pedocs
Open Access Erziehungswissenschaften

DIPF

Naumann, Alexander; Kuger, Susanne; Köhler, Carmen; Hochweber, Jan

Conceptual and methodological challenges in detecting the effectiveness of learning and teaching

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, S. 179-196. - (Zeitschrift für Pädagogik, Beiheft; 66)

Quellenangabe/ Reference:
Naumann, Alexander; Kuger, Susanne; Köhler, Carmen; Hochweber, Jan: Conceptual and methodological challenges in detecting the effectiveness of learning and teaching - In: 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, S. 179-196 - URN: urn:nbn:de:0111-pedocs-258732 - DOI: 10.25656/01:25873

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-258732
https://doi.org/10.25656/01:25873

in Kooperation mit / in cooperation with:

BELTZ JUVENTA

http://www.juventa.de

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.
This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

pedocs
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der Leibniz-Gemeinschaft

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

Die in der Zeitschrift veröffentlichten Beiträge sind urheberrechtlich geschützt.
Alle Rechte, insbesondere das der Übersetzung in fremde Sprachen, bleiben
dem Beltz-Verlag vorbehalten.
Kein Teil dieser Zeitschrift darf ohne schriftliche Genehmigung des Verlages in
irgendeiner Form – durch Fotokopie, Mikrofilm oder ein anderes Verfahren –
reproduziert oder in eine von Maschinen, insbesondere Datenverarbeitungsanlagen,
verwendbare Sprache übertragen werden. Auch die Rechte der Wiedergabe durch
Vortrag, Funk- und Fernsehsendung, im Magnettonverfahren oder auf ähnlichem
Wege bleiben vorbehalten. Fotokopien für den persönlichen oder sonstigen eigenen
Gebrauch dürfen nur von einzelnen Beiträgen oder Teilen daraus als Einzelkopie
hergestellt werden. Jede im Bereich eines gewerblichen Unternehmens hergestellte
oder genützte Kopie dient gewerblichen Zwecken gem. § 54 (2) UrhG und verpflichtet
zur Gebührenzahlung an die VGWort, Abteilung Wissenschaft, Goethestr. 49,
80336 München, bei der die einzelnen Zahlungsmodalitäten zu erfragen sind.

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

1. Auflage 2020

© 2020 Beltz Juventa
in der Verlagsgruppe Beltz · Weinheim Basel
Werderstraße 10, 69469 Weinheim
Alle Rechte vorbehalten

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

Themenblock V: Modellierung der Wirkungen von Unterrichtsqualität (englischsprachig)

Kommentare

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

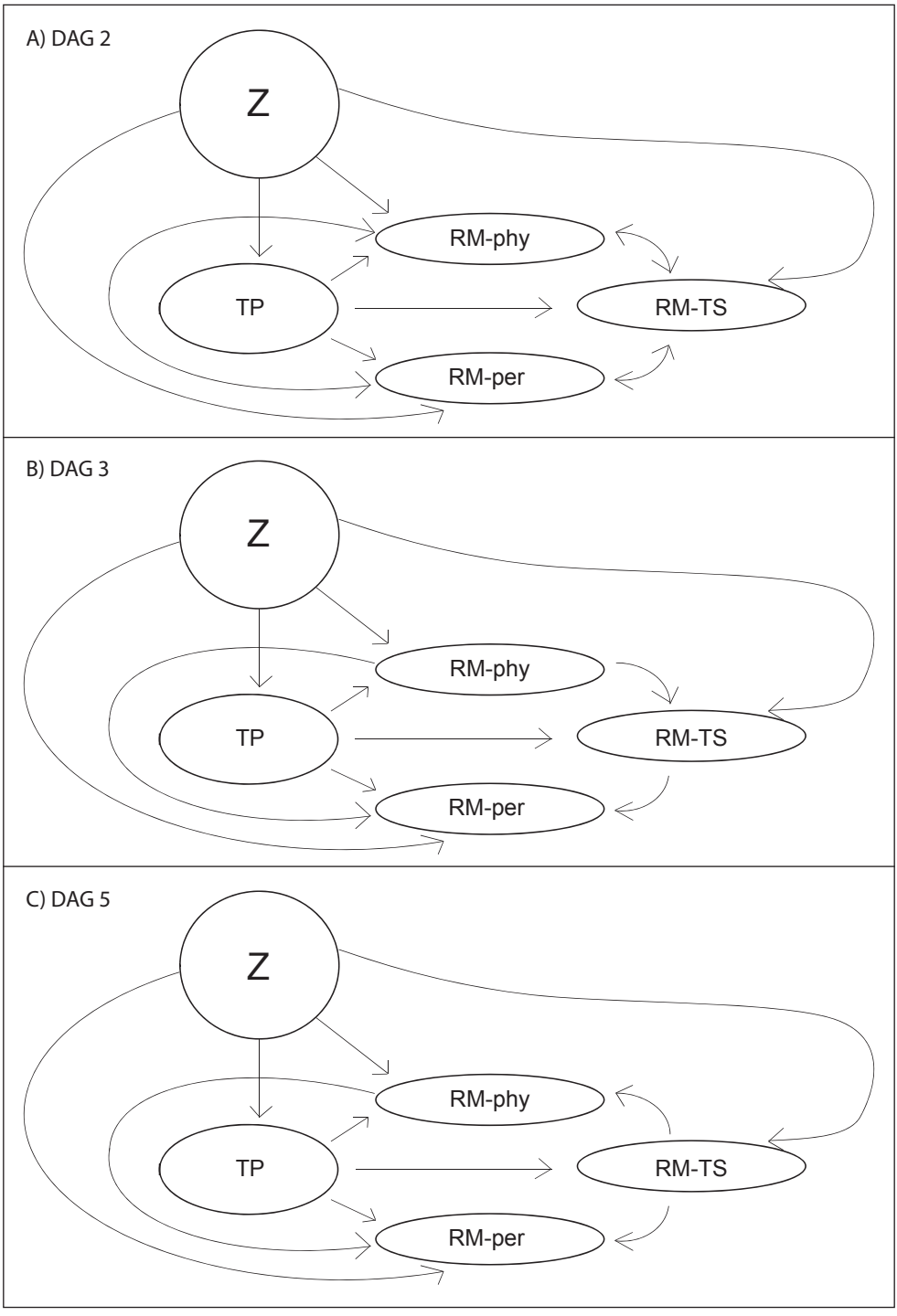

Fig. 1: Examples for DAGS (adapted from Boerebach et al., 2013)

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, 21*6(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