Köhler, Carmen; Hartig, Johannes; Naumann, Alexander

Detecting instruction effects. Deciding between covariance analytical and change-score approach

Educational psychology review 33 (2021) 3, S. 1191-1211

Educational Psychology Review (2021) 33:1191–1211
https://doi.org/10.1007/s10648-020-09590-6

REFLECTION ON THE FIELD

Detecting Instruction Effects—Deciding Between Covariance Analytical and Change-Score Approach

Carmen Köhler[1] · Johannes Hartig[1] · Alexander Naumann[1]

Accepted: 9 December 2020 / Published online: 27 January 2021

Abstract
The article focuses on estimating effects in nonrandomized studies with two outcome measurement occasions and one predictor variable. Given such a design, the analysis approach can be to include the measurement at the previous time point as a predictor in the regression model (ANCOVA), or to predict the change-score of the outcome variable (CHANGE). Researchers demonstrated that both approaches can result in different conclusions regarding the reported effect. Current recommendations on when to apply which approach are, in part, contradictory. In addition, they lack direct reference to the educational and instructional research contexts, since they do not consider latent variable models in which variables are measured without measurement error. This contribution assists researchers in making decisions regarding their analysis model. Using an underlying hypothetical data-generating model, we identify for which kind of data-generating scenario (i.e., under which assumptions) the defined true effect equals the estimated regression coefficients of the ANCOVA and the CHANGE approach. We give empirical examples from instructional research and discuss which approach is more appropriate, respectively.

Keywords Change-score model · Conditional model · Instruction effect · Multilevel SEM

✉ Carmen Köhler
carmen.koehler@dipf.de

Johannes Hartig
hartig@dipf.de

Alexander Naumann
naumanna@dipf.de

[1] DIPF | Leibniz Institute for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt, Germany

In educational research, the focus often lies on identifying features at the student, class, or school level that positively influence relevant student outcomes. A simplistic strategy to test the effects of a teaching, class, or school characteristic is to regress the outcome variable on this predictor. This, however, only informs about the current state of the relationship between two variables and not about the effectiveness of teaching, since such bivariate analyses fail capturing a process, and thus allow no assumptions regarding causality (Köhler et al. 2020b). More elaborate research designs allow taking threats against proper inferences from a study into account. For example, differences between groups that already existed prior to a treatment or differences before and after treatment that are due to maturation can be ruled out by including a measure prior to the treatment and including a control group. If these actions were taken when deciding on the study design, the respective measures can be included in the data analysis model. Given a specific study design, there exist still various options for an analysis model. For example, if a study entails two measurement occasions of the outcome variable, the measurement at the first time point (i.e., outcome at T1) can be included as a predictor in the regression model (see e.g., Holzberger et al. 2013; Lazarides and Buchholz 2019; Marsh et al. 2012; Morin et al. 2014; Spinath and Steinmayr 2012; Stipek and Valentino 2015; Vollet et al. 2017; Wagner et al. 2016). These so-called auto-regression or covariance-analytical models control for initial differences between the entities of measurement such as individuals or groups. We label this approach the ANCOVA approach, which can be written as

$$\text{outcome}T2_i = \beta 0 + \beta 1 X_i + \beta 2 \text{outcome}T1_i + e_i,$$

where i indexes an individual if the individual level is of interest or a group when the group level is of interest. Another approach to estimate the effect in such settings is to predict the change-score of the outcome variable (see, e.g., Ahmed et al. 2014; Hochweber and Vieluf 2016; Otis et al. 2005). Here, the change between the two measurement occasions serves as the dependent variable. We refer to this approach as the CHANGE approach, which can be expressed as

$$\text{outcome}T2_i - \text{outcome}T1_i = \beta 0 + \beta 1 X_i + e_i.$$

Note that throughout the article, we use the terms *outcome at T1* and *outcome at T2* instead of *pretest* and *posttest* to stress that the measurements of the variable of interest (i.e., the variable instruction has an effect on) needs to be on the same scale at T1 and T2 in order to even apply the CHANGE approach. Therefore, we label it the same: the outcome variable—measured at two different measurement occasions.

Both approaches aim at identifying the effect of the predicting variable, for example, student-perceived teaching quality or student competence beliefs, on relevant educational outcomes such as motivation, achievement, or competence. Several researchers demonstrated that both approaches can result in different conclusions regarding the reported effect (Allison 1990; Holland and Rubin 1983; Köhler et al. 2020a; Lord 1967; Maris 1998; Van Breukelen 2013; Wainer and Brown 2004; Wright 2006). This inconsistency is referred to as Lord's paradox, because Lord (1967) was the first to point it out. Köhler et al. (2020a) transferred Lord's paradox into the context of instructional research, pointing out discrepancies: Whereas Lord focused on the comparison between two groups who received a different treatment, in instructional research, the comparison concerns several classes that received different forms of instruction. Unless an intervention study in which randomly selected teachers received a specific form of training is performed, the form of instruction is teacher-inherent and not

randomly assigned. Köhler et al. (2020a) show that, in instructional research scenarios without random assignment of teachers to interventions, neither approach gives an unbiased estimate of the instruction effect. Under specific assumptions, either one or both approaches show unbiased effects. In the paper, it is argued that researchers need to debate—for each examined variable constellation separately—which assumptions are likely to hold in order not to produce artificial instruction effects. The authors also provide a syntax that allows practitioners to estimate—given their assumptions—the bias the methods produce.

The focus of the current paper is to discuss these different scenarios and provide a guideline for researchers who adjudicate on whether to use the ANCOVA or the CHANGE approach. We use empirical examples from instructional research to illustrate the necessary considerations for making an informed decision. These considerations concern the study design, the time points of the assessed variables, and possible time point–specific and time point invariant cofounders. Based on these considerations, assumptions regarding the hypothetical data-generating model can be formulated, which in turn allows identifying the more appropriate approach. The main aim of the manuscript is hence to assist educational and instructional researchers in making decisions regarding their analysis model.

The paper is set up as follows: We first provide an illustrative data example in which we conducted an analysis using both approaches, which led to deviating results. Subsequently, we give a summary of current recommendations for practitioners on when to apply which approach. Since these recommendations are not straightforward for most applied contexts of educational and instructional research (including our example), we use the idea of Köhler et al. (2020a) for an underlying data-generating model and apply it to various examples from practice, discussing for each which approach is more reliable. Based on this, we return to our data example to draw a conclusion, and finish with a general discussion.

Data Example

In this section, we demonstrate that the ANCOVA and CHANGE approach can produce differing results using an empirical example. We applied both approaches to data 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 (Beck and Klieme 2007). The sample size was $N = 10{,}985$; the number of classes was 427 (with a minimum of nine students and a maximum of 36 students in a class). The exemplary research question we aimed to answer concerns the effect of teacher supportiveness (Prenzel et al. 2006) on students' English listening comprehension skills. While student skills such as listening comprehension were tested at the beginning and at the end of the school year 2003/2004, the student questionnaire containing statements about the teacher was only provided at the second measurement occasion.

Teacher supportiveness was measured at the student level with four items, rated on a four-point Likert scale (1 = *Untrue*, 2 = *Somewhat untrue*, 3 = *Somewhat true*, 4 = *True*). The items inquired about the received support from the teacher, for example, "My English teacher gives me advice on how to improve". Listening comprehension was assessed by providing the students with six two audio texts and seven to ten corresponding multiple-choice items (Nold and Rossa 2007). In our analyses, we based the latent variable *listening comprehension* on the six texts and formed parcels by calculating the proportion of correct responses per text, using them as manifest indicators. Although students received different test booklets at T1 and T2, the simultaneous scaling of all students at all times points puts the items on the same scale.

Following the notation by Lüdtke et al. (2008), we ran two multilevel latent variable models with students nested in classes. In accordance with Marsh et al. (2009), our approach is doubly latent insofar that each latent factor at each level is measured using multiple indicators. A beneficial aspect to this approach is that it controls for both sampling and measurement error at both levels (Lüdtke et al. 2011; Marsh et al. 2009). To estimate the effect of teacher supportiveness on listening comprehension skills, we applied (1) the ANCOVA approach and (2) the CHANGE approach (see Fig. 1a and b, respectively).

Indicator loadings were restricted to be equal on both levels for both English listening comprehension skills and teacher supportiveness (cross-level invariance). Additionally, scalar invariance restrictions (i.e., identical loadings and intercepts) were applied to the measurement model of English listening comprehension skills between both time points. In the ANCOVA model, we included listening comprehension at T1 at both the within and the between level. We thus accounted for individual levels of previous achievement and class average levels of previous achievement (Morin et al. 2014). The teaching quality variable teacher support was also regressed on listening comprehension at T1 (at both levels), since individual student performance and average class performance might be related to the behavior of the teacher. In the CHANGE model, an additional latent change-score variable ΔLC was introduced at both levels (see Fig. 1b). At L1, the change-score represents the difference between a student's skill at T2 and at T1 relative to the class mean at the respective time points; at L2, it represents the difference between the average classroom skill level at T1 and the average classroom skill level at T2. All analyses were conducted using the software M*plus* 7.4 (Muthén and Muthén 1998-2015). Missing data were dealt with the full-information-maximum-likelihood (FIML) approach in M*plus*.

Results demonstrated that under the ANCOVA model, the regression coefficient when regressing listening comprehension at T2 on teacher supportiveness was not statistically significant ($\beta_{CA} = 0.034$, SE = 0.018, $p = .059$). In the change-score approach, however, the regression of the change-score on teacher supportiveness was about 30% larger and statistically significant ($\beta_{CS} = 0.090$, SE = 0.029, $p = .002$). Note that the standardized regression coefficients, which were $\beta^*_{CA} = 0.057$ and $\beta^*_{CS} = 0.259$, are not directly comparable, since they are not on the same scale (Köhler et al. 2020a). The question that arises is how these two differing results come about and which one can be trusted.

Both models have the same number of degrees of freedom and an identical global model fit ($\chi^2 = 715.5$; $df = 226$; RMSEA = 0.014; CFI = 0.963; TLI = 0.960). This illustrates that both models can be transformed into another, and the empirical model fit is no criterion for model selection. In the following, we summarize what the existing literature recommends thus far, and discuss these recommendations in light of educational and instructional research.

Existing Recommendations for Practice

The arguments for or against one of the approaches are based on a combination of the design of the study, the interrelations between the involved variables, and the underlying assumptions of the approaches. In the following, we summarize pieces of advice from different authors on the basis of different aspects. Some authors even argue that the approaches answer different research questions (Hand 1994; Hand and Taylor 1987; Holland and Rubin 1983; Köhler et al. 2020b; Wright 2006), which we discuss in the final sub-section.

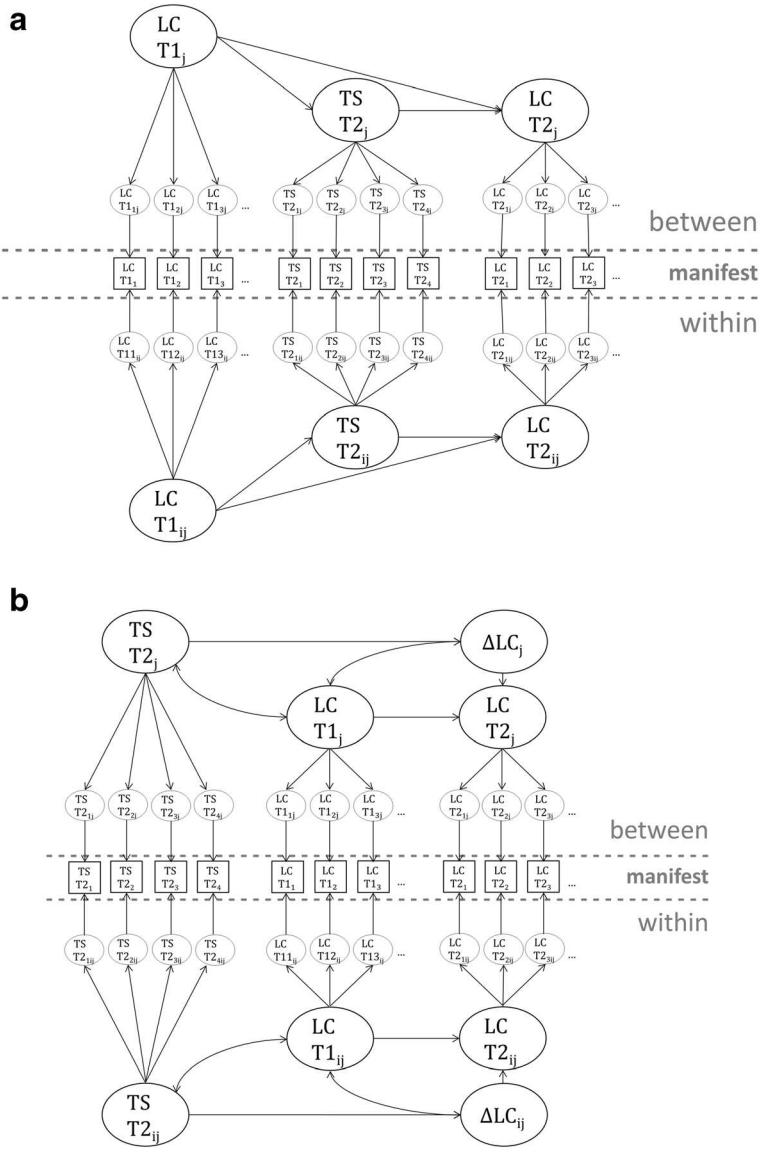

Fig. 1 Doubly-latent multilevel model estimating the effect of teacher support (TS) on listening comprehension (LS) using (**a**) the ANCOVA approach, controlling for listening comprehension at the first measurement occasion (T1) and (**b**) the CHANGE approach, calculating a latent change score for listening comprehension (ΔLC)

Selection into Treatment

One major distinction that is made with respect to the study design concerns the so-called *selection into treatment*. With randomized treatment assignment, both the ANCOVA and the CHANGE approach perform well and produce an unbiased treatment effect. Note that randomized treatment assignment will ensure that the outcome at T1 is independent of

treatment assignment. Several authors argue that the ANCOVA approach should be preferred in this setting, because it has more power than the CHANGE approach (Oakes and Feldman 2001; Wright 2006).

In nonrandomized settings, decisions for one or the other approach are less clear-cut. A number of authors recommend using the ANCOVA approach simply in general, arguing that it is appropriate more often (Maxwell and Delaney 2004; Senn 2006). Others advocate for it solely under specific circumstances, in some instances even contradicting each other. For example, whereas van Breukelen (2013) and Wright (2006) argue in favor of the ANCOVA method when selection into treatment depends on the outcome at T1, Jamieson (2004) states that the pretest should show no relation to the grouping variable. Several authors advocate for the CHANGE approach and demonstrate scenarios in which the CHANGE approach returns unbiased treatment effects (Kim and Steiner 2019; Liang and Zeger 2000; Wright 2006): For example, if outcome at T1 and T2 are related via a third variable and this variable is related to selection into treatment (Wright 2006).

Time-Invariant vs. Time-Specific Influences

We only give examples of the assumptions under which the approaches return unbiased effects because they turn even more complex when not only the variables themselves and their interrelations but also the trait- and state-specific components of the variables—or, put differently, time-invariant and time-specific parts of the construct—are considered (Allison 1990; Imai and Kim 2019). The distinction between trait- and state-specific components or time-invariant and time-specific influences stems from the considerations that the observed variance of a variable is a result of different factors/components/confounders, which can have an influence on the estimated treatment effect. For example, a time-specific influence that is independent of the treatment might cause a relation between the treatment variable and the outcome variable at T2 (e.g., an upcoming exam at the time of measurement), thus resulting in an overestimation of the treatment effect.

Many articles on the topic conclude that, in the end, the decision remains up to the researcher. Allison (1990), for example, states that the "decision should be based on our beliefs about which model better represents reality." (p. 105). More recent papers by Wainer and Brown (2004), Wright (2006) or van Breukelen (2013) declare that the justification for one or the other approach depends on untestable assumptions. Many of these researchers suggest applying both (or even multiple) approaches to the data and evaluate the consistency of effects across methods. At the same time, however, they all stress for readers to consider that both approaches might be wrong, meaning that even if they produced the same results, researchers can still not trust their results to be unbiased.

Besides the lack of consistent guidelines, another aspect that played a major part in previous studies on the topic was measurement error. Köhler et al. (2020a) were the first to address the comparison between the two approaches in a multilevel SEM framework. They point out that in a latent variable framework, one component that threatens unbiased inferences when using manifest variables, namely measurement error, needs not be dealt with. Using a theoretical underlying data-generating model, they outline which components are necessary to consider when deciding between the two approaches besides measurement error, and give recommendations for research designs that allow unbiased estimates of the instruction effect. The paper lacks examples from practice and clear guidelines regarding probable assumptions about the components and their interrelations.

Equivalence at the Baseline Measure

Xiao et al. (2019) took an empirical approach to uncovering reasons for inconsistencies between the two modeling options, and additionally investigated how a consideration of the multilevel structure that was present in their data influenced Lord's paradox. Their findings supported previous theoretical considerations that when groups have the same baseline measure, the approaches report almost identical results. When this is not the case, both the sign and the magnitude of the effect can differ. Considering the multilevel structure, the differences of the estimated effects of the approaches decreased, but there were still instances in which one approach would indicate a substantial effect whereas the other would not. Due to the unknown true effect, it was impossible to deduce which approach gave the more accurate estimate.

Equivalence of Research Question

Before describing how an underlying data-generating model can be helpful for deciding between the approaches and giving examples from practice, we want to point out a more substantial topic that was first brought up by Holland and Rubin (1983), who discussed Lord's paradox in detail, namely whether the two approaches, in fact, address the same research question. Wright (2006) gives a perfect example to demonstrate Lord's paradox and the potentially different research questions, which we reframe in light of an instructional research setting. Picture a study where ten classes, which are instructed by two teachers—five classes each—are given a test at the beginning of the school year. Suppose the teachers are identical twins who, in fact, do not treat the students any differently, and the students are retested at the end of the school year. Imagine that in the following year, you send your child to that school and are free to choose which twin becomes your child's teacher. Based on the data from the previous year, how could you decide without the knowledge that both instruct the kids in the exact same way? The overall question in this scenario is "does the teacher have an effect on learning outcome?" Say the mean outcome scores across all classes for teacher 1 were 30 in the pretest and 30 in the posttest; those for teacher 2 were 70 in the pretest and 70 in the posttest. Despite the fact that you might not want to send your child to either of the teachers at all anymore, we consider the conclusions from both approaches. In the CHANGE approach, which can be written as

$$\text{outcome}T2_c - \text{outcome}T1_c = \beta 0 + \beta 1\text{teacher}_x + e_c,$$

where c indexes the class and x indexes the teacher, the effect of the teacher is zero ($\beta 1 = 0$). Hence, the decision for either of the teachers makes no difference, since the occurring change is equal across both teachers. The ANCOVA approach, however, gives a different answer. The equation is given by

$$\text{outcome}T2_c = \beta 0 + \beta 1\text{teacher}_x + \beta 2\text{outcome}T1_c + e_c,$$

where $\beta 2$ captures the effect of preexisting differences in the pretest between the two teachers' classes, which was $\beta 2 = 0.5$ in the example. The effect of the teacher exceeds 0 ($\beta 1 = 20$): The expected score of your child is higher when you send it to a class from teacher 2 instead of a class from teacher 1. This seems odd, but essentially captures the paradox. The ANCOVA approach reports the effect of the teacher if all classes had started the school year at the same

level. For every child, regardless of the score at T1, the expected score at T2 is 20 points higher in a class taught by teacher 2. Put in statistical terms, the ANCOVA approach asks "whether the average gain, partialling out pre-scores, is different between the two groups" (Wright 2006, p. 666), whereas the CHANGE approach "asks whether the average gain in score is different for the two groups" (Wright 2006, p. 666). Note that the terms *change* or *gain* in conjunction with the ANCOVA approach appear misleading, as the approach involves no change-score (or gain score) in the same sense as the CHANGE approach. However, the ANCOVA approach is also applied with the interest of finding out whether groups or individuals have increased in score, and can be rewritten as

$$\text{outcome}T2_i - \text{outcome}T1_i = \beta 0 + \beta 1 X_i + (\beta 2 - 1)\text{outcome}T1_i + e_i.$$

The separation into two ways of formulating the research question is therefore artificial, and we argue that the question is identical. In the end, the primary interest lies on identifying the effect of instruction, and we would assume that $\beta 1$ in both approaches is an estimate of this effect. We explain in the remainder of the article why and under which circumstances $\beta 1$ differ between the approaches. In the example, it is relatively clear that the CHANGE approach gives us the presumed correct answer, namely that the child learns equally unsatisfactorily under both teachers. The ANCOVA approach would give us the same answer as the CHANGE approach if $\beta 2 = 1$ (see also Huck and McLean 1975; Wright 2006), which is hardly ever the case (Allison 1990) since it would indicate that the pretest perfectly predicts the posttest (i.e., the teachers are equally effective for all students). In our example, each individual would need to have the exact same pretest score as the posttest score.

The argument that $\beta 1$ in both approaches is an estimate of the effect of instruction and should thus be identical brings us to the conclusion that perhaps we need to define what is meant by *the correct effect* more closely. In the following, we therefore discuss the instruction effect in light of the true effect that exists in a hypothetical true model that generated the data. Note that we used the simple example as an illustrative tool to make Lord's paradox easily accessible. As mentioned previously, the context of instructional research is more complex than a comparison between two teachers. The subsequent considerations and examples refer to various classes with various teachers that differ on a teaching variable that is potentially continuous.

Underlying Data-Generating Model

In accordance with Köhler et al. (2020a), we use graphs to represent the state component, the trait component, their effects on the variables of interest (outcome at T1, outcome at T2, and instruction), and the instruction effect, denoted by δ (see Fig. 2). Such graphs are a frequently used tool to describe (causal) relationships between variables of a hypothetical data-generating model (cf. Allison 1990; Imai and Kim 2019; Kim and Steiner 2019; Pearl 2009; Steyer et al. 2015). The main difference between these graphs and graphs of structural equation models is that the former are assumed to contain all observed and unobserved variables that affect the outcome at T1 and T2 and the instruction/treatment variables. Using these graphs allows to display assumptions—based on apparent plausibility or empirical findings—about the underlying data-generating model, and help identify for which data-generating scenario the defined true effect (δ) will be mirrored by the estimated regression coefficients of the ANCOVA and

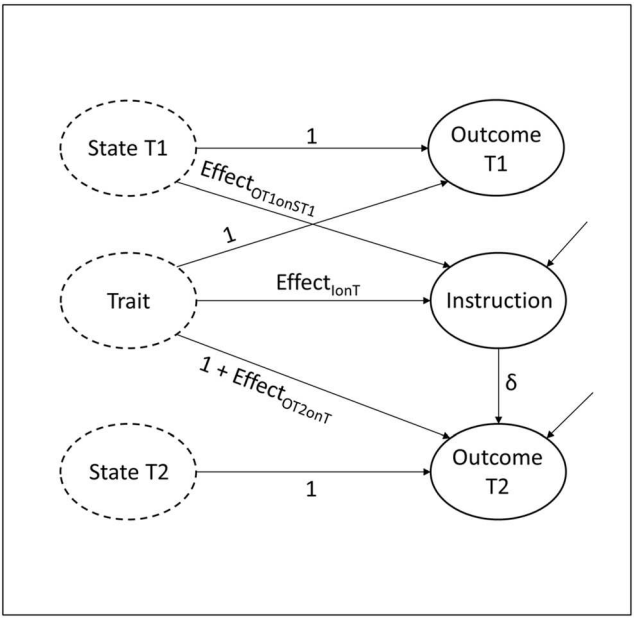

Fig. 2 Schematic representation of the underlying data-generating model

the CHANGE approach. We define bias as the difference between the true effect in the data generation and the estimated effects (i.e., the regression coefficients).

In our notation, solid circles represent the assessed outcome (T1 and T2) and instruction variables. In instructional research, where variables at the class/school level are often of interest, they might represent aggregates of information obtained at the individual level (e.g., the outcome of interest is the average reading literacy within classes, which is assessed via several manifest items at the student level). The same logic and inferences from the hypothetical data-generating model apply when solely the between level is of interest, since in the SEM framework, the total variance-covariance matrix is separated into a covariance matrix at the within level and a covariance matrix at the between level. The dashed circles represent unobserved trait- and state-specific components. The trait component includes variance that can potentially be explained by other variables that are stable across time and have the same effect on the outcome at all measurement occasions, for example, stable student characteristics such as general cognitive abilities (i.e., general intelligence), goal orientation, or class composition variables such as socioeconomic background. The stable differences might also be due to instruction prior to the first measurement occasion. The state-specific component represents time point–specific variation in the trait, which result in differences that are not stable across time. For example, if some of the tested classes have an exam around the same time the outcome at T2 is being assessed, and the exam covers a similar topic as the test, these classes might have an advantage. To give another example, if all schools were assessed simultaneously at the beginning of the school year (i.e., outcome at T1), but the summer break ended earlier in some schools, the time span between summer break and the assessment would differ between schools, which can have an effect on the ability level at T1. Clearly, researchers can aim to measure the time-invariant and time-variant variables that potentially explain some of the variance, include them in the regression model, and thus try to partial them

out. However, some degree of uncertainty will remain about whether all relevant variables were measured and no unobserved time-invariant and time-variant variables that explain preexisting relationships between the outcome at T1 and instruction exist. If they are unobserved and/or not included in the model, they become part of the estimated instruction effect, and might lead to a biased conclusion. Using the assumed data-generating model, we demonstrate in which scenarios this is the case.

The arrows in Fig. 2 indicate an existing influence of the component on the variable; the absence of an arrow represents the lack of an influence. Since the components might affect the variables to different extents, we included effects on the arrows. An effect of 1 means that the influence of a component on a variable is a fixed quantity. For example, part of the (mean) differences in the outcome at T1 are due to state-specific variables at T1 (e.g., motivation to perform well on the test), namely the differences due to the state component at T1 multiplied by 1. If the outcome at T1 and the instruction variable were assessed simultaneously, variance due to state-specific variables at T1 might also be present in the instruction variable to a certain degree (e.g., the teacher behaves more student-oriented or students simply perceive him/her as more student-oriented for the same reason they feel more motivated to perform well on the test, such as an above average result in an exam), which is represented by the effect of instruction on the state component T1. An effect of zero equals the absence of an arrow, that is, the assumption that the component does not affect the variable.

A central idea behind using these hypothetical models is that researchers can typically make assumptions on the presence of third variables, the direction of the relationship regarding these variables, and possibly also about the relative sizes of trait- and state-specific influences. These assumptions can stem from empirical findings of studies that involved the same or related constructs, or from purely theoretical considerations. Oftentimes, the design of the study and the knowledge regarding how the data were assessed informs about how likely the relationship between certain components is. For example, if the instruction variable were measured at T2, it is highly unlikely that state-specific variables from T1 have an effect on the instruction variable.

In our graphs, the treatment or rather the instruction effect is denoted by δ. It basically represents how much variance of the instruction variable is also present in the outcome at T2. What this means is that δ does not represent the pure causal effect of instruction itself, because the measured instruction variable might contain variance that is due to third variables. The effect can actually be a purely indirect effect. For example, a trait-specific component such as the average IQ level in a class will definitely affect the outcome at T1 if the variable of interest is a competence such as reading literacy. It might, however, also affect the instruction variable. Consider a variable such as *cognitive activation*: Teachers are likely to adapt the amount of cognitive activation to the average IQ level in a class. Hence, a positive effect of outcome at T2 on instruction is due to the average IQ level in a class, moderated by the teacher. The example nicely demonstrates what we mean when we say that the researcher needs to define exactly the *instruction effect*. Is it purely defined as the (direct causal) effect cognitive activation has if the classes were exactly identical and one teacher acted more cognitively activating than the other, or should the effect include adaptation of the teacher to the class level? If the teacher did not adapt to the class level, the effect of the outcome at T2 on instruction might be less. We argue that the effect should include indirect effects as well, since the teacher's adaptation to the class is part of his competence to improve instruction, which should be reflected in the instruction effect.

In the following sections, we use practical examples from educational and instructional research to discuss (a) the assumptions for the underlying data-generating model, (b) the likelihood that these assumptions hold, and (c) which of the approaches produces the accurate instruction effect. We cover four different scenarios: (1) A random experiment, (2) a study in which the instruction variable is affected by trait-specific variables, (3) a study in which the instruction variable is affected by state-specific variables, and (4) a study in which the instruction variable is affected by both trait- and state-specific variables. Note that for each of the examples, in order to make a statement about whether the approach would report a biased regression coefficient, we need to assume that the observed variables were measured without measurement error. In one of the subsequent sections, we discuss ways to attain error-free measurement especially in the context of instructional research, where variables at the classroom level are mostly of interest.

Examples from Practice

Random Experiment

In an experiment with random assignment to groups, prior influences are ideally equaled out, leading to similar group compositions. Certainly, this might not hold true for each individual study since randomization might, by chance, lead to inequalities at T1, which should be checked by comparing distributions of relevant variables between the groups. Given a successful randomization, neither stable nor time-specific components are associated with instruction. Note that the example for Lord's paradox as described above would probably not result from a random experiment, since teacher 1 taught classes that were, on average, less skilled at T1 than the classes taught by teacher 2. If the assignment of classes to teachers had been random, the average outcome at T1 would have been very similar, with only small random differences remaining. In a random experiment, instruction is independent of the outcome at T1 and it is also independent of the change-score, and the instruction effect can be estimated without bias under both approaches.

The four scenarios in Fig. 3 depict possible underlying data-generating models. To embed them in a substantial example, consider the outcome variable at T1 and T2 to be the solving of Raven's Matrices, with items of similar difficulty at both time points. Since this is a random experiment, a group of students is randomly selected to partake in an intervention, such as a specific training in how to solve these matrices. We now consider the different components. Students will have stable differences in their ability of solving these matrices (i.e., the trait part), which remain constant across time. The differences between students prior to the training might also be affected by time-point specific occurrences (i.e., the state at T1 part), such as teachers motivating specific students to perform well on the test. If this specific motivation is unrelated to selection into treatment, which will be the case in a random experiment, this has no influence on the estimated instruction effect.

In the first picture of Fig. 3, the change in performance from T1 to T2 is due to the training (δ), partly due to growth that cannot be explained by the treatment but is independent of the performance prior to the training (i.e., the error term, which captures unexplained variance in growth, for example, individual differences in students' mental development between T1 and T2), and time-specific influences at T2 (i.e., the state at T2 part, such as receiving back a math test and the obtained grade influences self-confidence in solving the matrices).

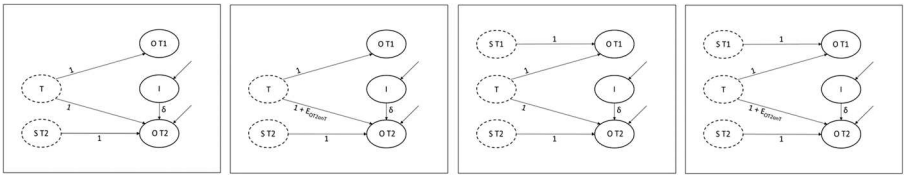

Fig. 3 Possible underlying data-generating models in random experiment

In the second picture, change is additionally allowed to depend on stable differences between students' abilities to solve the matrices (i.e., the effect of the trait part on the outcome at T2). This underlying data-generating model is more realistic, considering that changes such as mental development are likely to depend on stable student characteristics such as age or the degree of mental stimulation they experience in their day-to-day surroundings. These factors, however, have no influence on the estimated instruction effect—even when they are not considered in the estimation model—because they are not confounded with the instruction variable.

In the third picture, time point–specific occurrences (i.e., the state at T1 part) that influence the outcome at T1 but not the instruction variable come into play. The difference to picture four only lies in the trait part: In the third picture, the stable variables such as intelligence equally influence how well students solve the matrices at T1 and at T2, whereas in the fourth picture, the change from T1 to T2 is allowed to depend on stable variables. Again, these dependencies are not problematic for either of the approaches even if the variables are not measured and included in the model, because the variables are unrelated to instruction.

Nonrandom Study with Trait Influencing Instruction

Nonrandom studies are far more common in educational and instructional research than random experiments. In many cases, the variable of interest is not some form of intervention or treatment, but a behavior or characteristic of the teacher. Teachers are typically not randomly assigned to classes. To give a practical example, we discuss the study by Kunter et al. (2013) who aimed to investigate how teacher variables (beliefs, behavior, motivation, competence) influence student achievement and student motivation. One specific research question was whether mathematical achievement was influenced by pedagogical content knowledge (PCK) of the teacher. The achievement variable was assessed in grade nine (T1) and again in grade 10 (T2); PCK was assessed based on a test, which was given to the teachers at T2.

Consider a possible data-generating model for this practical example. Figure 4 depicts different data-generating models of a nonrandom study where the trait component (i.e., stable individual differences in mathematical skill) influences instruction. The third and fourth pictures are the most likely scenarios for our example, which is why we discuss these in more detail.

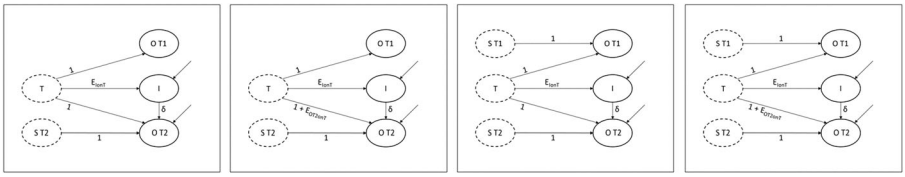

Fig. 4 Possible underlying data-generating models in nonrandom experiment with trait (T) influencing instruction (I)

In both pictures, the assumptions regarding the underlying data-generating model are that mathematical achievement at T1 and T2 is influenced by stable and time-varying variables, and that stable mathematical skill levels of classes influence the PCK of their teachers. This last assumption might, in fact, not necessarily be the case if teachers from this particular study were randomly assigned to classes, completely independent of their PCK. In the German school system, however, which is where the study was conducted, different school tracks exist, which typically lead to a confounding between teachers' skill levels and average class skill levels. In the practical example, it is therefore likely that teachers with higher PCK levels also teach classes that, on average, have higher mathematical skills. Therefore, the arrow between instruction and the trait is needed. In the fourth picture, an additional assumption we include is that the initial mathematical skill level influences the gain in mathematical achievement from T1 to T2. This assumption that the average change in achievement is related to baseline achievement is rather likely (see, e.g., Ahmed et al. 2014). Lastly, note that in none of the scenarios, an arrow between state T1 and instruction exists. The assumption that the assessment of the teacher variable PCK is independent of state-specific variance in the outcome variable at T1 is reasonable, since PCK was assessed at the teacher level at T2, and mathematical achievement was assessed at the student level at T1. We hence expect no influence of time-specific variables at T1 on instruction.

If we consider the data-generating model in the third picture as the model that is responsible for the variance that was observed in the variables of interest in the study, it is possible to deduce which approach reports an unbiased instruction effect (without including any other possible explanatory variables in the analysis model). In the third picture, where we do not assume that the average change in achievement is related to stable skill differences, the CHANGE approach gives the correct estimate of the instruction effect, because the effect of the stable skill differences is constant for the outcome at T1 and T2 and thus cancels out. The ANCOVA approach cannot deal with the confounding appropriately because the mixture of stable- and time-specific effects in the outcome at T1 leads to an imprecise amount of variance that is partialled out of the outcome at T2, leading to an overestimation of the instruction effect (in the case that in the true data-generating model, both the effect of the trait on instruction and δ are positive). If the data-generating model of picture four is the true model, neither of the approaches reports the correct instruction effect.

Nonrandom Study with State Influencing Instruction

To discuss the bias of the instruction effect for the two approaches when time-variant variables are part of the data-generating model, we consider a study conducted by Holzberger et al. (2013), who investigated reciprocal effects between teachers' self-efficacy and instructional quality in a setting with two measurement occasions. Teacher self-efficacy was assessed at the teacher level; instructional quality was assessed at both the teacher and the student level. All variables were measured at both time points, namely at the end of grade nine and the end of grade 10, with no teacher change across the grades.

Consider a likely data-generating model for the research question of whether teacher efficacy has an effect on the student-perceived learning support. It is likely that time-variant variables had an effect on the outcome at T1 and the instruction variable (see Fig. 5). If, for example, there were exams at the end of the school year in grade nine that were below average in some classes, the teachers in the respective classes might give lower ratings on the self-efficacy scales than usually, and students in the respective classes might feel less supported at

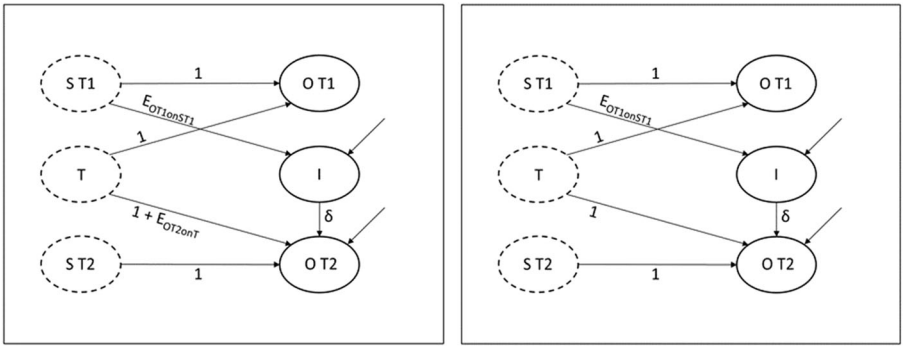

Fig. 5 Possible underlying data-generating models in nonrandom experiment with state at fist measurement occasion (S T1) influencing instruction (I)

that time. The difference between the two pictures in Fig. 5 is simply that in the second picture, stable student characteristics or class composition variables influence the gain in mathematics enjoyment from T1 to T2. Note that neither of the pictures includes an arrow from the trait part to *teacher self-efficacy*, which entails the assumption that differences between teacher self-efficacy ratings are not due to time-invariant variables.

Unfortunately, neither of the approaches reports the correct effect of teacher self-efficacy under the data-generating models in Fig. 5. In the ANCOVA approach, the inclusion of learning support at T1 as a predictor leads to too much variance that is being partialled out (i.e., variance that is due to an interaction of the state component at T1 and δ), therefore underestimating the effect of learning support at T2 on teacher efficacy. In the CHANGE approach, the change-score entails variance due to the state component at T1 and δ, which cannot be separated. When regressing the change-score on learning support, which also entails variance due to the state component at T1, the estimated instruction effect is biased. In the first picture, the size of the bias depends on the strength of the influence of time-varying variables on the teacher self-efficacy; in the second picture, it additionally depends on the effect of stable variables on the gain in perceived learning support from T1 to T2.

Nonrandom Study with Trait and State Influencing Instruction

Lazarides and Buchholz (2019) investigated how student-perceived teaching qualities relate to achievement emotions such as enjoyment. They used data from the Programme for International Student Assessment (PISA) 2009 and the German national extension of PISA in 2010. The outcome variables *student achievement emotions* were assessed in both grades, whereas the student-perceived teaching quality was only assessed in the ninth grade (i.e., T1). Only students who did not experience a teacher change between grades were included.

Again, consider a likely data-generating model for one of the investigated research questions in this study, for example, *mathematics enjoyment* as the outcome variable and *teacher support* as the instruction variable. Since the outcome at T1 and the instruction variable were both measured at the student level at T1, it is likely that time-variant confounders had an effect on them (i.e., the arrow from the state part to instruction in Fig. 6). In addition, time-invariant confounders that influence the average student mathematic enjoyment in a class plausibly influence the average student-perceived teacher support in a class (i.e., the arrow from the trait part to instruction). The difference between the two pictures in Fig. 6 is that in the second

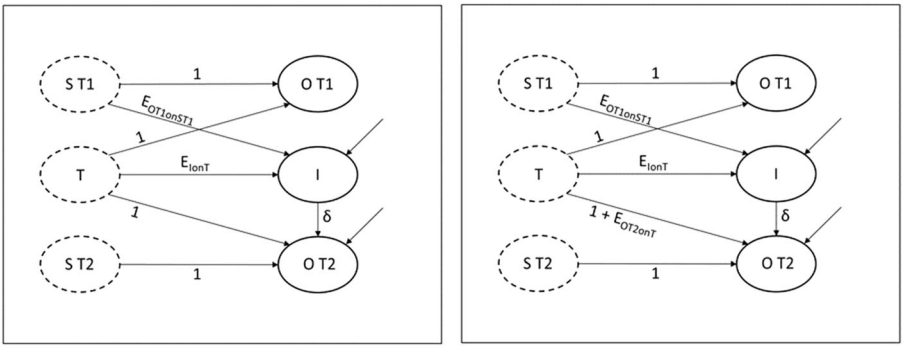

Fig. 6 Possible underlying data-generating models in nonrandom experiment with state at fist measurement occasion (S T1) and trait (T) influencing instruction (I)

picture, stable student characteristics or class composition variables influence the gain in perceived teacher support from T1 to T2.

When either of the possible underlying data-generating models in Fig. 6 hold, neither of the approaches report an unbiased instruction effect: Neither of the approaches can deal with the confounding due to unobserved stable and time-specific influences.

Additional Recommendations for Practice

The most straightforward conclusion from the previous examples is to conduct random experiments. These are especially feasible when investigating alternative teaching methods such as formative assessment. However, not every instruction variable is suitable to be randomly assigned to classes. Concepts such as a cognitively activating form of instruction, classroom management, or student orientation, teachers generally show to different degrees, which makes random assignment impossible. Clearly, such concepts can be specifically trained in an intervention study, which would lead to an increase of the independent variance of instruction (see Köhler et al. 2020a), but preexisting relations between the outcome at T1 and instruction would remain an issue. Another aspect that should be considered in intervention studies is that the effect one would obtain from a comparison between teachers who participated in the intervention and those who did not is confounded with whether or not the intervention is actually functional. Therefore, the intervention tools should be validated prior to investigating whether instruction has an effect on average motivational and cognitive outcomes in classes. If the teachers are randomly selected into the intervention and no differences exist between groups, both approaches will report the same regression coefficient. This effect can be interpreted as a gain due to a more cognitively activating form of instruction (or whichever focus the intervention had) in the classes of teachers who took part in the intervention.

In nonrandom studies, the answer for which approach to use is less clear-cut. Similar to our considerations in the previous examples, the researcher should consider, in each case, what the data-generating scenario probably looked like. The arguments for assumptions about the data-generating model can be based on apparent plausibility—for example, due to the design of the study—and on empirical findings—for example, when previous studies showed that pedagogical content knowledge is unrelated to the average mathematics enjoyment in classes, it is

unlikely that these variables share common trait variance. Once the most likely underlying model is established, in some of the scenarios to deduce which of the approaches is more appropriate (see, e.g., Fig. 4, pictures 2 and 3). In cases where both approaches report a biased effect, the data-generating model allows approximating the size and direction of the bias, thus also giving clues on which approach to use and how cautiously the results should be interpreted. For example, if the direction or the size of the empirically observed effect differs largely from the reported biased effect one would find with the approaches assuming the true effect in the data-generating model was zero, it is unlikely that the data underlying true effect was, in fact, zero. In the following section, we use our data example from the introduction to illustrate the procedure of using the data-generating models for drawing conclusions about the observed effect in more detail.

Data Example Revisited

In our data example from the DESI study, the effect of teacher supportiveness on listening comprehension skills was small and non-significant in the ANCOVA approach, and larger and significant in the CHANGE approach. The most plausible underlying data-generating model is the last picture of Fig. 4. State-specific occurrences at T1 such as class-specific activities are likely to effect the average listening comprehensions skills at T1 (i.e., the arrow from the state part at T1 to outcome T1); those state-specific occurrences cannot have affected teacher supportiveness (i.e., no arrow from the state part at T1 to instruction), since teacher supportiveness was measured at T2; time-invariant variables such as class composition are likely to affect the average listening comprehensions skills in classes at T1 and at T2, the average student-perceived teacher support in a class (i.e., the arrow from the trait part to instruction), and the improvement of listening comprehension skills from T1 to T2 (i.e., the effect of outcome at T2 on the trait part). In this data-generating model scenario, both approaches report a biased regression coefficient.

Using the syntax provided by Köhler et al. (2020a), we inserted likely values for all parameters of the data-generating model in the last picture of Fig. 4. We set the variance due to state variables to 0.2 and the variance due to trait variables to 0.8 (resulting in a total variance of 1 for the variable *listening comprehension at T1*), since a student's listening comprehension skills in a foreign language is a stable quality that is rather unaffected by a student's current (emotional or affectional) state. We assume the effect of trait variables on instruction to be small and negative (− .2), since classes with lower average general skills require more support from the teacher than classes that already show high levels of performance (see, e.g., Kuger et al. 2017). Lastly, we assume a small positive effect of trait variables on the gain in listening comprehension from T1 to T2 (.3), since classes with more competent students tend to show stronger increases compared to classes with less competent students (Dumont et al. 2013). Given these values for the data-generating model, we calculated the reported regression coefficients from the two approaches, β_{CA} and β_{CS}.

Given the data-generating model holds, both approaches would show a negative bias and report a negative instruction effect if the true effect δ was zero. Since we found positive regression coefficients in the data, it is unlikely that the true effect was zero (or negative). Positive effects would only be reported for $\delta > 0$. This means that, in all likelihood, teacher support does have an effect on student listening comprehension. Unfortunately, the exact size

and significance of this effect is impossible to determine, since the true parameter values are unknown.

General Discussion

This article aims to assist educational and instructional researchers in making decisions regarding their analysis model. We focus on multilevel data with two measurement occasions and one predictor, which is a data structure often used in educational and instructional research. We apply two common approaches, namely covariance-analytical (ANCOVA) approaches and latent change-score (CHANGE) models, and illustrate under which circumstances which model is more appropriate. Results show that in random experiments, both approaches produce an unbiased estimate of the true effect; in nonrandom experiments where the trait influences instruction, scenarios exist where either the ANCOVA or the CHANGE model perform better; in nonrandom experiments, where the state influences instruction, and also in nonrandom experiments where both influence instruction—which, in fact, represent the more realistic instructional research scenarios—neither approach gives an unbiased estimate of the instruction effect. This means that, under specific assumptions, either both, one, or none of the approaches show unbiased effects.

Beyond the purely technical aspect of unbiased estimates, it is important for researchers to understand the underlying logic of the two approaches and draw sensible conclusions. The ANCOVA approach balances the baseline measures of different groups and informs about whether the expected means at T2 differ, which might not necessarily answer the question of interest. In the example from the beginning regarding the two twin teachers with exactly the same teaching practices, the relevant question for the parents was "Will my child profit more from being in the class of teacher 1 compared to being in the class of teacher 2, which the answer to was "no". In this case, the reported beta coefficient does not reflect the effect of the teacher, but whether the expected group means differ at T2, which they do. In the example, the reported effect is due to stable differences between the groups and state-specific variance, not due to teaching quality. When calculating difference scores, the important aspects to consider are the following: (a) The pre- and post-measures need to be on the same scale, (b) the size of a change-score might have a different meaning at different skill levels, and (c) the change-score might be differently reliable at different skill levels (Xiao et al. 2019). In the hypothetical data-generating model, we assumed equal reliability and equal meaning for equal sizes of change scores.

In order to avoid artificial results, researchers need to debate—for each examined variable constellation separately—what the data-generating scenario probably looked like, and which assumptions are likely to hold. The arguments for assumptions about the data-generating model can be based on logical conclusions—for example, resulting from the design of the study—and on empirical findings. In cases where both approaches report a biased effect, the data-generating model allows approximating the size and direction of the bias, thus also giving clues on which approach to use and how cautiously the results should be interpreted. Although the size and direction of the bias can be approximated, the true size of the effect (i.e., the unbiased regression coefficient) and a test for its statistical significance remain inestimable due to too many unknown parameters. A more favorable tactic than finding post-hoc solutions to scenarios in which both approaches result in bias of the estimated effect is to a priori debate whether a study is adequate for answering a particular research question. If this is not the case,

the study design can still be adjusted. Köhler et al. (2020a) discuss various actions a researcher can take to minimize bias in the estimated instruction effect: conduct quasi-experimental intervention studies, increase the time span between the outcome measure at T1 and the instruction variable, avoid student reports to measure the instruction variable, only investigate teachers that are new to a class, and minimize time-specific variance between classes. We advocate a thorough preregistration of the assumed data-generating model, which will not only increase a deep reflection of the theory and help decide on an adequate modeling approach, it also has the potential of enhancing transparency and reproducibility.

Another limitation of the presented approach of using a hypothetical data-generating model is that the made assumptions regarding this model might not hold. Alternative data-generating scenarios cannot be ruled out, and researchers should consider as many as possible and compare them with their empirical findings. Furthermore, not all possible data-generating scenarios were considered in our model. For example, an interaction effect between instruction and the outcome at T2 is plausible: Classes that are more skilled might profit more from the treatment. Such additional effects would add to the complexity of the data-generating model and make it much less feasible. We argue that our model captures the main components influencing the estimated instruction effect, and other effects might be present but comparatively small. A more eminent aspect to consider, especially with respect to educational research, is how the choice of the model influences cross-level effects or effects of cross-level interactions, which also require data-generating models that are more complex than the ones presented here. In addition, model-misspecification or factors that threaten the reliability of model parameter estimates such as an insufficient number of classes or low intraclass correlations might affect the performance of the approaches to accurately retrieve the instruction effect (Lüdtke and Robitzsch 2020). Whether one of the approaches is more susceptible to any of these threats needs further investigation.

Another prospective approach to conceptualize variance compositions of empirically collected data is to think more in terms of probabilities of effects instead of certainties of effects (Wasserstein et al. 2019). Instead of using fixed values, we could use Bayesian inference methods and define priors for the parameters of the hypothetical data-generating model, which would result in a multivariate distribution of the components and hence a prior bias distribution for the ANCOVA approach and one for the CHANGE approach. Bayesian model averaging could then be used to select the better model (see, e.g., Hoeting et al. 1999). More research is needed to decide on criteria for model selection.

Lastly, we would like to point out that our considerations are based on ideal measurements of the variables. In practice, however, measurements are prone to error due to (a) the sampling of items that serve as indicators for the latent variables, and often additionally by (b) the sampling of students within classrooms when group-level constructs are formed from measurements at the individual level, as it is regularly the case in educational research (Marsh et al. 2009; Skrondal and Rabe-Hesketh 2004). While the former source of error can be adequately addressed by the application of latent variable models like structural equation models (SEM; e.g., Bollen 1989) or item response theory (e.g., Embretson and Reise 2000) using a sufficient number of indicators representative for the latent constructs (Blömeke et al. 2015), ways of treating the latter source of error depends on whether the group-level constructs are "shared" or "configural" variables (Stapleton et al. 2016). Both types of constructs are based on aggregation of individual-level variables. In shared variables, the classroom-level construct is a latent variable based on multiple individual-level indicators (e.g., student ratings on teaching). That is, each student rates the same classroom-level construct. Accordingly, the idiosyncratic

proportion within the student ratings can be seen as sampling error, which decreases the higher the agreement among students within the same classroom is. In contrast, configural variables refer to constructs based on individual-level measures that we expect to differ for students within a classroom (e.g., gender ratio). The precision of such variables depends on the proportion of students assessed per classroom. Therefore, configural variables can be considered free of sampling error if all students within a classroom are measured, and susceptible to sampling error if not (Lüdtke et al. 2011; Marsh et al. 2009). A sound way of dealing with both measurement and sampling error in shared and configural variables are latent measurement and latent aggregation using multilevel SEM (Lüdtke et al. 2008). Using maximum likelihood estimation, such a latent covariate approach provides unbiased estimates of group-level effects if the group-level sample size exceeds 50 units and the intraclass correlation (ICC) is higher than .10 (Lüdtke et al. 2011). Otherwise, manifest aggregation should be preferred. Alternatively, Bayesian estimation of the latent covariate approach using slightly informative priors may provide stable group-level estimates in scenarios where group-level sample size and/or ICC is small (Zitzmann et al. 2015).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-020-09590-6.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Ahmed, Y., Wagner, R. K., & Lopez, D. (2014). Developmental relations between reading and writing at the word, sentence and text levels: a latent change score analysis. *Journal of Educational Psychology, 106*(2), 419–434. https://doi.org/10.1037/a0035692 .

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20*, 93–114. https://doi.org/10.2307/271083 .

Beck, B., & Klieme, E. (Eds.) (2007). Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International) (Linguistic competencies: concepts and measurements. DESI-Study (German English student performances international). Beltz Verlag.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Approaches to competence measurement in higher education. *Zeitschrift für Psychologie, 223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000193 .

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Dumont, H., Neumann, M., Maaz, K., & Trautwein, U. (2013). Die Zusammensetzung der Schülerschaft als Einflussfaktor für Schulleistungen (the composition of the student body as a factor that influences school performance). *Psychologie in Erziehung und Unterricht, 60*(3), 163–183.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Erlbaum.

Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society, 157*(3), 317–338. https://doi.org/10.2307/2983526 .

Hand, D. J., & Taylor, C. C. (1987). *Multivariate analysis of variance and repeated measures* (5th ed.). Chapman & Hall.

Hochweber, J., & Vieluf, S. (2016). Gender differences in reading achievement and enjoyment of reading: the role of perceived teaching quality. *The Journal of Educational Research, 111*(3), 268–283. https://doi.org/10.1080/00220671.2016.1253536 .

Hoeting, J., Madigan, D., Raftery, A., & Volinsky, C. (1999). Bayesian model averaging: a tutorial. *Statistical Science, 14*(4), 382–401.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 3–35). Erlbaum.

Holzberger, D., Philipp, A., & Kunter, M. (2013). How teachers' self-efficacy is related to instructional quality: a longitudinal analysis. *Journal of Educational Psychology, 105*(3), 774–786. https://doi.org/10.1037/a0032198 .

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: a potentially confusing task. *Psychological Bulletin, 82*(4), 511–518. https://doi.org/10.1037/h0076767 .

Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science, 63*(2), 467–490. https://doi.org/10.1111/ajps.12417 .

Jamieson, J. (2004). Analysis of covariance (ANCOVA) with difference scores. *International Journal of Psychophysiology, 52*(3), 277–283. https://doi.org/10.1016/j.ijpsycho.2003.12.009 .

Kim, Y., & Steiner, P. M. (2019). Gain scores revisited: a graphical models perspective. *Sociological Methods & Research, 105*, 1–23. https://doi.org/10.1177/0049124119826155 .

Köhler, C., Hartig, J., & Schmid, C. (2020a). Deciding between the covariance analytical approach and the change-score approach in two wave panel data. *Multivariate Behavioral Research*, advance online publication, 1–12. https://doi.org/10.1080/00273171.2020.1726723 .

Köhler, C., Kuger, S., Naumann, A., & Hartig, J. (2020b). Multilevel models for evaluating the effectiveness of teaching: conceptual and methodological considerations. *Zeitschrift für Pädagogik Beiheft, 1*, 197–209. https://doi.org/10.3262/ZPB2001197 .

Kuger, S., Klieme, E., Lüdtke, O., Schiepe-Tiska, A., & Reiss, K. (2017). Mathematikunterricht und Schülerleistung in der Sekundarstufe: Zur Validität von Schülerbefragungen in Schulleistungsstudien (Mathematics teaching and student performance in secondary education: on the validity of student surveys in studies on school performance). *Zeitschrift für Erziehungswissenschaft, 20*(2), 61–98.

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

Lazarides, R., & Buchholz, J. (2019). Student-perceived teaching quality: how is it related to different achievement emotions in mathematics classrooms? *Learning and Instruction, 61*, 45–59. https://doi.org/10.1016/j.learninstruc.2019.01.001 .

Liang, K. Y., & Zeger, S. L. (2000). Longitudinal data analysis of continous and discrete responses for pre-post designs. *Biostatistics, 62*(1), 134–148.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*(5), 304–305. https://doi.org/10.1037/h0025105 .

Lüdtke, O., & Robitzsch, A. (2020). Commentary regarding the section 'Modeling the effectiveness of teaching quality': methodological challenges in assessing the causal effects of teaching. *Zeitschrift für Pädagogische Psychologie, 66*(1), 210–222.

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. https://doi.org/10.1037/a0012869 .

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16*(4), 444–467. https://doi.org/10.1037/a0024376 .

Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods, 3*(3), 309–327. https://doi.org/10.1037/1082-989X.3.3.309 .

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*(6), 764–802. https://doi.org/10.1080/00273170903333665 .

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

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiements and analysing data: a model comparison perspective* (2nd ed.). Erlbaum.

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: an illustration. *The Journal of Experimental Education, 82*(2), 143–167. https://doi.org/10.1080/00220973.2013.769412 .

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Nold, G., & Rossa, H. (2007). Hörverstehen Englisch (Listening comprehension in English). In B. Beck & E. Klieme (Eds.), Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International) (pp. 120–129). Beltz Verlag.

Oakes, J. M., & Feldman, H. A. (2001). Statistical power for nonequivalent pretest-posttest designs. The impact of change-score versus ANCOVA models. *Evaluation Review, 25*(1), 3–28. https://doi.org/10.1177/0193841X0102500101 .

Otis, N., Grouzet, F. M. E., & Pelletier, L. G. (2005). Latent motivational change in an academic setting: a 3-year longitudinal study. *Journal of Educational Psychology, 97*(2), 170–183. https://doi.org/10.1037/0022-0663.97.2.170 .

Pearl, J. (2009). *Causality: models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M.,. . . Schiefele, U. (Eds.) (2006). PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres (PISA 2003: Studies on competence development in the course of a school year). Waxmann.

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine, 25*(24), 4334–4344. https://doi.org/10.1002/sim.2682 .

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC.

Spinath, B., & Steinmayr, R. (2012). The roles of competence beliefs and goal orientations for change in intrinsic motivation. *Journal of Educational Psychology, 104*(4), 1135–1148. https://doi.org/10.1037/a0028115 .

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. https://doi.org/10.3102/1076998616646200 .

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits-revised. *Annual Review of Clinical Psychology, 11*(1), 71–98. https://doi.org/10.1146/annurev-clinpsy-032813-153719 .

Stipek, D., & Valentino, R. A. (2015). Early childhood memory and attention as predictors of academic growth trajectories. *Journal of Educational Psychology, 107*(3), 771–788.

Van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: the difference. *Multivariate Behavioral Research, 48*(6), 895–922. https://doi.org/10.1080/00273171.2013.831743 .

Vollet, J. W., Kindermann, T. A., & Skinner, E. A. (2017). In peer matters, teachers matter: peer group influences on students' engagement depend on teacher involvement. *Journal of Educational Psychology, 109*(5), 635–652. https://doi.org/10.1037/edu0000172 .

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, 108*(5), 705–721. https://doi.org/10.1037/edu0000075 .

Wainer, H., & Brown, L. M. (2004). Two statistical paradoxes in the interpretation of group differences. *The American Statistician, 58*(2), 117–123. https://doi.org/10.1198/0003130043268 .

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician, 73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913 .

Wright, D. B. (2006). Comparing groups in a before-after design: when t test and ANCOVA produce different results. *The British Journal of Educational Psychology, 76*(3), 663–675. https://doi.org/10.1348/000709905X52210.

Xiao, Z., Higgins, S., & Kasim, A. (2019). An empirical unraveling of Lord's paradox. *The Journal of Experimental Education, 87*(1), 17–32. https://doi.org/10.1080/00220973.2017.1380591 .

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*(6), 688–705. https://doi.org/10.1080/00273171.2015.1090899 .