Köhler, Carmen; Kuger, Susanne; Naumann, Alexander; Hartig, Johannes

Multilevel models for evaluating the effectiveness of teaching. Conceptual and methodological considerations

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. 197-209. - (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

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

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.

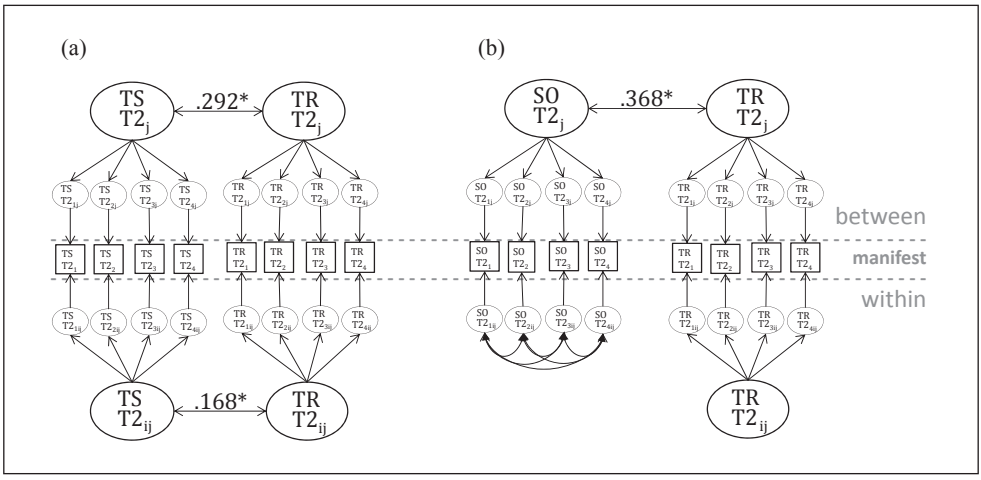

Fig. 1: Doubly-latent multilevel model estimating (a) the relationship between text reconstruction (TR) and the configural construct teacher supportiveness (TS) and (b) the relationship between text reconstruction (TR) and the shared construct student orientation (SO) at the second measurement occasion.

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.

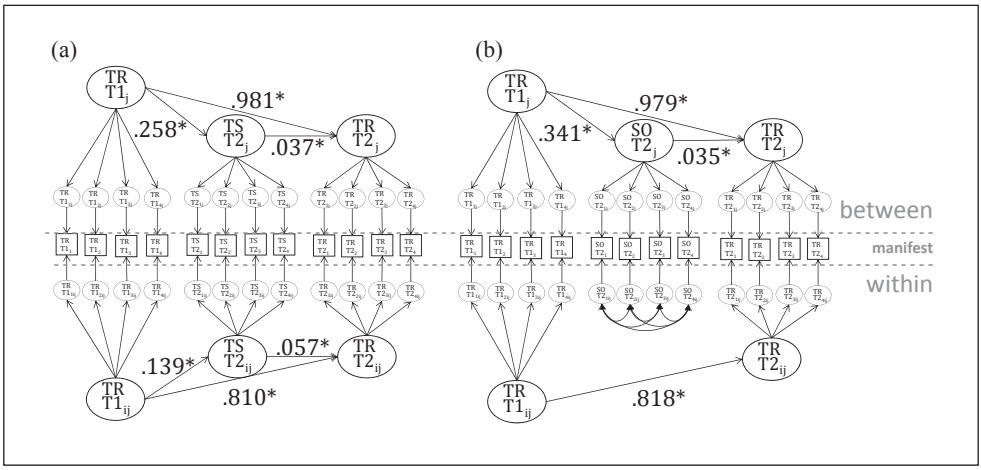

Fig. 2: Doubly-latent multilevel model estimating (a) the relationship between text reconstruction (TR) and the configural construct teacher supportiveness (TS) and (b) the relationship between text reconstruction (TR) and the shared construct student orientation (SO) at the second measurement occasion, 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.

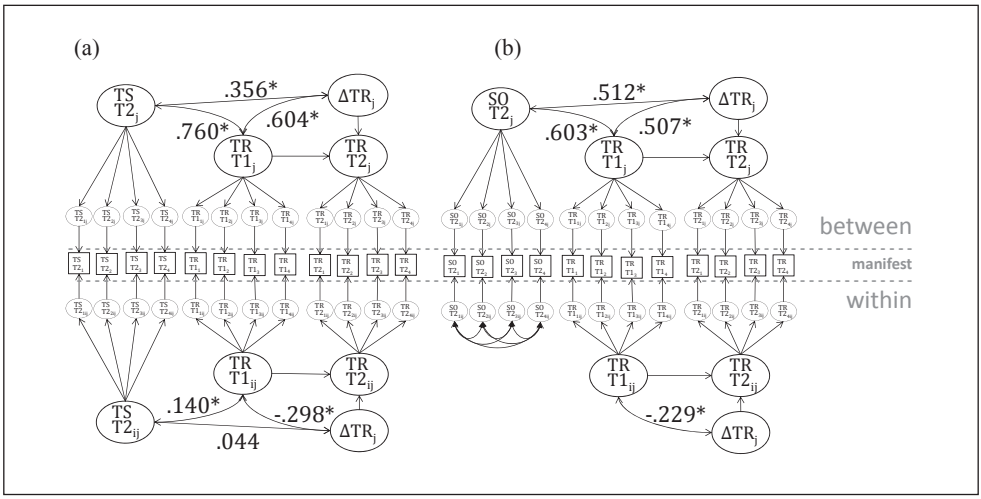

Fig. 3: Doubly-latent multilevel model estimating (a) the relationship between the text reconstruction change score (ΔTR) and the configural construct teacher supportiveness (TS) and (b) the text reconstruction change-score (ΔTR) and the shared construct student orientation (SO).

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