Educational Psychology Review (2025) 37:53
https://doi.org/10.1007/s10648-025-10029-z

REVIEW ARTICLE

A More Comprehensive, More Reliable Multilevel Approach for Assessing and Modeling Teacher Judgment Accuracy Using Latent Variables

Julian F. Lohmann[1] · **Nils Machts**[1] · **Jens Möller**[1] · **Steffen Zitzmann**[2]

Accepted: 12 May 2025 / Published online: 2 June 2025

Abstract

We propose a novel approach for modeling judgment accuracy that, for the first time, allows for simultaneously considering the rank, level, and differentiation component, the predominantly applied operationalization of teacher judgment accuracy. These components are conceptualized as latent, unobserved individual abilities. The model is introduced technically and its functionality is illustrated. Next, several model extensions are described that enhance the model's capabilities to address important research questions in teacher judgment accuracy research, such as concerning moderators of judgment accuracy or the effect of judgment accuracy on learning outcomes. We study the newly proposed model in three simulation studies demonstrating that our approach provides more accurate individual-level estimates of judgment accuracy components than the traditional, still widely applied person- and component-wise calculation (simulation study 1), that our approach yields more accurate standard errors of moderation effects of judgment accuracy components resulting in higher true positive rates across various typical sample size scenarios (simulation study 2), and that our approach yields lower parameter bias and higher coverage rates of predictive effects (simulation study 3). The findings underscore the model's potential for improving the assessment and modeling of judgment accuracy although the improvement over the person-score-based two-step approach was rather small in some conditions. We present two real data examples, including a step-by-step tutorial how to apply the newly proposed approach, and conclude by discussing the implications of our results and suggesting directions for future research. Easy-to-use R code is provided to simplify the application of the new judgment accuracy model.

Keywords Teacher judgments · Judgment accuracy · Latent variable model · Multilevel

Extended author information available on the last page of the article

Introduction

The question of whether people can accurately assess the characteristics of others has long been the subject of psychological research (e.g., Funder, & D.C., 1995; Jussim, 2012). A prominent application of this research pertains to teachers' judgments of student characteristics (Hoge & Coladarci, 1989; Jussim & Eccles, 1992; Machts et al., 2016; Südkamp et al., 2012). Accurately assessing student characteristics, such as cognitive ability, learning gains, motivation, or emotional states, is a crucial aspect of teaching, often referred to as *assessment competence* (also *diagnostic competence*, e.g., Koeppen et al., 2008; Weinert et al., 1990). Assessment competence is supposed to ensure fair grading (e.g., Brookhart, 2011; Glock et al., 2013) and effective instructional decisions (Artelt & Rausch, 2014; Lorenz & Artelt, 2009; Südkamp & Praetorius, 2017).

Building on the work of Cronbach (1955), Schrader and Helmke (1987) identified three components of judgment accuracy: the rank, level, and differentiation component. These components have since become central to operationalizing judgment accuracy in research on teacher assessments and are used as indicators of teachers' assessment competence (see, e.g., Leuders et al., 2018; Spinath, 2005; Urhahne & Wijnia, 2021). Although some researchers have attempted to unify these components into a more comprehensive model (e.g., Karst et al., 2017), no overarching model currently exists that captures all three components simultaneously. Due to the lack of a comprehensive model, either only single components can be examined, or three separate analyses have to be conducted, with the latter rarely occurring (see Artelt & Rausch, 2014; Urhahne & Wijnia, 2021). The present article aims to overcome this limitation by proposing a comprehensive multilevel latent variable modeling (ML-LVM) approach. We conceptualize the three components of judgment accuracy as raters' latent, unobserved abilities (Leuders et al., 2018), which are measured via a finite number of judgments (e.g., a teacher's assessments of students in a single class) assembled from a potentially infinite number of judgments (a teacher's assessments of "all" students in the population of students). The proposed ML-LVM accounts for the sampling error that occurs when only a small number of judgments are used (Lüdtke et al., 2011; Lüdtke et al., 2008; see also Zitzmann et al., 2022).

Much research on judgment accuracy is concerned with moderation effects. This is reflected by many theoretical models on judgment accuracy, where moderators of judgment accuracy play a crucial role (Funder, & D.C., 1995; Herppich et al., 2018; Loibl et al., 2020; Südkamp et al., 2012). In research on teacher judgment accuracy, one type of moderator of primary interest is teachers' characteristics (Urhahne & Wijnia, 2021). For instance, one research topic is the question whether teacher' content knowledge affects judgment accuracy (e.g., Jansen et al., 2021; Möller et al., 2022). Consequently, we extend our model in such a way that it allows to investigate moderators.

To sum up, we demonstrate that the ML-LVM approach offers several advantages over the traditional component-wise and person-wise modeling approach. Specifically, our approach (1) allows for the simultaneous assessment of the three

components of judgment accuracy, (2) provides more reliable estimates of person-level components by accounting for sampling error, (3) yields population-level estimates of the variability in accuracy components (i.e., between-person differences in diagnostic competence), and (4) offers more precise estimates of standard errors (*SE*s) and confidence intervals when analyzing moderators of judgment accuracy.

The article is structured as follows. First, we introduce the statistical details of our model and provide an illustrative introduction to its functionality. Next, three simulation studies are conducted that demonstrate the functionality and advantages of our judgment accuracy model as compared to the component- and person-wise calculation. Third, we present two empirical applications that illustrate how the ML-LVM can be used to investigate teacher judgment accuracy. Finally, we discuss limitations of the proposed approach and paths for future research.

The Multilevel Latent Variable Model of Judgment Accuracy

Investigating Judgment Accuracy

Researchers typically quantify judgment accuracy by comparing participants' judgments with an objective benchmark (e.g., Leuders et al., 2018). In this context, a *benchmark* refers to a theoretically or empirically grounded criterion that serves as the standard against which the accuracy of judgments is evaluated. In research on teacher judgment accuracy, for instance, teacher judgments of student characteristics have been compared with students' performance in achievement tests (Südkamp et al., 2012) or intelligence tests (Machts et al., 2016). For motivational or emotional student characteristics, teacher judgments have been compared with students' self-reports (Spinath, 2005). Other studies have utilized virtual classroom environments that simulate student performance (e.g., Kaiser et al., 2017; Machts et al., 2024; Südkamp et al., 2008) or provided teachers with a selection of student works (e.g., student essays) sampled from a large corpus of authentic student works that had been assessed by trained expert raters (e.g., Jansen et al., 2021; Möller et al., 2022).

Cronbach (1955) was the first to highlight that judgment accuracy consists of several largely independent components. He cautioned that conflating these components obscures the interpretation of the psychological mechanisms underlying judgment (in-)accuracy and can lead to artifacts. Based on Cronbach ideas, Schrader and Helmke (1987) separated three judgment accuracy components that are today the most widely used conceptualization of judgment accuracy in research on teacher judgments (Urhahne & Wijnia, 2021). These components are referred to as the rank, level, and differentiation component of judgment accuracy. Figure 1, adapted from Leuders et al. (2018), illustrates the three components.

The rank component, also referred to as *relative judgment accuracy* in the literature, is defined as the within-rater correlation between judgments and benchmarks. To obtain the population mean, the rater-specific correlations are averaged (and sometimes Fisher-z transformed, see, e.g., Urhahne & Wijnia, 2021) in a second step. The level component represents the difference between the rater-specific

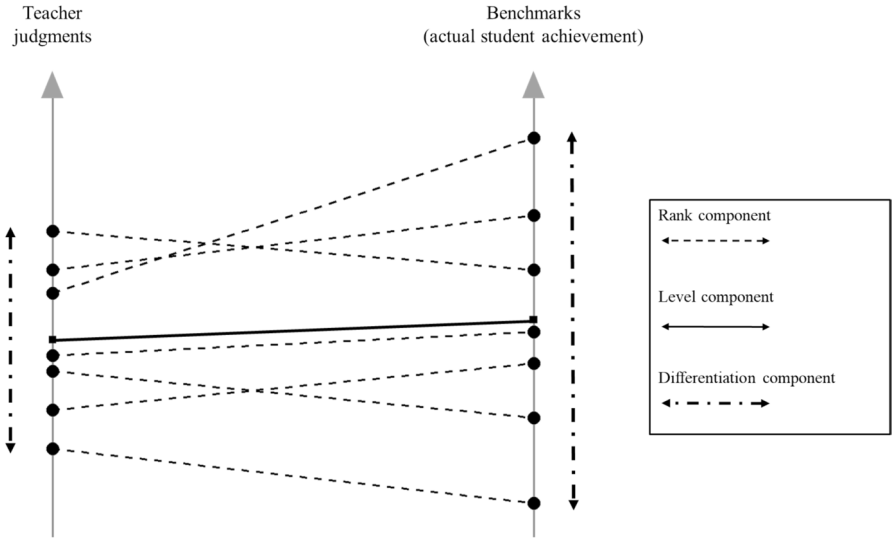

Fig. 1 Illustration ofthe rank, level, and differentiation component of judgment accuracy. *Note*. Adapted from *Diagnostic Competence of Mathematics Teachers: Unpacking a Complex Construct* by T. Leuders, T. Dörfler, J. Leuders, & K. Philipp, 2018, p. 10, Copyright 2018 by Springer International Publishing AG. Adapted with permission.

mean averaged across judgments and the mean of the corresponding benchmarks. This indicates a rater's tendency to over- or underestimate the target characteristic, with negative values implying underestimation and positive values implying overestimation. At the population level, the level component is calculated by averaging the rater-specific level components. Lastly, the differentiation component is the ratio of the variance of judgments and the variance of the corresponding benchmarks (Schrader, 2013; M. Zhu & Urhahne, 2014). A value of 1 indicates accurate assessment of the variability of the target characteristic, values below 1 indicate underestimation, and values above 1 indicate overestimation.

In recent years, researchers have aimed to model the population and individual level jointly by using multilevel approaches to judgment accuracy (e.g., Bonefeld et al., 2020; Dicke et al., 2012; Gnas et al., 2022; Karst et al., 2017; Kolovou et al., 2021). Some of these approaches also included two components of judgment accuracy into a joint model. For instance, Karst and colleagues (2017) proposed a special centering approach before applying a multilevel model. This approach allows representing the level and the rank component simultaneously. Others, such as Gnas and colleagues (2022), applied a two-step approach that operationalized the rank component as a regression coefficient in a hierarchical regression model and then used the standardized residuals from this model to calculate the level component in a separate step.

However, no current approach allows for the joint representation of all three components along with their within-rater and between-rater measures. Therefore, most

research on judgment accuracy has focused on only one component, often the rank component (Südkamp et al., 2012; Urhahne & Wijnia, 2021). To overcome this shortcoming, we propose a new model that captures all three components simultaneously. In using a multilevel latent variable framework (see, e.g., Feng & Hancock, 2024; Lüdtke et al., 2008; Rabe-Hesketh et al., 2004), our approach promises more reliable estimates of judgment components by accounting for sampling error due to a limited and typically only small number of judgments. Furthermore, this approach likely enhances the accuracy of standard errors and confidence intervals for moderations of judgment accuracy.

The ML-LVM Approach

In judgment accuracy research, one typically compares a vector of judgments $\boldsymbol{y}$, containing $i = 1, 2, \ldots, I$ judgments from $j = 1, 2, \ldots, J$ raters (e.g., teachers), with a corresponding vector of benchmark scores $\boldsymbol{x}$ (e.g., students actual abilities represented by test scores or simulated student ability in a virtual classroom environment). The three components of judgment accuracy concern the rater-specific means, variances, and covariance of the two vectors $\boldsymbol{x}_j$ and $\boldsymbol{y}_j$. These means, variances, and covariance can be represented by a multivariate normal distribution. To account for between-person differences in judgment accuracy, the parameters of the bivariate normal distribution are allowed to vary across raters, representing the latent (unobserved) variables of interest in our model. Specifically, we specify rater-specific bivariate normal distributions, which lie at the core of our model:

$$\begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_j^{(x)} \\ \mu_j^{(y)} \end{bmatrix}, \begin{bmatrix} \sigma^{2(x)}_j & \sigma_j^{(xy)} \\ \sigma_j^{(yx)} & \sigma^{2(y)}_j \end{bmatrix} \right) \tag{1}$$

with

$$\mu_j^{(x)} = \mu^{(x)} + u_j \tag{2}$$

$$\sigma^{2(x)}_j = \sigma^{2(x)} \times \exp(v_j) \tag{3}$$

where $\mu_j^{(x)}$ is a rater-specific mean and $\sigma^{2(x)}_j$ is a rater-specific variance of the benchmark scores. Similarly, $\mu_j^{(y)}$ is the rater -specific mean and $\sigma^{2(y)}_j$ is the rater-specific variance of the judgments. The term $\sigma_j^{(yx)}$ is the rater-specific covariance of the benchmark scores and judgments. The rater-specific mean of the benchmark scores is modelled as the sum of the overall population-level mean $\mu^{(x)}$ and a rater-specific deviation from this mean u_j (i.e., random intercept). These deviations are assumed to follow a normal distribution $u_j \sim N(0, \omega^{2(\mu)})$, where the between-rater variability in the benchmark mean levels is quantified via the variance component $\omega^{2(\mu)}$.

The rater-specific variance of the benchmark scores is modeled as the product of the overall population-level variance $\sigma^{2(x)}$ and a rater-specific deviation from this variance v_j (i.e., random variance; e.g., Feng & Hancock, 2024). To

avoid negative variances, we use a log-exponential transformation here (see, e.g., Bauer, 2017). The amount of between-person variability in the benchmark variance is also quantified via an additional variance component $\omega^{2(\sigma^2)}$ and, thus, the random effects are assumed to follow $v_j \sim N(0, \omega^{2(\sigma^2)})$. However, both variance components ω^2 can be fixed to zero when all raters are provided with the same judgment stimuli implying that $\mu^{(x)}$ and $\sigma^{2(x)}$ stay constant across participants in a given study.

To express the rank, level, and differentiation components, that is, the parameters of interest, we express the rater judgment parameters $\mu_j^{(y)}$ and $\sigma^{2(y)}_j$ as functions of the corresponding benchmarks as follows:

$$\mu_j^{(y)} = \mu_j^{(x)} + \text{level component}_j, \tag{4}$$

$$\sigma^{2(y)}_j = \sigma^{2(x)}_j \times exp(differentiationcomponent_j), \tag{5}$$

Subscript j indicates that we also allow individual differences for the differentiation and the level components. Again, we employ the log-exponential transformation to ensure that the variances for all raters are positive. To derive the (rater-specific) rank component we standardize the covariance with the well-known equation for determining the Pearson correlation coefficient:

$$\text{rank component}_j = \frac{\sigma_j^{(xy)}}{\sqrt{\sigma^{2(x)}_j \times \sigma^{2(y)}_j}} \tag{6}$$

The three judgment accuracy components follow a multivariate normal distribution:

$$\begin{bmatrix} \text{level component}_j \\ \text{differentiation component}_j \\ \text{rank component}_j \end{bmatrix} \sim MVN\left(\begin{bmatrix} \text{level component} \\ \text{differentiation component} \\ \text{rank component} \end{bmatrix}, \begin{bmatrix} \omega^{2(l)} & \omega^{(ld)} & \omega^{(lr)} \\ \omega^{(dl)} & \omega^{2(d)} & \omega^{(dr)} \\ \omega^{(rl)} & \omega^{(rd)} & \omega^{2(r)} \end{bmatrix} \right) \tag{7}$$

In this equation, $\omega^{(dl)}$ is the covariance between the individual differentiation and level components, $\omega^{(rl)}$ is the covariance between the individual rank and level components, and $\omega^{(rd)}$ is the covariance between the individual rank and differentiation components. The means (first part of the MVN) serve as population-level estimates of the components. The variances (diagonal elements of the second part of the MVN) quantify the amount of individual differences for each component and the covariances (off-diagonals of the second part of the MVN) represent the interrelation between components. In applied research, the correlation between multiple operationalizations of judgment accuracy is often of substantive interest, which is one key motivation to explicitly model the covariance structure of these components (e.g., Schrader, 1989; Südkamp et al., 2008). The person-specific

components $\begin{bmatrix} \text{level component}_j \\ \text{differentiation component}_j \\ \text{rank component}_j \end{bmatrix}$ are represented as individual deviations from the population means and are stacked row-wise into a $J \times 3$ matrix $\boldsymbol{\theta}$.

Illustration of the Newly Proposed ML-LVM

To illustrate how the newly proposed ML-LVM works, consider a scenario in which 20 teachers each evaluate the 30 students in their class. For simplicity, assume that all classes have the same size. The teachers are asked to predict how well their students will perform on a standardized test—a common approach in teacher judgment accuracy research (see Urhahne & Wijnia, 2021). Meanwhile, the students complete the test, and the teachers' predictions are later compared to the students' observed performance.

In this example, the test scores range from 0 (no correct answers) to 100 (all answers correct). The teachers are asked to predict each student's performance on a corresponding scale ranging from 0 to 100.

Descriptive statistics reveal that, on average, students score 60 points and that the test results vary across students with a standard deviation of 10. However, mean performance and variability also differs across classes. While the two quantities—mean $=60$ and $SD=10$—describe the population distribution of the test scores, which correspond to $\mu^{(x)}$ and $\sigma^{(x)}$ of the ML-LVM, teacher A's class yields its own true mean achievement and variability. For instance, teacher A's class might score an average of 70 points, which is 10 points above the overall mean; this deviation is captured in the ML-LVM by u_j. If the standard deviation of teacher A's class is 12—as opposed to the overall 10—this difference is represented by v_j. The parameters $\omega^{2(\mu)}$ and $\omega^{2(\sigma)}$ quantify the extent to which classes vary in their average achievement and variability.

A key step in teacher judgment accuracy research is relating students' actual scores to teachers' predictions. In the ML-LVM, three primary quantities describe this relationship: level, differentiation, and rank component. The level component examines whether teachers systematically overestimate or underestimate student performance. Suppose the average teacher rating is 65, whereas the overall mean score is 60. The level component of $+5$ indicates an overestimation. However, teachers may differ in this regard; if teacher A's mean rating is 66 compared to the class's actual mean of 70, teacher A's level component is -4. These teacher-specific deviations are denoted by level component_j, and the variance of these differences across teachers is captured by $\omega^{2(l)}$.

The differentiation component focuses on how accurately teachers judge the variability of abilities of the students in their class. If the average within-class *SD* of teacher ratings across the entire sample is 8 and thus deviates from the actual within-class test score *SD* of 10, the ratio $\frac{8^2}{10^2} = 0.64$ reflects a population-level differentiation accuracy that indicates that the teacher underestimates the variability in their class. As with the level component, teachers may vary in how well

they capture this variability. The teacher-specific differentiation is represented by differentiation component$_j$, with its variance across teachers denoted by $\omega^{2(d)}$.

Finally, the rank component captures the extent to which the relative ordering of students' predicted scores matches their actual performance ranking. If the average correlation between teacher ratings and test scores is 0.55, but teacher A's correlation is 0.45, it suggests that teacher A's rank accuracy is lower than average. In the ML-LVM, these teacher-specific correlations are represented by rank component$_j$ and $\omega^{2(r)}$ is the variance of rank accuracy across teachers.

Bayesian Estimation and the Choice for Prior Distributions

While varying intercept parameters are well established and frequently applied by researchers when using frequentist estimation, varying variances and covariances are less common. This is largely due to the fact that models with many so-called random effects can quickly lead to estimation problems in frequentist estimation, as high-dimensional numerical integration is required (Asparouhov & Muthén, 2012; Zyphur & Oswald, 2015). However, in Bayesian modeling and estimation, the specification of random effects is very common and implied by Bayesian theory (see, e.g., Gelman et al., 2013; Zitzmann et al., 2016). Moreover, sampling-based Bayesian estimation techniques can circumvent estimation problems, such as convergence issues or inadmissible solutions that arise when fitting highly complex models with optimization algorithms (e.g., Ulitzsch et al., 2023; Zitzmann et al., 2015, 2020; Zyphur & Oswald, 2015). Against this background, we decided to build our model in the Bayesian environment of Stan (Stan Development Team, 2024) and use Markov Chain Monte Carlo (MCMC) techniques for model estimation.

Regarding the prior specifications, we relied on standard choices, particularly normal distributions for the intercept parameters and means of the judgment components, half-Cauchy priors for the population-level variances, and the Lewandowski–Kurowicka–Joe prior (Lewandowski et al., 2009) for the (co)variances of all random effects (e.g., Depaoli, 2021; Gelman et al., 2013 and more details are provided in the Supplemental Material on OSF).

Several Extensions for the ML-LVM

Teacher judgment accuracy research often goes beyond quantifying accuracy across various judgment tasks (e.g., Südkamp et al., 2012) in order to also examine predictors and outcomes associated with individual differences in judgment accuracy (e.g., Herppich et al., 2018; Loibl et al., 2020). Substantive questions include whether higher judgment accuracy is linked to better instructional decisions or increased learning, and whether more experienced teachers demonstrate higher accuracy (Urhahne & Wijnia, 2021).

A notable advantage of the ML-LVM is its flexibility to integrate predictors and outcome variables directly into the model, facilitating the examination of whether and how these variables explain interindividual differences in judgment accuracy. This feature enables the testing of hypotheses derived from teacher judgment

theories without relying on conventional, suboptimal two-step approaches. In those two-step approaches, person-specific scores (e.g., individual estimates of judgment accuracy) are calculated in the first stage for each teacher separately and then treated as observed variables in subsequent analyses. Following Liu (2017), who discussed similar approaches in longitudinal modeling, we refer to this conventional person-wise approach as the PS method. However, such methods do not account for the uncertainty in the initial person-specific scores, which can have various consequences that reach beyond unreliable assessments of individual persons' judgment accuracy when these scores are used in further analyses. These potential consequences range from severe biases in the relations between judgment accuracy and other variables, false representations of uncertainty by standard errors, up to losses in the power to detect a relationship (e.g., Brose et al., 2022; Zitzmann & Orona, 2025).

Person-Specific Judgment Accuracy Components via Empirical Bayes Estimates

The person-specific judgment accuracy estimates provided by the ML-LVM, which are represented as individual deviations from the population mean, can be derived from the posterior means of the individual-level parameters in the $\boldsymbol{\theta}$ matrix. These estimates reflect a rater's judgment accuracy when the hierarchical structure of the data is considered. Technically, they represent empirical Bayes estimates, which are informed both by the individual's observed data and the overall population distribution. This approach allows for shrinkage of individual estimates toward the group mean, with the degree of shrinkage depending on the amount of information available for a given rater and the knowledge derived from the population distribution (for mathematical details, see, e.g., Lüdtke et a., 2008, Zitzmann et al., 2024).

Importantly, this aspect of our model enables the derivation of person-specific estimates and thus aligns with Brunswik's emphasis on adopting a more idiographic perspective on human judgment (see, e.g., Brunswik, 1955; Kaufmann et al., 2007). Studying interindividual differences is critical for understanding the factors that shape judgment accuracy (Südkamp et al., 2012). Thus, our approach aims to effectively integrate both the population (nomothetic) and individual (idiographic) perspective on judgment accuracy so that insights from each can inform the other.

Including Moderators of Judgment Accuracy Components

Theories and research on (teacher) judgment accuracy are concerned with moderators. In this section, we show how our proposed model can be extended to include rater-level moderators for the three components of judgment accuracy. To this end, we assume a rater-specific vector z_j of $m = 1,2,\ldots,M$ rater characteristics. To be able to assess their influence, we extended Eq. 7 as follows:

$$\begin{bmatrix} \text{level component}_j \\ \text{differentiation component}_j \\ \text{rank component}_j \end{bmatrix} \sim MVN\left(\begin{bmatrix} \text{level component} \\ \text{differentiation component} \\ \text{rank component} \end{bmatrix} + \mathrm{B}z_j, \begin{bmatrix} \sigma^{2(l)} & \sigma^{(ld)} & \sigma^{(lr)} \\ \sigma^{(dl)} & \sigma^{2(d)} & \sigma^{(dr)} \\ \sigma^{(rl)} & \sigma^{(rd)} & \sigma^{2(r)} \end{bmatrix} \right) \quad (8)$$

where z_j is a rater-specific vector of length M containing potential moderators and B is a $3 \times M$ matrix of moderation effects quantifying the influence of the moderators on the three judgment accuracy components. All other equations and parameters stay as described before.

Individual-Level Judgment Accuracy Components as Predictors of Learning Outcomes

Another aspect of research on judgment accuracy are the consequences of making more or less accurate judgments. For example, researchers have examined whether teachers with high judgment accuracy make better instructional decisions or whether their students show increased learning gains (Kolovou et al., 2024). The ML-LVM can be extended by adding outcome variables to address questions about the relationship between individual judgment accuracy components and various learning outcomes. Specifically, the following regression equation can be incorporated into the ML-LVM described in "The ML-LVM Approach" section:

$$z_j = \boldsymbol{\alpha} + \boldsymbol{\theta}_j \mathbf{B} + \boldsymbol{\varepsilon} \quad (9)$$

where z_j is a rater-specific vector, which includes $k = 1,2,\ldots,K$ outcome variables of interest. B is a $K \times 3$ matrix of regression weights interrelating individual judgment accuracy components and outcome variables. Additionally, $\boldsymbol{\alpha}$ is a vector of intercepts and $\boldsymbol{\varepsilon}$ is a vector of residuals, both of length K.

Simulation Studies

In this section, we present three simulation studies designed to examine the potential advantages of the newly proposed ML-LVM and its extensions. In each study, the ML-LVM served as the data-generating model so that the analysis model matched the data-generating model perfectly.

Simulation Study 1

In our first simulation study, we aimed to compare the population-level and individual-level judgment accuracy estimates between the proposed ML-LVM approach and the PS method.

Simulation Settings and Design Factors

To study the performance of the two approaches, we adopted a range of typical sample sizes and parameter settings. For this purpose, we distinguished fixed and random design factors (see, e.g., Brandt et al., 2023). Fixed design factors remained constant within simulation conditions and only varied across simulation conditions, while random design factors varied within and across simulation conditions.

For fixed design factors, we varied (1) the number of teachers ($N_{Level\text{-}2}$) and (2) the number of judgments per teacher ($N_{Level\text{-}1}$). The conditions were informed by the existing empirical literature (for an overview of sample sizes, see, e.g., Kaufmann, 2020; Urhahne & Wijnia, 2021). For the first fixed design factor, we chose the settings $N_{Level\text{-}2} = 50$, 100, and 200. For the second fixed design factor, we chose the settings $N_{Level\text{-}1} = 5$, 10, and 30. This resulted in a 3×3 simulation study design.

As random design factors, we varied the population means of the three judgment accuracy measures: rank, level, and differentiation component. The decisions for the simulation settings were informed by the existing literature (Machts et al., 2016; Südkamp et al., 2012; Urhahne & Wijnia, 2021). For the population mean of the rank component, we sampled a random value of the parameter space: [.3;.4;.5;.6;.7] for each simulated dataset. The same procedure was applied for the population mean of the level component: [−0.5; − 0.2;0;0.2;0.5] and the population mean of the differentiation component: [0.9;0.95;1;1.05;1.1].

Following first available findings (e.g., Bonefeld et al., 2020; Mack et al., 2023; M. Zhu & Urhahne, 2015), the *SD*s of the individual deviations from the population means were set to 0.1 for all three components and held constant across conditions. Thus, for each individual in a given simulated dataset, the three individual-level judgment accuracy components were sampled from normal distributions with population mean defined by the specified parameter spaces described above and a fixed $SD = 0.1$. For the rank component, we additionally set an upper bound of 0.95 for the individual parameters.

We simulated 250 datasets for each condition. To analyze the data, we used (1) the PS method for the calculation of the three Schrader components (see Eqs. 9–11 and, e.g., Urhahne & Wijnia, 2021) and (2) our proposed ML-LVM. To estimate the ML-VLM, we used MCMC estimation via Stan's NUTS sampler with 3 chains and 10,000 iterations per chain.[1]

Evaluation Metrics

To evaluate the performance of the approaches, we employed two common metrics. The first is parameter bias (Muthén & Muthén, 2002; in the following referred to as *bias*), which is calculated by averaging the differences between the estimated parameter and the simulated population parameter and dividing it by the population

[1] Runtime was up to 270 min per dataset in the largest sample size condition (running three chains in parallel).

parameter. Muthén and Muthén recommended that the relative bias should not exceed 10%. We always multiplied bias with factor 100 to yield percentage bias.

Bias only represents systematic deviations from the true values, and (potentially large) negative and positive deviations can cancel each other out. Therefore, it is necessary also to consult an additional measure, such as the root mean squared error (RMSE; Zitzmann et al., 2021). For population parameters, the RMSE is calculated for $k = 1, 2, \ldots K$ replications as follows:

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(predicted_k - true_k)^2}$$

For individual-level parameters, the RMSE is calculated for $k = 1, 2, \ldots K$ replications and $j = 1, 2, \ldots J$ individuals:

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{J}\sum_{j=1}^{J}(predicted_{kj} - true_{kj})^2}$$

The RMSE is typically interpreted as how far the errors are from zero on average in the same unit of the parameter (Kuhn & Johnson, 2013). It provides good scope for comparing simulation results under different conditions relative to each other. However, there is no absolute criterion that defines when a particular result is sufficiently good, as is the case when using bias.

Bias and RMSE on the individual level were calculated by first averaging the individual bias and RMSE for each dataset.

To determine whether a model had reached convergence, we used the Rhat statistic ($\widehat{R}$). In doing so, we employed the typical threshold of $\widehat{R} < 1.1$ (e.g., Brandt et al., 2023).

Results

Comparing Population-Level Judgment Accuracy Components The results from our simulation study regarding the population-level estimates of the three judgment accuracy components are presented in Table 1. Both approaches yielded similar results concerning RMSE for the level and rank component. The RMSE was smaller for both approaches with larger level-1 and level-2 sample sizes. This pattern was also evident with regard to the differentiation component. Additionally, the RMSE was considerably higher for the PS method than the ML-LVM in all conditions. These differences were most prominent in small sample size conditions.

Regarding bias, a systematic overestimation was observed for the differentiation component across conditions for both approaches. However, overestimation was much higher for the differentiation component when using the PS method. Bias even exceeded the acceptable threshold of 10% in the small and medium level-1 sample size conditions. For the rank and level components, negative and positive deviations

Table 1 Percentage biasand RMSEof population-level judgment accuracy components

Sample size/simulation condition	Rank component				Level component				Differentiation component				Convergence (%)
	Bias (%)		RMSE		Bias (%)		RMSE		Bias (%)		RMSE		
	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	
$N_{\text{Level-2}} = 50$													
$N_{\text{Level-1}} = 10$	− 0.20	− 1.26	0.045	0.045	− 0.29	− 0.19	0.038	0.038	25.85	0.89	0.260	0.072	94
$N_{\text{Level-1}} = 20$	2.20	1.21	0.035	0.034	− 0.76	− 0.79	0.031	0.031	12.92	1.57	0.048	0.049	97
$N_{\text{Level-1}} = 30$	0.49	− 0.26	0.033	0.033	0.93	0.97	0.025	0.026	9.31	1.85	0.095	0.042	94
$N_{\text{Level-2}} = 100$													
$N_{\text{Level-1}} = 10$	0.64	0.01	0.036	0.036	− 0.02	− 0.04	0.028	0.028	24.27	1.17	0.243	0.045	88
$N_{\text{Level-1}} = 20$	1.34	0.52	0.027	0.027	0.07	0.03	0.021	0.021	11.98	0.99	0.120	0.034	82
$N_{\text{Level-1}} = 30$	1.12	0.28	0.026	0.025	− 0.38	− 0.37	0.018	0.018	8.95	1.67	0.090	0.032	91
$N_{\text{Level-2}} = 200$													
$N_{\text{Level-1}} = 10$	0.04	− 0.49	0.025	0.025	0.41	0.38	0.018	0.018	24.97	1.18	0.250	0.033	52
$N_{\text{Level-1}} = 20$	0.63	− 0.06	0.024	0.024	0.28	0.28	0.014	0.014	11.73	0.93	0.117	0.026	63
$N_{\text{Level-1}} = 30$	0.65	− 0.02	0.022	0.022	0.11	0.12	0.012	0.012	8.79	1.90	0.024	0.029	78

The convergence rate informs whether all estimates for the population parameters reached the criterion of $\hat{R} < 1.1$

varied across conditions and were very small, indicating no systematic bias. The percentage bias did not exceed the critical range of 10% for level and rank components and both approaches.[2]

The convergence rates indicated a relatively low likelihood for convergence in the $N_{\text{Level-2}} = 200$ conditions. However, it turned out that many runs just missed the necessary $\widehat{R}$ for single parameters and very closely. We are confident that these cases would have reached the convergence criterion when running for more iterations. In our simulations, we fixed the number of iterations across conditions to 10,000 to limit the run time.

Comparing Individual Judgment Accuracy Components Table 2 presents the evaluation measures for the individual-level parameters of the PS method and the ML-LVM. Bias indicated that both approaches consistently overestimate individual differentiation components. This tendency was much more pronounced for the PS method. Here, the percentage bias exceeded the acceptable threshold of 10% for the small and medium level-1 sample size conditions when using the PS method. For the individual level and rank components, findings imply very small, negligible bias for both the PS method and the ML-LVM.

The RMSE was considerably higher for the PS method across conditions and judgment accuracy components than for the ML-LVM. The RMSE was particularly high when the level-1 sample size was low. With higher level-1 sample sizes, the differences between both approaches became smaller. However, the RMSEs were still clearly smaller for the ML-LVM under the highest sample size condition.

Discussion Simulation Study 1

For the ML-LVM, we found that bias and RMSE for population-level estimates of the three components were small and consistently below the threshold of 10% across conditions. The PS method performed similarly well except for the differentiation component, where RMSEs were considerably higher than the ML-LVM. Additionally, bias indicated overestimation of the population-level differentiation component for the PS method that exceeded the acceptable threshold, especially in small sample size conditions.

On the individual level, results showed considerably higher RMSEs across the three components for the PS method than the ML-LVM. With higher sample sizes, the differences between the approaches diminished but were still evident. Systematic bias was again only present for the PS method and the differentiation component.

This simulation study highlighted the advantages of the ML-LVM of judgment accuracy for the individual-level (i.e., teacher-specific) parameter estimates of the rank, level, and differentiation components taking sampling error into account. However, if population estimates of the rank and level component are only of interest to

[2] The relative bias was calculated using the average difference between simulated and estimated parameters per component and condition and dividing it by the respective simulated average true parameter (i.e., 0.28 for the level, 1 for the differentiation, and 0.5 for the rank component).

Table 2 Parameter bias and RMSE of individual-level judgment accuracy components

	Rank component				Level component				Differentiation component				
	Bias (%)		RMSE		Bias (%)		RMSE		Bias (%)		RMSE		
Sample size	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	Convergence (%)
$N_{\text{Level-2}} = 50$													
$N_{\text{Level-1}} = 10$	− 0.61	− 1.66	0.161	0.012	− 1.99	− 1.99	0.104	0.012	25.90	0.97	1.161	0.018	89
$N_{\text{Level-1}} = 20$	2.36	1.38	0.083	0.011	− 0.55	− 0.60	0.052	0.011	12.76	1.42	0.309	0.013	96
$N_{\text{Level-1}} = 30$	0.64	− 0.47	0.058	0.011	0.78	0.82	0.035	0.010	9.11	1.67	0.184	0.012	89
$N_{\text{Level-2}} = 100$													
$N_{\text{Level-1}} = 10$	0.94	0.31	0.164	0.012	0.27	0.25	0.100	0.011	24.16	1.07	0.853	0.013	88
$N_{\text{Level-1}} = 20$	1.13	0.33	0.085	0.011	0.55	0.51	0.049	0.010	11.78	0.82	0.292	0.012	81
$N_{\text{Level-1}} = 30$	1.11	0.27	0.059	0.010	− 0.11	− 0.09	0.033	0.010	9.02	1.75	0.186	0.011	90
$N_{\text{Level-2}} = 200$													
$N_{\text{Level-1}} = 10$	0.19	− 0.34	0.157	0.011	0.15	0.12	0.101	0.010	24.88	1.09	0.962	0.012	52
$N_{\text{Level-1}} = 20$	0.61	− 0.09	0.083	0.010	0.17	0.17	0.050	0.010	11.74	0.99	0.299	0.011	63
$N_{\text{Level-1}} = 30$	0.68	0.01	0.058	0.010	− 0.11	− 0.11	0.034	0.010	8.78	1.88	0.189	0.011	77

The convergence rate informs whether all parameter estimates including individual components ones reached the criterion of $\hat{R} < 1.1$

researchers, our study shows that the PS method perform well and on par with the ML-LVM.

Simulation Study 2

In our second simulation study, we aimed to explore how the conventional PS method and our proposed ML-LVM perform with respect to the detection of moderation effects. To this end, we extended our simulation setup by incorporating rater-level moderators of judgment accuracy.

Simulation Settings and Design Factors

For simulation study 2, we chose a 2 × 3 design with the same fixed design factors and the following conditions: $N_{Level\text{-}2} = 50$ and 200; and $N_{Level\text{-}1} = 10$, 30, and 50. This time, the population-level judgment components were held constant across conditions, with parameters set to rank component = 0.5, level component = 0.2, and differentiation component = 1. These are parameters that frequently occur in empirical research (Machts et al., 2016; Südkamp et al., 2012; Urhahne & Wijnia, 2021). Furthermore, we simulated a predictor variables for each dataset. As a random design factor, we sampled for each dataset which judgment component was moderated. When no effect was present, the moderation effect was set to 0. When a moderation effect was present, we sampled whether the effect was negative or positive. The *SD*s of the random effects were again set to 0.1 for all components. The size of the moderation effect was set to a large effect (± 0.5) (see, e.g., Arend & Schäfer, 2019).

We simulated 500 datasets for each condition. To examine the moderation effects (1) with the PS method, we first calculated each judgment accuracy component for each individual (see Eqs. 9–11). In doing so, we received three vectors of rater-specific components. Second, we regressed each of these three vectors on the two moderators using the linear model (lm-function) from base R. In doing so, we ran a separate model for each outcome component. To examine the moderation effects (2) with the proposed ML-LVM, we used MCMC estimation with 3 chains and 5000 iterations per chain.

Evaluation Metrics

Again, we evaluated bias and RMSE as introduced in simulation study 1. Bias was calculated in relation to the average absolute moderation effect when present (0.5). Furthermore, we studied rates of true positives (i.e., power) and false positives (i.e., type 1 error) for the moderation effects. While the false-positive rate is typically 5%, we considered a true-positive rate of 80% as sufficient (Muthén & Muthén, 2002). We used *p*-values provided by the lm function from base R to determine the significance of a given moderation effect on the $\alpha = 0.05$ level. For our ML-LVM model, we relied on the 95% credible intervals provided by Stan.

Results

Table 3 depicts bias and RMSE of the moderation effects under study. Biases turned out to be consistently below one percent for both approaches and thus negligible.

The RMSEs indicated smaller bias for the moderation effects of all components with growing sample sizes. Small advantages were present for the ML-LVM approach considering the level and differentiation component, yielding smaller RMSEs. The advantages of the ML-LVM over the PS method tended to be smaller when the level-1 sample size was high.

Figure 2 presents the overall and component-specific true-positive and false-positive (i.e., type 1 error) rates. In the low level-2 sample size condition ($N_{\text{Level-2}} = 50$), true-positive rates were consistently below the desired threshold of 80% for all components and both approaches. For the $N_{\text{Level-2}} = 200$, $N_{\text{Level-1}} = 10$ condition, only the true-positive rate for the rank component using the ML-LVM approach was above 80%. For the two high sample size conditions ($N_{\text{Level-2}} = 200$, $N_{\text{Level-1}} = 10$ and $N_{\text{Level-2}} = 200$, $N_{\text{Level-1}} = 50$), both approaches reached the desired overall true positive of at least 80%. However, the detection rate of the moderation effects for the differentiation component was still below this threshold for both approaches and conditions. Another finding is that the desired power of 80% is reached for the rank components with lower sample size conditions than for the level and differentiation components—with advantages for the ML-LVM approach. Uncovering moderation effects for the differentiation component is most demanding, missing the threshold of 80% for all conditions analyzed in the present simulation study but should be reached with lower sample sizes for the ML-LVM than the PS method.

Comparing the PS method and the ML-LVM, consistent performance advantages in terms of true-positive rates occurred for the ML-LVM approach across conditions and components except in the first condition.

Discussion Simulation Study 2

The results of simulation study 2 revealed that neither our proposed ML-LVM approach nor the PS two-step approach systematically underestimated or overestimated moderation effects. The RMSEs tended to be smaller for the ML-LVM approach, especially when the level-1 sample size was low promising more accurate point estimates of moderation effects.

Regarding the true-positive rates, findings showed slight advantages of the ML-LVM approach except in the smallest condition. The power to uncover moderation effects was consistently higher across conditions and components. However, results showed that the low level-2 sample size conditions were insufficient to reliably detect moderation effects for both approaches. For medium sample size conditions, the ML-LVM approach should be preferred over the PS approach. Comparing the different sample size conditions on both levels, the compensating effect of level-1 and level-2 sample size as known from other multilevel models, such as longitudinal modeling (see, e.g., Hecht & Zitzmann, 2021), was apparent. In the two highest sample size conditions, both approaches performed well, with true-positive rates for moderation effects over 80% across components. However, this threshold was still

Table 3 Parameter bias of moderation effects

	Rank component				Level component				Differentiation component				
	Bias (%)		RMSE		Bias (%)		RMSE		Bias (%)		RMSE		
Sample size	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	Convergence (%)
$N_{\text{Level-2}} = 50$													
$N_{\text{Level-1}} = 10$	− 0.56	− 0.56	0.041	0.037	0.09	0.08	0.048	0.047	0.49	0.15	0.128	0.074	100
$N_{\text{Level-1}} = 30$	0.04	− 0.01	0.026	0.025	− 0.19	− 0.17	0.032	0.031	− 0.04	− 0.09	0.055	0.049	99
$N_{\text{Level-1}} = 50$	0.09	0.10	0.022	0.021	− 0.02	− 0.00	0.024	0.024	− 0.31	− 0.23	0.039	0.035	98
$N_{\text{Level-2}} = 200$													
$N_{\text{Level-1}} = 10$	− 0.12	− 0.08	0.020	0.019	− 0.17	− 0.15	0.024	0.024	0.09	0.14	0.064	0.040	99
$N_{\text{Level-1}} = 30$	− 0.07	− 0.04	0.011	0.011	− 0.09	− 0.08	0.015	0.014	− 0.09	− 0.05	0.026	0.023	99
$N_{\text{Level-1}} = 50$	0.05	0.06	0.011	0.011	− 0.05	− 0.05	0.012	0.012	− 0.06	− 0.06	0.020	0.019	100

The convergence rate informs whether all estimates for the population parameters including moderation effects reached the criterion of $\widehat{R} < 1.1$

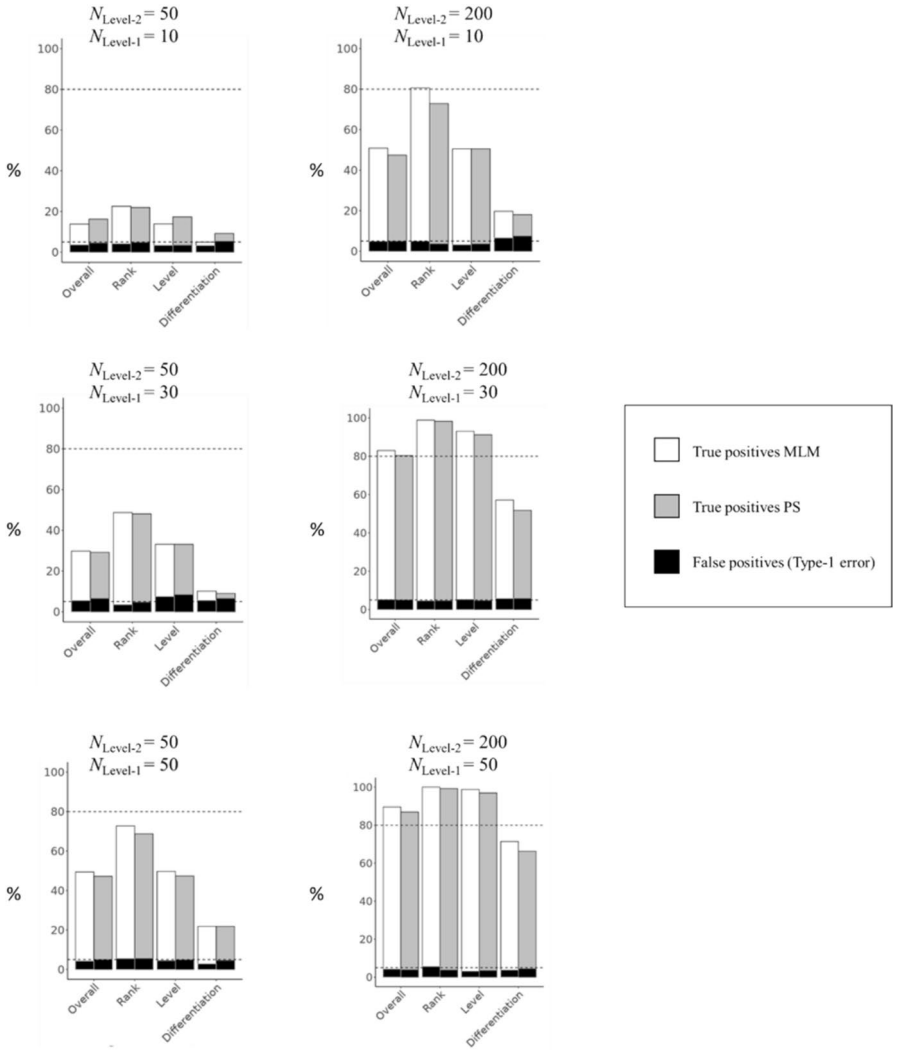

Fig. 2 True-positive and false-positive rates of the ML-LVM and the PS approaches

not reached for the differentiation component but advantages for the ML-LVM were present.

Simulation Study 3

In the third simulation study, we aimed to explore the ML-LVM's potential to relate judgment accuracy to learning outcome measures.

Simulation Settings and Design Factors

The data-generating model was again the ML-LVM so that there was a match between the analysis and the data-generating models. This time, we added three outcome variables to each simulated dataset. These outcome variables were simulated in such a way that each judgment accuracy component was a substantial predictor for one of these outcomes with $\beta = 0.5$ (i.e., representing a strong effect). We tested the following four sample size conditions: ($N_{Level\text{-}2} = 50$, $N_{Level\text{-}1} = 50$), ($N_{Level\text{-}2} = 200$, $N_{Level\text{-}1} = 10$), ($N_{Level\text{-}2} = 200$, $N_{Level\text{-}1} = 30$), ($N_{Level\text{-}2} = 200$, $N_{Level\text{-}1} = 50$).

Evaluation Metrics

As in the previous simulation studies, we analyzed parameter bias and RMSE. In addition, we also explored the coverage of the predictor effects—the proportion of replications in which the 95% confidence interval contained the true regression weight. We included coverage as an additional evaluation criterion here because the two-step PS method does not account for the uncertainty of the individual judgment accuracy scores when calculating regression weights in the outcome regression. Consequently, we anticipated the ML-LVM would offer a particular advantage in terms of coverage. In line with the recommendation by Muthen and Muthen (2000), coverage values between 91 and 98% were considered acceptable.

Results

Table 4 presents the parameter bias and the RMSE for the regression weights relating the judgment accuracy components and the outcome variables for the four tested sample size conditions. Similar to the results from simulation studies 1 and 2, bias and RMSE were smaller when level-1 and level-2 sample size were higher. Furthermore, bias and RMSE were considerably higher for the regression weights quantifying the effect of the differentiation component on the outcome variable.

Comparing the ML-LVM and the two-step PS method, smaller parameter bias was present for the ML-LVM approach for almost all conditions and components. The bias exceeded the threshold of $\pm 10\%$ except for the ML-LVM method and the regression weights of the rank component (all sample size conditions) and the level component (the two largest sample size conditions). For the two-step PS method, the parameter bias was consistently above this threshold and negative, implying that the regression weights were strongly underestimated.

Comparing both methods in terms of the RMSE, small differences in favor of the ML-LVM (i.e., smaller RMSE) were present except for the smallest sample size condition.

Figure 3 shows the coverage rates of the regression weights for the four simulation conditions. As assumed, the two-step PS method showed consistently smaller coverage rates indicating that the true values of the simulated regression weights were less frequently included in the model-implied 95% confidence intervals than for the ML-LVM.

Table 4 Parameter biasand RMSEof regression weights relating judgment accuracy components and outcome variables

	Rank component				Level component				Differentiation component				
	Bias (%)		RMSE		Bias (%)		RMSE		Bias (%)		RMSE		
Sample size $N_{\text{Level-2}}/N_{\text{Level-1}}$	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	PS	ML-LVM	Convergence (%)
50/50	− 11.11	− 2.39	0.16	0.17	− 14.07	− 10.87	0.17	0.19	− 18.97	− 21.53	0.22	0.24	100
200/10	− 22.22	− 11.17	0.20	0.16	− 24.12	− 21.16	0.22	0.21	− 29.99	− 26.43	0.26	0.25	95
200/30	− 13.07	1.81	0.14	0.13	− 16.92	− 2.65	0.17	0.15	− 23.83	− 17.27	0.22	0.22	98
200/50	− 10.97	0.91	0.12	0.11	− 14.90	2.44	0.14	0.13	− 22.27	− 11.54	0.20	0.19	99

The convergence rate informs whether all estimates for the population parameters including moderation effects reached the criterion of $\hat{R}$< 1.1

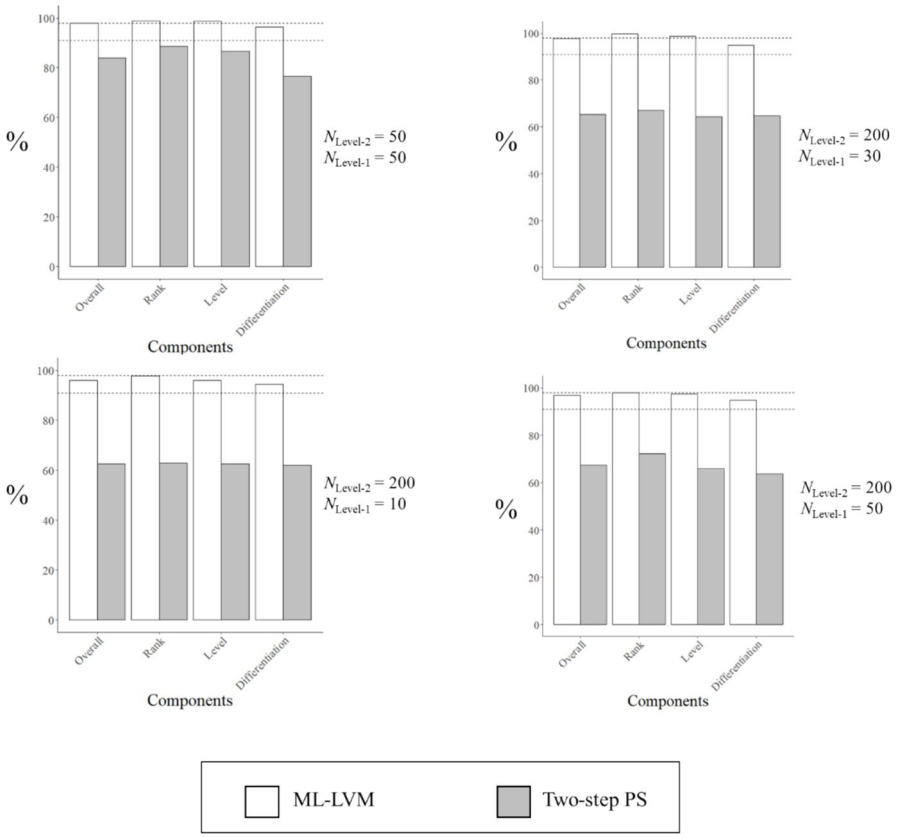

Fig. 3 Coverage rates of estimated regression weights

Interestingly, the differences between the ML-LVM and the PS method were less pronounced when the level-2 sample size was small. In addition, coverage rates were almost constant across different level-1 sample sizes for the PS method while they increased with level-1 sample size for the ML-LVM. Here, the disadvantage of the two-stage approach of the PS method becomes apparent. The uncertainty concerning the calculation of individual-level components is not taken into account in the second step (regression model) of the PS method, while the one-step approach of the ML-LVM considers this uncertainty.

Concerning the ML-LVM approach, Fig. 3 shows that coverage was in the desired range for all components when level-1 sample size was 50. However, when level-2 sample size was small (i.e., 50) there was even a tendency for over-coverage indicated by very broad confidence intervals.

Discussion Simulation Study 3

Simulation study demonstrated several advantages of the ML-LVM over the PS method in examining outcome regression where individual judgment accuracy components act as predictors of outcome variables. The ML-LVM approach exhibited less biased parameter estimates and confidence intervals that more often covered the simulated true regression weights.

Empirical Application 1

In this section, we provide an empirical application illustrating the functionality of our proposed model and comparing it with the rater- and component-specific calculation (i.e., the PS method). The substantive goals of the presented study were (1) to measure the judgment accuracy of pre-service Chemistry teachers (i.e., teacher students) and (2) to examine the influence of teacher characteristics as potential moderators of judgment accuracy components. The data was collected in 2020 at a German university and has never been published before.

For this empirical investigation, the simulated classroom environment was used (Kaiser et al., 2013; Südkamp et al., 2008). In this virtual environment, pre-service teachers interacted with virtual students on a computer by selecting and posing questions, and asking students to respond. The questions and answers—three false answers and one correct answer per question—covering four competence domains of Chemistry had been developed by the university's Chemistry department. One example item is "Describe the material and chemical properties that characterize elements that belong to a main group." The simulation settings allowed for the adaptation of student abilities (i.e., student-specific probabilities of providing correct answers to questions), with ability levels ranging from 0 to 100.

Fig. 4 The simulated classroom environment. *Note*. For data protection reasons, AI-generated student pictures were used

The classroom consisted of 12 eighth-graders, each represented with a name and a picture (Fig. 4). When the teacher asked a question, the students virtually raised their hands to respond. Participants were not provided with the correct answers but were required to assess the correctness of the given responses themselves. The person-specific ability level was randomly sampled for each student and participating teacher. Therefore, the true ability level (i.e., the benchmarks) varied across participants.

Each participants had 30 min to interact with the students in the virtual classroom. After the virtual lesson, the participants were asked to rate the students' ability levels on a continuous scale ranging from 0 to 100.

Research Questions

The scope of this empirical application is to illustrate the application of the proposed ML-LVM to empirical data and to compare the results to the traditional PS-method for calculating the three judgment accuracy components. In Appendix 1, we also present a step-by-step tutorial on how to apply the ML-LVM in R. In the following, we present two research objectives that guided the analyses.

1. **Judgment Accuracy of Pre-service Chemistry Teachers**: Our first research goal was to determine the judgment accuracy of the pre-service teachers with respect to the rank, level, and differentiation components.
2. **Moderation Effects**: Next, we aimed to test potential moderation effects. We chose participants' status of education (Bachelor student = 0, Master student = 1) as a dichotomous predictor. Additionally, we tested whether the number of questions asked during the virtual class moderated judgment accuracy. This research goal was exploratory. However, if any moderation effects were present, we expected higher values to be associated with more accurate judgments.[3]

Method

Sample

The sample consisted of $N = 53$ pre-service Chemistry teachers (female = 58.5%, male = 41.5%). The mean age was $M = 24.8$ ($SD = 2.2$) and 74% were pursuing a Master's degree and 36% were pursuing a Bachelor's degree. All participants were preparing to teach at upper track secondary schools upon completing their university education.

[3] In this context, it is essential to mention that positive moderation effects necessarily imply more accurate judgments only for the rank component. For the level and differentiation component, the substantive interpretation of moderation effects depends on the components' baseline (population) measures. The optimal values of the level and differentiation component are 0 and 1, respectively. Individual components closer to these values imply higher rater-specific judgment accuracy.

Analyses

To address our research objectives, we applied our newly developed ML-LVM. We ran a separate model for each research objective. Model 1 aimed to derive the judgment accuracy components. Model 2 included the two teacher characteristics *education status* (Bachelor or Master student) and *number of questions raised* as two potential moderators of judgment accuracy. Finally, Model 3 analyzed the influence of the simulation conditions manipulating the class-average performance level. Two dummy variables indicated whether participants were in the low, middle, or high average class achievement conditions.

We used Stan's No-U-Turn Sampler (NUTS) with 50,000 iterations and 3 chains. Convergence was assessed using the parameter-specific $\widehat{R}$, which should be close to 1 and ideally less than 1.01 for all parameter estimates, and the effective sample size, which should be above 1000 for all parameter estimates (e.g., Gelman et al., 2013; Vehtari et al., 2021; Zitzmann & Hecht, 2019; Zitzmann et al., 2023). Normal distributions served as priors for the population means of the judgment components, half-Cauchy priors for the population-level variances, and the Lewandowski–Kurowicka–Joe prior (Lewandowski et al., 2009) for the covariances matrices.

To compare the judgment accuracy components as obtained from our ML-LVM with the PS method, we additionally calculated the three judgment accuracy components relying on the following equations (Urhahne & Wijnia, 2021):

$$\text{rank component}_j = cor(\boldsymbol{x}_j, \boldsymbol{y}_j) \tag{10}$$

$$\text{level component}_j = mean\left(\boldsymbol{y}_j\right) - mean(\boldsymbol{x}_j) \tag{11}$$

$$\text{differentiation component}_j = \frac{var\left(\boldsymbol{x}_j\right)}{var(\boldsymbol{y}_j)} \tag{12}$$

Here, $\boldsymbol{y}_j$ and $\boldsymbol{x}_j$ are the individual judgments and corresponding benchmarks, respectively. Students' ability levels implied by the simulation and judged by the participants ranged from 0 to 100%. We transformed the judgments and benchmarks to the range of 0 to 1 before the analyses.

To analyze the moderation effects with the PS method, we used the typical two-step procedure. In the first step, we calculated the individual components as described in Eqs. (9) to (10). In the second step, we regressed these individual components on the moderators of interest. In doing so, we run component-specific linear regression models with the lm function from base R. The continuous predictor involved in RQ 2 (number of questions raised) was z-standardized.

For statistical inference, we used a significance level of 5% (two-sided) and corresponding 95% confidence intervals. The analysis code and the dataset is available on this OSF repository (https://osf.io/ydfaz/?view_only=fbaee809850f4a048bb6531a945fe328, *anonymized for review*).

Results

Judgment Accuracy of Pre-service Chemistry Teachers

The parameter estimates of our ML-LVM and the PS method are presented in Table 5. For the level and rank components, both approaches yield corresponding results. The rank component of 0.60 matches findings from the meta-analyses of Südkamp and colleagues (2012), indicating that the pre-service teachers possessed a high diagnostic competence regarding relative judgment accuracy. Unlike the rank component, the level component depends on the metric of the teacher judgments and benchmarks. In our case, the scale ranged from 0 to 1 (transformed before the analyses as described above). Importantly, the level component was not statistically significantly different from 0 in both approaches, suggesting no general tendency among the pre-service teachers to underestimate or overestimate student achievement. This result implies high judgment accuracy with respect to the level component as well. Finally, for the differentiation component, both approaches implied considerable underestimation. As introduced above, values below 1 indicate an underestimation of achievement differences across students. The confidence interval for both approaches indicated that the differentiation component is significantly different from 1, suggesting that the participants perceived the students as more homogeneous than they actually were. However, while the differentiation component was 0.80 using the ML-LVM, it was 0.83 with the PS method—thus slightly higher using the PS method.

By default, the ML-LVM provides additional information on (1) the benchmarks (i.e., grand mean, *SD* of the individual deviations from the mean, variance on the individual level, and *SD* of the individual deviations from the individual variance), (2) the *SD* of the individual judgment components, and (3) correlations between the individual judgment components. To compare the additional information, that is, the *SD*s of the individual components and the correlations of the individual components, we further analyzed the individual scores derived from the PS method (Eqs. 9–11). This, however, requires a two-step procedure associated with the shortcomings discussed in "Several Extensions for the ML-LVM" section. Therefore, we anticipated differences in *SD*s of the judgment accuracy components and *SE*s of the correlations because the PS method (1) does not consider sampling error and (2) the individual judgment accuracy components are unreliable. Therefore, we anticipated differences in *SD*s of the judgment accuracy components and *SE*s of the correlations because the PS method (1) does not consider sampling error and (2) the individual components are uncertain.

Results revealed that the *SD* of the random effect implied by the ML-LVM were much smaller than for the PS method, illustrating a strong effect of the error correction. The reverse effect was present for the *SE* of the correlation coefficients, where much higher *SE* and confidence intervals yielded by the ML-LVM estimates reflect the consideration of the uncertainty in the random effects, while the two-step approach assumes the point estimates of the individual parameters as the true components (measured without error).

Table 5 Parameter estimates of the PS method and the multilevel latent variable model analyzing data of pre-service Chemistry teachers

	PS method				ML-LVM				Corresponding ML-LVM parameters[1]
			95% CI				95% CI		
	Mean	*SE*	LL	UL	Mean	*SE*	LL	UL	
Fixed effects									
Level component	0.03	0.02	− 0.01	0.06	0.03	0.02	− 0.01	0.06	Level component
Differentiation component	0.83	0.05	0.73	0.94	0.80	0.06	0.70	0.92	Diff. component
Rank component	0.60	0.03	0.53	0.66	0.60	0.03	0.54	0.65	Rank component
Mean benchmark	–	–	–	–	0.48	0.02	0.44	0.52	$\mu^{(x)}$
Within-level variance benchmark	–	–	–	–	0.07	0.00	0.06	0.08	$\sqrt{\omega^{2(x)}}$
Variance components									
SD random effects benchmark	–	–	–	–	0.14	0.02	0.11	0.19	$\sqrt{\omega^{2(\mu^{(x)})}}$
SD random effects within-level variance benchmark	–	–	–	–	0.05	0.04	0.01	0.14	$\sqrt{\omega^{2(\sigma^{2(x)})}}$
SD random effects level component	0.12	–	–	–	0.10	0.02	0.07	0.14	$\sqrt{\omega^{2(l)}}$
SD random effects differentiation component	0.37	–	–	–	0.14	0.06	0.04	0.28	$\sqrt{\omega^{2(d)}}$
SD random effects rank component	0.23	–	–	–	0.10	0.04	0.01	0.18	$\sqrt{\omega^{2(r)}}$
Correlations									
Level-differentiation	− 0.32	0.13	− 0.55	− 0.06	− 0.26	0.30	− 0.77	0.37	$\omega^{(ld)}$
Level-rank	− 0.09	0.14	− 0.35	0.18	− 0.17	0.27	− 0.66	0.38	$\omega^{(lr)}$
Differentiation-rank	0.22	0.14	− 0.06	0.46	0.23	0.33	− 0.49	0.76	$\omega^{(dr)}$

CI = confidence interval. *LL* = lower limit; *UL* = upper limit. [1]Model parameters as introduced in Eqs. (1)–(7)

The *SD*s of the individual components ($\sqrt{\omega^{2(l)}}$, $\sqrt{\omega^{2(d)}}$, $\sqrt{\omega^{2(r)}}$) considerably deviated for the differentiation ($SD_{PS} = 0.37$; $SD_{ML-LVM} = 0.14$) and the rank components ($SD_{PS} = 0.23$; $SD_{ML-LVM} = 0.10$) while being close for the level component ($SD_{PS} = 0.12$; $SD_{ML-LVM} = 0.10$). The *SD*s were consistently higher for the PS method compared to the ML-LVM.

The correlations of the individual-level components and differentiation components, as well as for the individual level components and rank components, were negative for both approaches. The correlation of the individual rank components and differentiation components was positive for both approaches. However, the *SE*s were considerably higher for the ML-LVM. This also resulted in the correlation between differentiation and level components being statistically significantly different from zero using the PS method but not significant using the ML-LVM. The other two correlations were not significant in either of the approaches.

Teacher Characteristics as Moderators of Judgment Accuracy

Our second substantive research objective was to explore the influence of the level of academic degree and teaching behavior on judgment accuracy. More specifically, we analyzed (1) whether Bachelor and Master students differed in their judgment accuracy and (2) whether judgment accuracy was moderated by the number of questions asked by the pre-service teachers during class. Table 6 presents the parameter estimates of the moderation effects (Model 1; for all parameters of the ML-LVM, see Appendix 2 Table A1).

Table 6 Moderations of judgment accuracy components

	PS method				ML-LVM			
			95% CI				95% CI	
Model 1	Mean	*SE*	LL	UL	Mean	*SE*	LL	UL
Components (fixed effects)								
Level component	0.03	0.02	− 0.01	0.07	0.03	0.02	− 0.01	0.07
Differentiation component	0.88	0.06	0.76	0.99	0.83	0.02	0.71	0.97
Rank component	0.60	0.03	0.53	0.67	0.60	0.03	0.52	0.66
Moderation effects								
Outcome: rank								
Status (BA = 0, MA = 1)	− 0.00	0.04	− 0.08	0.07	0.01	0.03	− 0.06	0.08
Number of questions	0.03	0.03	− 0.04	0.09	0.01	0.03	− 0.05	0.07
Outcome: level								
Status (BA = 0, MA = 1)	− 0.01	0.02	− 0.05	0.03	− 0.01	0.02	− 0.05	0.03
Number of questions	0.01	0.02	− 0.03	0.04	0.01	0.02	− 0.03	0.04
Outcome: differentiation								
Status (BA = 0, MA = 1)	− 0.09	0.06	− 0.21	0.02	− 0.06	0.08	− 0.21	0.10
Number of questions	0.07	0.05	− 0.03	0.18	0.08	0.07	− 0.05	0.21

CI = confidence interval; *LL* = lower limit; *UL* = upper limit

Our analysis revealed that none of the moderation effects were statistically significantly different from 0 for both approaches. Thus, the level of academic degree nor teaching behavior, in terms of the number of questions raised during class, moderated judgment accuracy.

Discussion Empirical Application 1

As was expected, we found that the population-level estimates of judgment accuracy components between our ML-LVM approach and the PS method corresponded. A small difference was only present concerning the differentiation component, but the difference did not change substantive conclusions.

Furthermore, both approaches revealed no significant moderation effects for teacher characteristics. However, significant moderations of the level component were detected with both approaches when studying grading on the curve effects. The effect of higher class-average achievement on individuals' judgment accuracy in terms of the level component was negative, implying contrast effects. Thus, teachers tended to underestimate students in high achieving classes and tended to overestimate student achievement in low achieving classes.

Comparing the PS method and the ML-LVM in more detail, we recognized that *SD*s of the individual judgment accuracy components were considerably higher using the PS method compared to the ML-LVM. We hypothesized that these differences were due to the sampling error. While the ML-LVM adjusts for sampling error, this is not the case with the PS method. One goal of simulation study 1 is to test this hypothesis.

Furthermore, the *SE*s for some moderation effects and correlations of individual parameters, particularly those involving the ML-LVM method and the differentiation component, were relatively high, resulting in broad confidence intervals. Although our sample size conditions broadly matched many previous empirical studies (see, e.g., Südkamp et al., 2012; Urhahne & Wijnia, 2021), we speculated that these issues could be due to the relatively small sample size ($N_{\text{Level-1}} = 53$, $N_{\text{Level-1}} = 12$) and thus insufficient statistical power. One goal of simulation study 2 is to shed light on this speculation.

Empirical Application 2—A Reanalysis of Förster et al. (2022) with the ML-LVM

To incorporate real student data and demonstrate how the ML-LVM can be used to relate judgment accuracy to learning outcomes, we reanalyzed data from Förster et al. (2022). The original study investigated how teachers' initial judgments of their students' reading fluency and comprehension—collected at the beginning of the school year—affected the students' subsequent learning progress. The authors tested multiple hypotheses, including the accuracy hypothesis, the overestimation hypothesis, and the Matthew effect. In our reanalysis, we employ our ML-LVM to obtain the three judgment accuracy components and their predictive relationships with students' learning progress.

Research Question

1. **Teacher Judgment Accuracy**: How accurate are teachers in judging students reading fluency and reading comprehension?
2. **Predictive Effects:** Are teacher judgment accuracy components predictive for students' learning gains in reading fluency and reading comprehension?

Method

Sample

In Förster et al.'s (2022) study, data from three longitudinal investigations conducted in Germany examining reading fluency and comprehension among third- and fourth-grade students were analyzed integratively. All three studies administered standardized tests and collected teacher ratings at the beginning of the school year. In addition, eight brief computer-based reading assessments were administered at 3-week intervals to track students' learning progress throughout the year. After data cleaning and applying exclusion criteria, the final sample comprised 2880 students and 145 teachers (see Förster et al., 2022 for details).

Measures

At the start of each study, reading fluency and comprehension were measured using two standardized tests. Fluency was assessed with the Salzburger Reading Screening (SLS) for Grades 1–4. Comprehension was evaluated with a short version of the Hamburger reading test (HAMLET-S).

Teachers, unaware of the actual test scores, evaluated each student's reading fluency and comprehension using the same scale as the standardized tests. For fluency, they estimated the number of sentences a student could correctly answer within 3 min on the SLS. For comprehension, they estimated how many questions the student answered correctly on the HAMLET-S.

To measure students' learning progress, students completed eight 10-min computer-based reading tests every 3 weeks throughout the school year. Each test began with a reading fluency task, where every seventh word was masked and replaced with three options. After the task, students answered multiple choice comprehension questions based on the passage with one correct answer and three distractors each. Accuracy and response time were combined into overall efficiency scores, which served as indicators for both reading fluency and comprehension. The tests were similar for third and fourth graders across studies, differing only in the number of items and text content. For more details, see Förster et al. (2022).

Analyses

In the present application, we reanalyzed the data published by Förster et al. on their OSF repository. We integrated the estimation of the learning growth-curve model with the calculation of teachers' judgment accuracy into a single model (one-step approach). In doing so, we added a growth-curve model accounting for between-teacher differences in intercepts and slopes into the ML-LVM described in Section Individual-Level Judgment Accuracy Components as Predictors of Learning Outcomes (the respective stan code can be found on our OSF repository). Because estimating a direct multiple regression between the judgment accuracy components and individual growth parameters led to convergence issues, we instead focused on estimating the correlations between these components and students' growth parameters. We suspect these convergence problems stem from the combination of high model complexity and a relatively small sample size—judgments by only 80 teachers were available for each outcome. In addition, as indicated in simulation study 3, we already anticipated that we would only have a limited power to detect predictive effects with this data, although for applied teacher judgment accuracy research, this is already a rather large sample.

Before analyzing the data, we standardized the achievement test scores. The longitudinal data were grand-mean centered and standardized, and the time variable was scaled so that time point 1 corresponded to zero. Teacher judgments of students' abilities were also standardized using the means and standard deviations from the actual test scores, so benchmarks and teacher judgments were still on the same metric after the transformation (see, e.g., Karst et al., 2017). Following the procedure of Förster et al. (2022), we estimated separate models for reading fluency and reading comprehension.

Results

Table 7 presents the population-level judgment accuracy components, the parameters of the growth-curve model, and the model-implied correlations between judgment accuracy components and learning growth. On average, teachers demonstrated relatively high rank accuracy, with rank components of 0.62 for reading comprehension and 0.71 for reading fluency (see, e.g., Südkamp et al., 2012). The level component indicates that the teachers tended to overestimate their students' abilities, a common pattern when teachers judge their own students (Urhahne & Wijnia, 2021). By contrast, the differentiation component varied across the two tasks: teacher generally underestimated the variability of their students reading comprehension, yet overestimated the variability in students' reading fluency. Finally, the standard deviations of the random effects suggest considerable between-teacher differences in all three judgment accuracy components.

Table 7 Parameter estimates of the ML-LVM to predict learning gains in the data of Förster et al. (2022)

	Comprehension				Fluency			
			95% CI				95% CI	
	B	*SE*	LL	UL	*B*	*SE*	LL	UL
Fixed effects								
Level component	0.32	0.06	0.20	0.44	0.26	0.08	0.11	0.41
Differentiation component	0.71	0.05	0.61	0.81	1.78	0.11	1.57	2.00
Rank component	0.62	0.02	0.58	0.66	0.71	0.01	0.68	0.74
Random effects (*SD*)								
Level component	0.51	0.05	0.42	0.61	0.68	0.06	0.57	0.80
Differentiation component	0.48	0.06	0.38	0.61	0.42	0.05	0.32	0.53
Rank component	0.08	0.03	0.02	0.13	0.04	0.02	0.00	0.09
Growth model								
Intercept	− 0.38	0.08	− 0.53	− 0.23	− 0.31	0.07	− 0.44	− 0.17
Linear growth	0.11	0.01	0.08	0.13	0.09	0.01	0.07	0.10
SD random effects intercept	0.64	0.05	0.55	0.75	0.60	0.05	0.51	0.70
SD random effects growth	0.09	0.01	0.07	0.11	0.07	0.01	0.06	0.09
Correlation with learning growth								
Level component	0.05	0.11	− 0.18	0.27	0.17	0.11	− 0.07	0.38
Differentiation component	− 0.24	0.14	− 0.52	0.04	0.02	0.14	− 0.26	0.28
Rank component	− 0.15	0.19	− 0.51	0.25	− 0.09	0.24	− 0.56	0.41

The estimated linear growth parameters, $B_{Lineargrowthcomprehension} = 0.11$ and $B_{Lineargrowthfluency} = 0.09$, indicate that, on average, students demonstrated a statistically significant increase in reading comprehension and reading fluency over the course of the school year. However, the correlations between the judgment accuracy components and students' learning gains were not statistically significant. Given the relatively small sample sizes for both datasets, the confidence intervals for these correlations were correspondingly wide.

Descriptively, the level component showed a positive (though non-significant) association with learning gains in both datasets, whereas the rank component was negatively (but again non-significantly). For the differentiation component, the pattern was inconsistent: it was negatively correlated with learning gains in reading comprehension and positively correlated with learning gains in reading fluency. None of these correlations were statistically significantly different from zero.

Discussion Empirical Application 2

Our reanalysis of the Förster et al. (2022) dataset yielded results consistent with the authors' original findings, showing that neither relative judgment accuracy (rank component) nor overestimation (level component) was significantly associated with

students' learning progress. The standard deviations of the random effects for the rank component were notably small, indicating that, for example, about 94% of the teachers had rank component values ranging between 0.62 and 0.79 for reading fluency. This small range reflects relatively little variation in teachers' rank accuracy for these data.

Our previous simulation study also suggests that the sample size in the present dataset may still be too small to reliably estimate and detect the targeted predictive effects. Furthermore, the estimated model is highly complex, which presents additional challenges when applied to a relatively small dataset (see simulation study 3). As an alternative, one could adopt a two-step approach that accounts for the uncertainty in judgment accuracy estimation by using plausible values drawn from the posterior. However, this strategy necessitates re-estimating the target models multiple times and applying pooling rules for final inference—a potentially demanding implementation process.

General Discussion

In the present article, we introduced the ML-LVM approach for analyzing judgment accuracy as operationalized via the rank, level, and differentiation component (Schrader & Helmke, 1987; Urhahne & Wijnia, 2021). Although these components are the most commonly used indicators in research on teacher judgment, our study is the first to integrate them within a single, comprehensive framework. By conducting three simulation studies, we assessed the performance of this new approach and tested its potential advantages relative to the commonly used component- and person-wise calculation (the PS method). Additionally, in two empirical applications, we illustrated how the ML-LVM can address substantive questions in teacher judgment accuracy research, offering a more comprehensive and potentially more robust alternative to existing methods.

The ML-LVM offers several potential advantages over the PS method. First, for the first time, all three components of judgment accuracy can be analyzed with one single model. This makes focusing on a single component—as often done in empirical studies—less likely (see, e.g., Urhahne & Wijnia, 2021). Second, the proposed approach models the population and individual levels simultaneously, providing information about average judgment accuracy across individuals and the amount of between-person differences in judgment accuracy. Third, our model provides more accurate individual-level estimates of judgment accuracy components, particularly under small sample size settings often present in empirical studies (simulation study 1). Fourth, our model provides more accurate standard errors and confidence intervals of potential moderators of the rank, level, and differentiation component (simulation study 2). Again, these advantages were most prominent under small and medium sample size conditions. Fifth, the ML-LVM provides less biased parameter estimates and higher coverage of regression weights relating judgment accuracy and learning outcomes (simulation study 3).

Practical Implications

Our study yields practical implication in at least two respects: for researchers conducting judgment accuracy research and for the diagnosis and training of pre- and in-service teachers' judgment accuracy.

First, our study has demonstrated the advantages of the proposed ML-LVM approach over the traditional PS method in terms of modeling and assessing judgment accuracy. The possibility of always modeling all three components of judgment accuracy in one model can help to overcome a one-sided focus on individual components in empirical studies, which have often concentrated on either the level or rank components (Urhahne & Wijnia, 2021). Additionally, the proposed method considers and provides information on other important aspects of teacher judgment accuracy research, such as quantifying the amount of interindividual differences in judgment accuracy across teachers (see, e.g., Bonefeld et al., 2020; Mack et al., 2023) and the interrelation of the different judgment accuracy components (Leuders et al., 2018). In addition, our study has highlighted that the ML-LVM is a suitable alternative to two-step approaches often employed for calculating moderation or predictive effects of judgment accuracy. These two-step approaches typically ignore the uncertainty of individual-level components calculated in the first step.

Another insight from our simulation studies concerns the sample sizes commonly used in empirical research (see, e.g., Urhahne & Wijnia, 2021), which often may be insufficient for reliably estimating or detecting moderation or predictive effects. Additionally, the simulation results suggest that the sample size requirements for obtaining accurate population- and individual-level judgment accuracy components—as well as for capturing moderation and predictive effects—differ across the three components of judgment accuracy. For instance, for using the ML-LVM simulation study 2 implies that a sufficient power is achieved with the sample size combination $N_{\text{Level-2}} = 200$ and $N_{\text{Level-1}} = 10$. However, this sample size was not sufficient to achieve adequate power for moderation effects of the level and differentiation component. The highest sample size requirements are associated with research questions concerning the differentiation component. For the moderation effect, even the highest sample size condition tested was not sufficient to reach a power of 0.80. Our findings, thus, also provide guidance for researchers in this respect. Based on their research questions, researchers can decide whether they are mainly interested in the rank component—then a smaller sample size might be sufficient—or whether all three components are relevant—then the sample size has to be adapted accordingly.

Another current challenge of teacher judgment accuracy research is the still lacking evidence for the importance of teacher judgment accuracy for student learning (Kolovou et al., 2024). Our proposed approach also yields methodical advantages over two-step approaches based on the PS method where judgment accuracy components act as predictors of students' learning outcomes. Simulation study 3 showed that parameter estimates of these effects are less biased and coverage is considerably higher when using the ML-LVM. These advantages should be further explored in future studies.

Second, recent efforts have been made to train teachers' assessment competencies and judgment accuracy (e.g., Keller et al., 2019; Thiede et al., 2018; C. Zhu

& Urhahne, 2018). This requires an accurate assessment of judgment accuracy at the individual level to determine training demands, provide feedback, and quantify improvements. Our study shows that the PS method falls short here. The application of ML-LVM seems thus promising and recommendable also for diagnostic purposes in teacher education.

Limitations and Future Directions

While our study provides valuable first insights into the properties and advantages of the proposed ML-LVM, several limitations must be acknowledged. These limitations also open avenues for future research to further refine and expand upon our findings.

First, despite the conceptual advantages of the ML-LVM, the observed advantages under some of the simulated conditions were relatively small to moderate at best. Nevertheless, the ML-LVM may encourage a more comprehensive examination of judgment accuracy—consistent with calls in the literature (see Urhahne & Wijnia, 2021) to report all relevant components in studies of teacher judgment accuracy.

Second, we have compared our newly developed model only to the widely applied PS method. However, other approaches such as hierarchical regression or structural equation models have been proposed and applied in judgment accuracy research (Bonefeld et al., 2020; Kolovou et al., 2021; Mack et al., 2023). Future research could compare these methods and our ML-LVM theoretically and regarding performance and capabilities. While regression models typically only consider the rank component (i.e., relative judgment accuracy; but see Karst et al., 2017 for simultaneously modeling the level component), our proposed approach considers all three components alongside interindividual differences of these components. Implicit assumptions of these alternative approaches, such as homoscedasticity or no interindividual differences in the differentiation component, might also affect the estimation of the other judgment accuracy component. In addition, the regression approach has also been employed to operationalize other targets of judgment accuracy research, such as halo effects (Schmidt et al., 2023; Vögelin et al., 2019) or reference-group effects. Future studies could explore how halo effects and reference-group effects can be represented using the ML-LVM.

Third, the proposed model includes teacher-level moderators. However, practical investigations have also been concerned with student-level moderators of teacher judgment accuracy (e.g., Bonefeld et al., 2020; Kaiser et al., 2013; Meissel et al., 2017; Südkamp et al., 2012). The proposed approach should therefore be extended to student-level moderators as well in future model development.

Fourth, our proposed approach takes advantages of Bayesian estimation techniques by using MCMC. While these methods come with advantageous features when fitting more complex models (see, e.g., Ulitzsch et al., 2023; Zitzmann et al., 2020; Zyphur & Oswald, 2015), runtime can be high. The runtime in our experiments ranged from a few minutes to several hours depending on the sample size of the respective dataset. When fitting the model to large datasets, runtime can be very

high. Practitioners should therefore start with small number of iterations of the sampler (e.g., 500) per chain to explore whether everything is set up properly. When the first results seem plausible, a model with more iterations can be run for final inference (see also the Stan User's Guide[4] for recommendations).

Fifth, the ML-LVM relies on the assumption that the analyzed variables follow approximately normal distributions. The model is designed for continuous metric data, which may limit its applicability when ratings are categorical or highly skewed. However, in most studies on teacher judgment accuracy, both teacher ratings and student achievement measures are typically collected on metric or quasi-metric scales (e.g., grades) and tend to approximate normal distributions. Thus, while this assumption may limit the model's flexibility for some data types, it should be met in the majority of applications within this research field (see Urhahne & Wijnia, 2021). Future work could explore potential extensions of the ML-LVM to accommodate categorical or non-normally distributed data, possibly by employing generalized linear modeling.

Conclusion

Our study introduced and evaluated the ML-LVM approach for analyzing judgment accuracy. By conducting empirical applications and simulation studies, we demonstrated the advantages of our approach over the PS method. Key findings indicate that the ML-LVM approach not only provides comprehensive and simultaneous analyses of all three judgment accuracy components, but also yields more accurate rater-level estimates and more accurate standard errors and confidence intervals for moderators of judgment accuracy, particularly under sample sizes typically present in teacher judgment accuracy research. These benefits are crucial for enhancing the robustness and trustworthiness of judgment accuracy research.

Appendix 1 Data analysis tutorial

The first step to replicate the results presented in "Empirical Application 1" section is to download the folder "R_code" from our OSF repository. Then, open the R-script called "ML-LVM Application 1" in R-Studio. The required packages to run our analyses are rstan and dplyr which can be installed and applied via the following lines of code:

```
install.packages("rstan")
install.packages("dplyr")
library(rstan)
library(dplyr)
```

[4] https://mc-stan.org/docs/2_35/stan-users-guide-2_35.pdf

To apply the ML-LVM and its different extensions, all what is needed is to load the functions provided in the file "helperfunctions" included in the folder downloaded from OSF. These functions can be accessed via: source("helperfunctions.Rdata"). An overview of the different configurations to estimate the different models is provided in the excel file "overview JAM function."

The core part of this collection of functions is the JAM function that allows for the following arguments:

```
> Jam1 <- JAM(data=df, # provide your data
+         benchmark="x", # column name in the dataframe containing the benchmarks
+         ratings="y", # column name in the dataframe containing participants' ratings
+         person_identifier = "person", # column name of the cluster id
+         iter=2000, # number of iterations per chain
+         chains=3, # number of chains
+         cores=3, # run chains in parallel to speed up the analyses
+         L1moderators=NULL, # if no L2 moderators are to be analyzed (default = NULL)
+         L2moderators=NULL, # provide column names of L1 moderators varying within raters (default = NULL)
+         outcomes=NULL, # # provide column names of L2 moderators non-varying across raters (default = NULL)
+         same_benchmarks=TRUE, # did participants judge the same targets (i.e., same benchmarks apply to all of them) or not?
+         just_return_stancode=FALSE, # to TRUE if the stan code is requested (otherwise the respective model is estimated
+         indiv_pars=FALSE # whether individual level parameters (ind. deviations from the population means) should be returned
+         )
```

For our application, we import the to be analyzed dataset "Chemistry_teachers" (read.csv("Chemistry_teachers.csv"). The column containing the benchmarks as occurred in the simulated classroom is called "emp.benchmark." The column containing teacher judgments is called "rating." The column "id" identifies different pre-service teachers that participated in this study. To address the first research question we, therefore, specify the following model:

```
> JAM_fitobj1 = JAM(data=df, # provide your data
+               benchmark="emp.benchmark", # column with benchmarks
+               ratings="rating", # column with participants' ratings
+               person_identifier = "id", # column name of the cluster ids
+               same_benchmarks = FALSE)
```

As new student-specific ability values were simulated for each participant, we have to set same_benchmarks to FALSE. For an initial try, we can set the iter argument, which essentially defines the number of iterations done for Bayesian estimation, to a small value (e.g., 1000) to just look whether everything runs fine. For final inferences, iter typically has to be raised to reach common convergence criteria for Bayesian estimation. Such convergence criteria, for instance, require that for all parameters $\widehat{R}$ and N_eff should be below 1.01 and above 1000, respectively (see Zitzmann et al., 2021[5]). Thus, iter has to be adapted accordingly, which can sometimes imply time-consuming attempts in practice.

[5] Note that Zitzmann et al. (2021) recommended using much stricter Bayesian convergence criteria for substantive real-data application while lower convergence criteria can be sufficient for results of simulation studies—as also applied in the simulation studies presented in this article. This issue arises because the evaluation of Bayesian estimation procedures in simulation studies is additionally always a trade-off between sufficient convergence and run-time.

To access the parameter estimates, the function summary.JAM also included in the helperfile can be applied to the fit object: `summary.JAM(JAM_fitobj1)`

The resulting output in the present application looks as follows.

```
> summary.JAM(JAM_fitobj1)
JAM summary$`Population means`
                    mean    sd   2.5% 97.5%  Rhat   n_eff
Benchmark mean         0.477 0.022  0.435 0.518 1.003 272.240
Benchmark SD           0.067 0.004  0.061 0.075 0.998 635.794
Level component        0.026 0.017 -0.007 0.059 1.006 289.424
Differentiation component 0.797 0.054  0.691 0.902 0.997 810.920
Rank component         0.603 0.030  0.541 0.660 1.004 382.201

$`SD of random effects`
                       mean    sd  2.5% 97.5%  Rhat   n_eff
SD Benchmark mean         0.145 0.018 0.113 0.180 1.002 339.667
SD Benchmark SD           0.079 0.051 0.011 0.189 1.371   6.866
SD Level component        0.106 0.014 0.082 0.133 1.007 324.167
SD Differentiation component 0.189 0.096 0.040 0.387 1.183  23.927
SD Rank component         0.108 0.034 0.044 0.175 1.029  80.679

$`Correlations of the random effects`
                          mean    sd   2.5%  97.5%  Rhat   n_eff
Level~~Differentiation       -0.339 0.257 -0.770  0.207 1.018 177.975
Level~~Rank                  -0.097 0.247 -0.589  0.390 1.016 196.803
Rank~~Differentiation         0.183 0.312 -0.452  0.710 1.024 181.910
Benchmark mean~~Level        -0.609 0.108 -0.790 -0.383 0.999 588.552
Benchmark mean~~Rank          0.225 0.257 -0.264  0.674 1.006 203.388
Benchmark mean~~Differentiation -0.049 0.247 -0.497  0.434 1.011 233.052
Benchmark SD~~Level           0.011 0.349 -0.634  0.657 1.008 124.011
Benchmark SD~~Rank            0.011 0.378 -0.717  0.729 1.034 117.451
Benchmark SD~~Differentiation  0.007 0.357 -0.651  0.689 1.000 136.043
Benchmark SD~~Benchmark mean  -0.124 0.324 -0.707  0.514 1.002 160.189
```

The five population means represent the parameters $\mu^{(x)}$, $\sigma_j^{2(x)}$ and the three population-level judgment accuracy components. The five *SD*s of the random effects quantify the amount of between-person variability for these five key model parameters. The interindividual differences can be subject of further analyses, for instance concerning moderators explaining variability or using interindividual differences to predict consequences.

The correlation coefficients of the random effects indicate how strongly subject-specific parameters of the five model parameters are related. The subject specific-level of the model (i.e., empirical Bayes estimates of the person-specific components) can be accessed through setting the argument indiv_pars of the JAM function to TRUE:

```
> JAM_fitobj1 = JAM(data=df, # provide your data
+                   benchmark="emp.benchmark",
+                   ratings="rating",
+                   person_identifier = "id",
+                   iter=5000,
+                   same_benchmarks=FALSE,
+                   indiv_pars=TRUE)
> JAM_fitobj1[[2]][grepl("theta", rownames(JAM_fitobj1[[2]])), c("mean", "sd", "2.5%",
"97.5%")] %>% round(3)
           mean    sd   2.5%  97.5%
theta[1,1]   0.045 0.066 -0.077  0.180
theta[1,2]  -0.054 0.060 -0.167  0.064
theta[1,3]   0.015 0.094 -0.196  0.208
theta[1,4]   0.100 0.193 -0.230  0.604
theta[1,5]  -0.073 0.089 -0.287  0.070
theta[2,1]  -0.156 0.069 -0.291 -0.017
theta[2,2]   0.050 0.054 -0.056  0.158
theta[2,3]  -0.008 0.091 -0.209  0.167
theta[2,4]   0.004 0.176 -0.360  0.408
theta[2,5]   0.018 0.085 -0.156  0.187
theta[3,1]  -0.099 0.063 -0.218  0.019
theta[3,2]   0.032 0.052 -0.069  0.134
theta[3,3]   0.008 0.081 -0.145  0.194
theta[3,4]  -0.013 0.146 -0.300  0.329
theta[3,5]  -0.019 0.084 -0.213  0.136
theta[4,1]  -0.123 0.070 -0.253  0.010
theta[4,2]   0.031 0.054 -0.071  0.132
theta[4,3]   0.022 0.090 -0.149  0.260
(...)
```

Row indices in the theta matrix identify different persons (e.g., teacher IDs). The column IDs indicate model parameters with 1 = benchmark mean, 2 = level component, 3 = benchmark *SD*, 4 = differentiation component, and 5 = rank component. Thus, theta[3,4] represents the differentiation component for teacher 3.

Appendix 2

Table 8 Teacher characteristics moderating judgment accuracy components

			95% Credible interval	
	Mean	*SE*	LL	UL
Fixed effects				
Mean benchmark	0.47	0.03	0.42	0.52
Level component	0.03	0.02	− 0.01	0.07
Differentiation component	0.83	0.07	0.71	0.97
Rank component	0.60	0.03	0.52	0.66
Variance components				
Within-level variance benchmark	0.07	0.00	0.06	0.08
SD random effects benchmark	0.14	0.02	0.11	0.18
SD random effects within-level variance benchmark	0.07	0.05	0.00	0.19
SD random effects level component	0.11	0.01	0.08	0.14
SD random effects differentiation component	0.19	0.10	0.01	0.36
SD random effects rank component	0.10	0.04	0.01	0.18
Correlations				
Level-differentiation	− 0.34	0.29	− 0.79	0.34
Level-rank	− 0.09	0.25	− 0.57	0.42
Differentiation-rank	0.13	0.33	− 0.54	0.7
L2-Moderators	Level component			
Status (BA = 0, MA = 1)	− 0.01	0.02	− 0.05	0.03
Number of questions	0.01	0.02	− 0.03	0.04
	Differentiation component			
Status (BA = 0, MA = 1)	− 0.06	0.08	− 0.21	0.1
Number of questions	0.08	0.07	− 0.05	0.21
	Rank component			
Status (BA = 0, MA = 1)	0.01	0.03	− 0.06	0.08
Number of questions	0.01	0.03	− 0.05	0.07

LL = lower limit, *UL* = upper limit (of the 95% confidence interval)

Author Contribution JL: formal analysis, methodology, software, writing—original draft, writing—review and editing, visualization. JM: conceptualization, supervision, validation, writing—review and editing. SZ: conceptualization, methodology, supervision, validation, writing—review and editing. MH: methodology, software, validation, writing—review and editing. CN: conceptualization, data curation, funding acquisition, investigation, project administration, supervision, writing—review and editing.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was funded by a grant from the German Research Foundation (DFG) to Jens Möller (grant number: MO 648/25–2).

Data Availability Data, analysis code, and materials of the present study are available on OSF (https://osf.io/ydfaz/?view_only=fbaee809850f4a048bb6531a945fe328).

Declarations

Ethics Approval All procedures involving human participants were performed in accordance with the ethical standards of the institutional or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Consent to Participate Participation in the study was voluntary and based on active and informed consent of the participants.

Conflict of Interest The authors declare no competing interests.

References

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods, 24*(1), 1–19. https://doi.org/10.1037/met0000195

Artelt, C., spsampsps Rausch, T. (2014). Accuracy of teacher judgments. In S. Krolak-Schwerdt, S. Glock, spsampsps M. Böhmer (Eds.), *Teachers' professional development* (pp. 27–43). SensePublishers. https://doi.org/10.1007/978-94-6209-536-6_3

Asparouhov, T., & Muthén, B. O. (2012). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. *Annual Meeting of the National Council on Measurement in Education*.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507–526. https://doi.org/10.1037/met0000077

Bonefeld, M., Dickhäuser, O., & Karst, K. (2020). Do preservice teachers' judgments and judgment accuracy depend on students' characteristics? The effect of gender and immigration background. *Social Psychology of Education, 23*(1), 189–216. https://doi.org/10.1007/s11218-019-09533-2

Brandt, H., Chen, S. M., & Bauer, D. J. (2023). *Bayesian penalty methods for evaluating measurement invariance in moderated nonlinear factor analysis*. Advance online publication. https://doi.org/10.1037/met0000552

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Brose, A., Neubauer, A. B., & Schmiedek, F. (2022). Integrating state dynamics and trait change: A tutorial using the example of stress reactivity and change in well-being. *European Journal of Personality, 36*(2), 180–199. https://doi.org/10.1177/08902070211014055

Cronbach, L. J. (1955). Processes affecting scores on understanding of others and assumed similarity. *Psychological Bulletin, 52*(3), 177–193. https://doi.org/10.1037/h0044919

Depaoli, S. (2021). *Bayesian structural equation modeling*. *Methodology in the social sciences*. The Guilford Press.

Dicke, A.-L., Lüdtke, O., Trautwein, U., Nagy, G., & Nagy, N. (2012). Judging students' achievement goal orientations: Are teacher ratings accurate? *Learning and Individual Differences, 22*(6), 844–849. https://doi.org/10.1016/j.lindif.2012.04.004

Feng, Y., & Hancock, G. R. (2024). A structural equation modeling approach for modeling variability as a latent variable. *Psychological Methods, 29*(2), 262–286. https://doi.org/10.1037/met0000477

Förster, N., Humberg, S., Hebbecker, K., Back, M. D., & Souvignier, E. (2022). Should teachers be accurate or (overly) positive? A competitive test of teacher judgment effects on students' reading

progress. *Learning and Instruction, 77*, Article 101519. https://doi.org/10.1016/j.learninstruc.2021.101519

Funder, & D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*(4), 652–670. https://doi.org/10.1037/0033-295x.102.4.652

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. *Chapman and Hall/CRC*. https://doi.org/10.1201/b16018

Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2013). Beyond judgment bias: How students' ethnicity and academic profile consistency influence teachers' tracking judgments. *Social Psychology of Education, 16*(4), 555–573. https://doi.org/10.1007/s11218-013-9227-5

Gnas, J., Mack, E., & Preckel, F. (2022). When classmates influence teacher judgment accuracy of students' cognitive ability: Studying frame-of-reference effects in primary school. *Contemporary Educational Psychology, 69*, Article 102070. https://doi.org/10.1016/j.cedpsych.2022.102070

Hecht, M., & Zitzmann, S. (2021). Sample size recommendations for continuous-time models: Compensating shorter time series with larger numbers of persons and vice versa. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(2), 229–236. https://doi.org/10.1080/10705511.2020.1779069

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., Klug, J., Hetmanek, A., Ohle, A., Böhmer, I., Karing, C., Kaiser, J., & Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education, 76*, 181–193. https://doi.org/10.1016/j.tate.2017.12.001

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297–313. https://doi.org/10.3102/00346543059003297

Jansen, T., Vögelin, C., Machts, N., Keller, S. D., Köller, O., & Möller, J. (2021). Judgment accuracy in experienced versus student teachers: Assessing essays in English as a foreign language. *Teaching and Teacher Education, 97*, Article 103216. https://doi.org/10.1016/j.tate.2020.103216

Jussim, L. (2012). Social perception and social reality. *Oxford University Press*. https://doi.org/10.1093/acprof:oso/9780195366600.001.0001

Jussim, L., & Eccles, J. S. (1992). Teacher expectations: II. Construction and reflection of student achievement. *Journal of Personality and Social Psychology*, *63*(6), 947–961. https://doi.org/10.1037/0022-3514.63.6.947

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28*, 73–84. https://doi.org/10.1016/j.learninstruc.2013.06.001

Kaiser, J., Südkamp, A., & Möller, J. (2017). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology, 109*(6), 871–888. https://doi.org/10.1037/edu0000156

Karst, K., Hartig, J., Kaiser, J., & Lipowsky, & F. (2017). Mehrebenenmodelle als Werkzeuge zur Analyse diagnostischer Kompetenz von Lehrkräften - ein lineares Mischmoell (LMM) und seine Anwendung in R [Multilevel models as a tool for analyzing teachers' diagnostic]. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen* (pp. 153–174). Waxmann.

Kaufmann, E. (2020). How accurately do teachers' judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology, 63*, Article 101902. https://doi.org/10.1016/j.cedpsych.2020.101902

Kaufmann, E., Sjödahl, L., & Mutz, R. (2007). *The idiographic approach in social judgment theory: A review of components of the lens model equation components*.

Keller, S. D., Vögelin, C., Jansen, T., Machts, N., & Möller, J. (2019). Can an instructional video increase the quality of English teachers' assessment of learner essays? *Research in Subject-Matter Teaching and Learning*, *2*(1), 140–161. https://doi.org/10.23770/rt1829

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift Für Psychologie / Journal of Psychology, 216*(2), 61–73. https://doi.org/10.1027/0044-3409.216.2.61

Kolovou, D., Hochweber, J., & Praetorius, A.-K. (2024). Does teacher judgment accuracy matter? How judgment accuracy, teaching quality, and student achievement development are related. *Teaching and Teacher Education, 144*, Article 104555. https://doi.org/10.1016/j.tate.2024.104555

Kolovou, D., Naumann, A., Hochweber, J., & Praetorius, A.-K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education, 100*, Article 103298. https://doi.org/10.1016/j.tate.2021.103298

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. *Springer, New York*. https://doi.org/10.1007/978-1-4614-6849-3

Leuders, T., Dörfler, T., Leuders, J., spsampsps Philipp, K. (2018). Diagnostic competence of mathematics teachers: Unpacking a complex construct. In T. Leuders, K. Philipp, spsampsps J. Leuders (Eds.), *Diagnostic competence of mathematics teachers* (pp. 3–31). Springer International Publishing. https://doi.org/10.1007/978-3-319-66327-2_1

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Liu, S. (2017). Person-specific versus multilevel autoregressive models: Accuracy in parameter estimates at the population and individual levels. *The British Journal of Mathematical and Statistical Psychology, 70*(3), 480–498. https://doi.org/10.1111/bmsp.12096

Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teaching and Teacher Education, 91*, Article 103059. https://doi.org/10.1016/j.tate.2020.103059

Lorenz, C., & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. *Zeitschrift Für Pädagogische Psychologie, 23*(34), 211–222. https://doi.org/10.1024/1010-0652.23.34.211

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

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*(3), 203–229. https://doi.org/10.1037/a0012869

Machts, N., Chernikova, O., Jansen, T., Weidenbusch, M., Fischer, F., & Möller, J. (2024). Categorization of simulated diagnostic situations and the salience of diagnostic information. *Zeitschrift Für Pädagogische Psychologie, 38*(1–2), 3–13. https://doi.org/10.1024/1010-0652/a000364

Machts, N., Kaiser, J., Schmidt, F. T., & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review, 19*, 85–103. https://doi.org/10.1016/j.edurev.2016.06.003

Mack, E., Gnas, J., Vock, M., & Preckel, F. (2023). The domain-specificity of elementary school teachers' judgment accuracy. *Contemporary Educational Psychology, 72*, Article 102142. https://doi.org/10.1016/j.cedpsych.2022.102142

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education, 65*, 48–60. https://doi.org/10.1016/j.tate.2017.02.021

Möller, J., Jansen, T., Fleckenstein, J., Machts, N., Meyer, J., & Reble, R. (2022). Judgment accuracy of German student texts: Do teacher experience and content knowledge matter? *Teaching and Teacher Education, 119*, Article 103879. https://doi.org/10.1016/j.tate.2022.103879

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*(2), 167–190. https://doi.org/10.1007/BF02295939

Schmidt, F. T. C., Kaiser, A., & Retelsdorf, J. (2023). Halo effects in grading: An experimental approach. *Educational Psychology, 43*(2–3), 246–262. https://doi.org/10.1080/01443410.2023.2194593

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung fuer die Gestaltung von Effektivitaet des Unterrichts*. *Europaeische Hochschulschriften. Reihe 6: Psychologie: Vol. 289*. P. Lang.

Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen [Diagnostic competence of teachers]. *Beiträge zur Lehrerbildung*, *31*(2), 154–165. https://doi.org/10.25656/01:13843

Schrader, F.-W., & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen [Diagnostic competence of teachers: Components and effects]. *Empirische Pädagogik, 1*(1), 27–52.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz [Accuracy of teacher judgments of student characteristics and the construct of diagnostic competence]. *Zeitschrift Für Pädagogische Psychologie, 19*(1/2), 85–95. https://doi.org/10.1024/1010-0652.19.12.85

Stan Development Team (2024). Stan Modeling Language Users Guide and Reference Manual, Version 2.35. https://mc-stan.org

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum. *Zeitschrift Für Pädagogische Psychologie, 22*(34), 261–276. https://doi.org/10.1024/1010-0652.22.34.261

Südkamp, A., & Praetorius, A.-K. (2017). *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen* [Diagnostic competence of teachers: Theoretical and methodological developments]. Waxmann.

Thiede, K. W., Brendefur, J. L., Carney, M. B., Champion, J., Turner, L., Stewart, R., & Osguthorpe, R. D. (2018). Improving the accuracy of teachers' judgments of student learning. *Teaching and Teacher Education, 76*, 106–115. https://doi.org/10.1016/j.tate.2018.08.004

Ulitzsch, E., Lüdtke, O., & Robitzsch, A. (2023). Alleviating estimation problems in small sample structural equation modeling-A comparison of constrained maximum likelihood, Bayesian estimation, and fixed reliability approaches. *Psychological Methods, 28*(3), 527–557. https://doi.org/10.1037/met0000435

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review, 32*, Article 100374. https://doi.org/10.1016/j.edurev.2020.100374

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved Rˆ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-ba1221

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing, 39*, 50–63. https://doi.org/10.1016/j.asw.2018.12.003

Weinert, F. E., Schrader, F.-W., & Helmke, A. (1990). Educational expertise. *School Psychology International, 11*(3), 163–180. https://doi.org/10.1177/0143034390113002

Zhu, C., & Urhahne, D. (2018). The use of learner response systems in the classroom enhances teachers' judgment accuracy. *Learning and Instruction, 58*, 255–262. https://doi.org/10.1016/j.learninstruc.2018.07.011

Zhu, M., & Urhahne, D. (2014). Assessing teachers' judgements of students' academic motivation and emotions across two rating methods. *Educational Research and Evaluation, 20*(5), 411–427. https://doi.org/10.1080/13803611.2014.964261

Zhu, M., & Urhahne, D. (2015). Teachers' judgements of students' foreign-language achievement. *European Journal of Psychology of Education, 30*(1), 21–39. https://doi.org/10.1007/s10212-014-0225-6

Zitzmann, S., Bardach, L., Horstmann, K. T., Ziegler, M., & Hecht, M. (2024). Quantifying individual personality change more accurately by regression-based change scores. *Structural Equation Modeling: A Multidisciplinary Journal, 31*(5), 909–922. https://doi.org/10.1080/10705511.2023.2274800

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian Estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(4), 646–661. https://doi.org/10.1080/10705511.2018.1545232

Zitzmann, S., Helm, C., & Hecht, M. (2020). 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*, Article 611267. 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*(6), 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, Miocevic, and van de Schoot (2020). *Structural Equation Modeling: A Multidisciplinary Journal, 28*(1), 40–50. https://doi.org/10.1080/10705511.2020.1752216

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 661–679. https://doi.org/10.1080/10705511.2016.1207179

Zitzmann, S., & Orona, G. A. (2025). Why we might still be concerned about low Cronbach's alphas in domain-specific knowledge tests. *Educational Psychology Review*, *37*(2). https://doi.org/10.1007/s10648-025-10015-5

Zitzmann, S., Wagner, W., Hecht, M., Helm, C., Fischer, C., Bardach, L., & Göllner, R. (2022). 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. *Educational Psychology Review, 34*(2), 511–536. https://doi.org/10.1007/s10648-021-09635-4

Zitzmann, S., Weirich, S., & Hecht, M. (2023). Accurate standard errors in multilevel modeling with heteroscedasticity: A computationally more efficient jackknife technique. *Psych, 5*(3), 757–769. https://doi.org/10.3390/psych5030049

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference. *Journal of Management, 41*(2), 390–420. https://doi.org/10.1177/0149206313501200

Authors and Affiliations

Julian F. Lohmann[1] · Nils Machts[1] · Jens Möller[1] · Steffen Zitzmann[2]

✉ Julian F. Lohmann
lohmann@ipn.uni-kiel.de

[1] Institute for Psychology of Learning and Instruction, Kiel University, Olshausenstrasse 75, 24118 Kiel, Germany

[2] Medical School Hamburg, Hamburg, Germany