Educational Psychology Review (2022) 34:511–536
https://doi.org/10.1007/s10648-021-09635-4

REVIEW ARTICLE

How Many Classes and Students Should Ideally be Sampled When Assessing the Role of Classroom Climate via Student Ratings on a Limited Budget? An Optimal Design Perspective

Steffen Zitzmann[1,2] · Wolfgang Wagner[1] · Martin Hecht[1] · Christoph Helm[3] · Christian Fischer[1] · Lisa Bardach[1] · Richard Göllner[1]

Accepted: 16 July 2021 / Published online: 6 September 2021

Abstract

A central question in educational research is how classroom climate variables, such as teaching quality, goal structures, or interpersonal teacher behavior, are related to critical student outcomes, such as students' achievement and motivation. Student ratings are frequently used to measure classroom climate. When using student ratings to assess classroom climate, researchers first ask students to rate classroom climate characteristics and then aggregate the ratings on the class level. Multilevel latent variable modeling is then used to determine whether class-mean ratings of classroom climate are predictive of student outcomes and to correct for unreliability so that the relations can be estimated without bias. In this article, we adopt an optimal design perspective on this specific strategy. Specifically, after briefly recapping a prominent model in climate research, we show and explain (a) how statistical power can be maximized by choosing optimal numbers of classes and students per class given a fixed budget for conducting a study and (b) how the budget required to achieve a prespecified level of power can be minimized. Moreover, we present an example from research on teaching quality to illustrate the procedures and to provide guidance to researchers who are interested in studying the role of classroom climate. Also, we present a Shiny App that can be used to help find optimal designs for classroom climate studies. The app can be accessed at https://psychtools.shinyapps.io/optimalDesignsClassroomClimate

Keywords Classroom climate · Multilevel modeling · Latent variable · Optimal design · Bayes

✉ Steffen Zitzmann
steffen.zitzmann@uni-tuebingen.de

Extended author information available on the last page of the article.

Organizational research often seeks to identify organization characteristics that positively affect the people in the organization. In a similar vein, educational research looks for specific learning environment characteristics that positively predict and might foster relevant outcomes, such as students' achievement, motivation, and emotions (e.g., Emmer & Stough, 2001; Pianta et al., 2008; Reyes et al., 2012; Wang & Degol, 2016). Typical characteristics of the learning environment include variables that can be subsumed under the umbrella term *classroom climate* (Marsh et al., 2012). Over the last several decades, different theoretical frameworks for studying classroom climate constructs have evolved, such as theories involving teaching quality (e.g., the Three Basic Dimensions of teaching quality; e.g., Praetorius et al., 2018; Göllner et al., in press; Fauth et al., 2014), goal structures from achievement goal theory (e.g., Bardach et al., 2020; Kaplan et al., 2002; Rolland, 2012), interpersonal teacher behavior in terms of agency and communion (e.g., Mainhard et al., 2011; Rimm-Kaufman et al., 2015; Patrick et al., 2007), and teacher autonomy support versus control from self-determination theory (e.g., Vansteenkiste et al., 2012). Researchers' basic motivation for studying these variables is the assumption that the context in which students learn can play a significant role in students' development (see Seidel & Shavelson, 2007). Recently, Wang et al. (2020) systematically reviewed and synthesized 61 studies from classroom climate research to achieve integration along with a more robust understanding of the relations between classroom climate variables and achievement, motivation, emotions, and other relevant outcomes. Overall, the authors found small to medium-sized effects across the different outcomes. Still, there was considerable variation in the relations, thus justifying further analyses. In addition to the overall confirmation that classroom climate affects relevant outcomes, the authors found that the sizes of the relations varied with the methodology employed in the different studies, such as how the study was designed and which approach was applied to assess classroom climate.

To investigate how classroom climate is related to the outcomes, researchers typically rely on sampling designs that yield data with a multilevel structure in which students are nested within classrooms. Multilevel modeling (e.g., Raudenbush and Bryk, 2002; Snijders & Bosker, 2012) is then used to regress the outcome variables on classroom climate constructs (and on reasonable covariates). Multilevel modeling is the method of choice because it allows researchers to distinguish between different levels of the analysis and to investigate relations between variables across these different levels. For example, features of the classroom climate (i.e., class-level variables) can be related to students' achievement and positive self-beliefs (i.e., student level variables). As Morin et al. (2014) showed in their study, classroom climate (in terms of classroom mastery goal structure, challenge, and teacher caring) positively predicted students' achievement and academic self-efficacy. More recently, progress has been made in integrating multilevel modeling with latent variable modeling (e.g., Bollen, 1989) as the standard, and numerous influential articles have promoted the use of multilevel latent variable modeling in classroom climate research. For example, Marsh et al. (2012) described how sophisticated multilevel latent variable models can be used to assess the role of classroom climate (see also Bardach et al. 2020; Morin et al. 2014; Wagner et al. 2013).

One efficient and widely used way to assess classroom climate is to ask students to rate a characteristic of the learning environment (e.g., Downer et al., 2015; Fauth et al., 2014; Patrick et al., 2007; Stornes and Bru, 2011). Thereby, the referent of such classroom climate ratings is usually the classroom or the teacher (e.g., "In this class, we should… "; "Our teacher tells us… ") rather than some characteristic of the individual student (Marsh et al., 2012). Individual student ratings of the classroom climate are then aggregated on the class level (i.e., averaged across the students in a classroom). The class-mean rating reflects the shared perception of the students in a class with regard to the classroom climate characteristic (corrected for individual idiosyncrasies). This method of assessing a classroom climate construct is often employed in research on teaching quality as one example (Lüdtke et al., 2006). For instance, Kunter et al. (2013) used it in their widely cited study on the role of teachers' teaching quality and teachers' professional competence in students' math achievement and enjoyment (see also Lazarides and Buchholz, 2019). However, this strategy is not without problems. As Kunter et al. (2013) reported, reliabilities of class-mean ratings turned out to be around .80 and were thus not very high when the intraclass correlation that accounts for the number of students per class (i.e., the ICC(2); Bliese, 2000) was used as the reliability coefficient (see also Baumert et al., 2010).[1] Although values at this level are usually interpreted as indicating acceptable reliability (LeBreton & Senter, 2008), such values can nevertheless bias the results for the role of the classroom climate. This kind of situation occurs when a predictor's lack of reliability is ignored in the analysis, subsequently biasing the relation between the predictor and the outcome variable (e.g., Buonaccorsi, 2010; Fuller, 1987). This issue led (Lüdtke et al., 2009) to argue that reliability should always be checked before computing the relations between class-mean ratings and outcome variables. As a remedy, Lüdtke et al. (2008) suggested an approach in which a latent variable (i.e., the latent class mean) is used in place of the not very reliable manifest class-mean rating (latent aggregation). When this latent aggregation approach—which the authors names the *multilevel latent covariate model*—is taken, relations between classroom climate and outcome variables are estimated without bias (see also Asparouhov and Muthén, 2007; Croon & van Veldhoven, 2007; Shin & Raudenbush, 2010). Before we go on, we want to emphasize once more that this latent variable approach was specifically developed for individuals' environment ratings that are latently aggregated on the group level to form the climate construct. This kind of construct is also referred to as a *reflective* group-level construct in order to distinguish it from formative group-level constructs (e.g., the percentage of girls in a class), which are analyzed in contextual studies and do not require this specific latent variable approach (Marsh et al., 2012). Also, classroom climate as assessed by student ratings should be distinguished from other group-level constructs, such as those that rely on teachers' self-ratings, for which a different analytical approach should generally be taken.

The adoption of the latent variable approach in classroom climate research is complicated by the fact that this approach is relatively expensive (i.e., it requires rather

[1] The ICC(2) does not assess reliability that is due to classical measurement error.

large samples), and a classroom climate study should be sufficiently powered so that the study is likely to detect an existing relation between the classroom climate and the outcome variable and thus to allow for a substantial conclusion. In such studies, the statistical power to detect such a relation is a function of two sample sizes—the number of classes and the number of students per class—and two intraclass correlations. However, these sample sizes are usually limited by the study's budget, which is why it is preferable to choose numbers in such a way that the highest possible level of power can be achieved given the budget or, alternatively, that the budget can be minimized given a certain prespecified level of power. These goals play a significant role in study planning and in writing grant proposals. The aim of optimal design research is to determine how these goals can be achieved. There is extant literature on power and optimal designs in multilevel research, and some software has been developed to help researchers design their studies. For example, there is the OD (Optimal Design) program, which is a stand-alone software developed by Raudenbush and his team and which provides useful visualizations (e.g., it plots power against sample sizes), with many more options. Other examples are ML-DEs (Cools et al., 2008) and MLPowSim, both of which provide R scripts that create macros for running simulation studies with the special purpose software MLwiN. Yet another example is *PowerUp!* (Dong & Maynard, 2013), which is capable of tackling even three-level models with (cross-level) moderations and mediations. However, the literature primarily focuses on experiments (e.g., cluster randomized trials) or correlational studies (e.g., contextual studies) in which the predictors are observed variables (e.g., Donner & Klar, 2000; Maas & Hox, 2005; Raudenbush, 1997; Rhoads, 2011; Snijders, 2005; van Breukelen & Candel, 2015). Also, except for PowerUp!, all the programs deal with relatively prototypical multilevel models. To the best of our knowledge, optimal designs for studies with latent class means as predictors have not yet been discussed. Therefore, with the present article, our goals are to fill this gap by developing optimal designs for these studies and to provide guidance to researchers who are interested in studying the role of classroom climate.

The article is organized as follows. After briefly recapping (Lüdtke et al.'s 2008) multilevel latent covariate model, we show and explain (a) how, given a fixed budget, power can be maximized by choosing optimal numbers of classes and students per class and (b) how, given a prespecified level of power, a study's budget can be minimized. Next, we provide an example from research on teaching quality to illustrate the procedures. We then present a newly developed Shiny App and show how it can be used to help find optimal designs in classroom climate research. Finally, we discuss possible directions for future research, and we provide an outlook on how the maximum level of power can be further increased and the minimum required budget can be further decreased by employing a Bayesian estimator—an approach that lends itself well to situations in which a study's budget is very limited.

Multilevel Analysis of Classroom Climate

In classroom climate research, the general procedure involves fitting a multilevel model in which, at the class level, an outcome variable is regressed on the class-mean rating of classroom climate. Importantly, the class mean is of primary interest in these studies rather than the individual students' ratings because classroom climate is first and foremost a class-level construct (Marsh et al., 2012), and the class mean captures this construct best. The class mean reflects the students' shared perception of the classroom (or the teacher), whereas the students' class-mean-centered ratings reflect individual deviations from this shared perception (Lüdtke et al., 2009). Because these deviations reflect perceptions that are not shared across the students in a class, and the class mean reflects what is shared, the relations of these two different measures with an outcome variable can differ (Snijders & Bosker, 2012). Often, the outcome variable is also regressed on the students' (class-mean-centered) ratings at the student level. However, as Cronbach (1976) already noted, studying these deviations might be interesting, but this question is relatively unrelated to questions about the class-level construct and its role in relevant outcome variables (Cronbach, 1976; but see Göllner et al., 2018, for the argument that individual idiosyncrasies can also be a valuable source of information).

A special variant of this general procedure of relating classroom climate variables to outcomes is the *multilevel latent covariate model*, which was suggested by Lüdtke et al. (2008) and published in Psychological Methods, a journal with a much broader readership that includes researchers from organizational research for which climate variables are also of interest. This model is considered the method of choice because it allows researchers to obtain unbiased estimates of the relations between classroom climate and outcome variables even when the ICC(2) values are not very high. Although other methods exist (e.g., Grilli and Rampichini, 2011; Croon & van Veldhoven, 2007; Zitzmann, 2018; see also Zitzmann & Helm, 2021), we focus on this model because it is well-known (the model has been cited more than 600 times; Google Scholar, April 2021) and has also often been applied by researchers. To explain the model and its use in classroom climate research in a manner that is easy to understand, we use an example from research on teaching quality. This research has established the notion that student achievement is positively related to classroom management (e.g., Kounin, 1970; Matheny & Edwards, 1974; Lewis, 2001; Helmke et al., 1986; Praetorius et al., 2018; Kunter et al., 2007). Classroom management consists of two core components: identifying and strengthening desirable student behaviors and preventing undesirable ones (Hochweber et al., 2014). Teachers scoring high on classroom management communicate clear rules and effectively prevent disruptions while they are teaching in order to maximize the amount of time that the students are engaged in learning, thereby promoting students' learning progress. To assess classroom management, students are asked, for example, whether they agree that their teacher does not have to wait a long time for students to quiet down (see, e.g., Wagner et al., 2016; Göllner et al., 2018).

Instead of averaging these ratings across the students in the same class to compute the class-mean rating, the ratings can be subjected to latent variable software that uses the latent class mean instead of the class-mean rating (latent aggregation). One such type of software is M*plus* (Muthén & Muthén, 2012), which performs the latent aggregation of the students' ratings by default when the multilevel module is used unless this option is overwritten (see also the R package lavaan; Rosseel, 2012). More specifically, M*plus* decomposes the students' individual ratings of classroom management into two components: the latent class mean, which is referred to as the *between* part in M*plus* because this component varies only between classes (or teachers), and the individual deviation from the latent class mean, which is the *within* part because it varies only within classes (Asparouhov & Muthén, 2007). More formally, the decomposition can be written as:

$$CMgmt_{ij} = CMgmt_{between,j} + CMgmt_{within,ij} \tag{1}$$

for students $i = 1, \ldots, n$ in classes $j = 1, \ldots J$. *CMgmt* is an abbreviation for classroom management. When applying (Raudenbush & Bryk's 2002) notation for multilevel models, the following student-level regression shows the relation between student achievement with the within part of a student's rating:

$$\text{Student Level: } Ach_{ij} = \beta_{0j} + \beta_{within} \cdot CMgmt_{within,ij} + \varepsilon_{ij} \tag{2}$$

where β_{within} is the within slope, which assesses the relation between achievement (Ach) and the within part, and ε_{ij} are residuals, which exist because, as is true for all regression type models, it is not reasonable to assume that the outcome variable can be perfectly predicted by a single predictor. Because the within part of a student's rating reflects only that student's idiosyncratic perception (i.e., what is not shared with other students), the within slope is typically not of much interest in classroom climate research. What is more important is how achievement is associated with the between part, which reflects the students' shared perception and thus captures the climate construct classroom management best. To assess this relation, the intercept β_{0j}, which varies between classes (i.e., the between part of achievement), is regressed on the between part at the class level:

$$\text{Class Level: } \beta_{0j} = \alpha + \beta_{between} \cdot CMgmt_{betwenn,j} + \delta_j \tag{3}$$

where α is the overall intercept, and $\beta_{between}$ is the between slope, which assesses the relation between achievement and classroom management. δ_j are residuals. Combining the class-level regression with the student-level regression yields the multilevel latent covariate model:

$$Ach_{ij} = \alpha + \beta_{within} \cdot CMgmt_{within,ij} + \beta_{between} \cdot CMgmt_{betwenn,j} + \delta_j + \varepsilon_{ij}\ . \tag{4}$$

Because classroom climate research is primarily interested in whether and to what extent the classroom climate construct is related to achievement, the most interesting parameter in the model is the between slope. Hence, the optimal design should address how this parameter can be obtained in an optimal and cost-efficient way.[2] By this, we mean (a) how the power to detect a relation between classroom management and achievement can be maximized by choosing sample sizes that are optimal given a budget for conducting the study and (b) how the study's budget can be minimized given a prespecified level of power. Next, we present procedures that address these questions.

[2] This solution might not be optimal for other parameters, for example, the within slope. See the "Discussion" section for a cautionary note on this limitation.

Optimal Designs

The concept of power is closely related to the variability of the results from estimating the between slope in repeated samples. These results can be very different, particularly when the numbers of classes and the number of students per class are not very large. The less variable the results are, the narrower the confidence intervals (CIs) for these results are because the width of the confidence interval is a direct function of the standard error, which is an estimate of the scatter of the results. The narrower the CIs, the less often these CIs will include zero, and the more often the relation between the classroom climate and the outcome variable will be detected. Thus, to achieve a high level of power, the variability needs to be small.

As demonstrated by Grilli and Rampichini (2011) and, more recently, Zitzmann et al. (2021), see also Zitzmann et al. (2021), the variability of the results for the between slope in Lüdtke et al.'s (2008) model can be computed by using the first-order Taylor expansion. If we assume that the students' ratings and their achievement are standardized variables (standardized at the student level), an approximation is given by:

$$\mathrm{Var} \approx \frac{1}{J-1} \cdot \left\{ \left[\frac{\mathrm{ICC}(1)_{Ach}}{\mathrm{ICC}(1)_{CMgmt}} + \frac{1-\mathrm{ICC}(1)_{CMgmt}}{n \cdot \mathrm{ICC}(1)_{CMgmt}} \cdot \left(\frac{\mathrm{ICC}(1)_{Ach}}{\mathrm{ICC}(1)_{CMgmt}} + \frac{1-\mathrm{ICC}(1)_{Ach}}{1-\mathrm{ICC}(1)_{CMgmt}} \right) \right] + \left[-1 - \frac{2 \cdot \left(1-\mathrm{ICC}(1)_{CMgmt}\right)}{n \cdot \mathrm{ICC}(1)_{CMgmt}} \cdot \frac{\beta_{within}}{\beta_{between}} \right] \cdot \beta_{between}^{2} \right\} . \tag{5}$$

This equation is insightful with regard to the various quantities on which the variability depends. First and foremost, it shows that the variability is a function of the sample size, which means that the variability will decrease when the number of classes (J) gets larger and all other quantities are held constant. This is an example of the well-known effect of sample size on the variability of the results from repeated samples: The larger the sample size, the less variable and thus the more similar the results. The same will be true when the number of students per class (n) gets larger. Second, the intraclass correlation of student achievement ($\mathrm{ICC}(1)_{Ach}$) influences the variability. This ICC(1) assesses the amount of variance between students that can be attributed to differences between the classes (e.g., Snijders & Bosker, 2012) The variability of the results will increase when the ICC(1) gets higher. Third, the variability also depends on the ICC(1) of the students' ratings of classroom management ($\mathrm{ICC}(1)_{CMgmt}$) in such a way that when this ICC(1) gets higher, the variability will decrease. It is interesting to note that the two ICC(1) values have different effects on the variability: Whereas the ICC(1) of the students' ratings *de*creases the variability, the ICC(1) of the achievement variable *in*creases it. Fourth, another quantity that influences the variability is the between slope ($\beta_{between}$). The larger this slope, the smaller the variability of the results for this slope will be. Fifth, the within slope (β_{within}) also influences this variability, with larger values for this slope leading to smaller values for the variability of the results for the between slope.

It would be preferable if the between slope was assessed with the smallest possible variability and thus with maximum power. This requires the variability to be minimized, which is the topic of the next section.

Maximizing Power Given a Fixed Budget

As Eq. 5 shows, the variability in the results for the between slope critically depends on the number of classes and the number of students per class. To decrease this variability—and thus to increase the power to detect the relation between classroom management and achievement—the sample sizes must be increased. However, sampling additional classes and students per class imposes further costs. Because the study's budget is typically limited, it is thus preferable to choose numbers in such a way that the variability will be as small as possible. In order to obtain these optimal sample sizes, a constrained optimization problem must be solved. The goal is to find the two sample sizes that minimize the variability subject to a constraint.

One way in which a researcher can profit from optimal design research is when the study's budget is fixed and the researcher's aim is to maximize power. Thus, the variability is the objective function of the optimization problem in this case. As the constraint under which the variability is minimized, we consider the following simple cost function, which determines the relations between the fixed budget to be spent on data collection and the numbers of classes and students per class:

$$\text{budget} = J \cdot \text{costs per class} + N \cdot \text{costs per student} \tag{6}$$

where $N = n \cdot J$ is the overall number of students. Note that the budget in this cost function includes only costs that depend on one or the other of the two sample sizes. That is, the budget does not include the salary of the principal investigator, for example. A great deal of the costs per class are composed of the wages of the people who administer the questionnaires and tests in schools (personnel costs) and the costs of getting these people there (travel costs). Examples of the costs per student are the printing costs for the student's test booklet, the costs for scanning, and the costs for coding the student's responses. Once the objective function and the constraint are defined, the optimization problem can be solved.

An efficient way to find the numbers of classes and students per class needed to minimize the variability subject to the cost function is, first, to rearrange the cost function in such a way that the number of students per class (n) is expressed as a function of the number of classes (J): $n = \frac{\text{budget}-J\cdot\text{costs per class}}{J\cdot\text{costs per student}}$. Then, this expression is substituted for n in the formula for the variability. As a consequence, the objective function is no longer a function of the number of students per class, and thus, the optimal number of classes can be found via *uni*dimensional optimization using R's (R Development Core Team, 2016) general-purpose optimizer optim( ), which uses the Nelder-Mead algorithm (Nelder & Mead, 1965). The optimal number of students

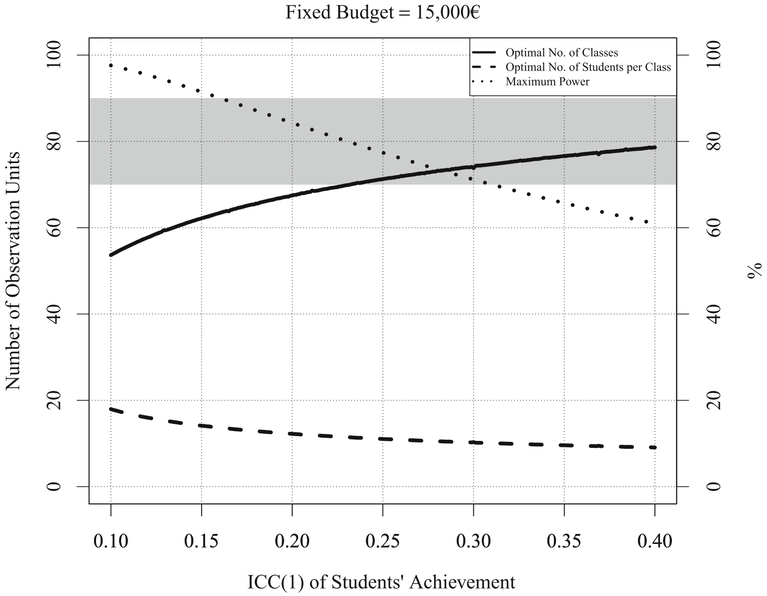


Fig. 1 Optimal numbers of classes and students per class and the maximum power to detect the relation between classroom management and achievement as a function of the ICC(1) of student achievement for a fixed budget of 15,000€, costs per class of 100€, and costs per student of 10€. The ICC(1) of the students' ratings of classroom management is set to .2, and the standardized between and within slopes are both .2

per class is obtained by inserting the optimal number of classes into the expression for n.[3]

To apply this procedure in study planning, researchers must first think about values for the quantities that influence the variability (and thus the power). More specifically, they must consider how large the ICC(1)s of the two variables will be. Fortunately, ICC(1)s are well-studied in educational research, and thus, the ICC(1)s can be set equal to the values that previous studies have reported for the variables. Recently, Stallasch et al. (2021) provided an overview of ICC(1)s for a broad array of student outcome variables, and Baumert et al. (2010), Kunter et al. (2013), Lazarides and Buchholz (2019), and Marsh et al. (2012), and many more scholars reported ICC(1)s of students' ratings of classroom climate variables.

To sensitize researchers to the consequence that their choice of the ICC(1) of student achievement will have, Fig. 1 illustrates the effect of this ICC(1) on the optimal sample sizes and the maximum power for a given budget of 15,000€ (see the figure's caption for detailed information about the values assumed for the other quantities

[3] To assess the relation of interest with maximum power, we seek to determine the lowest possible variability of the between slope. Technically speaking, this is an example of A-optimality, which is generally achieved by minimizing the trace of the parameters' covariance matrix. A-optimality is to be distinguished from D-optimality, which minimizes the determinant of this matrix. In optimal design research, D-optimality is often preferred over A-optimality (van Breukelen, 2013). However, when the covariance matrix consists of only one entry as in our case (i.e., only one variance), A- and D-optimality will yield identical solutions. Thus, our solution is D-optimal as well.

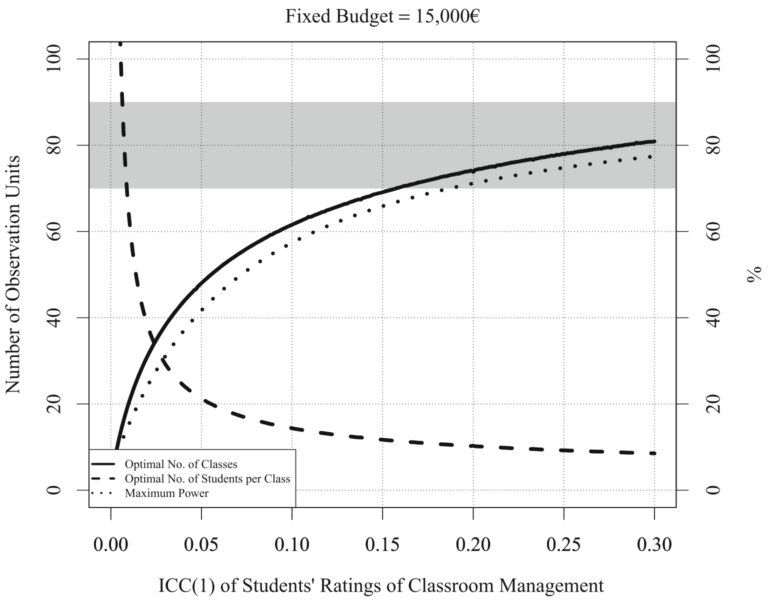


Fig. 2 Optimal numbers of classes and students per class and the maximum power to detect the relation between classroom management and achievement as a function of the ICC(1) of students' ratings of classroom management for a fixed budget of 15,000€, costs per class of 100€, and costs per student of 10€. The ICC(1) of student achievement is set to .2, and the standardized between and within slopes are both .2

and the costs in this illustration).[4] The solid and dashed lines are the optimal numbers of classes and students per class, respectively, and the dotted line represents the maximum power to detect the between slope. As can be seen, the higher the ICC(1) is (when all other quantities are held constant), the larger the optimal number of classes will be, and the smaller the optimal number of students per class will be. This means that given the abovementioned budget and the costs, a researcher would be well-advised to allocate the budget primarily to the classes when the ICC(1) of the outcome variable and thus the amount of unexplained variance in this variable at the class level are expected to be rather large. Furthermore, a higher ICC(1) will reduce the power to detect the between slope (because the amount of variance explained at the class level will decrease).

The impact of the ICC(1) of the students' ratings of classroom management is shown in Fig. 2. Again, when a higher ICC(1) is selected, the optimal number of classes will increase and the optimal number of students per class will decrease. However, unlike the ICC(1) of student achievement, a higher ICC(1) of students' ratings of classroom management will increase the power to detect the between slope.

Besides the ICC(1)s, researchers must consider the sizes of the two slopes. Compared with the ICC(1)s, less research has been conducted on the slopes of the

[4] These results were generated from a formula for the variance of the between slope that provided an approximation that is more precise than the one presented in Eq. 5 because this formula also included terms involving higher order factors, such as $\frac{1}{n^2(n-1)}$ or $\frac{1}{n^2}$.

classroom climate variables so far. See Bardach et al. (2020), Kunter et al. (2013), and Lazarides and Buchholz (2019), and Lüdtke et al. (2006) for examples of such research. Moreover, different procedures have been applied to standardize the slopes, thus further complicating the adoption of values from previous research. For example, in M*plus*, the between slope is standardized with respect to the between variances of the variables. In contrast to M*plus*, Marsh et al. (2009) suggested that the between slope be standardized with respect to the total variance of the outcome variable but only the between variance of the classroom climate variable—a suggestion that we adopted here.

Figure 3 shows how the size of the standardized between slope (standardized according to Marsh et al.'s (2009) formula) affects the optimal sample sizes and the maximum power to detect the between slope. The figure shows that the larger this slope is chosen to be, the smaller the optimal number of classes will be, and the larger the optimal number of students per class will be. Moreover, a larger standardized between slope will increase the power to detect the between slope.

Figure 4 shows the influence of the size of the standardized within slope. Unlike the standardized between slope, the larger this slope is, the larger the optimal number of classes will be, and the smaller the optimal number of students per class will be. It is interesting to note that the size of the standardized within slope tends to have

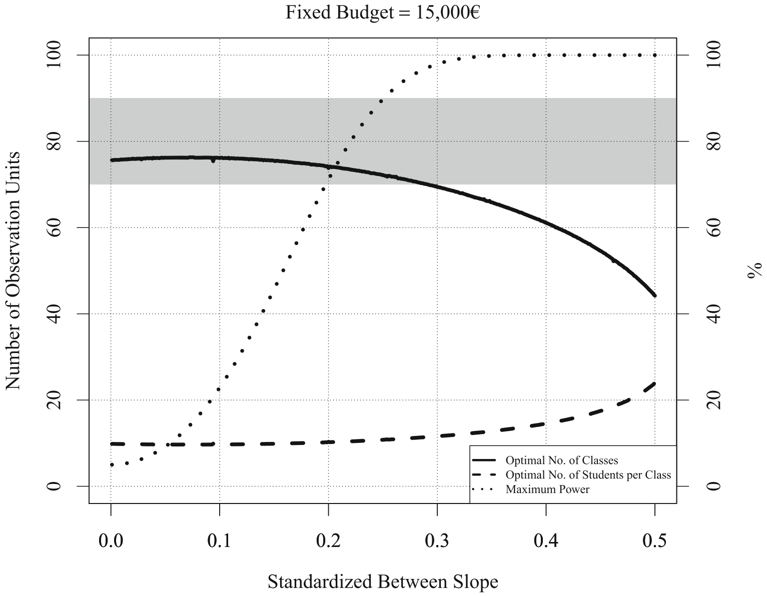


Fig. 3 Optimal numbers of classes and students per class and the maximum power to detect the relation between classroom management and achievement as a function of the standardized between slope for a fixed budget of 15,000€, costs per class of 100€, and costs per student of 10€. The ICC(1) of student achievement and the ICC(1) of students' ratings of classroom management are both set to .2, and the standardized within slope is .2

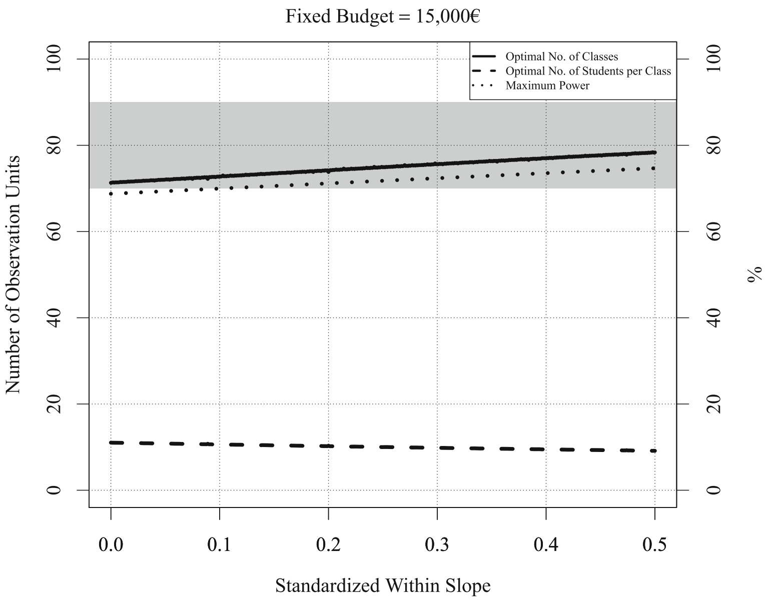


Fig. 4 Optimal numbers of classes and students per class and the maximum power to detect the relation between classroom management and achievement as a function of the standardized within slope for a fixed budget of 15,000€, costs per class of 100€, and costs per student of 10€. The ICC(1) of student achievement and the ICC(1) of students' ratings of classroom management are both set to .2, and the standardized between slope is .2

a positive effect on the power to detect the between slope, indicating the potential advantage of a larger standardized within slope.

To summarize so far, we explained how the two sample sizes can be chosen in such a way that, *given a fixed budget*, the variability of the between slope will be as small as possible, and thus, the power to detect the relation between classroom management and achievement will be as high as possible. Also, we discussed and illustrated the consequences that the choice of values for the two ICC(1)s and the standardized slopes will have on these optimal sample sizes and on the power. We adopted the perspective that the study's budget is fixed, and the researcher wants to find the design with the maximum level of power. In the following section, we change the perspective and discuss another application scenario.

Minimizing a Study's Budget Given a Prespecified Level of Power

The maximum power might not be sufficient for a given fixed budget, which makes it necessary to reconsider the budget. Optimal design research can help researchers find the numbers of classes and students per class that minimize the budget required to achieve a certain level of power to detect the relation between classroom management and achievement. These numbers constitute the most cost-effective design with the desired level of power. Thus, in this optimization problem, Eq. 6 is the objective function, and the constraint under which the budget is minimized is the power:

$$\%\text{Pow} = 100 \cdot \left[1 - F\left(1.96 - \frac{\beta_{between}}{\sqrt{\text{Var}}}\right) + F\left(-1.96 - \frac{\beta_{between}}{\sqrt{\text{Var}}}\right)\right] \quad (7)$$

where F denotes the cumulative normal distribution function (see, e.g., Kelcey et al., 2017). Again, the optimal sample sizes can be found with the help of optim(), and we briefly discuss the consequences that the choice of the values for the different quantities will have on the optimal sample sizes and on the smallest budget that can be achieved.

The effect of the ICC(1) of student achievement is illustrated in Fig. 5. Again, the solid and dashed lines represent the optimal numbers of classes and students per class, respectively. The dotted line is the smallest budget needed to achieve a power of 80% to detect the relation between classroom management and achievement. Eighty percent is a typical choice in study planning. The figure shows that the higher the ICC(1) is (and thus, the larger the amount of unexplained variance at the class level is), the larger the optimal number of classes will be, and the smaller the optimal number of students per class will be when all other quantities are held constant. Thus, when a researcher wants to achieve a power of 80% and expects the ICC(1) of the outcome variable to be rather large, he or she should spend the budget primarily on the classes. Also, with a higher ICC(1), the minimum budget needed to achieve a power of 80% will increase.

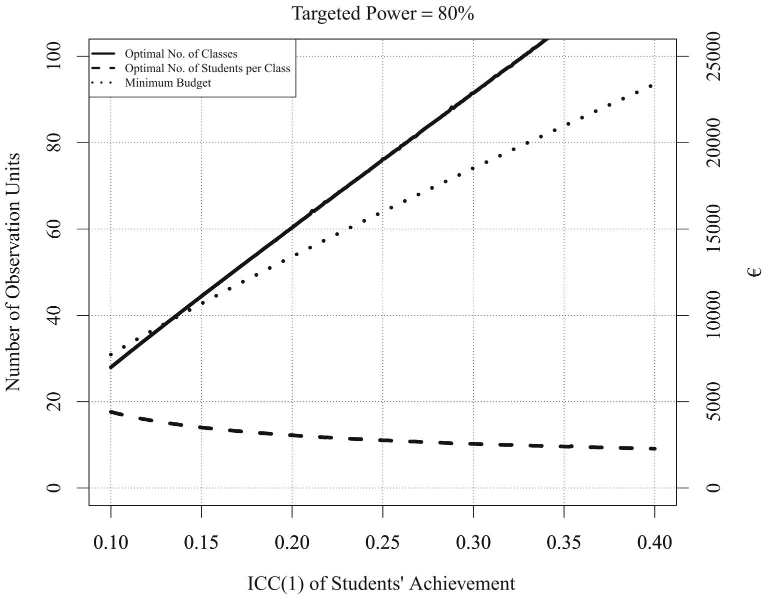


Fig. 5 Optimal numbers of classes and students per class and the minimum budget required for a power of 80% to detect the relation between classroom management and achievement as a function of the ICC(1) of student achievement for costs per class of 100€ and costs per student of 10€. The ICC(1) of students' ratings of classroom management is set to .2, and the standardized between and within slopes are both .2

Figure 6 shows the impact of the ICC(1) of students' ratings of classroom management. As can be seen, selecting a higher ICC(1) will decrease both the optimal numbers of classes and students per class and the minimum budget required to achieve a power of 80%.

A slightly different pattern emerges when a larger standardized between slope is selected. The impact of the size of this slope on the optimal sample sizes and the minimum required budget is shown in Fig. 7. The choice of a larger standardized between slope will decrease the optimal number of classes but will increase the optimal number of students per class. Moreover, the larger this slope is, the lower the minimum budget required to achieve a power of 80% will be.

How the size of the standardized within slope affects the optimal sample sizes and the minimum required budget is shown in Fig. 8. A larger slope will slightly decrease the optimal numbers of classes and students per class as well as the minimum budget required to achieve a power of 80%, indicating once more the possible advantage of a larger standardized within slope.

To summarize, we mentioned how the two sample sizes can be chosen in such a way that *given a certain level of power to detect the relation between classroom management and achievement*, the budget required to achieve this power will be as small as possible. We also illustrated how the two ICC(1)s and the standardized slopes will change these optimal sample sizes and the minimum budget. Next, we show how the different procedures for obtaining optimal designs can be applied in research on teaching quality.

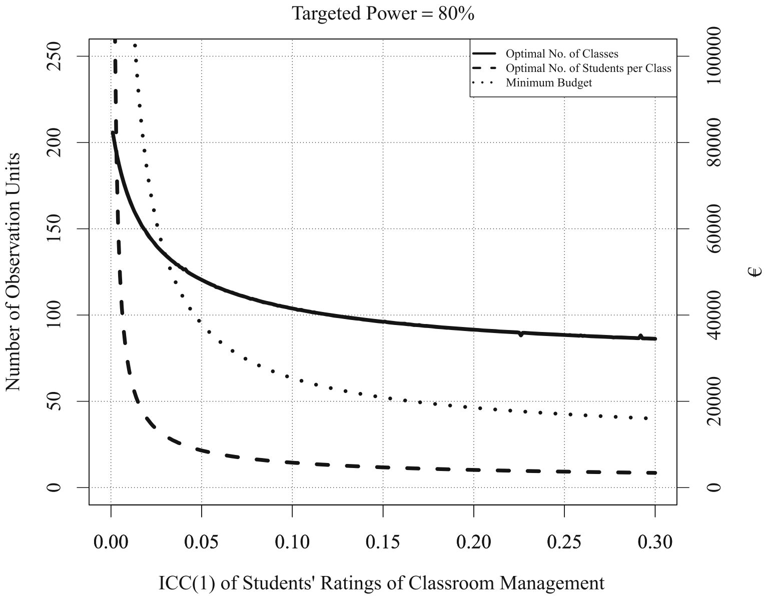


Fig. 6 Optimal numbers of classes and students per class and the minimum budget required for a power of 80% to detect the relation between classroom management and achievement as a function of the ICC(1) of students' ratings of classroom management for costs per class of 100€ and costs per student of 10€. The ICC(1) of student achievement is set to .2, and the standardized between and within slopes are both .2

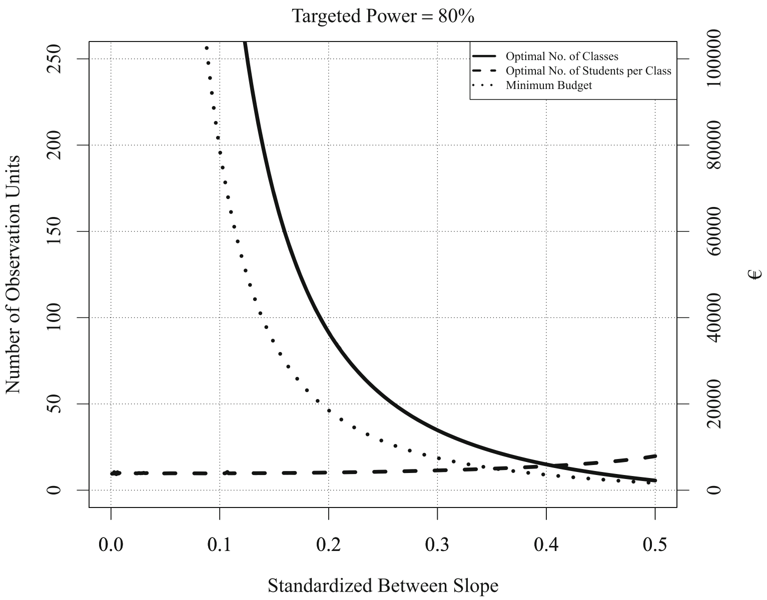


Fig. 7 Optimal numbers of classes and students per class and the minimum budget required for a power of 80% to detect the relation between classroom management and achievement as a function of the standardized between slope for costs per class of 100€ and costs per student of 10€. The ICC(1) of student achievement and the ICC(1) of students' ratings of classroom management are both set to .2, and the standardized within slope is .2

Illustrative Example

To illustrate the application, we consider the study by Arens and Morin (2016) and make use of their results. The sample used by the authors was the German sample from the Progress in International Reading Literacy Study 2006 (PIRLS), which consisted of 414 classes with 18 students per class. Arens and Morin (2016) investigated the role that teachers' support plays in students' reading achievement. The concept of support captures different ways to support students' learning, such as individualized feedback and encouragement, and these forms of support have been shown to be positively related to achievement (Hamre & Pianta, 2005; Hughes et al., 2008; Klem & Connell, 2004; Kunter et al., 2013). The extent to which a teacher supports his or her students was assessed by asking the students whether the teacher gives advice to students on how to do better, for example (see also Morin et al., 2014). Along with students' scores on a reading achievement test, these ratings were then subjected to a multilevel latent variable model in M*plus*, which performs the latent aggregation as described above and computes the relation between the teachers' support and the students' achievement (while controlling for other variables).

We first consider what the optimal design would have been for a fixed budget. Because we do not know the PIRLS' budget and costs, we assume a fixed budget of 60,000€ for the sampling of classes and students and costs of 68.20€ and 7.05€ per class and student, respectively. The costs correspond with the costs in one of our own

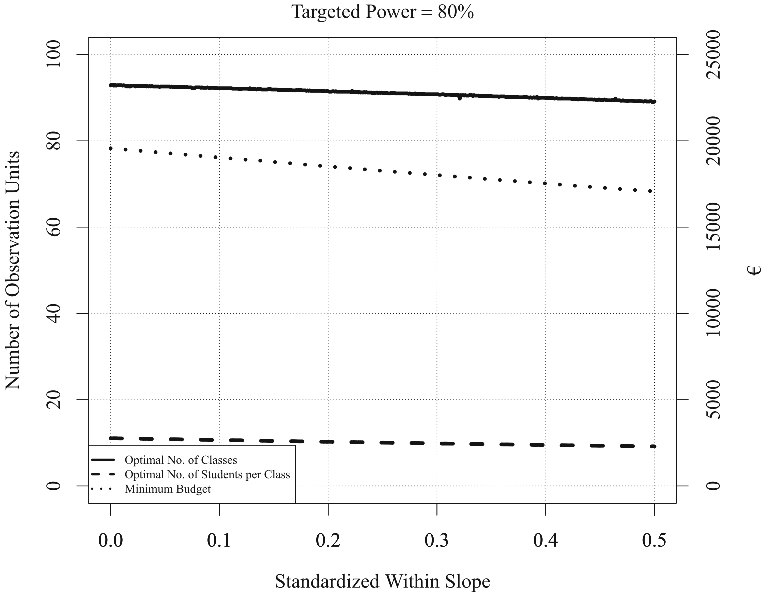


Fig. 8 Optimal numbers of classes and students per class and the minimum budget required for a power of 80% to detect the relation between classroom management and achievement as a function of the standardized within slope for costs per class of 100€ and costs per student of 10€. The ICC(1) of student achievement and the ICC(1) of students' ratings of classroom management are both set to .2, and the standardized between slope is .2

large-scale studies and involve personnel costs and travel costs as well as printing, scanning, and coding costs. Also, it is necessary to consider particular values for the quantities that influence the variability (and thus the power) of the between slope. Arens and Morin (2016) reported ICC(1)s of the achievement test and the students' ratings of support of .32 and .12, respectively. Thus, we chose these values for the ICC(1)s. Also, the authors reported the standardized within slope, which was 0.159 (see their Table 2).[5] Arens and Morin (2016) found that the standardized between slope, which describes the relation between support and achievement, was nonsignificant. Thus, for the standardized between slope to be detected in the study, we chose .1 (i.e., one-tenth of a standard deviation), which is rather small but is in line with the authors' reasoning. Applying our procedure for determining the design with the maximum power given a fixed budget to these values provided optimal numbers of 395 classes and 12 students per class. Thus, the optimal design had 19 classes less and 6 students per class less than the actual sample sizes used in the German PIRLS. However, a study with this optimal design would have a power of only 73% to detect a

[5] This is the conditional slope (i.e., it was adjusted for other predictors and covariates). However, because the study did not report the value for the unconditional slope, we used the value of the conditional slope here, although these two values could have been different.

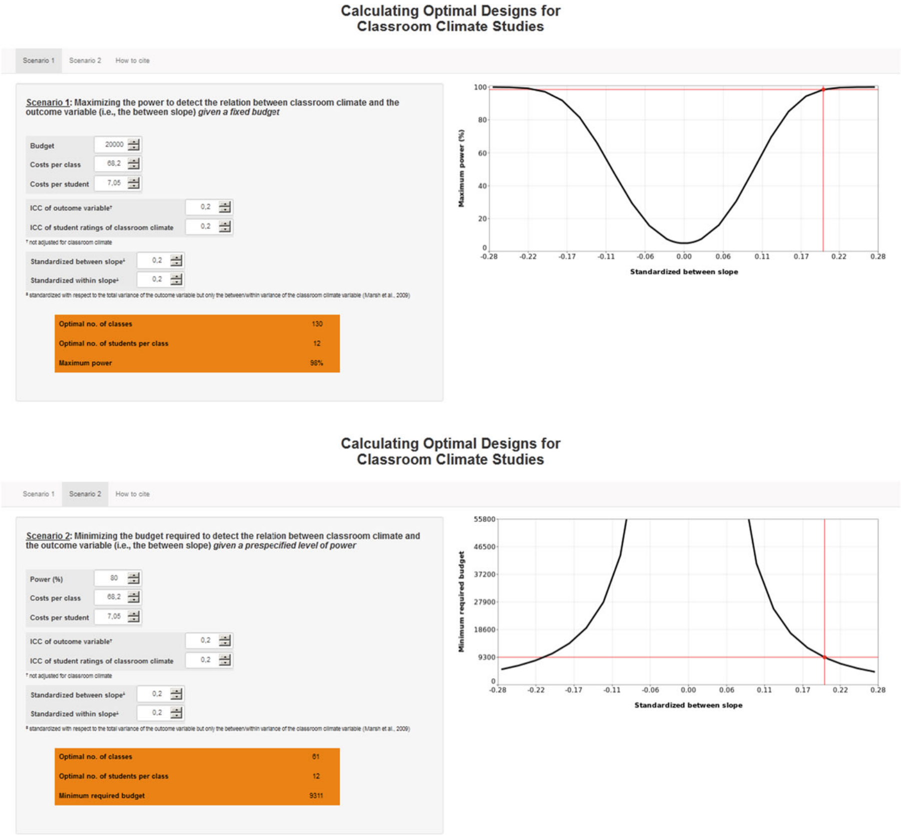
standardized between slope of .1, which would not be deemed sufficient. Therefore, next, we ask what design with a sufficiently high level of power would have been the most cost-efficient one.

We chose a power of 80%, which is considered sufficiently high. All other quantities were the same as before. Under this specification, we obtained an optimal design with 469 classes and 12 students using our procedure for finding the most cost-efficient design with a power of 80%. The required budget was 71,250€. It is interesting to note that the optimal number of students per class did not change. Only the optimal number of classes increased by 74, indicating that in order to achieve higher power, more classes must be sampled. The optimal design differed from the actual design of the PIRLS, which raises the question of whether the actual design was sufficiently powered. The answer is a clear yes because the power was 83%.

Shiny App

We created an interactive web application (Shiny App) to help researchers plan classroom climate studies and learn about optimal designs and the impact of the quantities that need to be considered when planning such studies. Thus, our app has a twofold function. First and foremost, it is a tool for finding optimal designs. Second, this app allows researchers to examine how optimal designs will change depending on the choice of the values for the two ICC(1)s and the standardized slopes. A correct understanding of these aspects is central to planning a study, and the app can help researchers develop this understanding.

Figure 9 shows screenshots of the app. The app consists of two tabs, which correspond to the two application scenarios: The first one involves finding the sample sizes that maximize the power to detect the relation of interest when the study's budget is fixed, and the second one involves finding the design that minimizes the budget required to achieve a given level of power. When the app is accessed, it defaults to some reasonable values. In Scenario 1, a fixed budget of 20,000€ is assumed. The costs per class and costs per student are set to 68.20€ and 7.05€, respectively. The ICC(1)s of students' achievement and students' ratings of classroom climate are both set to .2, and the standardized between and within slopes are also both .2. For these values, the app returns an optimal number of 130 classes, an optimal number of 12 students per class, and a power of 98%. In Scenario 2, a power of 80% is assumed instead of a fixed budget, and all other quantities are the same as those used in Scenario 1. The output of the app shows optimal numbers of classes and students per class of 61 and 12, respectively, and a required budget of 9,311€. The users can adjust the initial values by modifying the input fields. Each time a new value is entered, the output is updated immediately by R, which runs in the background of the app. For example, when the standardized between slope is increased by inputting 0.3 in Scenario 1, the output changes to 116 classes, 15 students per class, and a power of 100%. Similarly, when the slope is increased in Scenario 2, the output changes to 21 classes, 14 students per class, and a budget of 3,526€.



Calculating Optimal Designs for Classroom Climate Studies

Scenario 1: Maximizing the power to detect the relation between classroom climate and the outcome variable (i.e., the between slope) *given a fixed budget*

Budget	20000
Costs per class	68,2
Costs per student	7,05
ICC of outcome variable†	0,2
ICC of student ratings of classroom climate	0,2

† not adjusted for classroom climate

Standardized between slope‡	0,2
Standardized within slope‡	0,2

‡ standardized with respect to the total variance of the outcome variable but only the between/within variance of the classroom climate variable (Marsh et al., 2009)

Optimal no. of classes	130
Optimal no. of students per class	12
Maximum power	98%

Calculating Optimal Designs for Classroom Climate Studies

Scenario 2: Minimizing the budget required to detect the relation between classroom climate and the outcome variable (i.e., the between slope) *given a prespecified level of power*

Power (%)	80
Costs per class	68,2
Costs per student	7,05
ICC of outcome variable†	0,2
ICC of student ratings of classroom climate	0,2

† not adjusted for classroom climate

Standardized between slope‡	0,2
Standardized within slope‡	0,2

‡ standardized with respect to the total variance of the outcome variable but only the between/within variance of the classroom climate variable (Marsh et al., 2009)

Optimal no. of classes	61
Optimal no. of students per class	12
Minimum required budget	9311

Fig. 9 Shiny App

Discussion

Classroom environment is an important context not only for learning but also for other aspects of students' development, such as motivation and emotions (Wang et al., 2020). It is therefore imperative that researchers identify positive predictors of students' development, such as classroom climate. In this article, we discussed how classroom climate research can profit from optimal designs. In particular, we extended the standard of knowledge by describing (a) how the power to detect a relation between the classroom climate and an outcome variable can be maximized by choosing optimal sample sizes given a fixed budget for conducting a study and (b) how the study's budget can be minimized given a certain prespecified level of power. After we explained the procedures, we presented a Shiny App, which is useful to researchers who are interested in studying the role of classroom climate. In the following, we present and discuss further theoretical and practical considerations.

Statistical Considerations

Even though we focused primarily on the multilevel latent covariate model and its application in classroom climate research, it should be mentioned that there are further developments that are routinely applied in this field (e.g., the doubly latent model; Marsh et al., 2009). However, although these models are more complex, the multilevel latent covariate model still provides their building block. Moreover, one particular challenge associated with the application of doubly latent models is that estimation problems can occur (see Lüdtke et al., 2011), and this is why less complex models, such as the multilevel latent covariate model, can be an even better choice (see Zitzmann & Helm, 2021).

An optimal design can only be obtained with exact values for the quantities on which the variability depends. However, these exact values are usually not known for sure in the planning phase of a study, and thus, the optimal design is only "locally optimal," meaning that it is optimal only for the specific values that the researcher specified for the quantities and can be suboptimal for other values. This issue is known as the *local optimality problem*. One workaround for this problem involves choosing the maximin design, which maximizes the minimum variability (and thus minimizes the maximum power) across a range of reasonable values for the quantities (e.g., van Breukelen & Candel, 2015).

Our app is based on the evaluation of approximate formulas that were developed elsewhere (see Grilli & Rampichini, 2011; Zitzmann et al., 2021). Of course, optimal designs can also be found by running tailored computer simulations. However, these studies are computationally very demanding and time-consuming, and often, an expert from the field of statistics is needed to conduct them. By contrast, the application of our app does not require such experts, and the app is easy to use once the underlying model and the different quantities involved are understood. In educational research, it is often deemed important that an empirical study be well-designed, particularly when the implementation of the study depends on the success of a grant proposal. The app can help researchers design their studies, and it allows them to evaluate studies that were already conducted by determining whether the designs were close to optimal. Furthermore, the app can be used as a didactic tool to teach students about optimal designs in classroom climate research.

A Cautionary Note

In the multilevel latent covariate model, students' ratings of their classroom (or their teacher) are latently aggregated on the class level, and this latent class mean is related to the outcome variable. Because student ratings are used to assess the climate, and the class mean reflects the shared perception of the students in a class, the latent class mean may be referred to as a reflective class-level construct. Such constructs should not be confused with group-level constructs that are assessed by asking teachers to rate themselves. Reflective class-level constructs were discussed in depth by Lüdtke et al. (2008), who distinguished them from formative group-level constructs (e.g., the percentage of girls in a class), which are used in context research. Thus, the distinction between reflective and formative group-level constructs is closely related to Marsh et al.'s (2012) distinction between climate and contextual variables. Although the terminology differs, a similar distinction is made in organizational psychology (e.g., Bliese, 2000; Kozlowski and Klein, 2000).

Perhaps the most important difference between reflective and formative group-level constructs is that in reflective group-level constructs, the individuals act as indicators of a latent variable (i.e., the latent group mean; Croon & van Veldhoven, 2007), and thus, like the indicators in domain sampling theory (e.g., Ghiselli et al., 1981), the individuals can be considered a sample from a potentially infinitely large number of individuals per group (e.g., the number of potential raters of a learning environment), whereas in formative group-level constructs, the individuals stem from a finite number of individuals per group (e.g., the size of the class; Lüdtke et al., 2008). Hence, in reflective group-level constructs, the sampling ratio approaches zero by definition, whereas in formative constructs, this ratio can range up to 100% (i.e., when all individuals are sampled). The multilevel latent covariate model is most appropriate for reflective class-level constructs as discussed in the present article (Lüdtke et al., 2008), and the analytical approach will generally differ when formative class-level constructs are analyzed. This means that our app, which is based on this specific multilevel model, is a convenient tool for finding optimal designs when the predictor in the model is a reflective class-level construct. However, even though we generally recommend that users avoid using the app when the predictor is a formative construct, the app can still be a valid tool in this case when a certain additional condition is met. A central assumption of the app is that the sampling ratio approaches zero. Thus, the app can be a valid tool even when a formative class-level construct is analyzed, provided that only a small number of students are sampled per class. This argument is supported by extensive simulation work showing that when the sampling ratio is very small, the multilevel latent covariate model can perform reasonably well (Lüdtke et al., 2008). Because the app is based on this model, this finding speaks for the usefulness of the app also when a formative class-level construct is analyzed and the sampling ratio is very small.

In this article, we focused on the between slope of the multilevel latent covariate model, which is most important to researchers interested in studying the role of classroom climate because it assesses the relation between the classroom climate and the outcome variable. Our procedures find designs that are optimal for estimating the between slope. However, these procedures might not provide optimal solutions for other parameters, for example, when the focus is on the within slope, which assesses the role of students' idiosyncratic perceptions and might reflect dyadic effects between students and teachers to some extent (Göllner et al., 2018).

As with most optimal design research, some caution should be exercised when applying the results in small sample contexts because some of the assumptions that are made might not be realistic in these contexts. For example, to compute the power (Eq. 7), we implicitly assumed that we were estimating the between slope without bias. However, whereas this is true in large samples ($N > 1,000$), results tend to be biased when sample sizes are small (e.g., McNeish, 2017). As a consequence, the power formula and thus the optimal design found under a prespecified level of power can deviate from what is really optimal.

A practical challenge is the decision of how to proceed when the optimal number of students per class exceeds the typical class size of 25 students. A simple rule of thumb for coping with this issue is to set this number to 25 and then increase the number of classes to compensate for the loss of power. This natural boundary points to the need to take into account not only the optimality of a design when planning the study but also the suboptimality of possible boundary conditions.

Future Research and Conclusions

We did not consider additional covariates and their influence on optimal designs in classroom climate research—a limitation that is worth addressing in future research. For example, it can make sense theoretically to also include a lagged outcome to control for students' previous achievement (e.g., Köhler et al., 2021). Moreover, we did not consider data with a three-level structure (e.g., when students are nested in classes, and classes are nested in schools). To our knowledge, a closed form expression for the variability of the results for the between slope at the school level (i.e., the relation between the school climate and the outcome) has not yet been derived. However, this expression is a necessary prerequisite for our procedures for obtaining optimal designs, which are very time-efficient (i.e., results are delivered in less than 1 s). Thus, it would be a valuable task to obtain the expression so that the procedures can be extended to yield optimal designs also for the three-level extension of the multilevel latent covariate model. Finally, we focused on traditional Maximum Likelihood (ML) estimation, but as Hamaker and Klugkist (2011) pointed out, Bayesian estimation can offer a promising alternative. Zitzmann et al. (2015) even showed how this type of estimation can be fruitfully applied to Lüdtke et al.'s (2008) model to assess the role of the classroom climate, which was also the model of interest here. The main feature of Bayesian estimation is the so-called prior distribution, a vehicle that can be used to adjust results in an advantageous way. What is most important for the present work is that the prior can be specified in such a way that the variability of the results for the between slope will be small. Zitzmann et al. (2021) showed how a weak prior for the slope can achieve this goal. More formally, when this prior is specified, the variability will be w^2Var—a factor times the variability in Eq. 5. Because this factor is less than 1, the variability will be reduced in comparison with the variability in Eq. 5. As a consequence, the maximum level of power will be increased given a fixed budget, and the minimum budget required to achieve a prespecified level of power will be reduced. Thus, Bayesian estimation in combination with a weak prior appears to be an attractive option, particularly when a study's budget is very limited. It would thus be very interesting to further investigate the benefits of Bayesian estimation in optimal designs in future research.

To conclude, it is widely assumed that the learning context plays a significant role in students' development, and a number of studies have already been conducted to address the role that classroom climate plays in students' achievement, motivation, and emotions. However, as Wang et al. (2020) pointed out, although a strong theory exists for the notion that classroom climate influences relevant outcomes, empirical evidence for this causal claim is still thin, thus calling for more studies. The present article addressed the important question of how the role of classroom climate can be studied optimally and cost-efficiently. To assist researchers with study planning, we developed a Shiny App, which can be accessed online via the following link: https://psychtools.shinyapps.io/optimalDesignsClassroomClimate. It is our hope that the app will further contribute to the widespread use of optimal designs in educational research. In closing, we want to emphasize once more that the application of the app is not limited to the educational context, but it may also be useful to researchers from other areas of research in which climate variables are of interest, such as organizational research focusing on organizational climate and other climate variables.

Acknowledgements We acknowledge support from the University of Tübingen's Open Access Publishing Fund.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Arens, A.K., & Morin, A.JS. (2016). Relations between teachers' emotional exhaustion and students' educational outcomes. *Journal of Educational Psychology*, *108*, 800–813. https://doi.org/10.1037/edu0000105.

Asparouhov, T., & Muthén, B.O. (2007). Constructing covariates in multilevel regression (Mplus web notes no. 11, version 2). http://www.statmodel.com/download/webnotes/webnote11.pdf.

Bardach, L., Oczlon, S., Pietschnig, J., & Lüftenegger, M. (2020). Has achievement goal theory been right? A meta-analysis of the relation between goal structures and personal achievement goals. *Journal of Educational Psychology*, *112*, 1197–1220. https://doi.org/10.1037/edu0000419.

Bardach, L., Yanagida, T., Klassen, R.M., & Lüftenegger, M. (2020). Normative and appearance performance approach goal structures: Two-level factor structure and external linkages. *The Journal of Experimental Education*. Advance online publication. https://doi.org/10.1080/00220973.2020.1729081.

Bardach, L., Yanagida, T., & Lüftenegger, M. (2020). Studying classroom climate effects in the context of multi-level structural equation modelling: An application-focused theoretical discussion and empirical demonstration. *International Journal of Research & Method in Education*, *43*, 348–363.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*, 133–180. https://doi.org/10.3102/0002831209345157.

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: Foundation, extensions, and new directions* (pp. 349–381). San Francisco: Jossey-Bass.

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

Buonaccorsi, J.P. (2010). *Measurement error: Models, methods, and applications*. New York: CRC Press.

Cools, W., van den Noortgate, W., & Onghena, P (2008). Ml-des: A program for designing efficient multilevel studies. *Behavior Research Methods*, *40*, 236–249. https://doi.org/10.3758/BRM.40.1.236.

Cronbach, L.J. (1976). *Research on classrooms and schools: Formulations of questions, design and analysis*. Stanford: Stanford Evaluation Consortium.

Croon, M.A., & van Veldhoven, M.JPM. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*, 45–57. https://doi.org/10.1037/1082-989X.12.1.45.

Dong, N., & Maynard, R. (2013). Powerup!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*, 24–67. https://doi.org/10.1080/19345747.2012.673143.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. New York: Wiley.

Downer, J.T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E (2015). Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. *The Journal of Early Adolescence*, *35*, 722–758. https://doi.org/10.1177/0272431614564059.

Emmer, E.T., & Stough, L.M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist*, *36*, 103–112. https://doi.org/10.1207/S15326985EP3602_5.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Buttner, G (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001.

Fuller, W.A. (1987). *Measurement error models*. New York: Wiley.

Ghiselli, E.E., Campbell, J.P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W. Freeman & Co.

Göllner, R., Fauth, B., & Wagner, W (in press). Student ratings of teaching quality dimensions: Empirical findings and future directions. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.) *Student feedback in schools – Using perceptions for the development of teaching and teachers*. Berlin: Springer.

Göllner, R., Wagner, W., Eccles, J.S., & Trautwein, U (2018). Students' idiosyncratic perceptions of teaching quality in mathematics: A result of rater tendency alone or an expression of dyadic effects between students and teachers?. *Journal of Educational Psychology*, *5*, 709–725. https://doi.org/10.1037/edu0000236.

Grilli, L., & Rampichini, C. (2011). The role of sample cluster means in multilevel models. *Methodology*, *7*, 121–133. https://doi.org/10.1027/1614-2241/a000030.

Hamaker, E.L., & Klugkist, I. (2011). Bayesian estimation of multilevel models. In J. J. Hox, & J. K. Roberts (Eds.) *Handbook of advanced multilevel analysis* (pp. 137–161). New York: Routledge.

Hamre, B.K., & Pianta, R.C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure?. *Child Development*, *76*, 949–967. https://doi.org/10.1111/j.1467-8624.2005.00889.x.

Helmke, A., Schneider, W., & Weinert, F.E. (1986). Quality of instruction and classroom learning outcomes: The German contribution to the IEA classroom environment study. *Teaching and Teacher Education*, *2*, 1–18. https://doi.org/10.1016/0742-051X(86)90002-8.

Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology*, *106*, 289–300. https://doi.org/10.1037/a0033829.

Hughes, J.N., Luo, W., Kwok, O.-M., & Loyd, L.K. (2008). Teacher-student support, effortful engagement, and achievement: A 3-year longitudinal study. *Journal of Educational Psychology*, *100*, 1–14. https://doi.org/10.1037/0022-0663.100.1.1.

Kaplan, A., Gheen, M., & Midgley, C. (2002). Classroom goal structure and student disruptive behaviour. *British Journal of Educational Psychology*, *72*, 191–211. https://doi.org/10.1348/000709902158847.

Kelcey, B., Dong, N., Spybrook, J., & Shen, Z. (2017). Experimental power for indirect effects in group-randomized studies with group-level mediators. *Multivariate Behavioral Research*, *52*, 699–719. https://doi.org/10.1080/00273171.2017.1356212.

Klem, A.M., & Connell, J.P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *The Journal of School Health*, *74*, 262–273. https://doi.org/10.1111/j.1746-1561.2004.tb08283.x.

Köhler, C., Hartig, J., & Naumann, A. (2021). Detecting instruction effects – Deciding between covariance analytical and change-score approach. *Educational Psychology Review*. Advance online publication. https://doi.org/10.1007/s10648-020-09590-6.

Kounin, J.S. (1970). *Discipline and group management in classrooms*. New York: Holt, Rinehart & Winston.

Kozlowski, S.WJ., & Klein, K.J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein, & S. W. J. Kozlowski (Eds.) *Multilevel theory, research, and methods in organizations* (pp. 3–90). San Francisco: Jossey-Bass.

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction*, *17*, 494–509. https://doi.org/10.1016/j.learninstruc.2007.09.002.

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*, *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.

LeBreton, J.M., & Senter, J.L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852. https://doi.org/10.1177/1094428106296642.

Lewis, R. (2001). Classroom discipline and student responsibility: The students' view. *Teaching and Teacher Education*, *17*, 307–319. https://doi.org/10.1016/S0742-051X(00)00059-7.

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*, 444–467. https://doi.org/10.1037/a0024376.

Lüdtke, O., Marsh, H.W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B..O (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229. https://doi.org/10.1037/a0012869.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modelling. *Educational Psychology*, *34*, 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research*, *9*, 215–230. https://doi.org/10.1007/s10984-006-9014-8.

Maas, C.JM., & Hox, J.J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 85–91. https://doi.org/10.1027/1614-2241.1.3.85.

Mainhard, M.T., Brekelmans, M., den Brok, P., & Wubbels, T. (2011). The development of the classroom social climate during the first months of the school year. *Contemporary Educational Psychology*, *36*, 190–200. https://doi.org/10.1016/j.cedpsych.2010.06.002.

Marsh, H.W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A.JS., 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*, 106–124. https://doi.org/10.1080/00461520.2012.670488.

Marsh, H.W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. O., & 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. https://doi.org/10.1080/00273170903333665.

Matheny, K.B., & Edwards, C.R. (1974). Academic improvement through an experimental classroom management system. *Journal of School Psychology*, *12*, 222–232.

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research*, *5*, 661–670. https://doi.org/10.1080/00273171.2017.1344538.

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

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

Nelder, J.A., & Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, *7*, 308–313.

Patrick, H., Ryan, A.M., & Kaplan, A. (2007). Early adolescents' perceptions of the classroom social environment, motivational beliefs, and engagement. *Journal of Educational Psychology*, *99*, 83–98. https://doi.org/10.1037/0022-0663.99.1.83.

Pianta, R.C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F.J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*, 365–397. https://doi.org/10.3102/0002831207308230.

Praetorius, A.K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM*, *50*, 407–426. https://doi.org/10.1007/s11858-018-0918-4.

R Development Core Team (2016). R: A language and environment for statistical computing. http://www.R-project.org.

Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173–185. https://doi.org/10.1037/1082-989X.2.2.173.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods*. *Advanced quantitative techniques in the social sciences*, 2nd edn. Thousand Oaks, CA: Sage.

Reyes, M.R., Brackett, M.A., Rivers, S.E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology*, *104*, 700–712. https://doi.org/10.1037/a0027268.

Rhoads, C.H. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics*, *36*, 76–104. https://doi.org/10.3102/1076998610379133.

Rimm-Kaufman, S.E., Baroody, A.E., Larsen, R.A., Curby, T.W., & Abry, T. (2015). To what extent do teacher-student interaction quality and student gender contribute to fifth graders' engagement in mathematics learning?. *Journal of Educational Psychology*, *107*, 170–185. https://doi.org/10.1037/a0037252.

Rolland, R.G. (2012). Synthesizing the evidence on classroom goal structures in middle and secondary schools: A meta-analysis and narrative review. *Review of Educational Research*, *82*, 396–435. https://doi.org/10.3102/0034654312464909.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02.

Seidel, T., & Shavelson, R.J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*, 454–499. https://doi.org/10.3102/0034654307310317.

Shin, Y., & Raudenbush, S.W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, *35*, 26–53. https://doi.org/10.3102/1076998609345252.

Snijders, T.A.B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt, & D. C. Howell (Eds.) *Encyclopedia of Statistics in Behavioral Science* (pp. 1570–1573). Chichester: Wiley.

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

Stallasch, S.E., Lüdtke, O., Artelt, C., & Brunner, M. (2021). Multilevel design parameters to plan cluster-randomized intervention studies on student achievement in elementary and secondary school. *Journal of Research on Educational Effectiveness*. Advance online publication. https://doi.org/10.1080/19345747.2020.1823539.

Stornes, T., & Bru, E. (2011). Perceived motivational climates and self-reported emotional and behavioral problems among Norwegian secondary school students. *School Psychology International*, *32*, 425–438. https://doi.org/10.1177/0143034310397280.

van Breukelen, G.J.P. (2013). Optimal experimental design with nesting of persons in organizations . *Zeitschrift für Psychologie*, *221*, 145–159. https://doi.org/10.1027/2151-2604/a000143.

van Breukelen, G.J.P., & Candel, M.J.J.M. (2015). Efficient design of cluster randomized and multicentre trials with unknown intraclass correlation. *Statistical Methods in Medical Research*, *24*, 540–556. https://doi.org/10.1177/0962280211421344.

Vansteenkiste, M., Sierens, E., Goossens, L., Soenens, B., Dochy, F., Mouratidis, A., Aelterman, N., Haerens, L., & Beyers, W. (2012). Identifying configurations of perceived teacher autonomy support and structure: Associations with self-regulated learning, motivation and problem behavior. *Learning and Instruction*, *22*, 431–439. https://doi.org/10.1016/j.learninstruc.2012.04.002.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, *28*, 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003.

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

Wang, M.T., & Degol, J.L. (2016). School climate: A review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review*, *28*, 315–352. https://doi.org/10.1007/s10648-015-9319-1.

Wang, M.-T., Degol, J.L., Amemiya, J., Parr, A., & Guo, J. (2020). Classroom climate and children's academic and psychological wellbeing: A systematic review and meta-analysis. *Developmental Review*, *57*, 1–21. https://doi.org/10.1016/j.dr.2020.100912.

Zitzmann, S. (2018). A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivariate Behavioral Research*, *53*, 612–632. https://doi.org/10.1080/00273171.2018.1469086.

Zitzmann, S., & Helm, C. (2021). Multilevel analysis of mediation, moderation, and nonlinear effects in small samples, using expected a posteriori estimates of factor scores. *Structural Equation Modeling*. Advance online publication. https://doi.org/10.1080/10705511.2020.1855076.

Zitzmann, S., Helm, C., & Hecht, M. (2021). Prior specification for more stable Bayesian estimation of multilevel latent variable models in small samples: A comparative investigation of two different approaches. *Frontiers in Psychology*, *11*, 1–11. https://doi.org/10.3389/fpsyg.2020.611267.

Zitzmann, S., Lüdtke, O., & Robitzsch, A. (2015). A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivariate Behavioral Research*, *50*, 688–705. https://doi.org/10.1080/00273171.2015.1090899.

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Hecht, M. (2021). On the performance of Bayesian approaches in small samples: A comment on Smid, McNeish, Miočević, and van de Schoot (2020). *Structural Equation Modeling*, *28*, 40–50. https://doi.org/10.1080/10705511.2020.1752216.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Steffen Zitzmann[1,2] · **Wolfgang Wagner**[1] · **Martin Hecht**[1] · **Christoph Helm**[3] · **Christian Fischer**[1] · **Lisa Bardach**[1] · **Richard Göllner**[1]

[1] University of Tübingen, Tübingen, Germany

[2] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, 72072 Tübingen, Germany

[3] Johannes Kepler University Linz, Linz, Austria