leadership,* 253–255.

Kennedy, M. (2016). Parsing the practice of teaching. *Journal of Teacher Education, 67*(1), 6–17.

Kiper, H. (2006). Rezeption der psychologischen Didaktik. 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. 74–85). Bern: h. e. p.

Klafki, W. (1964). *Studien zur Bildungstheorie und Didaktik.* Weinheim/Basel: Beltz.

Klieme, E. (2011). Standards der Unterrichtsqualiät – Kann es das geben? (Referat an der *Tagung „Unterrichtsforschung und Unterrichtspraxis: Innovation und Transfer"* anlässlich des 60. Geburtstages von Kurt Reusser, Universität Zürich).

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

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge: Cambridge University Press.

Lehtinen, E., Hannula-Sormunen, M., McMullen, J., & Gruber, H. (2017). Cultivating mathematical skills: From drill – and – practice to deliberate practice. *ZDM Mathematics Education, 49(4),* 625–636.

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

Mead, G. H. (1969). *Philosophie der Sozialität: Aufsätze zur Erkenntnisanthropologie.* Frankfurt a. M.: Suhrkamp.

Messner, H., & Reusser, K. (2000). Berufliches Lernen als lebenslanger Prozess. *Beiträge zur Lehrerbildung, 18*(3), 277–294.

Messner, R., & Reusser, K. (2006). Aeblis Didaktik auf psychologischer Grundlage im Kontext der zeitgenössischen Didaktik. 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. 52–73). Bern: h. e. p.

Meyer, H. (2004). *Was ist guter Unterricht?* Berlin: Cornelsen Scriptor.

Pauli, C., & Reusser, K. (2011). Expertise in Swiss mathematics instruction. In Y. Li & G. Kaiser (eds.), *Expertise in mathematics instruction. An international perspective* (pp. 85–108). Berlin: Springer.

Pauli, C. (2012). Merkmale guter Unterrichtsqualität im mathematisch-naturwissenschaftlichen Unterricht aus der Perspektive von Lernenden und Lehrpersonen. In R. Lazarides & A. Ittel (Hrsg.), *Differenzierung im mathematisch-naturwissenschaftlichen Unterricht: Implikationen für Theorie und Praxis* (S. 13–34). Bad Heilbrunn: Klinkhardt.

Pauli, C., Reusser, K., & Grob, U. (2007). Teaching for understanding and/or self-directed learning? A video-based analysis of reform-oriented mathematics instruction in Switzerland. *International Journal of Educational Research, 46,* 294–305.

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*(1), 2–12.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal, 45,* 206–230.

Reusser, K. (1983). Die kognitive Wende in der Psychologie: Eine Annäherung an phänomenologische und geisteswissenschaftliche Problemstellungen. In L. Montada, K. Reusser & G. Steiner (Hrsg.), *Kognition und Handeln* (S. 169–188). Stuttgart: Klett.

Reusser, K. (2008). Empirisch fundierte Didaktik – didaktisch fundierte Unterrichtsforschung. In M.A. Meyer, M. Prenzel & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik* (9. Sonderheft der Zeitschrift für Erziehungswissenschaft, S. 219–238). Wiesbaden: Verlag für Sozialwissenschaften.

Reusser, K. (2019). Unterricht als Kulturwerkstatt in bildungswissenschaftlich-psychologischer Sicht. In U. Steffens & R. Messner (Hrsg.), *Unterrichtsqualität – Konzepte und Bilanzen gelingenden Lehrens und Lernens* (Grundlagen der Qualität von Schule 3, S. 129–166). Münster: Waxmann.

Reusser, K., & Pauli, C. (1999). Unterrichtsqualität: Multideterminiert und multikriterial. Anforderungen an einen Unterrichtsqualitätsbegriff als Grundlage videobasierter Unterrichtsforschung (Schriftliche Fassung der *Keynote, gehalten an der 7. Tagung Pädagogische Psychologie der Deutschen Gesellschaft für Psychologie am 14. September 1999 in Freiburg/CH*).

Reusser, K., & Pauli, C. (2013). Verständnisorientierung in Mathematikstunden erfassen. Ergebnisse eines methodenintegrativen Ansatzes. *Zeitschrift für Pädagogik, 59*(3), 308–335.

Reusser, K., & Pauli, C. (2015). Co-constructivism in educational theory and practice. In J.D. Wright (ed.), *International encyclopedia of the social & behavioral sciences* (Vol. 3, pp. 913–917). Oxford: Elsevier.

Reusser, K., Pauli, C., Stebler, R., & Grob, U. (2015). perLen: Personalisierte Lernkonzepte in heterogenen Lerngruppen. Methodologische Herausforderungen eines Praxisforschungsprojektes *(Referat am Forschungsforum der DGfE-Sektion Schulpädagogik, Göttingen, September 2015).*

Stebler, R., Pauli, C., & Reusser, K. (2018). Personalisiertes Lernen – Zur Analyse eines Bildungsschlagwortes und erste Ergebnisse aus der perLen-Studie. *Zeitschrift für Pädagogik, 64*(2), 159–178.

Tyack, D., & Tobin, W. (1994). The „grammar" of schooling: Why has it been so hard to change? *American Educational Research Journal, 31*(3), 453–479.

von Wright, H. (1974). *Erklären und Verstehen.* Frankfurt: Athenäum.

Wittmann, E. (1974). *Grundfragen des Mathematikunterrichts.* Braunschweig: Vieweg.

Abstract: In this commentary, the contributions in this supplement are discussed from a cognitive psychological and teaching perspective. Introductory remarks on the status of teaching quality in the broader educational discussion are followed by remarks focused on the five theoretical strands presented in this volume, and their connection to each other. This is followed by more specific comments, in particular in relation to the following topics: the three basic dimensions of teaching quality (as examples of the in-depth quality of teaching), the so-called opportunity-use model, and the perspective-dependency of perceptions of teaching quality. The commentary closes with remarks on desiderata for empirical-quantitative research on learning and teaching, especially against the background of an observable, significant change in the grammar of schooling (Tyack & Tobin, 1994).

Keywords: Teaching Quality, Didactics (in the German Tradition), Pedagogical Theory of Teaching, Research on Teaching, Instruction

Anschrift des Autors

Prof. em. Dr. Kurt Reusser, Universität Zürich,
Freiestrasse 36, 8032 Zürich, Schweiz
E-Mail: reusser@ife.uzh.ch

Anke Lindmeier/Aiso Heinze

Die fachdidaktische Perspektive in der Unterrichtsqualitätsforschung: (bisher) ignoriert, implizit enthalten oder nicht relevant?

Zusammenfassung: Dieser Beitrag diskutiert die Artikel des Beihefts „Empirische Forschung zu Unterrichtsqualität" aus einer fachdidaktischen Perspektive. Dazu wird zunächst die eingenommene Sicht genauer spezifiziert, indem insbesondere der normative fachdidaktische Aspekt von Unterrichtsqualität betont wird. Auf dieser Basis wird dann die Rolle fachdidaktischer Merkmale bei den bisher betrachteten Indikatoren für Unterrichtsqualität untersucht. Anschließend wird diskutiert, inwieweit die in der Unterrichtsforschung betrachteten eher kurzen Ausschnitte von Unterricht eine Erfassung von genuin fachdidaktischen Qualitätsindikatoren ermöglichen. Der Beitrag schließt mit ergänzenden, möglicherweise relevanten fachspezifischen Faktoren der Unterrichtsqualität, die in diesem Beiheft nicht betrachtet wurden.

Schlagworte: Unterrichtsqualität, Unterrichtsforschung, Forschungsperspektiven, Fachdidaktik, Qualität von Fachunterricht

1. Einleitung

Aus Sicht der Fachdidaktiken scheint die Unterrichtsqualitätsforschung einen paradoxen Charakter zu haben. So verfolgt sie einerseits das Ziel, das komplexe Konstrukt ‚Unterrichtsqualität' *fachunabhängig* zu konzeptualisieren und relevante Qualitätsindikatoren zu bestimmen, und hat dabei andererseits den Anspruch, die Qualität von *Fachunterricht* messen zu können. Dem Wilhelm von Ockham zugeschriebenen Forschungsprinzip der Parsimonie folgend kann dieses fachunabhängige Vorgehen zweifellos als sinnvoll angesehen werden, da ein Qualitätskonstrukt für Unterricht, das ohne fachspezifische Merkmale auskommt, äußerst sparsam wäre. Allerdings wirft so ein Vorgehen viele Fragen auf, die für die Fachdidaktik geradezu existenzielle Bedeutung haben: Wenn Unterrichtsqualität ohne Bewertung fachlicher und fachdidaktischer Merkmale des Unterrichtsgeschehens auskommt,

- warum erhalten Lehrkräfte eine fachliche und fachdidaktische Ausbildung anstelle nur einer allgemeindidaktischen Ausbildung?
- warum werden für den Fachunterricht fachdidaktische Instruktionsansätze und Lehrmaterialien für bestimmte Themengebiete entwickelt anstatt nur Inhalte entlang der Fachlogik zu lehren?
- wozu gibt es überhaupt die Fachdidaktiken?

Wenn man das Konstrukt ‚Unterrichtsqualität' in der Konzeptualisierung wirklich fachunabhängig versteht und entsprechend misst, dann müsste man die Qualität von Unterrichtsstunden unterschiedlicher Fächer, beispielweise von Mathematik- und Deutschunterricht, vergleichen können. Die einzige uns dazu bekannte Studie in Deutschland von Praetorius, Vieluf, Saß, Bernholt und Klieme (2016) vergleicht Deutsch- und Englischunterricht, also zwei sprachliche und sich nicht gänzlich fremde Fächer bezüglich der als fachunabhängig angesehenen Qualitätsdimensionen „Klassenführung" und „Motivationale Unterstützung". Es ergab sich bereits hier für die Motivationale Unterstützung ein nicht zu vernachlässigender Einfluss des Faches, und die Vermutung ist, dass dieser bei der Dimension „Kognitive Aktivierung" noch stärker ausfallen dürfte.

Empirische Evidenz für den Einfluss fachspezifischer Merkmale auf die Effektivität von Fachunterricht liefert zusammenfassend die Metaanalyse von Seidel und Shavelson (2007). Reusser (2009) verweist auf das didaktische Dreieck, das die verschiedenen Einflussfaktoren für Unterrichtsqualität vereint und das neben Lehrkräften und Schülerinnen und Schülern auch die Rolle des Faches explizit enthält. Auch diverse Beiträge in diesem Beiheft enthalten Hinweise, dass das Fach und das Verständnis von Unterrichtsqualität nicht vollständig voneinander zu separieren sind. So enthält beispielsweise der Vorschlag des revidierten Angebots-Nutzungs-Modells in Vieluf, Praetorius, Rakoczy, Kleinknecht und Pietsch (in diesem Heft) das Fach als relevanten Kontext zur Beschreibung von Unterricht; Hess und Lipowsky (in diesem Heft) beziehen bei der Bewertung des kognitiven Frageniveaus als Tiefenmerkmal des Leseunterrichts die „inhaltliche Relevanz" der Frage für die Leseübung explizit mit ein, und Decristan, Hess, Holzberger und Praetorius (in diesem Heft) sehen das Verhältnis von Fachlichkeit und Tiefenstrukturen als ein gegenwärtig zu klärendes Forschungsfeld an.

Die Frage, ob die fachdidaktische Perspektive in der Unterrichtsqualitätsforschung bisher ignoriert wurde, implizit enthalten ist oder keine Relevanz hat, bleibt komplex. Die Antwort wird auch davon abhängen, ob es um das theoretische Konstrukt Unterrichtsqualität oder um dessen Messung geht. Oder kurz: Ob die Fachspezifität bereits mit der Konzeptualisierung oder erst mit der Operationalisierung beginnt (vgl. Vehmeyer, 2009, S. 169).

Dass das Fach den Fachunterricht prägt, wird kaum bestritten werden. Im Englischunterricht wird englische Grammatik eine Rolle spielen, und der Chemieunterricht wird nicht ohne die Behandlung chemischer Elemente und Reaktionen stattfinden. Empirisch lässt sich zeigen, dass die *fachlichen* Inhalte im Sinne eines implementierten Curriculums Einfluss auf die Schülerleistung haben (z. B. auf Basis von Schülerberichten: Kuger, Klieme, Lüdtke, Schiepe-Tiska & Reiss, 2017; auf Basis der Schulbuchinhalte: Sievert, van den Ham, Niedermeyer & Heinze, 2019). Ob genuin fach*didaktische* Aspekte neben fachlichen, allgemeindidaktischen und lehr-lern-psychologischen Aspekten des Unterrichts relevant sind, ist nicht mehr so klar. Dahinter steckt auch die Frage, was die Fachdidaktiken ausmacht und ob sie als eigene Disziplin angesehen werden können. So stellte etwa Paul Kirschner in Bezug auf die Lehrerprofessionsforschung infrage, dass das Konstrukt *fachdidaktisches Wissen* notwendig sei, da Lehrkräfte für das Unterrichten nur Fachwissen und darauf angewendetes bildungswissenschaftliches Wissen

benötigen würden (Kirschner, Verschaffel, Star & Van Dooren, 2017). Auf die Unterrichtsqualitätsforschung übertragen hieße dies, dass sie ohne eine genuin fachdidaktische Perspektive auskäme und Unterrichtsmerkmale aus einer Kombination von fachlichen sowie allgemeindidaktischen und lehr-lern-psychologischen Kriterien zu werten wären, die in Teilen aufeinander zu beziehen sind. Dass dieses ‚Aufeinanderbeziehen' jedoch komplexer ausfallen dürfte, als es vielleicht klingt, wird unten noch thematisiert.

Betrachtet man generell die Forschung in den Fachdidaktiken, so können hier grob zwei Bereiche unterschieden werden. Zum einen beschäftigen sich die Fachdidaktiken mit deskriptiven und explikativen Fragestellungen, d. h. wie die Realität im Fachunterricht ist und welche Mechanismen hinter dem fachlichen Lernen stecken. Diese Sichtweise korrespondiert mit der der empirischen Unterrichtsforschung. Zum zweiten erarbeiten die Fachdidaktiken normativ-präskriptive Vorschläge für den Unterricht, d. h. wie die Realität im Zusammenhang mit dem Fachunterricht sein sollte. Auch dieses zweite Ziel spielt für die Unterrichtsqualitätsforschung eine zentrale Rolle. So beschreibt Berliner (2005) die Notwendigkeit von zwei Arten von normativ festgelegten Kriterien in Bezug auf Unterrichtsqualität: Kriterien zur unmittelbaren Bewertung der Qualität von Unterricht („good teaching") und Kriterien, die sich auf den Effekt von Unterricht beziehen und Unterrichtsqualität damit indirekt über den Unterrichtseffekt bewerten („effective teaching"). Auch andere Autorinnen und Autoren betonen, dass die Unterrichtsforschung nicht frei von normativen Vorstellungen ist (z. B. Kunter & Ewald, 2016; Openshaw & Clarke, 1970). Wie bei Praetorius et al. (in diesem Heft) bereits herausgearbeitet, stehen Forschende im Bereich Unterrichtsqualität vor dem Dilemma, Wertneutralität anzustreben, während gleichzeitig das soziale Phänomen Unterricht per se durch geteilte Wertvorstellungen und Normen geprägt ist. Ebenso wie normative Vorstellungen zu Unterrichtsansätzen einfließen (z. B. Kunter & Ewald, 2016, zur positiven Konnotation des entdeckenden Lernens), ist in Bezug auf Fachunterricht davon auszugehen, dass in vielen Studien normative fachdidaktische Aspekte sowohl für die Unterrichtsmerkmale (z. B. Funktion des Experiments in der Chemiestunde) als auch – falls genutzt – für die Unterrichtseffekte (z. B. Konzeptualisierung von Kompetenz in Chemie als Outcome-Maß) relevant sind. Die jeweiligen normativen Setzungen werden aber in den verwendeten Modellen nicht immer expliziert, was zu Problemen bei der Interpretation der Messergebnisse führen kann. Diese Feststellung gilt im Übrigen nicht nur aus fachspezifischem Blick, sondern allgemeiner. Etwa zeigt die Analyse der Genese verschiedener Angebots-Nutzungs-Modelle von Vieluf et al. (in diesem Heft) deutlich unterschiedliche Grundsetzungen. Implizite Normdifferenzen können sich dabei generell auf den verschiedenen Kontext-Ebenen (z. B. Fach, Schularten, Bundesländer, siehe Vieluf et al., in diesem Heft) ergeben.

2. Die Rolle von normativen Vorstellungen in der Unterrichtsforschung

Die Rolle von Normen wird aus unserer Sicht in der aktuellen Literatur nur punktuell berücksichtigt, und man könnte kritisch anmerken, dass es häufig unklar ist, ob sich die Wissenschaftlerinnen und Wissenschaftler dieser normativen Setzungen bewusst sind. Im Folgenden soll an verschiedenen Beispielen die Rolle von Normen aufgezeigt werden, wobei ein Fokus auf die fachdidaktische Perspektive gelegt wird.

2.1 Die Rolle von fachspezifischen Normen bei der Konzeptualisierung und Operationalisierung von Unterrichtsqualität

In den bisherigen Studien wurde die generische Auffassung von Unterrichtsqualität bei der Operationalisierung von einzelnen Konstrukten – etwa kognitive Aktivierung – jeweils für das betrachtete Fach mehr oder weniger adaptiert. Fachdidaktische Qualitätsmerkmale wurden dabei selten explizit benannt, und das Verhältnis von Fachlichkeit und Unterrichtsqualitätsmerkmalen bleibt oft unklar (vgl. Decristan et al., in diesem Heft). Oft findet die Beurteilung hochinferent statt, wobei fachdidaktische Aspekte möglicherweise implizit bleiben. So geben Kunter und Ewald (2016, Tab. 1) einen Überblick zur Erfassung der kognitiven Aktivierung in verschiedenen Fächern und nennen als Beispiele u. a. die Indikatoren „Fragen, die zu langen und inhaltlichen Antworten anregen", „Fragen, die zum Nachdenken anregen" oder die „Vermittlung von Lesestrategien durch die Lehrkraft", die im jeweiligen Fach beim Rating fachdidaktisch interpretiert werden können – aber nicht müssen. So können Fragen, die fachliche Fehlvorstellungen bestärken, ebenfalls zum Nachdenken anregen und lange inhaltliche Antworten induzieren, und nicht jede beliebige Lesestrategie wird das Lesen lernen nachhaltig unterstützen. Vereinfacht gesagt, stellt sich die Frage, inwieweit die kognitive Aktivierung zielgerichtet den fachlichen Lernprozess unterstützt und ihn nicht be- oder verhindert. Kriterien auf Basis fachdidaktischer Erkenntnisse dürften dabei eine zentrale Rolle spielen.

Studien zur Unterrichtsqualität, die explizit die Fragen der Fachlichkeit adressieren, sind eher rar. Exemplarisch liegt eine Studie aus der Fachdidaktik Mathematik vor (Brunner, 2018), für die eine Unterrichtsstunde mit drei Erhebungsinstrumenten zur Unterrichtsqualität beurteilt wurde, darunter mit dem stark fachbezogen ausgerichteten Instrument TRU (Schoenfeld et al., 2013). Es ergaben sich je nach verwendetem Instrument für die gleiche Mathematikstunde deutlich unterschiedliche Qualitätsbewertungen, wobei insbesondere das Einbeziehen der fachlichen Korrektheit als Kriterium eine Ursache für die verschiedenen Ergebnisse war. Ein weiteres prominentes Beispiel für die Erfassung fachdidaktischer Unterrichtsmerkmale ist das Pythagoras-Projekt (Lipowsky, Drollinger-Vetter, Klieme, Pauli & Reusser, 2018). Hier wurden zunächst auf fachdidaktischer Basis theoretische ‚Verstehenselemente' für den Satz des Pythagoras herausgearbeitet und diese dann für die fachdidaktische Bewertung von Unterricht genutzt (Drollinger-Vetter, 2011). Anzumerken ist, dass es sich dabei um themenspezifische inhaltsbezogene Qualitätsmerkmale handelt, die in bestimmten Phasen von Unter-

richtsstunden (sog. Theoriephasen) erfasst wurden. Es zeigte sich in weiteren Analysen, dass sich ein über Verstehenselemente konzeptualisiertes Konstrukt für fachdidaktische Qualität des Unterrichts zum Satz des Pythagoras von den drei Basisdimensionen empirisch abgrenzen lässt (Lipowsky et al., 2018). Auf den ersten Blick mag es nicht verwundern, dass in beiden Studien der Einsatz anderer oder zusätzlicher Instrumente andere Bewertungsaspekte betont. Dies kann im Falle divergierender Ergebnisse der Qualitätsmessung aber zu einem Problem führen, etwa wenn fachliche Korrektheit in Studien nicht erfasst wird und die Ergebnisse normbildend rezipiert werden.

Offen bleibt an dieser Stelle die interessante Frage, ob das Modell der drei Basisdimensionen durch eine fachspezifische vierte Dimension erweitert werden sollte oder nicht. Die Ergebnisse von Lipowsky et al. (2018) deuten auf eine empirische Trennbarkeit einer fachdidaktischen Qualitätsdimension hin, beziehen sich aber auf einen sehr speziellen Unterrichtsinhalt. Schlesinger, Jentsch, Kaiser, König & Blömeke (2018) stellen ein Instrument vor, das sich global auf Mathematikunterricht bezieht und neben Skalen für die drei Basisdimensionen zwei weitere fachspezifische Skalen enthält *(subject-related quality, teaching-related quality)*. Unabhängig von der noch ausstehenden empirischen Trennbarkeit möglicher fachspezifischer Basisdimensionen stellt sich vor allem die Frage der theoretischen Begründung. So können diverse Items der vorgeschlagenen fachspezifischen Skalen theoretisch auch der Basisdimension „Kognitive Aktivierung" zugeordnet werden (z. B. using multiple representations, dealing with mathematical errors of students, relevance of mathematics for students). Damit wäre die Frage nicht, ob die drei Basisdimensionen durch fachspezifische Dimensionen ergänzt werden sollten, sondern eher, ob die Dimension „Kognitive Aktivierung" nicht fachspezifisch ausdifferenziert werden muss. Letzteres würde explizit machen, was bisher als implizite fachdidaktische Operationalisierung (s. o.) bereits vorliegt. Inwieweit diese Ausdifferenzierung sinnvoll und notwendig ist, dürfte von der Forschungsfrage abhängen, die mit dem Instrument beantwortet werden soll – und damit von der in der jeweiligen Studie benötigten Auflösung des Konstrukts Unterrichtsqualität.

2.2 Inwieweit ist die Operationalisierung der Qualität von Fachunterricht kulturell geprägt?

Die zuvor angesprochene Rolle der fachspezifischen Normen bei der Untersuchung von Unterrichtsqualität wird komplexer, wenn kulturell geprägte Normen zum Fachunterricht einbezogen werden. Wie in der Debatte von Heid (2013) und Klieme (2013) herausgearbeitet wurde, muss im Kern zwischen der Deskription von Unterrichtsmerkmalen und deren Bewertung unterschieden werden. Erfasst man beispielsweise die Anzahl an offenen Lehrerfragen pro Unterrichtsstunde, so ist dies ein deskriptiver Akt. Qualifiziert man einen höheren Anteil offener Lehrerfragen als besser, so ist das der bewertende Akt. Bei genauerem Blick lassen viele der aktuell genutzten Unterrichtsqualitätsindikatoren ihren Ursprung in der Idealvorstellung eines konstruktivistisch geprägten Unterrichts erkennen, der mit einer gewissen Auffassung vom Lernen, der Rolle der

Lehrkräfte und der Rolle von Diskurs, Wissen oder Übung für individuelle Lernprozesse aufgeladen ist (vgl. auch Kunter & Ewald, 2016). Gleichzeitig weiß man aus der kulturvergleichenden Sozialpsychologie, dass solche Werthaltungen und Normen zwischen verschiedenen Kulturen stark differieren können (z. B. Hofstede, 2001). So können beispielsweise im Gegensatz zu individualistischen Gesellschaften (z. B. in Europa und Nordamerika, den sog. westlichen Ländern) in kollektivistischen Gesellschaften (z. B. den asiatischen Ländern der sog. *confucian heritage culture*) Indikatoren wie ein hoher Anteil individueller Schüleräußerungen oder das Eingehen einer Lehrkraft auf individuelle Verständnisschwierigkeiten negativ bewertet werden (z. B. Clarke, 2013a zur Rolle von Schüleräußerungen im Mathematikunterricht). Diese Normunterschiede machen die Grenzen kulturübergreifender Forschung im Bereich Unterrichtsqualität deutlich. In den internationalen Vergleichsuntersuchungen werden deswegen Konsensverfahren genutzt und strittige Indikatoren eliminiert, womit allerdings Validitätsprobleme auftreten können (*validity-comparability compromise,* Clarke, 2013b).

Der Kontrast zwischen stark unterschiedlichen Kulturen ist besonders geeignet, um das potenzielle Problem unterschiedlicher Normvorstellungen deutlich zu machen. Dieses Phänomen kann aber auch bei größerer Nähe der Kulturen auftreten. Praetorius und Charalambous (2018) stellen beispielsweise beim Vergleich von 12 generischen und mathematikspezifischen Instrumenten für Unterrichtsqualität aus dem westlichen Kontext fest, dass es auffällige Unterschiede gibt: Während die Indikatoren meist konstruktivistische Auffassungen von Lehren und Lernen spiegeln (wie z. B. kognitive Aktivierung), werden in manchen Instrumenten auch behavioralere Aspekte (wie z. B. Üben) abgebildet. Dabei ist unstrittig, dass die Wahl der Indikatoren und ihrer qualitätskonstitutierenden Ausprägungen keine vollständig objektivierbaren Prozesse sein können. Es liegen entsprechend auch viele unterschiedliche Operationalisierungen vor, die allerdings die Vergleichbarkeit der Ergebnisse beinträchtigen.

Vielleicht lassen sich einige ‚blind spots' dadurch erklären, dass ohne kontrastierende Zugänge die häufig impliziten eigenen Normen und Wertvorstellungen kaum der Reflektion zugänglich sind. Deswegen wird beispielsweise im „lexicon project" (Mesiti & Clarke, 2017) daran gearbeitet, Begriffe zur Beschreibung von (Mathematik-)Unterricht in verschiedenen Sprachen zu sammeln und zu sortieren. Ein Ziel ist, für zentrale Terme eine standardisierte Übersetzung zur Erleichterung komparativer Forschung zu gewinnen. Ein zweites Ziel ist, die genuinen Entwicklungen der verschiedenen Kulturen wechselseitig zugänglich zu machen. Dabei folgt das Projekt der Erkenntnis, dass ein Unterrichtsmerkmal, für das es in einer Sprache keinen Begriff gibt, in der zugehörigen Kultur auch nicht leicht zugänglich ist und entsprechend in Operationalisierungen kaum berücksichtigt werden kann (Clarke, 2013b). Dieses Grundprinzip trifft auch für Unterschiede in der Unterrichtskultur zwischen den Fächern oder zwischen einem allgemeinen und fachspezifischen Blick auf Unterricht zu, beispielsweise wenn generische Messinstrumente nicht zwischen verschiedenen Erklärqualitäten unterscheiden. Für die Schülerfrage: „Warum hat 4:0 kein Ergebnis?" könnten etwa die zwei Erklärungen „Weil das eine Division durch 0 ist und das ist nicht definiert" und „Weil geteilt durch 0 ja bedeuten würde, man verteilt 4 Dinge auf 0 Personen und das geht nicht" aus

generischer Sicht ähnlich bewertet werden, während die Mathematikdidaktik ihnen sehr unterschiedliche Qualitäten in Bezug auf das für den Mathematikunterricht wichtige Kriterium der Verständnisorientierung zuweisen würde.

Vergleiche von verschiedenen Unterrichtsqualitätsmessinstrumenten (Praetorius & Charalambous, 2018), die aus unterschiedlichen Kontexten mit unterschiedlichen Zielsetzungen vor unterschiedlichen theoretischen Rahmungen entstanden sind, stellen einen ersten Ansatz dar, um die Vielfältigkeit der Operationalisierungen sichtbar zu machen. Im nächsten Schritt wäre es notwendig zu untersuchen, woher die Unterschiede stammen. Das kann zur Begriffsschärfung beitragen, da beispielsweise verschiedene Operationalisierungen desselben Konstrukts (etwa für verschiedene Fächer) von Unterschieden im Verständnis der Konstrukte (etwa zwischen Fächern oder Kulturen) differenziert werden könnten.

2.3 Fachspezifische Effektivität als Kriterium für Unterrichtsqualität

Wie in der Einleitung erwähnt, kann Unterrichtsqualität nicht nur unmittelbar bewertet werden, sondern auch mittelbar über den Unterrichtseffekt. Erkenntnisse zur Qualität von Unterricht nach diesem zweiten Kriterium hängen in hohem Grad von der Wahl der Forschungsfragen ab (siehe auch Seidel, in diesem Heft). So zeigen sich bei Studien zu kognitiven Outcomes andere Qualitätsmerkmale als prädiktiv als für nicht-kognitive Outcomes. Aus fachdidaktischer Sicht besteht die Schwierigkeit, dass die Wahl des Messinstruments für bzw. von Daten zu Unterrichtserfolg häufig von der Verfügbarkeit abhängt. Dies führt einerseits dazu, dass die Effekte von Unterrichtsqualität bevorzugt für Fächer untersucht werden, die als ‚leichter' messbar angesehen werden (wie z. B. Mathematik im Falle von kognitiven Outcomes). Andererseits stellt sich die Frage, ob das jeweilig eingesetzte Instrument für den intendierten Zweck als valide anzusehen ist. Wird beispielsweise die PISA-Mathematikskala als Outcome-Maß eingesetzt, so ist zu beachten, dass die Aufgaben aus fachdidaktischer Sicht eine klare normative Schwerpunktsetzung aufweisen (*mathematical literacy* ohne Anspruch auf curriculare Validität). Inwieweit die für internationale Vergleiche von Bildungssystemen entwickelten PISA-Aufgaben inhaltlich geeignet sind, um zur Unterrichtsqualitätsforschung im Fach Mathematik beizutragen, wäre also zu diskutieren (vgl. etwa die unterschiedliche Veränderungssensitivität der PISA-Skala und der Bildungsstandards-Skala in Lehner et al., 2017). Die Wahl des Outcome-Maßes als externes Kriterium für Unterrichtsqualität kann deutliche Auswirkungen auf die Ergebnisse haben. Systematische Prüfungen, inwiefern die Ergebnisse der Unterrichtsqualitätsforschung von der Wahl des Faches oder des Outcome-Messinstruments abhängen, fehlen bisher.

2.4 Wer erkennt was und warum? Zur Rolle von Raterinnen und Ratern bei der Messung von Unterrichtsqualität

Während sich die zuvor diskutierten Problempunkte hauptsächlich auf die Frage beziehen, *was* beobachtet wird, tritt bei hochinferenten Ratings während der Ausführung eine weitere Herausforderung hinzu, die aus fachspezifischer Sicht Aufmerksamkeit erregt. Dabei handelt es sich um die Frage, *wer* die Unterrichtsqualitätsmerkmale im Ratingprozess beurteilt. In der Lehrerkompetenzforschung wurde in den letzten Jahren betont, dass die professionelle Wahrnehmung, dazu gehört das Erkennen und Interpretieren von relevanten Unterrichtsmerkmalen, eine Funktion von professionellem Wissen ist. Im sog. advokatorischen Ansatz (Oser, Heinzer & Salzmann, 2010) werden beispielsweise Noticing-Fähigkeiten von Lehrkräften als Indikator für deren Expertise verwendet. In den Beiträgen von Vieluf et al. (in diesem Heft) und Fauth, Göllner, Lenske, Praetorius und Wagner (in diesem Heft) wird die Rolle von Vorerfahrungen und Wissen sowie eigenen Prädispositionen für die Wahrnehmung des Unterrichtsangebots durch Lernende, Lehrende und durch Forschende (an-)diskutiert. Es stellt sich die Frage, welche Kompetenzen die Beurteilenden aufweisen müssen, um hochinferente Ratings mit (explizitem oder impliziten) fachlichem Bezug durchführen zu können, beispielsweise bei der Einschätzung, ob eine Frage im Fachunterricht kognitiv aktivierend ist. Beurteilende können die Lernenden selbst, die betroffenen Lehrkräfte oder aber externe Ratende sein, wobei das Phänomen nicht-korrespondierender Urteile zwischen verschiedenen ‚Beurteilergruppen' als „perspektivenspezifische Unterrichtsqualität" bezeichnet wird. Fauth et al. (in diesem Heft) reflektieren dafür verschiedene Ursachen, von den gewählten Methoden und Indikatoren bis hin zur Formulierungsebene, und arbeiten so die damit verbundenen Herausforderungen auf. Wir möchten hier darüber hinaus einen bisher eher wenig betrachteten Aspekt innerhalb einer Perspektive adressieren.

Effekte unterschiedlicher externer Ratender[1] werden in der Regel mit Blick auf die Objektivität und Reliabilität der Urteile untersucht. Die in diesem Beiheft vorliegenden empirischen Beiträge berichten entsprechende Qualitätssicherungsmaßnahmen, womit sichergestellt wird, dass geschulte Personen ein ausgearbeitetes Ratingmanual möglichst intersubjektiv reliabel anwenden können. Unklar bleibt dabei, welchen Einfluss Wissen, Normen, Werthaltungen oder gar die Expertise der Beurteilenden selbst haben. Können Raterinnen und Rater mit einem bestimmten Hintergrund, beispielsweise in Erziehungswissenschaften oder Psychologie, fachliche Fehlvorstellungen von Schülerinnen und Schülern im (Chemie-, Deutsch-, Geschichts-, Mathematik-, Physik-)Unterricht erkennen[2] und valide einschätzen, ob Lehrkräfte damit fachdidaktisch adäquat um-

1 Bei Beurteilungen durch Lehrende selbst ist die Frage nach der Intersubjektivität hinfällig (Ausnahme: team-teaching). Bei der Beurteilung durch Lernende kann die Kohärenz der Urteile in einer Klasse als Indikator für die Beurteilungsgüte gesehen werden (z. B. Göllner, Fauth, Lenske, Praetorius & Wagner, in diesem Heft), in anderen Studien wird aber Varianz innerhalb der Klasse als Hinweis auf differenziell wirkenden Unterricht interpretiert.

2 Noch anspruchsvoller und gravierender sind fachliche Fehler der Lehrkräfte im Unterricht.

gehen oder die Fehlvorstellungen ggf. noch verstärken? Konkret geht es um die Frage, bis zu welchem Grad in Schulungen neben intersubjektiv übereinstimmenden Ratings auch kriterial suffiziente Beurteilungen der jeweiligen Unterrichtsqualitätsmerkmale erreicht werden können. Ausgehend von den oben erwähnten Ergebnissen zu Noticing-Fähigkeiten von Lehrkräften ist dies ein fachdidaktischer Aspekt der Messung von Unterrichtsqualität, der nicht zu verlässigen ist, bislang empirisch aber nicht in den Blick genommen wird.

3. Thin Slices, Unterrichtseinheit oder Unterrichtssequenz: Wann sind fachdidaktische Qualitätskriterien beobachtbar?

Beobachtungsstudien zur Erfassung von Unterrichtsqualität beschränken sich aufgrund des hohen Aufwands notgedrungen auf wenige Unterrichtsstunden (oft sogar nur eine Stunde pro Klasse). Aufgrund der hohen Kosten sind die Forschenden bemüht, möglichst ökonomische Methoden zur Qualitätsmessung von Unterricht zu entwickeln, um bei gleichem Aufwand mehr Unterrichtsstunden untersuchen zu können. Ein interessanter Ansatz dieser Optimierungsversuche stellt das Thin-Slices-Verfahren bei der videobasierten Erfassung von Unterrichtsqualität dar. Beispielsweise wurden in der Studie von Begrich, Fauth, Kunter und Klieme (2017) von jeder der 37 Sachunterrichtsstunden drei zufällig ausgewählte 10-sekündige Ausschnitte von ungeschulten Beurteilenden in Bezug auf (generische) Qualitätsmerkmale bewertet. Es ergab sich eine hohe Interrater-Übereinstimmung, und die gemessene Unterrichtsqualität wies eine angemessene prognostische Validität im Hinblick auf den Lernerfolg auf. Auch wenn die Betrachtung von fachdidaktischen Merkmalen der Unterrichtsqualität im Thin-Slices-Verfahren noch aussteht, so kann vor dem Hintergrund der Diskussion im vorherigen Abschnitt mit aller Vorsicht vermutet werden, dass eine fachdidaktische Qualitätseinschätzung auf Basis von drei 10-sekündigen Ausschnitten durch fachdidaktisch ungeschulte Ratende herausfordernd sein dürfte. Diese Annahme beruht auf dem Hintergrund, dass eine Einschätzung fachdidaktischer Aspekte von Lehr-Lern-Prozessen in kurzen und damit situationalen Ausschnitten von Unterricht nur schwer möglich ist.

Die meisten Unterrichtsstudien betrachten eine Unterrichtsstunde. Damit liegt eine zeitliche Beobachtungseinheit vor, die der Planungseinheit der Lehrkraft für den Fachunterricht entspricht und somit eine weitergehende fachdidaktische Bewertung erlaubt, die Fischer, Reyer, Wirz, Bos und Höllrich (2002, S. 132) als den Versuch, „den ‚latenten Plan' des Lehrers für den Unterricht im Verlauf zu interpretieren" beschrieben haben. Dabei geht es um die fachspezifische Tiefenstruktur des Fachunterrichts, die Fischer und Kollegen für den Physikunterricht operationalisiert haben, indem sie die Basismodelle von Oser und Patry (1990) für das unterrichtliche Physiklernen adaptierten (s. a. QuIP-Projekt, Fischer, Labudde, Neumann & Viiri, 2014; vgl. auch Decristan et al., in diesem Heft).

Aber auch die Betrachtung einer Unterrichtsstunde ermöglicht noch keine Bewertung von Angebotsstrukturen, die sich über mehrere Unterrichtsstunden entwickeln.

Praetorius, Pauli, Reusser, Rakoczy und Klieme (2014) konnten zeigen, dass insbesondere die Basisdimension der kognitiven Aktivierung, die am stärksten durch fachdidaktische Aspekte beeinflusst ist (s. Abschnitt 2.1), ein instabiler Merkmalsbereich ist und für verlässliche Aussagen eine Beobachtung von mindestens neun Unterrichtsstunden benötigt wird. Geht man davon aus, dass die fachdidaktische Planung von Unterricht idealtypisch nicht in Form von isolierten Unterrichtsstunden erfolgt, sondern als Unterrichtssequenz mit aufeinander aufbauenden Stunden unterschiedlicher inhaltlicher Zielsetzung, so sollte eine Betrachtung von themenbezogenen Unterrichtssequenzen das Potenzial für die Erfassung fachdidaktischer Qualitätsmerkmale erhöhen. Würden längerfristige Angebotsstrukturen betrachtet, so könnte man die Stabilität von Konstrukten wie der kognitiven Aktivierung (vgl. auch Seidel, in diesem Heft) und gleichzeitig die fachdidaktische Angebotsstruktur eines Themenbereichs untersuchen. Fachdidaktische Konzepte wie etwa die instruktionale Kohärenz eines inhaltlichen Unterrichtsangebots, die als Kohärenz der Lernziele, Kohärenz innerhalb von (mehrstündigen) Unterrichtseinheiten und Kohärenz über mehrere, zeitlich voneinander getrennte Unterrichtseinheiten konzeptualisiert wird (Shwartz, Weizman, Fortus, Krajcik & Reiser, 2008), bieten hier Ansätze zur Entwicklung von fachdidaktischen Qualitätsskalen. Da die Einschätzung der fachdidaktischen Qualität der Angebotsstruktur eine entsprechende Expertise voraussetzt, ist eine Erfassung auf Basis von Schülereinschätzungen nur schwer vorstellbar. Um aufwändige Unterrichtsbeobachtungen oder Videoaufzeichnungen zu vermeiden, wären beispielsweise Selbstberichte von Lehrkräften über ihre Unterrichtsplanung, die verwendeten Materialien, Aufgaben oder Arbeitsblätter eine ökonomische Variante zur Erhebung von Rohdaten.

4. Zusammenfassung und Ausblick

Zusammenfassend ergibt sich durch die Diskussion der Unterrichtsforschung aus fachdidaktischer Perspektive ein durchaus komplexes Bild. Normen beeinflussen die Unterrichtsqualitätsforschung bei der Fokussierung des Forschungsinteresses, der Operationalisierung und dem eigentlichen Messvorgang. Normen unterscheiden sich zwischen Kulturen, aber auch zwischen anderen sozialen Kontextsystemen, beispielsweise Schulformen und Fächern, da unterschiedliche Zielsetzungen in unterschiedlichen pädagogischen Traditionen verfolgt werden. Bisher findet eine explizite Reflektion dieser implizit wirkenden Normen (ob generisch oder fachspezifisch) nur in geringem Ausmaß statt, sodass die Reichweite bzw. Generalisierbarkeit der vorhandenen Ergebnisse nicht eingeschätzt werden kann (vgl. auch Praetorius et al., in diesem Heft). Vergleichende oder kontrastierende Studien, die explizit den Einfluss von Normen adressieren, wären hier wünschenswert (z. B. analog zu Dreher, Lindmeier, Wang & Hsieh, 2018). Zur Erfassung und Bewertung von fachlichen und fachdidaktischen Merkmalen der Unterrichtsqualität (ob implizit oder explizit) sollten Personen mit entsprechender fachlicher Expertise einbezogen werden, da schon der Austausch zwischen den Disziplinen erhellend für die eigene Kultur wirken und so die Validität der Vorgehensweisen erhöhen kann.

Darüber hinaus ist anzudenken, fachdidaktische Aspekte der Unterrichtsqualität über die Betrachtung längerfristiger Angebotsstrukturen zu erfassen und ihre Prädiktivität für den Lernerfolg zu untersuchen.

Den bisherigen Ergebnissen der Unterrichtsforschung folgend und aufgrund der Integration generischer und fachspezifischer Aspekte in der Praxis des Unterrichts gehen wir nicht davon aus, dass es genügt, generische und fachspezifische Merkmale gesondert zu betrachten. Hingegen bietet sich die Hierarchisierung von Qualitätsmerkmalen des Fachunterrichts an. So wurde in Praetorius et al. (in diesem Heft) mit Bezug auf Openshaw und Clarke (1970) bereits erwähnt, dass die Klassenführung Voraussetzung für die konstruktive Lernerunterstützung und die kognitive Aktivierung sei. Brunner (2018) schlug in ihrem Beitrag ein erweitertes hierarchisches Modell von Qualitätsmerkmalen vor, bei dem neben der Klassenführung auch die fachliche Fundierung des Unterrichts (inkl. der fachlichen Korrektheit) die Voraussetzung für weitere generische und fachdidaktische Merkmale eines effektiven Fachunterrichts darstellt. Vor dem Hintergrund der in der Einleitung erwähnten Erkenntnisse, dass das implementierte Curriculum einen Einfluss auf die Effektivität von Unterricht hat, ist dieser Vorschlag plausibel.

Abschließend sei noch auf einen weiteren Ansatz für mögliche Studien hingewiesen: Die bisher nur geringe Rolle von fachdidaktischen Merkmalen in der Literatur zur Unterrichtsforschung könnte auch darauf zurückzuführen sein, dass die fachdidaktische Qualität im realen Unterricht – wenn sie untersucht wird – gering ist. Damit würden Skalen zu fachdidaktischen Qualitätsmerkmalen wenig Varianz aufweisen bzw. nicht ‚funktionieren'. Um dies als mögliche Erklärung zu untersuchen, könnten Studien interessant sein, in denen idealtypische, d. h. fachdidaktisch optimierte Stundensequenzen umgesetzt und dann mit den entsprechenden Ratingverfahren bewertet werden. Eine ähnliche Idee wurde in der IGEL-Studie (Fauth, Decristan, Rieser, Klieme & Büttner, 2014) mit generischen Unterrichtselementen verfolgt (z. B. *peer assisted learning, formative assessment*). Auch in der Fachdidaktik gibt es Modelle für das fachliche Lernen, auf deren Basis idealtypische *best practice*-Stunden umgesetzt werden könnten (vgl. den Verweis von Seidel, in diesem Heft auf die *best practice*-Forschung in den USA).

Bleibt zum Schluss die Antwort auf die Frage im Titel dieses Diskussionsbeitrags: Die fachdidaktische Perspektive in der Unterrichtsqualitätsforschung: (bisher) ignoriert, implizit enthalten oder nicht relevant? Einerseits lässt sich feststellen, dass die fachdidaktische Perspektive in der Unterrichtsqualitätsforschung bisher weder vollständig ignoriert wurde noch dass sie als irrelevant angesehen wird. Andererseits muss aber auch konstatiert werden, dass die fachdidaktische Perspektive selten explizit einbezogen wurde, sondern eher ein implizites Dasein im Rahmen fachdidaktischer Interpretationen von Qualitätsindikatoren fristete. Dies mag mit einer bisher stark psychologischen und erziehungswissenschaftlichen Prägung der Unterrichtsforschung zu tun haben, die mit einem speziellen Blick auf Unterricht und damit verbundenen Fragestellungen und Forschungszugängen einhergeht. Eine Erweiterung dieses Blicks um fachdidaktische Fragestellungen wäre aus unserer Sicht ein naheliegender und wichtiger nächster Schritt.

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 Erziehungswissenschaft, 20* (Suppl. 1), 23–47.

Berliner, D.C. (2005). The near impossibility of testing for teacher quality. *Journal of teacher education, 56*(3), 205–213.

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *Journal für Mathematik-Didaktik, 39*(2), 257–284.

Clarke, D.J. (2013a). Contingent conceptions of accomplished practice: The cultural specificity of discourse in and about the mathematics classroom. *ZDM, 45*(1), 21–33.

Clarke, D.J. (2013b). The validity-comparability compromise in crosscultural studies in mathematics education. In *Proceedings of the Eighth Congress of the European Society for Research in Mathematics Education* (S. 1855–1864).

Dreher, A., Lindmeier, A., Wang, T-Y., & Hsieh, F-J. (2018). Teacher Noticing in Taiwan und Deutschland – Wie stark prägen kulturelle Normen das Verständnis von Unterrichtsqualitätsmerkmalen? In *Beiträge zum Mathematikunterricht 2018* (S. 461–464). Münster: WTM.

Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit: fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht.* Münster: Waxmann.

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.

Fischer, H.E., Reyer, T., Wirz, C., Bos, W., & Höllrich, N. (2002). Unterrichtsgestaltung und Lernerfolg im Physikunterricht. In M. Prenzel & J. Doll (Hrsg.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen* (S. 124–138). Weinheim: Beltz.

Fischer, H.E., Labudde, P., Neumann, K., & Viiri, J. (Hrsg.) (2014). *Quality of instruction in physics: Comparing Finland, Switzerland and Germany.* Münster: Waxmann.

Heid, H. (2013). Logik, Struktur und Prozess der Qualitätsbeurteilung von Schule und Unterricht. *Zeitschrift für Erziehungswissenschaft, 16*(2), 405–431.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations.* Thousand Oaks: Sage publications.

Kirschner, P.A., Verschaffel, L., Star, J., & Van Dooren, W. (2017). There is more variation within than across domains: An interview with Paul A. Kirschner about applying cognitive psychology-based instructional design principles in mathematics teaching and learning. *ZDM, 49*(4), 637–643.

Klieme, E. (2013). Qualitätsbeurteilung von Schule und Unterricht: Möglichkeiten und Grenzen einer begriffsanalytischen Reflexion – ein Kommentar zu Helmut Heid. *Zeitschrift für Erziehungswissenschaft, 16*(2), 433–441.

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, 33,* 61–98.

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). Münster: Waxmann.

Lehner, M.C., Heine, J.-H., Sälzer, C., Reiss, K., Haag, N., & Heinze, A. (2017). Veränderung der mathematischen Kompetenz von der neunten zur zehnten Klassenstufe. *Zeitschrift für Erziehungswissenschaft, 33,* 7–36.

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

Mesiti, C., & Clarke, D. (2017). The international lexicon project: Giving a name to what we do. In R. Seah, M. Horne, J. Ocean, & C. Orellana (Hrsg.), *Proceedings of the Mathematical Association of Victoria annual conference* (S. 31–38). Brunswick, Vic.: Mathematical Association of Victoria.

Openshaw, K., & Clarke, S. C. T. (1970). General teaching theory. *Journal of Teacher Education, 21*(3), 403–416.

Oser, F., & Patry, J.-L. (1990). Choreographien unterrichtlichen Lernens: Basismodelle des Unterrichts. *Berichte zur Erziehungswissenschaft* (Nr. 89). Freiburg (CH): Pädagogisches Institut der Universität Freiburg.

Oser F., Heinzer S., & Salzmann P. (2010), Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten. Chancen und Grenzen des advokatorischen Ansatzes. *Unterrichtswissenschaft, 38*(1)1, 5–28.

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM, 50*(3), 535–553.

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

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

Reusser, K. (2009). Empirisch fundierte Didaktik – didaktisch fundierte Unterrichtsforschung. In *Perspektiven der Didaktik* (S. 219–237). VS Verlag für Sozialwissenschaften.

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *ZDM, 50*(3), 475–490.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM, 45*(4), 607–621.

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.

Shwartz, Y., Weizman, A., Fortus, D., Krajcik, J. S., & Reiser, B. J. (2008). The IQWST experience: Using coherence as a design principle for a middle school science curriculum. *Elementary School Journal, 109*(2), 199–219.

Sievert, H., van den Ham, A-K., Niedermeyer, I., & Heinze, A. (2019,). Effects of mathematics textbooks on the development of primary school children's adaptive expertise in arithmetic. *Learning and Individual Differences, 74,* 1–13.

Vehmeyer, J. K. (2009). *Kognitiv anregende Verhaltensweisen von Lehrkräften im naturwissenschaftlichen Sachunterricht: Konzeptualisierung und Erfassung*. Dissertation Universität Münster. https://miami.uni-muenster.de/Record/74b36e17-38c8-4130-8e9c-f14834595217 [08. 10. 2019].

Abstract: The paper discusses the articles on teaching quality in this special issue from the perspective of subject-specific educational research. First, we specify the view adopted by focusing on normative aspects of teaching quality. Against this background, we then analyse the role of subject-specific aspects for indicators of teaching quality considered so far. Subsequently, we discuss to what extent the common practice of analysing short time segments of classroom teaching can be used to measure teaching quality with criteria derived from subject-specific educational research. Finally, we address aspects of teaching quality that are not considered in the articles of this special issue, but may be relevant for a subject-specific perspective.

Keywords: Teaching Quality, Instructional Research, Research Perspectives, Subject-Specific Educational Research, Quality of Subject Teaching

Anschrift der Autor_innen

Prof. Dr. Anke Lindmeier, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik Kiel,
Olshausenstraße 62, 24118 Kiel, Deutschland
E-Mail: lindmeier@leibniz-ipn.de

Prof. Dr. Aiso Heinze, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik Kiel,
Olshausenstraße 62, 24118 Kiel, Deutschland
E-Mail: heinze@leibniz-ipn.de

Zeitschrift für Pädagogik

Begründet durch:
Fritz Blättner, Otto Friedrich Bollnow, Josef Dolch, Wilhelm Flitner, Erich Weniger

Fortgeführt von:
Cristina Allemann-Ghionda, Dietrich Benner, Herwig Blankertz, Hans Bohnenkamp, Wolfgang Brezinka, Josef Derbolav, Andreas Flitner, Carl-Ludwig Furck, Georg Geissler, Oskar Hammelsbeck, Ulrich Herrmann, Diether Hopf, Walter Hornstein, Wolfgang Klafki, August Klein, Doris Knab, Andreas Krapp, Martinus J. Langeveld, Achim Leschinsky, Ernst Lichtenstein, Peter-Martin Roeder, Wolfgang Scheibe, Hans Scheuerl, Tina Seidel, Hans Schiefele, Franz Vilsmeier

Herausgeber_innen:
Sabine Andresen (Frankfurt), Marcelo Alberto Caruso (Berlin), Kai S. Cortina (Michigan), Reinhard Fatke (Zürich), Werner Helsper (Halle), Eckhard Klieme (Frankfurt), Roland Merten (Jena), Jürgen Oelkers (Zürich), Sabine Reh (Berlin), Roland Reichenbach (Zürich), Petra Stanat (Berlin), Heinz-Elmar Tenorth (Berlin), Ewald Terhart (Münster), Rudolf Tippelt (München)

Die Zeitschrift für Pädagogik wird in folgenden Datenbanken und bibliografischen Diensten ausgewertet:
– CIJE (Central Index to Journals in Education, Phoenix, USA)
– ERIC (Educational Resources Information Center, Washington D.C., USA)
– ERIH PLUS (European Reference Index for the Humanities, Bergen, Norwegen)
– FIS Bildung (Fachinformationssystem Bildung, Frankfurt a. M.)
– PSYINDEX (Zentralstelle für Psychologische Information und Dokumentation, Trier)
– SSCI (Social Sciences Citation Index, Institute for Scientific Information, Philadelphia, USA)
– SOLIS (Informationszentrum Sozialwissenschaften, Bonn)