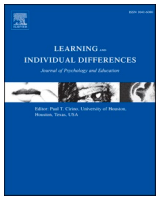
Learning and Individual Differences 124 (2025) 102812

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif



Beyond linear regression: Statistically modeling aptitude-treatment interactions and the differential effectiveness of educational interventions☆

Peter A. Edelsbrunner [a,b,*], Leonard Tetzlaff [c,d], Katharina M. Bach [a], Denis Dumas [e], Sarah I. Hofer [a], Carmen Köhler [c], Zoya Kozlova [a], Julia Moeller [f,g], Frank Reinhold [h], Garrett J. Roberts [i], Marie-Ann Sengewald [j,l], Sarah Bichler [a,k]

[a] LMU Munich, Germany
[b] ETH Zurich, Switzerland
[c] DIPF, Germany
[d] Centre for International Student Assessment (ZIB), Germany
[e] University of Georgia, United States of America
[f] Leipzig University, Germany
[g] University of Erfurt, Germany
[h] University of Education Freiburg, Germany
[i] University of Denver, United States of America
[j] Leibniz Institute for Educational Trajectories (LIfBi), Germany
[k] Universität Passau, Germany
[l] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

ARTICLE INFO

Keywords:
Aptitude-treatment interaction
differential effectiveness
latent profile analysis
additive model
Bayesian multilevel modeling

ABSTRACT

Research on aptitude-treatment interactions and the differential effectiveness of educational interventions faces statistical challenges that may contribute to sparse findings and unclear replicability. These challenges include the presence of nonlinear-, floor-, or ceiling effects, underpowered samples, and the multivariate nature of learner aptitudes. Linear regression, which prevails as the typical statistical approach in this research area, lacks the flexibility to meet these challenges. As alternatives, we present three statistical approaches: (1) Additive regression models to capture and control nonlinear or floor/ceiling effects, (2) Bayesian multilevel modeling, which can improve statistical power and allows for more complex models, and (3) clustering multivariate constellations of learner aptitudes via latent profile analysis. We demonstrate these three approaches on a motivating dataset from a scientific reasoning training, discussing their relative (dis-)advantages and how these and further models may aid research into differential effectiveness across different research topics and designs. *Educational relevance statement:* In educational interventions, researchers and practitioners are often interested in knowing for whom an intervention works best or worst. We present three statistical models that can help examine this question and overcome issues that have long bugged this field. We discuss how these approaches can help research across multiple areas, for example to examine the effects of educational technologies (augmented & virtual reality).

1. Introduction

Educational researchers increasingly recognize that while the question of "what works".

(i.e., which interventions are effective) is foundational, it does not provide the nuanced insights needed to understand learning (or instruction) in all its complexities. As a result, the focus has shifted towards asking: What works for whom, under which conditions, and for which outcomes (e.g., Faddar & Kjeldsen, 2022; Scherer & Nilsen, 2019)? This approach aims to uncover the contextual boundaries of educational (or instructional) interventions and ultimately tailor them to meet the diverse needs of learners, their learning contexts, and

☆ This article is part of a Special issue entitled: 'Future of Learning' published in Learning and Individual Differences.

* Corresponding author at: LMU Munich, Department of Psychology, Leopoldstraße 13, 80802, Munich, Germany.
E-mail address: peter.edelsbrunner@lmu.de (P.A. Edelsbrunner).

https://doi.org/10.1016/j.lindif.2025.102812
Received 29 December 2024; Received in revised form 29 June 2025; Accepted 4 October 2025
Available online 17 October 2025

educational goals.

Two research lines work on similar questions in this regard. Research on aptitude-treatment interactions (ATIs; Cronbach, 1957; Cronbach & Snow, 1981) investigates whether the effectiveness of an intervention targeting a specific educational outcome differs depending on learner characteristics such as variation in cognitive abilities or affective-motivational variables. The second research line, differential effectiveness research, has broadened the scope of this question. Here, not only experimental variations (i.e., random assignment to multiple conditions) but also observed variation (e.g., observed teaching quality) in instruction is considered. In addition, ATI research focuses on identifying disordinal interactions, meaning that one intervention condition works better for one group of learners but another condition works better for another group. Differential effectiveness research also considers ordinal interactions, in which the magnitude of an effect varies, but not its direction (Hunt, 1975). This more comprehensive conceptualization also allows for examining whether the effects of interventions vary across different kinds of educational outcomes (e.g., standardized achievement tests vs. school grades, cognitive vs. affective-motivational outcomes; Scherer & Nilsen, 2019), or across teachers, schools, and countries (Faddar & Kjeldsen, 2022; Kokkinou & Kyriakides, 2022; Yeager et al., 2019). The key distinction between the two approaches is ATI research's narrow focus on interactions of treatment conditions with learner characteristics (aptitudes), whereas differential effectiveness research also considers interactions between treatment and outcome type, or treatment and (school) context.

Both approaches have in common that they acknowledge potential heterogeneity in treatment effects (i.e., effect heterogeneity). In contrast, study designs that do not consider effect heterogeneity yield estimates of average effects that neglect the possibility of systematic effect variation. Besides neglecting the relevant information that is needed to tailor individualized interventions, unmodelled effect heterogeneity may seemingly decrease the replicability of findings. Specifically, studies that are conducted on different samples, contexts, or outcomes may find variation in results that is unexpected if it this variation is not part of the theoretical and statistical model (Bryan et al., 2021).

Here, we use the broader term *differential effectiveness,* which subsumes the ATI perspective. Our examples mostly focus on interactions between learner characteristics and the effects of experimental intervention conditions, but we will discuss the applicability of our ideas to broader questions of differential effectiveness.

We focus on a specific issue in this research area that likely contributes to prevailing difficulties in robustly finding theoretically expected effects: The difficulty of statistically modeling differential effectiveness. Since the 1970s, a general pattern has emerged: Some key findings are replicable, but further theoretically expected instances of differential effectiveness are frequently difficult to identify (Cronbach & Snow, 1981; Tetzlaff et al., 2023). For example, in a special issue on ATIs in special education research, a large part of the published analyses yielded non-significant effects (Fuchs & Fuchs, 2019). The finding that treatment effects depend on learners' level of domain expertise is perhaps the only instance of differential effectiveness that tends to generalize across different subjects, age groups, and learning outcomes (Tetzlaff et al., 2025). In its most pronounced form, this *expertise-reversal effect* (Kalyuga, 2007) manifests as a disordinal interaction between treatment and aptitude: Learners with less domain expertise benefit from more guidance and structure, whereas those with greater expertise are hindered by these same treatments and benefit instead from self-guidance and open learning settings.

Despite this unsatisfying empirical picture, all approaches that tailor instruction to individuals or groups of learners presume in principle that the effectiveness of instructional parameters depends on learner characteristics. This includes approaches under the labels of adaptive teaching, personalized or individualized instruction, and precision education (Bach, Hofer, & Bichler, 2025; Bernacki et al., 2021; Bernard et al., 2019; Plass & Pawar, 2020; Tetzlaff et al., 2021; Reinhold et al., 2020).

Specific applications of this principle include cognitive or intelligent tutors (e.g., Aleven et al., 2016) and technology- and data-based personalized learning programs (e.g., Boninger et al., 2020; Nörenberg et al., 2022). These approaches use computer-based automatization and artificial intelligence to dynamically adapt instruction to learners' progress and data. To adapt the instruction across multiple parameters and thus optimize learning processes and outcomes, educators need solid theory and empirical evidence.

Here, we argue that statistical modeling is a bottleneck contributing to the unclear picture regarding differential effectiveness (Tetzlaff et al., 2023). Differential effectiveness research addresses questions of statistical interaction, that is, how characteristics of learners, the learning setting, or the outcome moderate the effects of instruction. To identify such interactions, researchers require statistical models that can accommodate specific issues typically arising for these questions. As we will outline, researchers typically remain with linear regression models when modeling questions of differential effectiveness. These models make strict assumptions that, as we argue, do not hold for typical research on this topic and may fail to identify effect heterogeneity that is really present. To aid researchers in identifying effect heterogeneity, we outline three statistical approaches that are better equipped to meet statistical challenges that typically arise in this research area. We first outline the challenges to then explain how the three approaches - additive regression, Bayesian multilevel modeling, and latent profile analysis - can help to meet these challenges. Our aim is to provide a conceptual explanation of how these approaches work in research on differential effectiveness, exemplified on a data set to demonstrate their specific advantages and disadvantages. This paper is not meant to provide a manual or outline detailed steps for applying each of the methods.

2. Key statistical challenges in research on differential effectiveness

2.1. Issue 1: Ceiling effects

Ceiling effects arise when learners are close to the minimum or maximum of a scale. For example, if they have either little or a lot of knowledge (Ziegler et al., 2021), if they are a high- ability sample approaching the limit of a cognitive scale, if they show very little interest in a topic, or if measurement scales were designed inappropriately so that they do not cover the whole range of student variation (Grimm et al., 2023). In an analysis interested in interactions, floor- or ceiling effects can drastically affect conclusions (van Doorn et al., 2023). For example, if learners in one experimental condition get closer to the ceiling of a scale at posttest than those in a comparison condition, then their regression line must become flatter to accommodate the ceiling. This implies a bias in the modeled interaction between treatment condition and the moderator variable. A statistical interaction may then be an artifact, or the scale restriction can mask an interaction (Rohrer & Arslan, 2021). Consequently, we either need measures that capture the entire distribution of a construct such that students do not bottom- or top-out the scale, which is not always feasible or useful (e.g., when items already cover all aspects of a construct or when mastery is the goal), or we require statistical modeling techniques that aid against bias caused by floor or ceiling effects.

2.2. Issue 2: Nonlinearity

Floor and ceiling effects are a special case of the larger problem of nonlinearity as a method artifact, that is, being caused by sample or test design. In other cases, the studied phenomenon itself can be inherently nonlinear, in which case the nonlinear effects become actually of substantive interest to the researcher, such as the nonlinear growth in a knowledge outcome over time (e.g., Dumas et al., 2020). While, in such

studies, student background variables typically predict growth parameters linearly, this is not the case when a learner or treatment characteristic is shown to be associated with an altogether different functional form of growth over time. For example, in a nonlinear dosage-response study of reading interventions in early elementary school, Roberts et al. (2022) found that one-on-one interventions and small group interventions resulted in a completely different shape to the growth function over the course of the intervention. Other studies have reported nonlinear interactions of instructional conditions with intelligence (Ziegler et al., 2021), and working memory capacity (Grimm et al., 2023). Even without the presence of nonlinear interaction effects, nonlinear main effects can wrongly indicate interactions if they are not adequately modeled (Belzak & Bauer, 2019). In a typical implementation of linear regression, interaction effects are only implemented in a linear fashion. For example, for the commonly reported expertise reversal effect, it may be assumed that the effect of an intervention condition in comparison to a control condition becomes linearly weaker, decreasing from a positive to a zero value and eventually to a negative estimate, with increasing prior knowledge (Kalyuga, 2007). But if such interactions were nonlinear, this would remain unmodelled and thus unseen in a linear regression. Consequently, either substantive theory is required for ruling out nonlinear relationships, or statistical models need to be able to capture nonlinear effects.

2.3. Issue 3: Limited statistical power

A third challenge in differential effectiveness research is obtaining sufficient statistical power. In this case, this means that we have sufficiently large sample sizes to identify interaction effects with high statistical power (i.e., a high probability to correctly distinguish such effects from sampling error). Early ATI researchers already cautioned that obtaining high power for analyses of aptitude-treatment interactions generally requires more than 100 learners within each experimental condition (Cronbach & Snow, 1981). Typically, the sample size requirements for detecting interaction effects are much larger than for the detection of main effects. In multilevel modeling, which has become the norm in many educational research settings (Brauer & Curtin, 2018; Köhler et al., 2021), these requirements might be even higher. Consequently, results need to be carefully interpreted with regard to sample size requirements, and efficient modeling approaches are required to detect interaction effects.

2.4. Issue 4: Multivariate learner aptitudes

The fourth and final challenge in this research area is the multivariate nature of learner aptitudes. How much a learner benefits from specific instructional parameters may depend on a multitude of learner characteristics and their interplay, including cognitive, personality, behavioral, affective-motivational and other characteristics (Ackerman, 2003; Cronbach & Webb, 1975; Snow & Farr, 2021; Tetzlaff et al., 2023; Bichler et al., 2020; Schwaighofer et al., 2017; Hofer & Reinhold, 2025). Current research on differential effectiveness typically focuses on one isolated aptitude, leaving the relative importance of multiple aptitudes, as well as their interactions among each other and with the treatment, unexplored (Bichler et al., 2020; Schwaighofer et al., 2017). In particular, the combination of cognitive and non-cognitive variables is understudied (Bach et al., 2025; Hofer & Reinhold, 2025; Cronbach & Snow, 1981; Sternberg & Grigorenko, 1997). This means that the current approach to modeling ATIs or differential educational effects is unable to capture the full complexity of learning, especially learning in authentic settings where a) multiple learner characteristics operate at the same time, b) engagement with the intervention also determines its impact, and c) situational demands exert their additional effects (Bichler et al., 2022; Bichler et al., 2025). Consequently, we require statistical models that can capture interactions between multiple characteristics of learners, treatments, and potentially even context characteristics and learning outcomes.

3. Three suggested statistical methods to address key challenges

Research on differential effectiveness rarely applies statistical methods that are capable of addressing the above-mentioned challenges. The most commonly applied and recommended model is a (multilevel) linear regression with interaction terms that capture the interactions between learner characteristics and intervention variables (Hayes & Rockwood, 2020; Preacher & Sterba, 2019; Tetzlaff et al., 2023). For example, in a journal special issue on ATI research, all contributions but one implemented this approach (Fuchs & Fuchs, 2019). We argue that given the outlined statistical challenges, understanding what works for whom under which circumstances requires a more diverse statistical toolbox.

Here, we present three statistical methods that are not new but have seldom or never been used in this research area, despite their potential for tackling the outlined challenges. The approaches are additive models (Wood, 2017), which are great for modeling non-linearity and thereby capturing floor- or ceiling effects, Bayesian multilevel modeling (Bürkner, 2017), which can improve statistical power and offer many extensions to fully capture differential effectiveness, and mixture modeling (Hickendorff et al., 2018), which can accommodate multivariate learner characteristics and their interactions as moderator variables. We select these approaches because they are conceptually close to linear regression, such that researchers can build on their knowledge of this method. The methods we propose are better suited than linear regression to tackle these challenges by relaxing or overcoming some of its stringent assumptions.. While all three are better suited than regression, each method has their specific strengths and weaknesses compared to each other. In the following, we apply these approaches to an example data set to demonstrate their utility and compare their strengths and weaknesses.

We illustrate the three approaches based on empirical data from Peteranderl et al. (2023).

The authors conducted a training of the control-of-variables strategy (CVS), that is, understanding that in an informative experiment, only one thing is varied at a time. The study encompassed fifth- and sixth-graders, of which we use a subset of $N = 593$ from 38 school classes. The full dataset would consist of 618 students, but measures of moderation variables are missing for some students because they were absent on the school day when the data was collected. One half of the students received an explicit training on CVS (intervention condition). The training took place over three lessons in which students progressed from observing teacher demonstration of experiments including explicit explanations of the strategy, to guided practice, and finally to autonomous practice (setting up and discussing experiments in small groups). The other half of the students received an active control training (control condition) in which students engaged in self-guided inquiry without explicit training of the strategy. The students were randomized within classrooms, such that within each school class, half of the learners received the intervention and the other half the control training. We use this dataset because the study focused on establishing aptitude-treatment interactions and it used a sample size apt for demonstration of our approaches. Our conclusions regarding the different approaches arise from the different statistical assumptions they make, which do not depend on this dataset. We use this data set to exemplify the general strengths and weaknesses arising from the different assumptions that the approaches make.

As variables to inspect differential effectiveness, the authors gathered data on learners' skills at pretest (prior knowledge), reasoning ability, and reading comprehension, which we all used as z-standardized scores (a solution with alternative scale normalization is provided in the online supplementary materials). As the dependent measure, we use a z-standardized score of learners' achievement across four skills amounting to the control-of-variables strategy. For details on the measures,

descriptive statistics, and theoretical rationales for inclusion of the three moderator variables, see Peteranderl et al. (2023). As z-standardization can affect the interpretation of latent profile analysis (Moeller, 2025), we provide a robustness check against alternative scaling approaches, together with the analytic data and scripts, in the supplementary materials under https://osf.io/cd5v9/?view_only=0ef4056d33aa4d6bb8e667a89c17a16a. In the following, we use these data to first implement the traditional approach of multiple regression including interaction terms and then compare its results to those obtained with our three proposed approaches.

3.1. Traditional approach: Multiple regression with interaction term

To mimic the traditional approach described by Tetzlaff et al. (2023), we set up a multilevel linear regression model with a random intercept across school classes. Peteranderl et al. (2023) conducted an a priori power analysis, indicating that they would obtain power > 0.80 to find moderation effects using this approach. We included fixed effects of the treatment condition (0 = control condition as the baseline, 1 = intervention condition) and fixed effects as well as the interaction terms of the treatment condition and the three moderators (prior knowledge, reasoning ability, reading comprehension), fitting the model in the R package lme4 (Bates et al., 2014).

The results from this approach are presented in Table 1. As visible from this table, the traditional approach indicated a significant interaction effect of treatment condition with reasoning ability. The positive estimate indicated a stronger positive effect of the intervention condition in comparison to the control condition for learners with better reasoning ability. The other two interaction terms indicated a negative yet non-significant interaction of the treatment condition with prior knowledge, and a positive yet non-significant interaction with reading comprehension. The estimated interactions are depicted in Fig. 1, demonstrating how the effect of the intervention condition in comparison to the control condition depended on the moderator variables according to the traditional approach.

From these results, we would infer that learners with higher reasoning ability benefit more from the intervention, compared to those with lower reasoning ability. In addition, we would remain unsure whether the effect of the intervention is weaker for learners with more prior knowledge and stronger for those with better reading comprehension. As these effects are statistically non-significant, these hypotheses would have to be further examined in future research (Edelsbrunner & Thurn, 2024).

3.2. Additive regression model

Next, we estimated the same model, but within the modeling framework of the general additive mixed model (Wood, 2017). This means that in addition to random effects, which we again covered through a random intercept across school classes, this model uses additive effects instead of linear effects for regression terms. Additive effects, also called smooths, adapt to the data allowing nonlinearity in the effects of the moderators on the dependent variable. This can also capture floor- or ceiling effects, while avoiding overfitting (i.e., adapting overly to apparent nonlinearities) through a cross-validation procedure (Wood, 2017).

The results of the model are presented in Table 2 and Fig. 2. For each moderator, we have an empirical distribution function, indicating how much wiggliness (i.e., non-linearity) there is in each regression parameter within each treatment condition. In addition, we conducted model comparisons via the AIC (Dziak et al., 2020; Edelsbrunner et al., 2023) to examine whether including interaction terms between the treatment condition and the moderators improved the model fit. The model comparisons indicated that the effect of all three moderators should be allowed to differ between treatment conditions, amounting to differential effectiveness for all three moderators—in contrast to the traditional approach. Table 2 shows the model estimates, with values above 1 for the empirical degrees of freedom for smooth terms indicating nonlinearity in the relation of the respective moderator with the dependent variable within the respective treatment condition.

These results differ from those of the linear regression model. For Prior knowledge and reading comprehension, the estimated relations are nonlinear within the intervention condition, and for reasoning ability within the control condition. In contrast to the linear regression model, the additive model indicates a positive effect only within the higher range of reasoning in the control condition. In addition, in the intervention condition, reading comprehension only has a positive effect on the learning outcome in the lower range, indicating that some reading comprehension is required to benefit from the intervention, but the effect ceases in the higher range. Regarding prior knowledge, the model indicates a sigmoid curve in the intervention condition, with a strong effect in the middle range but weaker effects closer to the ceiling and floor of the scale.

Table 1
Results from traditional multilevel regression approach.

Parameter	Estimate	SE	95 % CI Lowe r	95 % CI Upper	*t*	*p*
Intercept	−0.21	0.03	−0.27	−0.14	−6.19	< 0.001
Condition	0.41	0.04	0.33	0.50	9.22	< 0.001
Prior knowledge	0.68	0.04	0.60	0.77	15.78	< 0.001
Reasoning ability	0.14	0.04	0.06	0.22	3.56	< 0.001
Reading comprehension	0.06	0.04	−0.01	0.14	1.60	0.111
Prior knowledge: Condition	−0.09	0.06	−0.20	0.03	−1.51	0.132
Reasoning ability: Condition	0.11	0.05	0.00	0.21	2.02	0.044
Reading Comprehension: Condition	0.07	0.05	−0.04	0.17	1.20	0.229

3.3. Bayesian multilevel model

Next, we set up the same model as for the traditional approach but used Bayesian estimation. Crucially, we specified parameter priors, that is, our prior expectations regarding each model parameter formalized as distributions. This is a requirement in Bayesian estimation of statistical models and can be advantageous for parameter estimation (van de Schoot et al., 2014). For example, following Peteranderl et al. (2023), we expected reading comprehension to have a more positive effect on the learning outcome in the intervention condition than in the control condition, because the verbal instructions in the intervention condition may require better verbal comprehension, which usually correlates substantially with reading comprehension. We expected the stronger guidance in the intervention condition to decrease the effect or reasoning ability compared to the control condition in accordance with Ziegler et al. (2021), and also a smaller effect of prior knowledge in the intervention condition. For full explanations and justifications of all prior settings, please see the online supplementary materials. Importantly, these prior specifications may affect the parameter estimates, which in turn may increase statistical power and decrease bias (although if priors are misinformed, they may have the opposite effects; van de Schoot et al., 2014). The results with Bayesian estimation are provided in Table 3 and Fig. 3.

The estimated interaction effects of prior knowledge and reading comprehension are stronger than in the traditional approach and their credible intervals exclude 0. The estimated interaction for reasoning ability on the other hand is smaller than in the traditional approach. Overall, whereas the interaction effects go in the same directions as in the traditional regression model, their estimated magnitudes and uncertainties around these (i.e., credible intervals) are different and

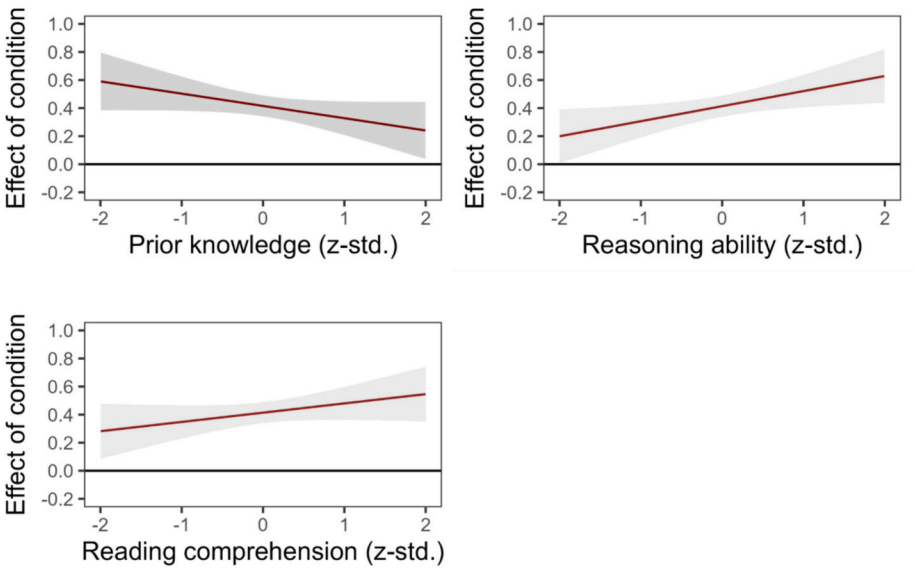


Fig. 1. Interaction effects of treatment condition with the three learner characteristics in the traditional multilevel regression model.

Table 2
Results from multilevel additive regression model.

Predictor	Estimate/edf	*t/F*	*p*
Intercept	0.02	0.72	0.474
s(Prior knowledge):Control	1	206.44	< 0.001
s(Prior knowledge):Intervention	3.55	47.18	< 0.001
s(Reasoning ability):Control	3.70	3.67	0.005
s(Reasoning ability):Intervention	1	30.40	< 0.001
s(Reading comprehension):Control	1	3.10	0.079
s(Reading comprehension):Intervention	2.92	8.32	< 0.001

Note. *s* indicate smooth (nonlinear) regression terms; edf = empirical distribution function estimate for smooth parameters. Intercept receives *t*-value, smooth terms *F*-values. Estimate/edf indicates estimated complexity of smooth term, with 1 indicating linear effect and higher estimates increasing nonlinearity.

indicate some diverging conclusions.

3.4. Bias-corrected latent profile analysis

Finally, we used latent profile analysis to investigate the differential effectiveness of the intervention dependent on the learner characteristics reading comprehension, prior knowledge, and reasoning ability. We used the three z-standardized moderator variables to build learner profiles and subsequently estimate effects of the invention condition in comparison to the control condition, as well as differences therein between the profiles (i.e., the moderation effects), on the learning

Table 3
Estimates from bayesian multilevel model.

Predictor	Estimate	SE	95 % CI Lower	95 % CI Upper
Intercept	−0.21	0.03	−0.28	−0.15
Condition	0.42	0.04	0.33	0.51
Prior knowledge	0.67	0.04	0.59	0.75
Reasoning ability	0.16	0.04	0.08	0.24
Reading comprehension	0.06	0.04	−0.01	0.13
Prior knowledge:Condition	−0.09	0.05	−0.19	0.02
Reasoning ability:Condition	0.08	0.05	−0.02	0.19
Reading comprehension: Condition	0.08	0.05	−0.02	0.18

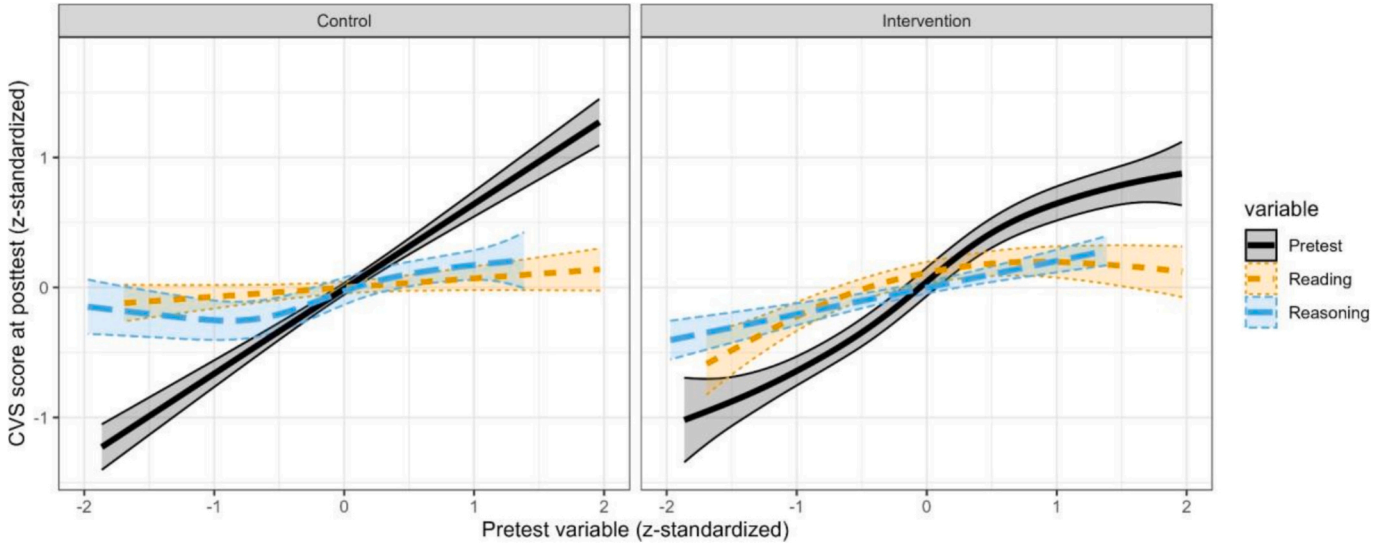


Fig. 2. Interaction effects of treatment condition with the three learner characteristics used as moderators in multilevel additive regression model.

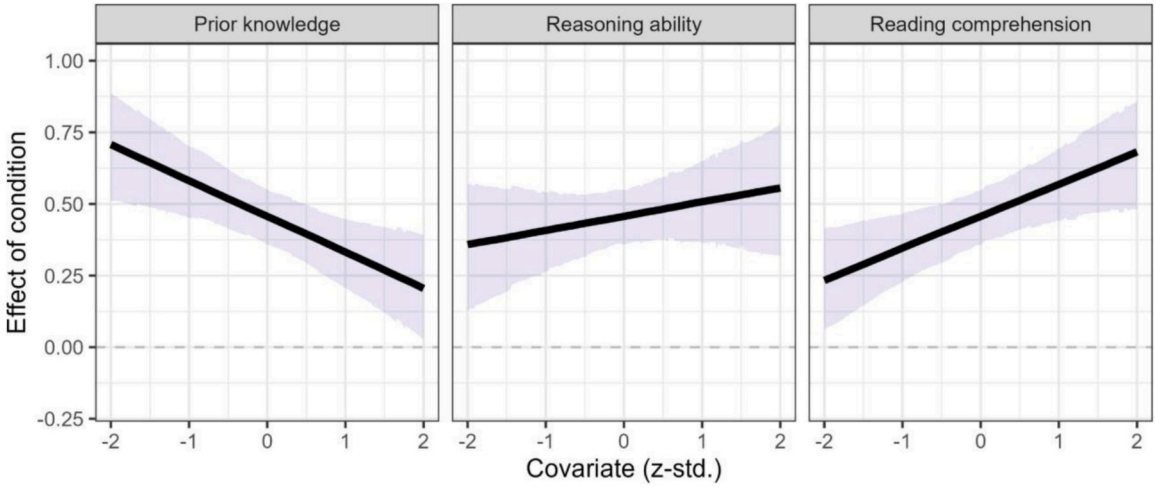


Fig. 3. Interaction effects of treatment condition with the three learner characteristics used as moderators in the bayesian multilevel model.

outcome. We followed the typical step-wise approach in which the number of profiles is first increased from one profile to two profiles and so on until the fit indices indicate that the correct number of profiles has been superseded or model estimation fails. We estimated models with one to eight profiles, with the resulting fit indices depicted in Fig. 4. As commonly observed (Edelsbrunner et al., 2023), the more stringent fit indices BIC and CAIC pointed towards the solutions with four or five profiles (indicated by the lowest fit estimates), whereas the less stringent AIC, sample-size adjusted BIC and AIC3 pointed to solutions with up to seven or eight profiles. We visually and numerically inspected all solutions within the range indicated by the different indices, considering in particular the sample-size adjusted BIC and the AIC3 which usually function well at our sample size (Edelsbrunner et al., 2023).

The solutions with six and seven profiles resulted in extremely small profiles (about 2 % of learners) that would be difficult to interpret substantively. We selected the solution with five profiles, which are depicted and labeled according to profile configurations in Fig. 5. We labeled the profiles in accordance with their levels on all moderators. One further profile was labeled as high achievers because these learners had high levels on the other moderators despite low reasoning ability. In the next step, we conducted a BCH-approach as suggested by Tetzlaff et al. (2023). This approach enables estimating structural equation models for learners within each profile, correcting for uncertainty in profile memberships.

As visible from Fig. 6, the model estimated that there were no visible effects of the intervention condition in comparison to the control condition for learners with low or high levels on all moderators. For those with moderate or good levels on all variables, as well as those who perform well in terms of prior knowledge and reading comprehension despite having weak reasoning ability, the effect of the treatment condition was clearly positive.

In comparison to the typical regression approach, two results stand out: First, only for learners with preconditions in the middle range, positive effects of the intervention condition were observed. This is in contrast to the typical approach, which by definition indicates linear relations, yet it appears consistent with the sigmoid effect estimate of prior knowledge (i.e., a positive effect on the outcome only in the middle range) in the additive model. Second, the high achievers profile (high prior knowledge and reading but low reasoning ability) showed a positive effect of the intervention condition despite being low on reasoning ability. This is in contrast with the typical regression approach, which indicated that the effect of the intervention becomes stronger with increasing reasoning ability and lower with increasing prior knowledge. Thus, since the traditional approach does not consider interactions with the other learner characteristics, it misses that the intervention has a positive impact for some learners with lower reasoning ability.

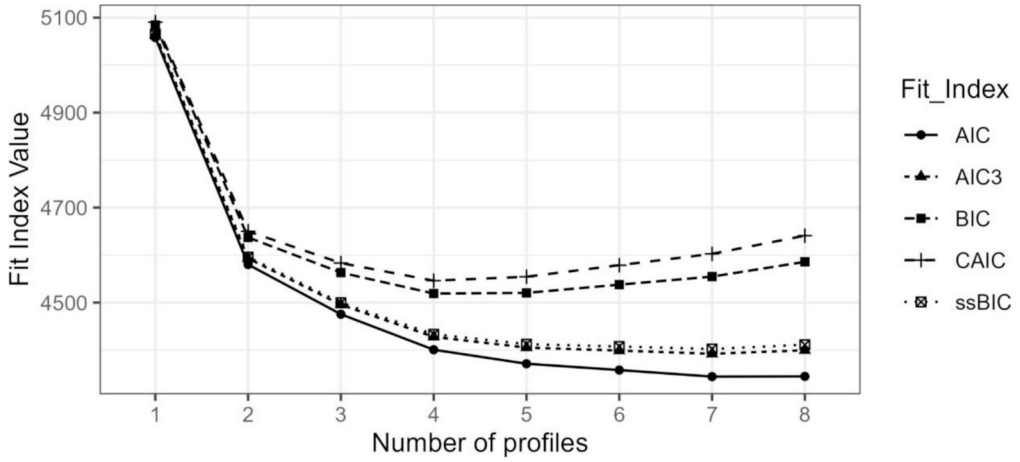


Fig. 4. Fit Indices from latent profile analyses with different numbers of profiles.
Note. *AIC* = Akaike Information Criterion; *AIC3* = Akaike Information Criterion with penalty term of three; *BIC* = Bayesian Information Criterion; *CAIC* = Consistent Akaike Information Criterion; *ssBIC* = sample-size Adjusted Bayesian Information Criterion. See Edelsbrunner et al. (2023) for explanations of these criteria.

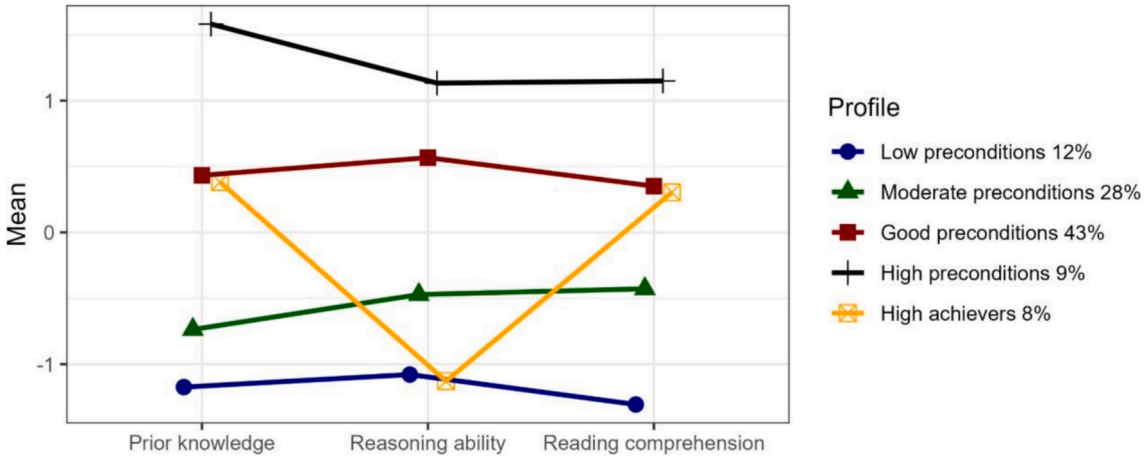


Fig. 5. Latent profiles based on the three learner characteristics.

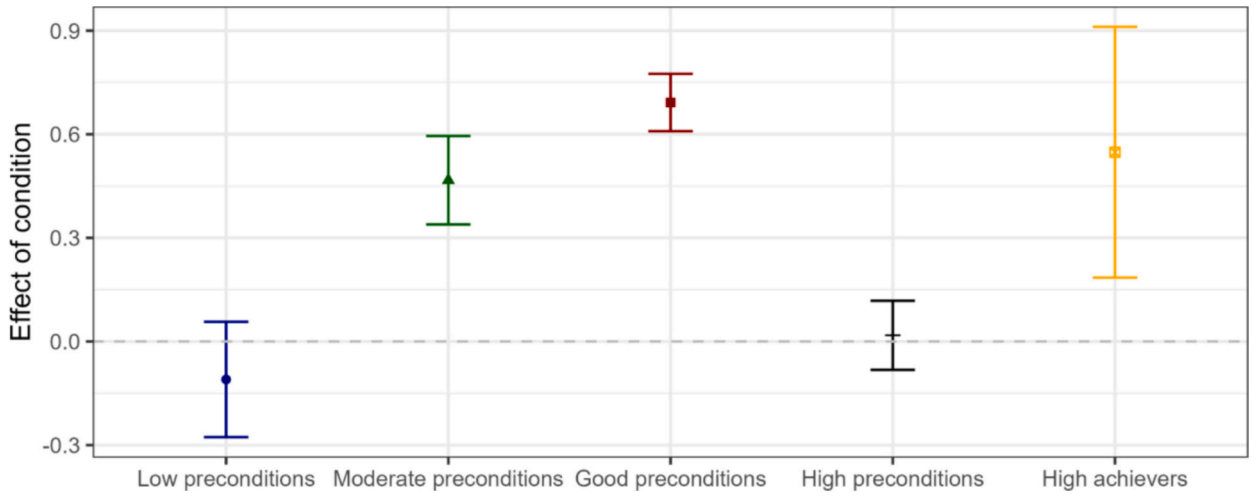


Fig. 6. Estimated effect of treatment condition on learning outcome within the different learner profiles.

3.5. Comparison of results across the four approaches

The traditional approach (multiple regression with interactions) descriptively indicated interactions with all three learner characteristics, but only the positive interaction effect with reasoning ability yielded a statistically significant effect estimate, despite Peteranderl et al. (2023) reporting a power simulation that predicted adequate statistical power to find interactions. From this approach, we might infer that better reasoning ability concurs with better effectiveness of the intervention condition in comparison to the control condition but remain unsure about the other effects.

The additive model in contrast indicated interactions of the condition with all three learner characteristics. Additionally, for each learner characteristic, the model indicated a nonlinear effect within one experimental condition. Effects of prior knowledge, reasoning ability, and reading comprehension all appear more nuanced than indicated in the traditional regression. In particular, the additive model indicated a sigmoidal (S-shaped) effect of prior knowledge in the intervention condition.

The Bayesian multilevel model with informative priors indicated stronger effects for prior knowledge and reading comprehension, but a weaker effect of reasoning ability.

The latent profile analysis similarly to the additive model yielded more nuanced effects than in the traditional regression model, particularly with regards to reasoning ability, showing that the intervention condition can be effective for learners with lower reasoning ability if they have higher levels on the other moderators.

4. Comparison of advantages and disadvantages of the modeling approaches: When should they be used?

The application of the four presented different methods to model differential effectiveness of treatments on an exemplary dataset showcases each method's advantages and disadvantages, which are summarized in Table 4. Based on this summary, we can now suggest in which situations each of the approaches may be most useful.

Before comparing the approaches, we note that researchers are not bound to select only one approach as the correct one given specific data and aims. Rather, prior research has shown that combining more than one statistical approach is in general a good idea to gain multiple perspectives on data, examine the robustness of results and conclusions, and combine different approaches` advantages (see Grimm et al., 2023; Hoogeveen et al., 2023; Marsh et al., 2009). For example, Grimm et al. (2023) combined latent profiles to model learners` prior knowledge and how it interacted with their working memory with visualizations from an additive regression model to be better able to interpret the results.

4.1. Traditional approach

The traditional approach (multiple regression with interactions)

Table 4
Contrasting and comparison of the different modeling approaches; issues in columns, proposed models in rows.

	Floor−/ceiling effects	Non-linearity	Statistical power	Multivariate characteristics	Further notable options of approach
Linear regression	–	–	–	–	Possibility of polynomial /e.g., quadratic) terms
Additive regression	Captured via smooth terms	Captured via smooth terms	Potential increase	–	Bayesian estimation (e.g., brms; Bürkner, 2017)
Bayesian multilevel modeling	Optionally via Tobit/ beta/hurdle/z ero−/ one-inflation	Optionally via smooth terms (Bürkner, 2017)	Optimized via informative priors	–	Additive (smooth) terms, multivariate models, response distributions (tobit, beta, hurdle, zero−/one-inflation, …), Model testing (loo/Bayes factor; see Bürkner, 2017), complex random effects structures incl. Effects of teachers, schools, items/ outcomes, distributional models (Haslbeck et al., 2024)
Latent profile analysis	Captured via profiles	Captured via profiles	Optimized via BCH- approach	Captured via profiles	Extension to moderated/mediated SEM (Costache et al., 2022)
Further notable approaches to handle issue	Tobit regression Beta regression Hurdle models Zero−/one-inflated models	Non-linear regression (e.g., sigmoidal effects; Preacher & Sterba, 2019) Machine learning (e. g., double ML; Knaus, 2021)	Structural equation modeling (Kline, 2023) Alpha error level increase (e. g., 10 %; Peteranderl et al., 2023) Covariate inclusion (Sengewald & Mayer, 2024)	Machine learning (e.g., double ML; Knaus, 2021)	

faces considerable methodological problems discussed at the beginning of this article and should only be chosen when the following conditions are met: Linearity, absence of floor- or ceiling effects, sufficient sample size to obtain high power regarding interactions, and the absence of dependence among moderator effects (i.e., no higher-order interactions). These preconditions may be difficult to evaluate, and we propose combining theoretical knowledge with exploratory data-driven approaches to decide whether these conditions are met. Of course, the traditional approach has the advantage of being well-known to many researchers. It also uses rather few model parameters, making it more parsimonious. Parsimony may be a desirable characteristic of scientific models (Vandekerckhove et al., 2015), but only so far as it still allows modeling the phenomena and questions of interest. This does not appear to be the case for typical differential effectiveness-questions for a simple model as linear regression. The stringent model assumptions, such as linearity of relations, hamper this approach to reliably answer questions of differential effectiveness and it will often be inadequate.

4.2. Additive regression

The additive regression model should be used if researchers suspect (e.g., based on theory or data inspection) that there may be non-linearities in the effects within all or specific conditions. The model should also be used if floor- or ceiling effects may be in play. When interpreting additive regression models, it is important to keep in mind that the graphs showing the nonlinear effects visualize the unique effect of one predictor but if this predictor is correlated with the other predictors, it is difficult to interpret each variable's independent contribution to the outcome (Baayen & Linke, 2020; Wieling, 2018). As in linear regression, nonlinear interaction terms are not automatically taken into account and need to be manually added, and complex regression paths or mediation paths cannot be modeled (Wieling, 2018). Researchers also need to be aware of concurvity, which is similar to co-linearity and may complicate the estimation of the model and the interpretation of the results (Baayen & Linke, 2020).

Another important consideration for researchers when choosing additive regression models is the extent to which the results provide actionable information for practitioners. It is probably not feasible to take the aptitude composition of each learner into account and adapt the instruction or intervention to their specific levels. If a relationship is not linear, a useful approach might be limiting the number of basis functions (determining the maximum complexity of the curve) to three or four. This would suffice to describe relations with the outcome in the low, medium, and high range of the predictor.

4.3. Bayesian multilevel model

The Bayesian multilevel approach may be particularly suited if some prior knowledge based on empirical data or theoretical models is available to guide the setup of informative prior distributions for key model parameters (see Browne & Draper, 2006; Sarma & Kay, 2020) to improve statistical power. This information does not need to be available for all model parameters; for those with little or no prior information, broad priors can be reflecting such lack of knowledge. For these parameters, the Bayesian approach then typically will not result in a visible difference compared to traditional non-Bayesian estimation (van de Schoot et al., 2014).

The concrete sample size planning can be quite challenging and requires knowledge about the size of the conditional (subgroup-specific) effects that can be specified from an interaction by investigating the treatment effects for specific values of a moderator (see e.g., Baranger et al., 2023). Consequently, estimating the required sample size is often not straightforward (Green & MacLeod, 2016). Further options to increase the power in an analysis are the correction for measurement error and the inclusion of additional variables that explain residual variance in the outcome (Table 4; see e.g., Cohen et al., 2003).

In addition, as indicated in Table 4, Bayesian multilevel modeling offers multiple opportunities that go beyond our simple demonstration and may benefit researchers within this area. For example, the Bayesian concept of effect size distributions, as opposed to the traditional axiom of the one and only true effect in terms of the classical test theory, has been proposed as a useful framework to understand and model treatment effect heterogeneity (Gelman, 2015).

Moreover, Bayesian multilevel models allow incorporating complex random effects structures that enable examining variation of effects of intervention conditions and their interactions across teachers, classes, or schools, as well as across outcomes of different types or across multiple items of tests or questionnaires (Donnellan et al., 2023; Haslbeck et al., 2025). In addition, Bayesian modeling in the *brms* package, which we used for our demonstration, can incorporate smooth terms like additive models and other options to handle floor- or ceiling effects such as Tobit regression and Beta-, zero- or one-inflated distributions (Bürkner, 2017; Haslbeck et al., 2024). Modeling such characteristics of data is called distributional modeling. This approach models effects of predictor

variables not only in predicting the expected mean of the outcome, but also its variation and further characteristics of the distribution such as the proportion of learners who achieve minimum or maximum scores (Haslbeck et al., 2025; Umlauf et al., 2018). Note, however, that specifying prior distributions for predictors in models that use nonlinear effect structure is challenging.

Another advantage of Bayesian modeling are recent advances that allow comparing the strength of evidence for different models via Bayes factors (Edinburgh et al., 2023; Gronau et al., 2020). A limitation that applies to both additive and Bayesian (multilevel) models is the difficulty of incorporating multivariate learner, school, or outcome variables in the manner of the latent profile analysis. Although this is in principle possible in Bayesian models, we are not yet aware of software packages and accompanying tutorials making this easily feasible.

4.4. Latent profile analysis

Latent profile analysis possesses the unique strength of being able to integrate multiple characteristics of learners or the learning context and their potential higher-order or nonlinear interactions (Tetzlaff et al., 2023). In defense of the other models, one could in principle add higher-order interactions between all moderator variables, but this approach usually results in an uninterpretable number of effects that moderate one another and risks (drastically) decreasing statistical power (Tetzlaff et al., 2023). In addition to profiling or clustering multiple learner characteristics, we suggest also considering modeling multivariate learning outcomes (Grimm et al., 2023) or interactions across multiple observed instructional variables by means of latent profile analysis. Latent profile analysis allows researchers to break down a complex variable space into a limited number of latent profiles, substantially improving interpretability and statistical power (Tetzlaff et al., 2023) while still capturing learners' complexity rather than reducing them to a single characteristic. In addition, multivariate profiles may capture nonlinear and higher-order interactions across multiple variables (Bauer & Shanahan, 2007). Through the data-driven identification of profiles, the resulting patterns avoid interpreting parameter areas that would represent variable levels in which learners do not realistically reside. If there is a multivariate mixture distribution (i.e., unobserved heterogeneity), then other models including the traditional linear regression models risk overlooking that fact. All other models described in this article examine and describe between-person variance. Only latent profile analyses reveal within-person patterns (so-called person-centered modeling; Hickendorff et al., 2018) and additionally quantify the frequency of each profile in the sample. The other models rely on the assumption that a single (one-size-fits-all) coefficient sufficiently describes the association between two, three, or more aptitudes or learner characteristics in the sample. Latent profile analysis is able to reveal that these associations differ between groups of learners, are positive for some and negative for others, or that one profile cluster shows high scores in the learner characteristics A and B, whereas another profile group shows high scores in learner characteristic A but low scores in learner characteristic B. In this way, latent profile analysis may uncover differential effects that would otherwise muddy average effects and yield results that appear non-replicable (Bryan et al., 2021).

Importantly, mixture analysis such as latent profile analysis serves as a test for the assumptions underlying all other models. The linear regression would assume multivariate unimodal distributions of all variables without testing that assumption, failing to provide trustworthy results if the assumptions are violated. A latent profile analysis reveals whether a mixture distribution is present in the data, whereas a linear regression just assumes that it is not. Thus, a latent profile analysis reveals crucial information even if it does not reveal a mixture distribution: In that case, it reveals that central assumptions of linear regressions are met. Unique information is even provided by latent profile analysis solutions in which all resulting profiles suggest linear associations among the included variables (i.e., models with profiles in which all variables are either all high, all moderate, or all low): These models reveal how high, and how low, the scores in the different profiles were in reference to the response scale and can therefore reveal co-endorsement, which is a very different information from the covariance examined in linear regressions (see Moeller, 2021; Moeller et al., 2018). In addition, the prevalences of such profiles are important information, as it is possible that the profile with all-high scores on all variables is relatively rare (e.g., 5 %), whereas the all-moderate and all-high profiles may be more frequent (e.g., 45 % and 50 %, respectively). All of this information is potentially crucial for research on differential intervention effects, and among the four methods introduced in this article, only latent profile analysis is capable of revealing it. Yet, due to its unknown replicability of profiles across different samples and populations of learners, the exploratory nature of latent profile analysis may not always be the preferred option. In addition, latent profiles may not always capture all relevant information of the indicator variables, in particular when the indicators cover a broad variety of variables rather than a narrow common construct (Daumiller et al., 2023).

As a result of latent profile analysis, instruction may be adapted to learners` profile by implementing differentiated tasks or activities for groups on learners with certain profiles. Alternatively, technology-supported solutions may implement tasks, activities, or instructional support on a more individual level, catering dynamically to several aspects of a learner's profile. In contexts in which individual students are in focus (e.g., those receiving one-on-one psychological services), comparing individual results to profiles of larger samples may help in designing individualized interventions.

4.5. Summary: It's all about assumptions

As described in our comparison of the four approaches, their strengths and limitations arise from the different statistical assumptions they make. Linear regression makes the well-known assumptions of independent, normally, and homoscedastic distributed residuals, as well as linearity in the predictor-outcome relation (with the multilevel extension removing the assumption of independence; Tabachnick et al., 2013). Additive regression relaxes (i.e., does not make) the linearity assumption by specifying smooth (i.e., non-linear) regression terms. In principle, Bayesian multilevel regression makes the same assumptions as linear regression. Yet, when residuals are non-normally distributed, this will not pose a major threat to valid inference in Bayesian estimation because this will be visible and accounted for in the posterior distributions, which are the prime source of inference in Bayes (McElreath, 2018). In addition, Bayesian estimation has outstanding capabilities to relax all assumptions that are easily available and typically converge without issues, in comparison to the often cumbersome traditional implementations (see Table 4 as well as Bürkner, 2017). Latent profile analysis makes the assumption of multivariate within each of the estimated profiles, but it relaxes that any assumptions must hold across the whole sample. This means that if an additive model or latent profile analysis indicate non-linearity, or that effects hold only within certain profiles, then the respective assumptions do not hold within linear regression. For each assumption that either model relaxes, it will become more complex, and researchers have to find the right balance they want for their model to be sufficiently informative while remaining well-interpretable for themselves and readers. This may be achieved by checking model assumptions and if either of these does not hold, researchers have to evaluate whether this affects their interpretations regarding differential effectiveness. If it does, a more complex model relaxing the respective assumption may be appropriate. Of note, the hypotheses that are tested by the different approaches are generally the same: Does an intervention show an interaction effect with a specific learner characteristic? Yet, latent profile analysis is the only approach in which multiple learner characteristics are modeled concurrently, expanding the tested hypothesis from a univariate to a multivariate question of interaction.

5. Using the proposed approaches to advance educational research across diverse topics and methods

An area where nuanced interaction effects among aspects of learners and learning contexts would be expected is student creative thinking. Creative ideas need to be maximally original and appropriate to the task at hand (Stein, 1953). In generating creative ideas, learners draw on their prior knowledge in the domain (Dumas et al., 2024), as well as their meta- cognitive beliefs about what the evaluator of their generated ideas (e.g., a human or machine rater; Acar et al., 2024) is likely to know (Lebuda & Benedek, 2023). Experimental or observed interventions would likely interact with these learner characteristics and these in turn with raters (Dumas & Kaufman, 2024; Scherbakova et al., 2024), opening up a vast exploration space of differential effectiveness to be modeled.

Another area in which latent profile analysis has shown potential in providing important information about the moderation effects of interventions are reading interventions for students with reading difficulties. Prior research found that many students are not adequately responding to small group evidence-based reading interventions (Case et al., 2014; Vaughn et al., 2019, 2020). Kulesz et al. (2024) found that responders to a year-long intervention could be distinguished from responders to a control condition by building latent profiles based on language, cognitive, and attention skills, while Tetzlaff et al. (2023) found that students respond differently to specific classroom instruction based on profiles across listening comprehension, decoding, and syntax comprehension. These findings suggest that there is promise in considering multiple learning prerequisites simultaneously to determine the appropriate customized intervention, particularly for those students who do not benefit from the offered instruction.

A different example of this can be found in the field of educational technologies, particularly Augmented Reality (AR). A recent review has shown that although researchers are commonly interested in the specific demands that learning with AR puts on learners with different learning prerequisites, appropriate study designs and in particular statistical methods to test differential effects of AR interventions are lacking (Kozlova et al., 2025). Only a limited number of studies have directly addressed the role of individual differences in learning with AR (as well as in virtual reality research; see Lawson et al., 2024). This gap can partly be explained by methodological challenges associated with AR research, particularly the issue of small sample sizes, which arise due to the technological complexity of AR. AR research often employs simple methods, such as *t*-tests to compare group means (Kozlova et al., 2025).

In this and other fields struggling with gathering appropriate sample sizes for analyses of differential effectiveness, Bayesian estimation, for instance, can improve statistical power by incorporating theoretical knowledge of effect sizes into prior distributions (McCarthy & Masters, 2005; van de Schoot et al., 2014). Another robust method already utilized in ATI research with AR is fuzzy set qualitative comparative analysis (fsQCA; (Ling et al., 2021). FsQCA bridges the gap between qualitative and quantitative methods and is especially valuable for studies with small sample sizes (as small as $N < 50$). It accommodates nonlinear relationships and asymmetric data patterns (Geremew et al., 2024) and handles multiple individual differences alongside a variety of learning outcomes. This may make fsQCA for analyses of differential effectiveness when sample sizes are too small for latent profile analysis, but multiple learner variables should be considered concurrently.

Independent of the specific field or application context, educational research that wants to provide deeper insights into how instruction can be tailored to individual learners—not only in terms of content but also in terms of potentially varying 'treatments'—it is important to clearly define what is meant by different treatments, as well as to identify the key factors that may influence their effectiveness (Reinhold et al., 2024). The treatment itself can, for example, be described as different combinations of instructional strategies and scaffolds, whereas crucial factors not only comprise diverse learner characteristics that may change over time but also varying complexity of the content that is to be learnt, as illustrated in the following: Over recent decades, research has identified several instructional strategies, e.g., learning with analogies, retrieval practice, problem solving prior to instruction, or comparing and contrasting solutions, that are grounded in widely accepted learning principles, aimed at optimizing cognitive, metacognitive, or motivational processes. Instructional strategies vary in how they are implemented (e.g., Schneider & Preckel, 2017), and students may not benefit equally due to individual differences in the cognitive, metacognitive, or motivational resources needed for learning (e.g., Hofer et al., 2018; Reinhold et al., 2020; Stern, 2017).

Other approaches that aim to accommodate individual learner needs, such as scaffolding that provides targeted cognitive, metacognitive, or motivational support, may as well require integrating information on multiple learner characteristics. For example, learners may not have the necessary prior (domain) knowledge or attentional resources to process dynamic visualizations that support learning with analogies, cases in which additional cognitive scaffolding, such as signaling, can help direct attention towards key features in dynamic visualizations. This approach of integrating multiple types of scaffolding is referred to as layered scaffolding (Hofer & Reinhold, 2025). Designing such layered scaffolding requires nuanced information about learner characteristics including information about more than one learner characteristic as well as information about how each of these characteristics varies dynamically over time or during the learning process. This underscores the importance of employing research designs and applying statistical methods that can capture, analyze and present such detailed insights.

The need of more appropriate statistical approaches for uncovering differential effectiveness is not just based on practical implications (designing treatments that maximize the potential learning gains for diverse students) but also highly relevant for advancing theory on interindividual differences in learning generally. By identifying specific learner characteristics that are predictive of learning under specific treatment parameters, we gain insight into the cognitive mechanisms and potential prerequisites (e.g. Breitwieser & Brod, 2021) that are at play to allow learners to make use of the offered treatment. Thus, establishing the use of statistical methods that increase the rigor of differential educational effects investigations and their results, will contribute equally to maturing individual differences in learning theory and providing practical design implications for adaptive instruction.

6. Future outlook

While education may have moved from assigning one and the same treatment to all learners to assigning specific treatments to specific learners, we believe that educational practice will have to move towards assigning a more or less general treatment to all learners and continuously and repeatedly adapting it to specific learners, accommodating learners changing needs during and in interaction with the learning process (Tetzlaff et al., 2021). This would not invalidate the use of cross-sectional modeling. Instead, this perspective moves the focus towards more fine-grained aptitudes (specific knowledge component vs. general prior domain knowledge; state motivation vs. trait interest) and treatments (situation specific scaffolding vs. general assistance/guidance), further exacerbating issues 3 (power) and 4 (higher-order interactions). Combining such a dynamic view of aptitudes with the affordances of log data (e.g. Goldhammer et al., 2017; Goldhammer & Zehner, 2017) allows specific solution processes to be part of the aptitude estimation. Instead of assigning a specific treatment for e.g. "high WMC learners" it is possible to adapt a given treatment for e.g. "exploration-based solvers". This is especially important in more complex learning environments that allow for self-regulated learner behavior or agency in the learning/solution process. Besides more fine-grained aptitude estimation, log data also allow for a more detailed analysis of treatment effects by looking at whether a treatment is actually utilized instead of just offered (Helmke & Weinert, 1997; Reinhold et al., 2024), or by

investigating which parts of an intervention actually mediate the learning outcome (e.g. Kristensen et al., 2024). This information can in turn be used to inform micro- adaptations for a specific learner at a specific time point within a given treatment (Plass & Pawar, 2020; Tetzlaff et al., 2021).

One challenge for future research on differential effectiveness is the need to disentangle psychological attributes of learners (e.g., prior knowledge, working memory, reasoning ability) from demographic attributes of learners (e.g., gender, race, age, immigrant status). Although both kinds of attributes might correlate if demographic attributes are related to different opportunities to develop psychologically and educationally, they should never be considered inherently connected. For instance, for decades, being male was associated with higher scores on a variety of educational outcomes (Rosser, 1989), but today, that gender gap has largely closed, and even reversed in U.S. schools (Reardon et al., 2019). Analogously, the current achievement gap across Black and White students in the U.S appears to be shrinking (e.g., Henry et al., 2020), and may (hopefully) close in the future. For this reason, the consideration of demographic variables like gender or race as in-themselves producing differences in aptitude for learning is neither scientifically sound, nor ethically appropriate. To put it another way, a mediational indirect pathway may exist where demographics can be related to educational opportunity, which is in-turn related to psychological aptitudes, but the direct pathway between demographics and aptitudes should ethically be assumed to be zero (Dumas & Mcneish, 2017).

In addition to the methods outlined here, we also suggest combining theory-guided modeling with machine learning approaches (Bosch, 2021). The traditional assumption of nomothetic, one-size-fits-all true treatment effects underlying much of the previous research does not fit to the focus on heterogeneity that is currently bringing new momentum to educational research (e.g., Bryan et al., 2021; Moeller, 2021). Recent innovations in personalized intervention research, such as personalized treatment plans (Montoya et al., 2023), dynamic treatment rules (Montoya et al., 2023) or sequential multiple assignment randomized trials (Almirall et al., 2014) have proposed methods of adapting interventions to individual characteristics. These innovations slowly lead to the dawning understanding that mechanisms of causal relations among learner characteristics, treatments, and outcomes, may be idiosyncratic, dependent on the interplay of numerous person-, time- and context-characteristics, and may require sophisticated data-driven machine learning procedures in support of theory-derived statistical models to be better understood (McConnell & Lindner, 2019).

In the absence of personalized interventions, differential effectiveness research aims for causal inference by using experimental designs controlling for all baseline differences between the treatment groups. Within this methodological context, it is important to carefully consider how to control for confounding. Causal inference in experimental designs relies on the principle of controlling for all baseline differences of the treatment groups. This is true for average effect estimates, as well as for conditional effects of moderators (e.g., Rubin, 2005; Steyer et al., 2014). In non-randomized comparisons of treatment conditions, as well as in randomized experiments with systematic missing data that invalidates randomization (e.g., Gomila & Clark, 2022; Rubin, 1976), confounding due to baseline differences can be present that results in selection bias in the effect estimates. Statistical modeling approaches are required that can include covariates (i.e., variables that describe baseline group differences *and* influence the outcome variable) in order to adjust for confounding factors. However, only observed covariates can be controlled and omitted variable bias can complicate the model specification (e.g., Sengewald & Pohl, 2019; Steiner & Kim, 2016). In general, researchers should aim at developing a model of the relations between moderator variables and potential confounders that allows including their main effects and, importantly, their interactions with intervention effects (Yzerbyt et al., 2004) in the model to reduce bias in inferences (Bailey et al., 2024). In this process, researchers should consider that moderator variables are usually observed (i.e., non-experimental) learner characteristics. Consequently, to draw conclusions about the role of learner variables for differential effectiveness, the validity of causal inference for these variables must be ensured (Bansak, 2021).

7. Conclusion

This manuscript focuses on the complexity of modeling interactions as a prime reason for knowledge gaps in educational research on differential effectiveness. By highlighting common methodological issues as well as how they can be tackled by specific modeling approaches, we hope to extend researchers' toolkit to continue pushing this field forward. We believe that taking up these methods, using them in an informed way, as well as paying attention to ethical considerations and taking care to reduce bias in causal inferences, educational theory and practice will continue to benefit greatly from research on differential effectiveness.

CRediT authorship contribution statement

Peter A. Edelsbrunner: Writing – original draft, Software, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Leonard Tetzlaff:** Writing – original draft, Conceptualization. **Katharina M. Bach:** Writing – original draft, Conceptualization. **Denis Dumas:** Writing – original draft, Conceptualization. **Sarah I. Hofer:** Writing – original draft, Conceptualization. **Carmen Köhler:** Writing – original draft, Conceptualization. **Zoya Kozlova:** Writing – original draft, Conceptualization. **Julia Moeller:** Writing – original draft, Conceptualization. **Frank Reinhold:** Writing – original draft, Conceptualization. **Garrett J. Roberts:** Writing – original draft. **Marie-Ann Sengewald:** Writing – original draft, Conceptualization. **Sarah Bichler:** Writing – original draft, Conceptualization.

Acknowledgments

This work resulted from collaboration at the conference "Modeling Individual Differences in Education" (MIDE) hosted at ETH Zurich in 2023 and supported by a grant of the Swiss National Fund under number 216450.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lindif.2025.102812.

References

Acar, S., Dumas, D., Organisciak, P., & Berthiaume, K. (2024). Measuring original thinking in elementary school: Development and validation of a computational psychometric approach. *Journal of Educational Psychology, 116*(6), 953–981. https://doi.org/10.1037/edu0000844

Ackerman, P. L. (2003). Aptitude complexes and trait complexes. *Educational Psychologist, 38*(2), 85–93. https://doi.org/10.1207/S15326985EP3802_3

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. In R. E. Mayer, & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 522–560). Routledge.

Almirall, D., Nahum-Shani, I., Sherwood, N. E., & Murphy, S. A. (2014). Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Translational Behavioral Medicine, 4*(3), 260–274. https://doi.org/10.1007/s13142-014-0265-0

Baayen, R. H., & Linke, M. (2020). An introduction to the generalized additive model. In S. T. Gries, & M. Paquot (Eds.), *A practical handbook of corpus linguistics* (pp. 563–591). Berlin: Springer.

Bach, K. M., Hofer, S. I., & Bichler, S. (2025). Adaptive learning, instruction, and teaching in schools: Unraveling context, sources, implementation, and goals in a systematic review. *Learning and Individual Differences, 124*, 102781. https://doi.org/10.1016/j.lindif.2025.102781

Bach, K. M., Reinhold, F., & Hofer, S. I. (2025). Unlocking math potential in students from lower SES backgrounds – Using instructional scaffolds to improve performance. *Sciene of Learning, 10*, 66. https://doi.org/10.1038/s41539-025-00358-7

Bailey, D. H., Jung, A. J., Beltz, A. M., Eronen, M. I., Gische, C., Hamaker, E. L., & Murayama, K. (2024). Causal inference on human behaviour. *Nature Human Behaviour, 8*(8), 1448–1459. https://doi.org/10.1038/s41562-024-01939-z

Bansak, K. (2021). Estimating causal moderation effects with randomized treatments and non- randomized moderators. *Journal of the Royal Statistical Society Series A: Statistics in Society, 184*(1), 65–86. https://doi.org/10.1111/rssa.12614

Baranger, D. A. A., Finsaas, M. C., Goldstein, B. L., Vize, C. E., Lynam, D., .R., & Olino, T. M. (2023). Tutorial: Power analyses for interaction effects in cross-sectional regressions. *Advances in Methods and Practices in Psychological Science, 6*(3). https://doi.org/10.1177/25152459231187531

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2014). Fitting linear mixed-effects models using lme4. In *arXiv preprint*. arXiv:1406.5823.

Bauer, D. J., & Shanahan, M. J. (2007). Modeling complex interactions: Person-centered and variable centered approaches. In T. Little, J. Bovaird, & N. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 255–283). Routledge.

Belzak, W. C., & Bauer, D. J. (2019). Interaction effects may actually be nonlinear effects in disguise: A review of the problem and potential solutions. *Addictive Behaviors, 94*(4), 99–108. https://doi.org/10.1016/j.addbeh.2018.09.018

Bernacki, M. L., Greene, M. J., & Lobczowski, N. G. (2021). A Systematic Review of Research on Personalized Learning: Personalized by Whom, to What, How, and for What Purpose(s)? *Educational Psychology Review, 33*(4), 1675–1715. https://doi.org/10.1007/s10648-021-09615-8

Bernard, R. M., Borokhovski, E., Schmid, R. F., Waddington, D. I., & Pickup, D. I. (2019). Twenty-first century adaptive teaching and individualized learning operationalized as specific blends of student-centered instructional events: A systematic review and meta- analysis. *Campbell Systematic Reviews, 15*(1–2). https://doi.org/10.1002/cl2.1017

Bichler, S., Schwaighofer, M., Stadler, M., Bühner, M., Greiff, S., & Fischer, F. (2020). How working memory capacity and shifting matter for learning with worked examples—A replication study. *Journal of Educational Psychology, 112*(7), 1320. https://doi.org/10.1037/edu0000433

Bichler, S., Stadler, M., Bühner, M., Greiff, S., & Fischer, F. (2022). Learning to solve ill-defined statistics problems: does self-explanation quality mediate the worked example effect? *Instructional Science, 50*, 335–359. https://doi.org/10.1007/s11251-022-09579-4

Bichler, S., Stadler, M., Bühner, M., Greiff, S., & Fischer, F. (2025). *Worked Examples, Cognitive Aptitudes and the Self-Explanation Mechanism – a Replication and an Exploration (preprint)*. Available at SSRN: https://ssrn.com/abstract=5201128 or. https://doi.org/10.2139/ssrn.5201128.

Boninger, F., Molnar, A., & Saldaña, C. (2020). *Big Claims, Little Evidence, Lots of Money: The Reality Behind the Summit Learning Program and the Push to Adopt Digital Personalized Learning Platforms*. Boulder, CO: National Education Policy Center. Retrieved Sep 2024 from http://nepc.colorado.edu/publication/summit-2020.

Bosch, N. (2021). Identifying supportive student factors for mindset interventions: A two-model machine learning approach. *Computers & Education, 167*, Article 104190. https://doi.org/10.1016/j.compedu.2021.104190

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods, 23*(3), 389–411. https://doi.org/10.1037/met0000159

Breitwieser, J., & Brod, G. (2021). Cognitive prerequisites for generative learning: Why some learning strategies are more effective than others. *Child Development, 92*(1), 258–272. https://doi.org/10.1111/cdev.13393

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1*(3), 473–514. https://doi.org/10.1214/06-BA117

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour, 5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Case, L., Speece, D., Silverman, R., Schatschneider, C., Montanaro, E., & Ritchey, K. (2014). Immediate and long-term effects of Tier 2 reading instruction for first-grade students with a high probability of reading failure. *Journal of Research on Educational Effectiveness, 7*(1), 28–53. https://doi.org/10.1080/19345747.2013.786771

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). London, UK: Erlbaum.

Costache, O., Edelsbrunner, P. A., Becker, E. S., Sticca, F., Staub, F. C., & Goetz, T. (2022). Growth trajectories of intrinsic value beliefs in mathematics and French: Relations with career orientations. *Zeitschrift fur Erziehungswissenschaft: ZfE, 25*(2), 269–291. https://doi.org/10.1007/s11618-022-01095-y

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671–684. https://doi.org/10.1037/h0043943

Cronbach, L. J., & Snow, R. E. (1981). *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. Ardent Media.

Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported aptitude * treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology, 67*(6), 717–724. https://doi.org/10.1037/0022-0663.67.6.717

Daumiller, M., Janke, S., Butler, R., Dickhäuser, O., & Dresel, M. (2023). Merits and limitations of latent profile approaches to teachers' achievement goals: A multi-study analysis. *PLoS One, 18*(4), Article e0284608. https://doi.org/10.1371/journal.pone.0284608

Donnellan, E., Usami, S., & Murayama, K. (2023). *Random item slope regression: An alternative measurement model that accounts for both similarities and differences in association with individual items*. Psychological Methods: Advance online publication. https://doi.org/10.1037/met0000587

Dumas, D., Forthmann, B., & Alexander, P. (2024). Using a model of domain learning to understand the development of creativity. *Educational Psychologist, 59*(3), 143–158. https://doi.org/10.1080/00461520.2023.2291577

Dumas, D., & Kaufman, J. C. (2024). Evaluation is creation: Self and social judgments of creativity across the four-C model. *Educational Psychology Review, 36*(3), Article 107. https://doi.org/10.1007/s10648-024-09947-1

Dumas, D., McNeish, D., & Greene, J. A. (2020). Dynamic measurement: A theoretical–psychometric paradigm for modern educational psychology. *Educational Psychologist, 55*(2), 88–105. https://doi.org/10.1080/00461520.2020.1744150

Dumas, D. G., & McNeish, D. M. (2017). Dynamic Measurement Modeling: Using Nonlinear Growth Models to Estimate Student Learning Capacity. *Educational Researcher, 46*(6), 284–292. https://doi.org/10.3102/0013189X17725747

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics, 21*(2), 553–565. https://doi.org/10.1093/bib/bbz016

Edelsbrunner, P. A., Flaig, M., & Schneider, M. (2023). A Simulation Study on Latent Transition Analysis for Examining Profiles and Trajectories in Education: Recommendations for Fit Statistics. *Journal of Research on Educational Effectiveness, 16*(2), 350–375. https://doi.org/10.1080/19345747.2022.2118197

Edelsbrunner, P. A., & Thurn, C. M. (2024). Improving the utility of non-significant results for educational research: A review and recommendations. *Educational Research Review, 42*, Article 100590. https://doi.org/10.1016/j.edurev.2023.100590

Edinburgh, T., Ercole, A., & Eglen, S. (2023). Bayesian model selection for multilevel models using integrated likelihoods. *PLoS One, 18*(2), Article 0280046. https://doi.org/10.1371/journal.pone.0280046

Faddar, J., & Kjeldsen, C. C. (2022). Perspectives on educational effectiveness in science and mathematics: The role of non-cognitive measures in TIMSS. Introduction to a special issue. *Studies in Educational Evaluation, 75*, Article 101218. https://doi.org/10.1016/j.stueduc.2022.101218

Fuchs, D., & Fuchs, L. S. (2019). On the importance of moderator analysis in intervention research: An introduction to the special issue. *Exceptional Children, 85*(2), 126–128. https://doi.org/10.1177/0014402918811924

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management, 41*(2), 632–643. https://doi.org/10.1177/0149206314525208

Geremew, Y. M., Huang, W. J., & Hung, K. (2024). Fuzzy-set qualitative comparative analysis as a mixed-method and analysis technique: a comprehensive systematic review. *Journal of Travel Research, 63*(1), 3–26. https://doi.org/10.1177/00472875231168619

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating Product Data to Process Data from Computer-Based Competency Assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Hrsg.) (Eds.), *Competence Assessment in Education: Research, Models and Instruments* (pp. 407–425). Springer International Publishing. S.

Goldhammer, F., & Zehner, F. (2017). What to Make Of and How to Interpret Process Data. *Measurement: Interdisciplinary Research and Perspectives, 15*(3–4), 128–132. https://doi.org/10.1080/15366367.2017.1411651

Gomila, R., & Clark, C. S. (2022). Missing data in experiments: Challenges and solutions. *Psychological Methods, 27*(2), 143–155. https://doi.org/10.1037/met0000361

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Grimm, H., Edelsbrunner, P. A., & Möller, K. (2023). Accommodating heterogeneity: The interaction of instructional scaffolding with student preconditions in the learning of hypothesis-based reasoning. *Instructional Science, 51*(1), 103–133. https://doi.org/10.1007/s11251-022-09601-9

Gronau, Q. F., Heathcote, A., & Matzke, D. (2020). Computing Bayes factors for evidence-accumulation models using Warp-III bridge sampling. *Behavior Research Methods, 52*(2), 918–937. https://doi.org/10.3758/s13428-019-01290-6

Haslbeck, J. M., Jover-Martínez, A., Roefs, A. J., Fried, E. I., Lemmens, L. H., Groot, E., & Edelsbrunner, P. A. (2025). Comparing likert and visual analogue scales in ecological momentary assessment. *Behavior Research Methods, 57*, 217. https://doi.org/10.3758/s13428-025-02706-2

Hayes, A. F., & Rockwood, N. J. (2020). Conditional process analysis: Concepts, computation, and advances in the modeling of the contingencies of mechanisms. *American Behavioral Scientist, 64*(1), 19–54. https://doi.org/10.1016/j.brat.2016.11.001

Helmke, A., & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Ed.), *Enzyklopädie der Psychologie. Band Vol. 3: Psychologie der Schule und des Unterrichts* (pp. 71–176). Göttingen: Hogrefe-Verlag. S.

Henry, D. A., Betancur Cortés, L., & Votruba-Drzal, E. (2020). Black–White achievement gaps differ by family socioeconomic status from early childhood through early adolescence. *Journal of Educational Psychology, 112*(8), 1471–1489. https://doi.org/10.1037/edu0000439

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences, 66*, 4–15. https://doi.org/10.1016/j.lindif.2017.11.001

Hofer, S. I., & Reinhold, F. (2025). Scaffolding of learning activities: Aptitude-treatment-interaction effects in math? *Learning and Instruction, 99*, Article 102177. https://doi.org/10.1016/j.learninstruc.2025.102177

Hofer, S. I., Schumacher, R., Rubin, H., & Stern, E. (2018). Enhancing physics learning with cognitively activating instruction: A quasi-experimental classroom intervention

study. *Journal of Educational Psychology, 110*(8), 1175–1191. https://doi.org/10.1037/edu0000266
Hoogeveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A. J., Allen, P. J., … Nilsonne, G. (2023). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior, 13*(3), 237–283. https://doi.org/10.1080/2153599X.2022.2070255
Hunt, D. E. (1975). Person-Environment Interaction: A Challenge Found Wanting Before it was Tried. *Review of Educational Research, 45*(2), 209–230. https://doi.org/10.3102/00346543045002209
Kalyuga, S. (2007). Expertise Reversal Effect and Its Implications for Learner-Tailored Instruction. *Educational Psychology Review, 19*(4), 509–539. https://doi.org/10.1007/s10648-007-9054-3
Kline, R. B. (2023). *Principles and Practice of Structural Equation Modeling*. Guilford Publications.
Knaus, M. C. (2021). A Double Machine Learning Approach to Estimate the Effects of Musical Practice on Student's Skills. *Journal of the Royal Statistical Society Series A: Statistics in Society, 184*(1), 282–300. https://doi.org/10.1111/rssa.12623
Köhler, C., Hartig, J., & Schmid, C. (2021). Deciding between the Covariance Analytical Approach and the Change-Score Approach in Two Wave Panel Data. *Multivariate Behavioral Research, 56*(3), 447–458. https://doi.org/10.1080/00273171.2020.1726723
Kokkinou, E., & Kyriakides, L. (2022). Investigating differential teacher effectiveness: Searching for the impact of classroom context factors. *School Effectiveness and School Improvement, 33*(3), 403–430. https://doi.org/10.1080/09243453.2022.2030762
Kozlova, Z., Bach, K. M., Edelsbrunner, P. A., & Hofer, S. I. (2025). Bringing learners into focus: A systematic review of learner characteristics in AR-supported STEM education. *Learning and Individual Differences, 122*, Article 102727. https://doi.org/10.1016/j.lindif.2025.102727.
Kristensen, J. K., Torkildsen, J. V. K., & Andersson, B. (2024). Repeated mistakes in app-based language learning: Persistence and relation to learning gains. *Computers & Education, 210*, Article 104966. https://doi.org/10.1016/j.compedu.2023.104966
Kulesz, P. A., Roberts, G. J., Francis, D. J., Cirino, P., Walczak, M., & Vaughn, S. (2024). Latent profiles as predictors of response to instruction for students with reading difficulties. *Journal of Educational Psychology, 116*(3), 363. https://doi.org/10.1037/edu0000832
Lawson, A. P., Martella, A. M., LaBonte, K., Delgado, C. Y., Zhao, F., Gluck, J. A., & Mayer, R. E. (2024). Confounded or controlled? A systematic review of media comparison studies involving immersive virtual reality for STEM education. *Educational Psychology Review, 36*(3). https://doi.org/10.1007/s10648-024-09908-8. Article 69.
Lebuda, I., & Benedek, M. (2023). A systematic framework of creative metacognition. *Physics of Life Reviews, 46*(5), 161–181. https://doi.org/10.1016/j.plrev.2023.07.002
Ling, Y., Zhu, P., & Yu, J. (2021). Which types of learners are suitable for augmented reality? A fuzzy set analysis of learning outcomes configurations from the perspective of individual differences. *Educational Technology Research and Development, 69*, 2985–3008. https://doi.org/10.1007/s11423-021-10050-3.
Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person-and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(2), 191–225. https://doi.org/10.1080/10705510902751010
McCarthy, M. A., & Masters, P. I. P. (2005). Profiting from prior information in Bayesian analyses of ecological data. *Journal of Applied Ecology, 42*, 1012–1019. https://doi.org/10.1111/j.1365-2664.2005.01
McConnell, K. J., & Lindner, S. (2019). Estimating treatment effects with machine learning. *Health Services Research, 54*(6), 1273–1282. https://doi.org/10.1111/1475-6773.13212
McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.
Moeller, J. (2021). Averting the next credibility crisis in psychological science: Within-person methods for personalized diagnostics and intervention. *Journal for Person-Oriented Research, 7*(2), 53–77. https://doi.org/10.17505/jpor.2021.23795
Moeller, J. (2025). Why and when you should avoid using z-scores in graphs displaying profile or group differences. *Journal of Person-Oriented Research, 11*(2), 58–78.
Moeller, J., Ivcevic, Z., Brackett, M. A., & White, A. E. (2018). Mixed emotions: Network analyses of intra-individual co-occurrences within and across situations. *Emotion, 18*(8), 1106–1121. https://doi.org/10.1037/emo0000419 [PRE-PRINT: https://osf.io/8mj84/]
Montoya, L. M., van der Laan, M. J., Luedtke, A. R., Skeem, J. L., Coyle, J. R., & Petersen, M. L. (2023). The optimal dynamic treatment rule superlearner: considerations, performance, and application to criminal justice interventions. *The International Journal of Biostatistics, 19*(1), 217–238. https://doi.org/10.1515/ijb-2020-0127
Nörenberg, L., Lazarides, R., Dietrich, J., & Moeller, J. (2022). *Quo vadis personalized learning?* Preprint available from: Aligning technological innovation with evidence-based research on learning and instruction. https://osf.io/preprints/osf/87hfz.
Peteranderl, S., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., & Stern, E. (2023). What skills related to the control-of-variables strategy need to be taught, and who gains most? Differential effects of a training intervention. *Journal of Educational Psychology, 115*(6), 813–835. https://doi.org/10.1037/edu0000799
Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education, 52*(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943
Preacher, K. J., & Sterba, S. K. (2019). Aptitude-by-Treatment Interactions in Research on Educational Interventions. *Exceptional Children, 85*(2), 248–264. https://doi.org/10.1177/0014402918802803
Reardon, S. F., Fahle, E. M., Kalogrides, D., Podolsky, A., & Zárate, R. C. (2019). Gender Achievement Gaps in U.S. School Districts. *American Educational Research Journal, 56*(6), 2474–2508. https://doi.org/10.3102/0002831219843824
Reinhold, F., Hofer, S. I., Hoch, S., Werner, B., Richter-Gebert, J., & Reiss, K. (2020). Digital support principles for sustained mathematics learning in disadvantaged students. *PLoS One, 15*(10), Article e0240609. https://doi.org/10.1371/journal.pone.0240609
Reinhold, F., Leuders, T., Loibl, K., Nückles, M., Beege, M., & Boelmann, J. M. (2024). Learning mechanisms explaining learning with digital tools in educational settings: A cognitive process framework. *Educational Psychology Review, 36*(1). https://doi.org/10.1007/s10648-024-09845-6. Article 14.
Roberts, G. J., Dumas, D. G., McNeish, D., & Coté, B. (2022). Understanding the Dynamics of Dosage Response: A Nonlinear Meta-Analysis of Recent Reading Interventions. *Review of Educational Research, 92*(2), 209–248. https://doi.org/10.3102/00346543211051423
Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science, 4*(2), 1–19. https://doi.org/10.1177/25152459211007368
Rosser, P. (1989). *The SAT Gender Gap: Identifying the Causes*. Center for Women Policy Studies. Retrieved Sep 2024 from https://eric.ed.gov/?id=ED311087.
Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581
Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331. https://doi.org/10.1198/016214504000001880
Sarma, A., & Kay, M. (2020). Prior setting in practice: Strategies and rationales used in choosing prior distributions for Bayesian analysis. In *In Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–12). , April.
Scherbakova, A., Dumas, D., Acar, S., Berthiaume, K., & Organisciak, P. (2024). Performance and perception of creativity and academic achievement in elementary school students: A normal mixture modeling study. *The Journal of Creative Behavior, 58*(2), 245–261. https://doi.org/10.1002/jocb.646
Scherer, R., & Nilsen, T. (2019). Closing the gaps? Differential effectiveness and accountability as a road to school improvement. *School Effectiveness and School Improvement, 30*(3), 255–260. https://doi.org/10.1080/09243453.2019.1623450
Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin, 143*(6), 565–600. https://doi.org/10.1037/bul0000098
Schwaighofer, M., Vogel, F., Kollar, I., Ufer, S., Strohmaier, A., Terwedow, I., & Fischer, F. (2017). How to combine collaboration scripts and heuristic worked examples to foster mathematical argumentation–when working memory matters. *International Journal of Computer-Supported Collaborative Learning, 12*, 281–305. https://doi.org/10.1007/s11412-017-9260-z
Sengewald, M. A., & Mayer, A. (2024). Causal effect analysis in nonrandomized data with latent variables and categorical indicators: The implementation and benefits of EffectLiteR. *Psychological Methods, 29*(2), 287–307. https://doi.org/10.1037/met0000489
Sengewald, M.-A., & Pohl, S. (2019). Compensation and amplification of attenuation bias in causal effect estimates. *Psychometrika, 84*(2), 589–610. https://doi.org/10.1007/s11336-019-09665-6
Snow, R. E., & Farr, M. J. (2021). Aptitude, Learning, and Instruction. In *, 3. Conative and Affective Process Analyses*. Taylor & Francis.
Stein, M. I. (1953). Creativity and Culture. *The Journal of Psychology, 36*(2), 311–322. https://doi.org/10.1080/00223980.1953.9712897
Steiner, P. M., & Kim, Y. (2016). The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of Causal Inference, 4*(2), 20160009.
Stern, E. (2017). Individual differences in the learning potential of human beings. *npj Science of Learning, 2*(1). https://doi.org/10.1038/s41539-016-0003-0
Sternberg, R. J., & Grigorenko, E. L. (1997). Are cognitive styles still in style? *American Psychologist, 52*(7), 700–712. https://doi.org/10.1037/0003-066X.52.7.700
Steyer, R., Mayer, A., & Fiege, C. (2014). Causal inference on total, direct, and indirect effects. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 606–631). Dordrecht: Springer.
Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics*. Boston, MA: Pearson.
Tetzlaff, L., Edelsbrunner, P., Schmitterer, A., Hartmann, U., & Brod, G. (2023). Modeling Interactions Between Multivariate Learner Characteristics and Interventions: A Person- Centered Approach. *Educational Psychology Review, 35*(4), 112. https://doi.org/10.1007/s10648-023-09830-5
Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing Personalized Education: A Dynamic Framework. *Educational Psychology Review, 33*(3), 863–882. https://doi.org/10.1007/s10648-020-09570-w
Tetzlaff, L., Simonsmeier, B., Peters, T., & Brod, G. (2025). A cornerstone of adaptivity–A meta-analysis of the expertise reversal effect. *Learning and Instruction, 98*, Article 102142. https://doi.org/10.1016/j.learninstruc.2025.102142
Umlauf, N., Klein, N., & Zeileis, A. (2018). BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond). *Journal of Computational and Graphical Statistics, 27*(3), 612–627. https://doi.org/10.1080/10618600.2017.1407325
Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development, 85*(3), 842–860.

van Doorn, J., Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2023). Bayes Factors for Mixed Models. *Computational Brain & Behavior, 6*(1), 1–13. https://doi.org/10.1007/s42113-021-00113-2

Vandekerckhove, J., Matzke, D., & Wagenmakers, E. J. (2015). Model comparison and the principle of parsimony. In J. R. In, Z. Busemeyer, J. T. Whang, & A. E. Townsend (Eds.), *The Oxford Handbook of Computational and Mathematical Psychology* (pp. 300–317). Oxford University Press.

Vaughn, S., Capin, P., Scammacca, N., Roberts, G., Cirino, P., & Fletcher, J. M. (2020). The critical role of word reading as a predictor of response to intervention. *Journal of Learning Disabilities, 53*(6), 415–427. https://doi.org/10.1177/0022219419891412

Vaughn, S., Roberts, G. J., Miciak, J., Taylor, P., & Fletcher, J. M. (2019). Efficacy of a word- and text-based intervention for students with significant reading difficulties. *Journal of Learning Disabilities, 52*(1), 31–44. https://doi.org/10.1177/0022219418775113

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics, 70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wood, S. N. (2017). Generalized Additive Models: An Introduction with R. In *Chapman and Hall/CRC* (Second Edition (2.Ed.)). https://doi.org/10.1201/9781315370279

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., … Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature, 573*(7774), 364–369. https://doi.org/10.1038/s41586-019-1466-y

Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology, 40*(3), 424–431. https://doi.org/10.1016/j.jesp.2003.10.001

Ziegler, E., Edelsbrunner, P. A., & Stern, E. (2021). The benefit of combining teacher-direction with contrasted presentation of algebra principles. *European Journal of Psychology of Education, 36*(1), 187–218. https://doi.org/10.1007/s10212-020-00468-3