Secondary Publication

Stallasch, Sophie E.; Lüdtke, Oliver; Artelt, Cordula; u. a.

Single- and Multilevel Perspectives on Covariate Selection in Randomized Intervention Studies on Student Achievement

Date of secondary publication: 18.11.2025

Version of Record (Published Version), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-111388x

Primary publication

Stallasch, Sophie E.; Lüdtke, Oliver; Artelt, Cordula; u. a. (2024): Single- and Multilevel Perspectives on Covariate Selection in Randomized Intervention Studies on Student Achievement, in: Educational psychology review, Dordrecht [u.a.]: Springer Science + Business Media B.V, Vol. 36, Nr. 4, 112, pp. 1–45, doi: 10.1007/s10648-024-09898-7.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

Educational Psychology Review (2024) 36:112
https://doi.org/10.1007/s10648-024-09898-7

META-ANALYSIS

Single- and Multilevel Perspectives on Covariate Selection in Randomized Intervention Studies on Student Achievement

Sophie E. Stallasch[1] · **Oliver Lüdtke**[2,3] · **Cordula Artelt**[4,5] · **Larry V. Hedges**[6] · **Martin Brunner**[1]

Accepted: 21 May 2024 / Published online: 25 September 2024
© The Author(s) 2024

Abstract

Well-chosen covariates boost the design sensitivity of individually and cluster-randomized trials. We provide guidance on covariate selection generating an extensive compilation of single- and multilevel design parameters on student achievement. Embedded in psychometric heuristics, we analyzed (a) covariate *types* of varying bandwidth-fidelity, namely domain-identical (IP), cross-domain (CP), and fluid intelligence (Gf) pretests, as well as sociodemographic characteristics (SC); (b) covariate *combinations* quantifying incremental validities of CP, Gf, and/or SC beyond IP; and (c) covariate *time lags* of 1–7 years, testing validity degradation in IP, CP, and Gf. Estimates from six German samples ($1868 \leq N \leq 10{,}543$) covering various outcome domains across grades 1–12 were meta-analyzed and included in precision simulations. Results varied widely by grade level, domain, and hierarchical level. In general, IP outperformed CP, which slightly outperformed Gf and SC. Benefits from coupling IP with CP, Gf, and/or SC were small. IP appeared most affected by temporal validity decay. Findings are applied in illustrative scenarios of study planning and enriched by comprehensive Online Supplemental Material (OSM) accessible via the Open Science Framework (OSF; https://osf.io/nhx4w).

Keywords Covariate selection · Design parameters · Individual participant data meta-analysis · Individually and cluster-randomized trials · Power analysis · Student achievement

What works to advance student learning? There is growing social and political call to answer this fundamental question based on sound evidence (Slavin, 2020). This

OSM A-G, all R scripts, and instructions for accessing the analyzed data can be retrieved from this study's OSF repository at https://osf.io/nhx4w. This project is accompanied by the multides R package (Stallasch, 2024).

Extended author information available on the last page of the article

has put a spotlight on randomized trials (RTs), which allow for causal inferences on the actual effects of educational interventions (Whitehurst, 2012). Individually randomized trials (IRTs) that randomly assign individual students to experimental conditions are imperative for conceiving and testing deliberate programs (e.g., Kelly et al., 2013). Validating a program's benefit in real-life schooling then requires upscaling and implementation in ecologically valid settings (Campbell, 1957; e.g., Gersten et al., 2015). Over half of educational RTs (Connolly et al., 2018) nowadays represent cluster-randomized trials (CRTs) involving random allocation of student groups, such as whole schools. CRT designs not only reflect the fact that educational interventions ought to reach a broader student body and/or operate at the group level by definition (Bloom, 2005), but also map the natural nesting of students within classrooms and schools in institutional contexts (Konstantopoulos, 2012).

Irrespective of whether individual or intact groups of students form the unit of randomization, one constitutive feature of a methodologically high-quality RT is an adequate design sensitivity (Lipsey, 1990), meaning sufficient statistical power $1-\beta$ to detect a treatment effect at significance level α with a high level of statistical precision (e.g., Dong & Maynard, 2013; Hedges & Rhoads, 2010; Raudenbush et al., 2007). This poses a key challenge—not exclusively, but especially—when planning CRTs: their inherent multilevel data structure often dramatically restricts power and precision, and thus often require large sample sizes (Schochet, 2008). For instance, Stallasch et al. (2021, p. 193) show that a CRT requires 4080 students (68 schools, each with 3 classrooms of 20 students) to detect an effect of $d=0.25$ on fourth graders' mathematics achievement ($\alpha=0.05$, $1-\beta=0.80$). An IRT, in stark contrast, requires only 504 students to detect the same effect. In other words, everything else being equal, CRTs are much more resource-intensive than IRTs.

A promising technique to raise sensitivity in RT designs without inflating the sample size is to statistically control for pre-treatment covariates (e.g., Bloom et al., 2007; Kahan et al., 2014; Porter & Raudenbush, 1987; Raudenbush, 1997; Raudenbush et al., 2007).[1] In the example above, a mathematics pretest that explained 40%/35%/76% of the variance between students/classrooms/schools could reduce the CRT's sample size requirements by almost two thirds to 1440 students (24 schools; Stallasch et al., 2021, p. 193). This scenario underpins that "well-chosen covariates do wonders for power" (Aberson, 2019, p. 135); yet, the effective value of a covariate is dictated by its prognostic performance. Scholars and agencies hence stress the importance of grounding the ideally preregistered decisions about covariate inclusion on a priori theoretical and empirical considerations that are tied to the specific research field in question (e.g., European Medicines Agency [EMA], 2015; Maxwell et al., 2017; Murray, 1998). Meanwhile, firm guidance on covariate choice is scarce (Pocock et al., 2002; Tafti & Shmueli, 2020), often not going beyond general recommendations for correlational thresholds (e.g., Bausell & Li, 2002, pp. 114–115; Cox & McCullagh, 1982; but see Bloom et al., 2007).

[1] This strategy is by no means a recent trend; in fact, it goes back to Fisher's (1932) original formulation of ANCOVA almost one century ago. In the field of agriculture, Fisher (1932, p. 158) evinced how "the precision of the comparison [between successive yields of tea crops] has been increased over six-fold" by adjusting for previously recorded yields. Likewise, other pioneers of modern experimental statistics advocated the use of covariates to increase power and precision in RTs (e.g., Campbell & Stanley, 1963; Cochran & Cox, 1957).

The overall aim of this two-part study is to build thorough empirical guidance on covariate selection to optimize design sensitivity in IRTs and CRTs on student achievement. To this end, we analyze single- and multilevel design parameters for a broad array of outcomes in grades 1–12 by capitalizing on large-scale assessment data from six German samples. Specifically, we quantify impacts of varying (a) covariate types (pretests in the outcome domain, a different domain, and fluid intelligence, as well as sociodemographic measures), their (b) combinations, and (c) time lags to the outcome (1–7 years), alongside the three psychometric heuristics of bandwidth-fidelity (Cronbach & Gleser, 1957), incremental validity (Sechrest, 1963), and validity degradation (Ghiselli, 1956; Humphreys, 1960). Our paper is organized as follows. The Introduction contains a quantitative research review in which we meta-analytically integrate the respective previous empirical evidence. In Part I, we estimate and meta-analytically integrate design parameters, and demonstrate their use in sample size and power computations. In Part II, we use the estimated design parameters in precision simulations to assess the actual covariate returns for the design sensitivity in RTs. This study is accompanied by an extensive OSF repository at https://osf.io/nhx4w. In addition to the full R code, it includes OSM A-G, with (A) expressions of single- and multilevel models; (B) methodology and results related to the quantitative research review; (C) methodology, further results, and manifold application scenarios of study planning related to Part I; (D) methodology and further results related to Part II, as well as interactive Excel workbooks compiling all (E) empirical, (F) meta-analytic, and (G) simulation results.

Statistical Underpinnings

Sufficient design sensitivity is a vital methodological quality criterion of rigorous research (American Psychological Association, 2020; Wilkinson & Task Force on Statistical Inference, 1999). It includes both statistical *power* and statistical *precision* (Zhang et al., 2023). Any RT should have an appropriate probability (commonly 80%, i.e., $1-\beta=0.80$; Cohen, 1988) to detect a true treatment effect.[2] The precision of an RT can be quantified by its minimum detectable effect size (MDES; Bloom, 1995, 2005) depicting the smallest possible significant (at α) standardized effect size (with $1-\beta$), given the sample size. Thus, a small MDES indicates high design sensitivity. The approximate MDES can be written as (Bloom, 2005, pp. 158–160; Dong & Maynard, 2013, pp. 31–32):

$$\text{MDES} = M_{df} SE\left(\overline{Y}_{\text{TG}} - \overline{Y}_{\text{CG}}\right)/\sigma_{\text{T}} \tag{1}$$

[2] Failure to do so may result in an underpowered study that is likely either to miss a meaningful effect or to inflate or even invert the estimate of the true population effect (Gelman & Carlin, 2014; Sims et al., 2022). This would make the RT "uninformative" (Lortie-Forgues & Inglis, 2019) at best and misleading at worst. An overpowered study, the other way around, may waste financial and human resources.

M_{df} reflects the *t*-distributions specific to α and 1−β, with *df* degrees of freedom. For a two-tailed test, $M_{df}=t_{\alpha/2}+t_{1-\beta}$, which converges to 2.8 when $df\geq 20$, given $\alpha=0.05$ and $1-\beta=0.80$ (Bloom, 2006). The term $SE\left(\overline{Y}_{\mathrm{TG}}-\overline{Y}_{\mathrm{CG}}\right)/\sigma_{\mathrm{T}}$ represents the treatment effect's $\overline{Y}_{\mathrm{TG}}-\overline{Y}_{\mathrm{CG}}$ standard error that is standardized by the (pooled) total student population's standard deviation σ_{T} of an achievement outcome *Y*, with TG and CG referring to the treatment and control group, respectively. For instance, MDES = 0.25 means that a standardized treatment effect of at least one quarter of a student-level *SD* in the applied achievement test would be significant under sufficient power (Bloom et al., 2007).[3]

As we show below, $SE\left(\overline{Y}_{\mathrm{TG}}-\overline{Y}_{\mathrm{CG}}\right)/\sigma_{\mathrm{T}}$ is a function of three factors[4]: (a) the sample size, (b) the allocation of the sample to the experimental conditions, and (c) so-called (multilevel) design parameters that quantify the unconditional (i.e., unadjusted) and conditional (i.e., covariate-adjusted) variance (components) in *Y*. Here, a relevant distinction in the assumptions about the (in)dependence of the underlying student sample between IRT and CRT designs is made that has important implications for the MDES.

A single-level IRT randomizes individual students, so that students are sampled independently of each other (i.e., regardless of e.g., school affiliation). Eq. (1) then transforms to (Bloom, 2006, Eq. 14; Dong & Maynard, 2013, p. 45):

$$\mathrm{MDES}_{\mathrm{IRT}}=M_{df}\sqrt{\frac{1-R_{\mathrm{T}}^2}{P_{\mathrm{T}}\left(1-P_{\mathrm{T}}\right)N}} \tag{2}$$

N is the total number of students (i.e., the sum of students *n* in TG and CG; $N=n_{\mathrm{TG}}+n_{\mathrm{CG}}$). Everything else being equal, the larger *N*, the smaller the MDES. P_{T} denotes the proportion of students assigned to TG (i.e., $P_{\mathrm{T}}=n_{\mathrm{TG}}\,/\,N$), where $P_{\mathrm{T}}=0.50$ (i.e., a balanced design; $n_{\mathrm{TG}}=n_{\mathrm{CG}}$) minimizes the MDES. The design parameter R_{T}^2 is of special interest in this study because it quantifies the amount of the total variance σ_{T}^2 in *Y* that can be explained by covariates C_{T}:

$$R_{\mathrm{T}}^2=\left(\sigma_{\mathrm{T}}^2-\sigma_{\mathrm{T}|\mathrm{C_T}}^2\right)/\sigma_{\mathrm{T}}^2 \tag{3}$$

$\sigma_{\mathrm{T}|\mathrm{C_T}}^2$ symbolizes the conditional total student population's variance of *Y*. $df=N-Q_{\mathrm{T}}-2$, where Q_{T} is the number of covariates C_{T}.

Unlike an IRT, a multilevel CRT randomizes groups of students (e.g., whole schools). Consider a two-level CRT (2L-CRT) with students at level (L) 1 nested within schools at L3, and a three-level CRT (3L-CRT) with students at L1 nested

[3] The treatment effect can be expressed by different parametrizations potentially implying different interpretations, for example, as a mean difference between TG and CG in (a) *final* posttest scores vs. (b) *change* from pre- to posttest scores. Whichever of these parametrizations with corresponding interpretation is chosen, the tested hypotheses and target estimands are equivalent under proper random assignment (Wan, 2021).

[4] For derivations, see e.g., Bloom (2005, 2006), Hedges and Rhoads (2010), and Raudenbush (1997).

within classrooms at L2 nested within schools at L3. This clustering implies dependencies among selected subjects—students within the same classroom or school tend to be (often much) more similar than students from distinct ones (Schochet, 2008; Stallasch et al., 2021). The degree of within-cluster similarity is typically expressed by the multilevel design parameters ρ_{L2} and ρ_{L3} (i.e., the intra-class correlation coefficients at L2 and L3), which are the proportions of σ^2_T in Y that is between classrooms within schools and between schools, respectively:

$$\rho_{L2} = \sigma^2_{L2}/\sigma^2_T \tag{4}$$

$$\rho_{L3} = \sigma^2_{L3}/\sigma^2_T \tag{5}$$

For a 2L-CRT, $\sigma^2_T = \sigma^2_{L1} + \sigma^2_{L3}$, and for a 3L-CRT, $\sigma^2_T = \sigma^2_{L1} + \sigma^2_{L2} + \sigma^2_{L3}$, where σ^2_{L1}, σ^2_{L2}, and σ^2_{L3} are the unconditional variances in Y between students within classrooms in schools, between classrooms within schools, and between schools, respectively.

For a 2L-CRT with randomization at L3, Eq. (1) then transforms to (Bloom, 2006, Eq. 21; Dong & Maynard, 2013, p. 33):

$$\mathrm{MDES}_{2L-CRT} = M_{df}\sqrt{\frac{\rho_{L3}(1-R^2_{L3})}{P_{L3}(1-P_{L3})K} + \frac{(1-\rho_{L3})(1-R^2_{L1})}{P_{L3}(1-P_{L3})Kn_{L3}}} \tag{6}$$

For a 3L-CRT with randomization at L3, Eq. (1) transforms to (Bloom et al., 2008, Eq. 3; Dong & Maynard, 2013, p. 52):

$$\mathrm{MDES}_{3L-CRT} = M_{df}\sqrt{\frac{\rho_{L3}(1-R^2_{L3})}{P_{L3}(1-P_{L3})K} + \frac{\rho_{L2}(1-R^2_{L2})}{P_{L3}(1-P_{L3})KJ_{L3}} + \frac{(1-\rho_{L3}-\rho_{L2})(1-R^2_{L1})}{P_{L3}(1-P_{L3})KJ_{L3}n_{L2}}} \tag{7}$$

n_{L2} and n_{L3} are the average numbers of students within classrooms and schools, respectively, J_{L3} is the average number of classrooms within schools, and K is the number of schools (i.e., the sum of schools K in TG and CG; $K = K_{TG} + K_{CG}$). Generally, K exerts greater impact on the MDES than n_{L2} or n_{L3} and J_{L3}: everything else being equal, the larger K, the smaller the MDES. P_{L3} is the proportion of schools assigned to the treatment condition (i.e., $P_{L3} = K_{TG} / K$) with $P_{L3} = 0.50$ minimizing the MDES. Further, everything else held constant, the larger ρ_{L2} and/or ρ_{L3}, the larger the MDES. Since ρ_{L2} and/or ρ_{L3} are fixed, the multilevel design parameters R^2_{L1}, R^2_{L2}, and R^2_{L3} are of particular importance in this study because they quantify the amounts of σ^2_{L1}, σ^2_{L2}, and σ^2_{L3} in Y that can be explained by covariates C_{L1} at the student, C_{L2} at the classroom, and C_{L3} at the school level, respectively[5]:

[5] C_{L1} and C_{L2} are group-mean centered (see e.g., Konstantopoulos, 2012). C_{L2} and C_{L3} may be either covariates directly assessed at L2 and L3 or classroom and school means of L1 covariates, respectively.

$$R^2_{L1} = \left(\sigma^2_{L1} - \sigma^2_{L1|C_{L1}}\right)/\sigma^2_{L1} \tag{8}$$

$$R^2_{L2} = \left(\sigma^2_{L2} - \sigma^2_{L2|C_{L2}}\right)/\sigma^2_{L2} \tag{9}$$

$$R^2_{L3} = \left(\sigma^2_{L3} - \sigma^2_{L3|C_{L3}}\right)/\sigma^2_{L3} \tag{10}$$

$\sigma^2_{L1|C_{L1}}$, $\sigma^2_{L2|C_{L2}}$, and $\sigma^2_{L3|C_{L3}}$ signify the conditional between-student, -classroom, and -school variances, respectively. $df = K - Q_{L3} - 2$, where Q_{L3} is the number of covariates C_{L3}.

Estimates of σ^2 can be obtained through (multilevel) regression (see OSM A). For both IRTs and CRTs, larger R^2 values generally result in smaller MDES values. Adjusting for highly prognostic baseline covariates is thus widely recognized and explicitly recommended to improve design sensitivity and to address chance covariate imbalance (e.g., Coens et al., 2020; EMA, 2015; Moerbeek & Teerenstra, 2016; Porter & Raudenbush, 1987; Raudenbush et al., 2007). Omitting factors which are strongly predictive but not equated between experimental groups may severely bias treatment effect estimates, impair power, and inflate type I error rates (Ciolino et al., 2019; Yang et al., 2020). At the same time, if not correctly done, covariate adjustment has some pitfalls in special cases: First, adjustment is worthless when the loss in *df* captured by each covariate (in CRTs, at the top hierarchical level) outweigh the gain in precision (Kahan et al., 2014; Moerbeek & Teerenstra, 2016). This situation, however, is very rare; the loss in *df* is most often without (practical) consequence unless the sample size is very small (Konstantopoulos, 2012; Maxwell et al., 2017, p. 501). Second, adjustment might be detrimental when the assumption of covariate-treatment orthogonality[6] is (severely) violated. This risk is amplified with covariates measured after randomization, which could therefore be affected by the treatment, as well as in (very) small-sized RTs, (highly) unbalanced designs (i.e., $n_{TG} \neq n_{CG}$ and/or in CRTs, unequal cluster sizes), and with (much) missing data on covariates (Kahan et al., 2014; Lin, 2013; Moerbeek, 2006; J. Wang, 2020). Violations of covariate-treatment orthogonality may be compensated in the RT design stage by imposing further balancing methods (e.g., matching, minimization, stratification; Moerbeek & Teerenstra, 2016), and in the RT analysis stage by modeling covariate-treatment interactions, optimally using a robust *SE* estimator (Lin, 2013; J. Wang, 2020). Either way, it is of utmost importance to exclusively control for carefully a priori selected *pre-treatment* covariates. Non-prognostic, poorly chosen, or post-treatment covariates likely act as "bad controls" that pose a threat to the validity of results (Cinelli et al., 2022; Kahan et al., 2014; Moerbeek, 2006; Montgomery et al., 2018; Porter & Raudenbush, 1987).

[6] Under this basic principle of randomization, covariates affect the *SE* but not the magnitude, direction, or meaning of the treatment effect (Maxwell et al., 2017; Porter & Raudenbush, 1987; Wan, 2021; see also Cohen et al., 2003, pp. 68–69).

Theoretical and Empirical Considerations on Covariate Selection

Well-founded decisions on pre-treatment covariates are key to designing strong RTs. Scholars and agencies agree that these ideally preregistered decisions should be justified by both substantive theory and empirical results (Committee for Proprietary Medicinal Products, 2004; Cook, 2005; EMA, 1998, 2015; Food and Drug Administration, 2021; Maxwell et al., 2017; Moerbeek & Teerenstra, 2016; Murray, 1998; Raab et al., 2000; Tafti & Shmueli, 2020; Wright et al., 2015). Following this recommendation, the present paper draws on prominent models of school learning (Haertel et al., 1983; M. C. Wang et al., 1993) as well as connects to and expands upon previous empirical studies that examine the impact of covariates on design sensitivity in RTs with student achievement as the target outcome (for an overview, see Stallasch et al., 2021): specifically, student achievement is a multifaceted, complex construct influenced by a myriad of cognitive and non-cognitive (e.g., motivational or sociodemographic) factors (see also Steinmayr et al., 2014; Winne & Nesbit, 2010), of which the following were highlighted as the most important. First, a measure of prior knowledge in the same domain as the outcome (e.g., previous mathematics skills predicting future mathematics skills), which we refer to as a domain-identical pretest (IP), is known to shape performance trajectories (e.g., Ausubel, 1968; Dochy et al., 1999). This view is rooted in the assumption that one's pre-existing knowledge base fundamentally molds input integration during knowledge acquisition (Brod, 2021; Woolfolk, 2020). Second, a measure of cognitive prerequisites in a certain domain may also explain achievement differences in another domain (e.g., previous language or reading skills predicting future mathematics skills; Peng et al., 2020; Ünal et al., 2023), which we refer to as a cross-domain pretest (CP). This idea is supported by the fact that scores from distinct achievement tests are often highly correlated (Baumert et al., 2009), reflecting the operation of a common cognitive capacity (often described as the g factor; Jensen, 1993) or the relevance of a specific ability to tasks in other domains (e.g., reading comprehension is needed to create a mental representation of mathematical problems; Kintsch, 1998). Third, there is broad consensus that fluid intelligence (Gf) is a powerful predictor of achievement in various domains (e.g., Cattell, 1987; Jensen, 1993; Neisser et al., 1996). Finally, sociodemographic characteristics (SC) such as gender, migration background, and socioeconomic status are also widely acknowledged as persistent precursors for academic success (e.g., Bradley & Corwyn, 2002; Stanat & Chistensen, 2006).

Importantly, educational RTs often address outcomes in multiple domains (Lortie-Forgues & Inglis, 2019; Morrison, 2020) that might need to be adapted or expanded during implementation (e.g., due to logistic or financial reasons, or political decisions; see Bloom et al., 2007), and often span several years (Connolly et al., 2018; Rickles et al., 2018). Moreover, apart from the rule that RTs should always be designed as parsimoniously as possible, they are usually subject to limited resources. Thus, in practice, researchers planning RTs often face the challenge of weighing the potential trade-offs between the different covariate types, their combinations, and time lags for design sensitivity. Three influential, albeit debated, psychometric

heuristics may help to derive predictions on the unique, relative, and incremental impacts of IP, CP, Gf, and SC: (a) the bandwidth-fidelity dilemma, (b) the incremental validity concept, and (c) the validity degradation principle.

In the following, we elaborate on each heuristic under both a theoretical and empirical lens. First, we briefly introduce the respective underlying conception. Figure 1 visualizes the implications for R^2 in student achievement. Second, we systematically review previous evidence on the links between standardized achievement tests and the covariate sets germane to each heuristic. For this purpose, we meta-analytically integrated R^2 as derived from (a) relevant studies providing single-level correlations r_T (i.e., not hierarchically decomposed between students, classrooms, and schools), which are informative for planning IRTs, and (b) available studies compiling multilevel design parameters, which are informative for planning CRTs. For each covariate type, combination, and time lag, we fitted a (multivariate) fixed-effect model with the R package metafor (Viechtbauer, 2010) to summarize the available effect sizes.[7] Figure 2 portrays the *Pooled* R^2 values discussed below (see OSM B for the listing of studies included per covariate set, and details on the methodology and results).

Covariate Types: Bandwidth-Fidelity

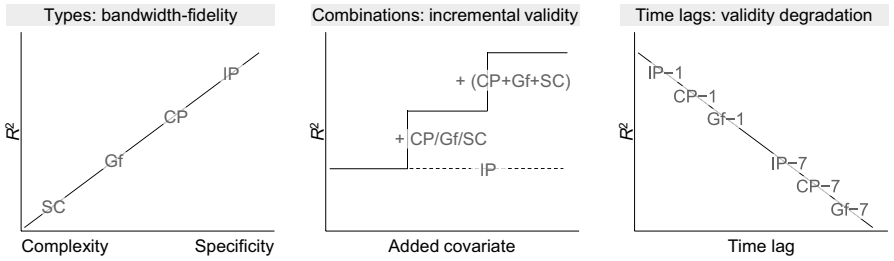
Theoretical Conception

The bandwidth-fidelity dilemma as originally introduced in psychometrics by Cronbach and Gleser (1957) describes an inherent compromise between the complexity (i.e., bandwidth) and the specificity (i.e., fidelity) of a covariate with respect to its predictive validity for an outcome (Hogan & Roberts, 1996; Salgado, 2017). The core idea is that maximal explanatory power requires the alignment of both the conceptual breadths and peculiarities between predictor and outcome (Hogan & Roberts, 1996; Salgado, 2017). Although the heuristic has primarily been discussed for cognitive and personality measures (see Cronbach & Gleser, 1957; Salgado, 2017), it is conceptually not limited to these constructs. Following the underlying rationale, when predicting a domain-specific achievement outcome, IP is expected to be superior to CP because the former matches the outcome domain; yet, as domain-specific cognitive measures, both should be covariates of high fidelity. CP is expected to outperform Gf, as Gf is a domain-general cognitive measure and should be a covariate of lower fidelity/broader bandwidth. Gf is expected to surpass SC, as SC are noncognitive measures and should be covariates of even broader bandwidth.

Previous Empirical Evidence

Single-Level Perspective The studies in our review demonstrated the high predictive power of IP for student achievement, explaining on average 56% of variance. Of

[7] Our analytic procedure implied that the sets of summarized effect sizes varied across meta-analytic models. This explains why *Pooled* R^2 for a covariate subset (e.g., IP+CP) could be sometimes larger than for the full set (i.e., IP+CP+Gf+SC).

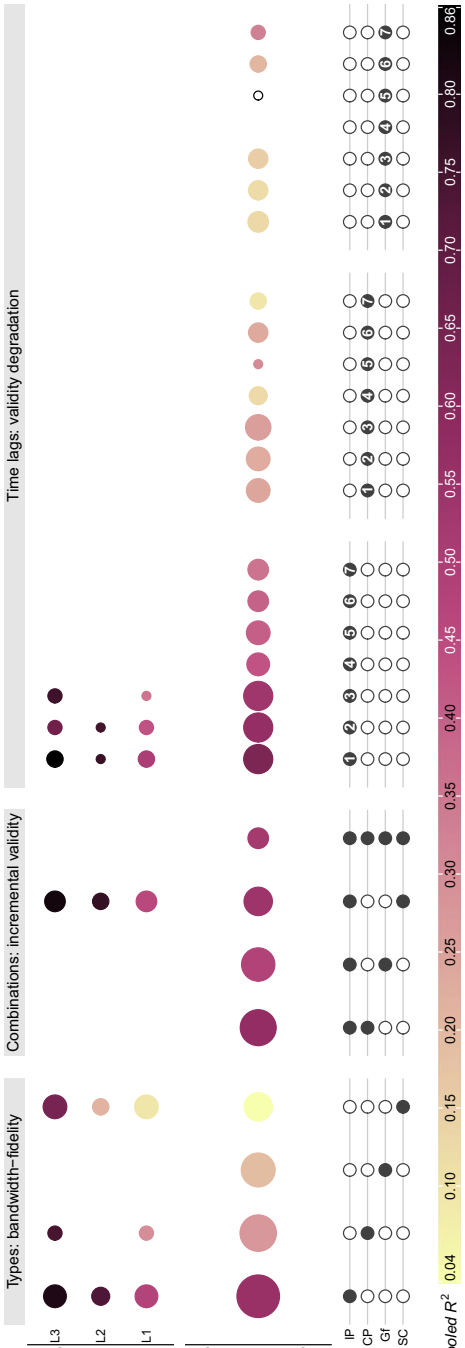


Note. IP = Domain-identical pretest. CP = Cross-domain pretest. Gf = Fluid intelligence pretest. SC = Sociodemographic characteristics. IP/CP/Gf−1 = IP/CP/Gf assessed 1 year before the outcome. IP/CP/Gf−7 = IP/CP/Gf assessed 7 years before the outcome.

Fig. 1 Schematic visualization of the theoretical predictions implied by psychometric heuristics for covariate impacts on R^2 in student achievement

note, while some found that these associations remained fairly stable across grades (e.g., Cole et al., 2011), others showed IP gains in relevance with older students (e.g., McCoach et al., 2017). CP was half as effective as IP (*Pooled* $R^2_{T|CP} = .28$). Gf turned out to be a significant predictor, with *Pooled* $R^2_{T|Gf}$ equaling 0.19. SC explained a meaningful but—relative to the cognitive covariates—small proportion of variance of about 4%. Importantly, for all covariate types, there was substantial between-study variation. For example, $R^2_{T|IP}$ ranged broadly from 0.17 to 0.73, due to variation across grade levels and/or domains and pre-posttest time lags. To conclude, the reviewed single-level evidence generally supports the theoretical predictions on the differential impacts of covariate types with varying bandwidth-fidelity.

Multilevel Perspective Across studies, IP appeared to be the most powerful covariate type, explaining an astonishing 73%/81% of achievement differences at L2/L3, and 48% at L1. In Hedges and Hedberg (2013), the prognostic value of IP at L3 strengthened throughout the school career, a trend that was replicated repeatedly (see Stallasch et al., 2021, Fig. 1). Despite domain mismatch, CP proved a highly robust predictor, particularly at L3: *Pooled* $R^2_{L3|CP}$ amounted to 0.74, whereas *Pooled* $R^2_{L1|CP}$ was 0.30. As far as we are aware, the predictive capacity of Gf has not yet been partitioned into its hierarchical variance components. SC exerted substantial predictive power at L3 (*Pooled* $R^2_{L3|SC} = 0.64$), but rather limited predictive properties at L1 (*Pooled* $R^2_{L1|SC} = 0.10$) and L2 (*Pooled* $R^2_{L2|SC} = 0.21$). For every covariate and at each hierarchical level, we recorded notable cross-study heterogeneity (e.g., $0.23 \leq R^2_{L1|IP} \leq 0.58$, $0.49 \leq R^2_{L2|IP} \leq 0.70$, and $0.54 \leq R^2_{L3|IP} \leq 0.83$). In sum, the available multilevel evidence fit the assumptions about the differential effects of covariate types of varying bandwidth-fidelity quite well. Yet, compared to the single-level findings, the respective differences in R^2 seemed far less pronounced, especially at the group levels.



Note. Multivariate fixed-effect meta-analysis with correlated effect sizes (with an assumed within-study correlation of $r = 0.90$). For single-level designs, we reviewed in total $S = 44$ studies, with $H = 53$ independent samples yielding $G = 1633$ correlations between all covariate sets and achievement outcomes which were transformed into R^2 effect sizes. Note that only Stern (2009) provided one single effect size for Gf–5, thus, no meta-analytic average could be computed. For multilevel designs, we reviewed in total $S = 12$ studies, with $H > 200$ independent samples yielding $G = 2394$ R^2 effect sizes for all covariate sets and achievement outcomes. See OSM B for details on studies, methodology, and results. On the x-axis, a filled/empty dot marks the in-/exclusion of a covariate, where a numbered dot specifies the pre-posttest time lag in years. IP = Domain-identical pretest. CP = Cross-domain pretest. Gf = Fluid intelligence pretest. SC = Sociodemographic characteristics.

Fig. 2 Previous research on covariate impacts: meta-analytic *Pooled* R^2 in student achievement for single- and multilevel designs

Covariate Combinations: Incremental Validity

Theoretical Conception

Incremental validity (Sechrest, 1963) refers to a measure's capacity to additionally explain variance in an outcome beyond what is explained by other prognostic factors (Haynes & Lench, 2003; Hunsley & Meyer, 2003) by contrasting a covariate combination with a subset (Haynes & Lench, 2003). As outlined above, IP is the best-known predictor of domain-specific student achievement. When planning RTs, an important question is therefore whether IP plus CP, Gf, and/or SC jointly explain more variance than IP alone.

Previous Empirical Evidence

Single-Level Perspective Averaged across the reviewed studies, CP contributed to the prediction of student achievement beyond IP, albeit to a small degree; the joint effect computed to *Pooled* $R^2_{\text{T|IP+CP}} = 0.57$. Overall, Gf showed no additional benefits over and above IP (*Pooled* $R^2_{\text{T|IP+CP}} = 0.48$). Combining IP and SC did also not lead to a general improvement over controlling for IP alone (*Pooled* $R^2_{\text{T|IP+SC}} = 0.55$). Taken together, the full covariate battery did not raise the amount of explained variance beyond IP (*Pooled* $R^2_{\text{T|IP+CP+Gf+SC}} = 0.52$). At the same time, for all covariate combinations, incremental returns occasionally reached meaningful thresholds, peaking at +15% of variance explanation over and above IP for the complete covariate array (Chu et al., 2018). The largest increments occurred consistently with elementary school samples, potentially implying that the incremental validities of CP, Gf, and/or SD might be stronger in younger than older students.

Multilevel Perspective We found no multilevel study quantifying incremental validities of CP or Gf, or their combination with SC over and above IP. Much more is known about the set of SC, which incrementally predicted student achievement after IP had been taken into account, although only at the group levels. Pooled across studies, the joint amounts of explained variance equaled 83% at L3, 77% at L2, and 46% at L1. Stallasch et al.'s (2021) analyses revealed that SC contributed around +21%/+13% to the prediction of L2/L3 achievement differences beyond IP, where additional returns appeared to be more pronounced in elementary than secondary school.

Covariate Time Lags: Validity Degradation

Theoretical Conception

The validity degradation principle (Ghiselli, 1956; Humphreys, 1960) implies that the amount of variance explained by a cognitive predictor steadily decreases with growing time lags to the outcome (Hulin et al., 1990; Keil & Cortina, 2001; Reeve

& Bonaccio, 2011). The developmental dynamics underlying validity degradation can be described as a simplex time series pattern (Humphreys, 1960). Accordingly, for domain-specific student achievement as outcome, the explanatory power of IP, CP, and Gf assessed 1 year ago should be higher than the explanatory power of IP, CP, and Gf assessed, for example, 7 years ago.[8]

Previous Empirical Evidence

Single-Level Perspective The vast majority of reviewed investigations suggest that $R^2_{T|IP}$ in student achievement decreases with greater pre-posttest time lags: values considerably dropped from *Pooled* $R^2_{T|IP-1} = 0.63$ to *Pooled* $R^2_{T|IP-7} = 0.36$. Of note, this trend holds true for all grade levels (e.g., McCoach et al., 2017). Analogous results—though far less striking—were reported for the predictive properties of CP: *Pooled* $R^2_{T|CP-1} = 0.24$ declined to *Pooled* $R^2_{T|CP-7} = 0.10$. In some studies, however, $R^2_{T|CP}$ barely diminished (e.g., Erbeli et al., 2021) or even increased with growing time lags (e.g., Träff et al., 2020). The few available studies on the potential validity degradation of Gf indicate fairly robust long-term impacts: pooled across studies, Gf−1 explained 13% and Gf−7 explained 33% of achievement differences. In their review, Reeve and Bonaccio (2011) concluded that the decay of Gf's predictive property is subtle at best, even across numerous years.

Multilevel Perspective The few existing studies on multilevel design parameters addressing the temporal validity degradation of covariates attested a notable decrement of explanatory power of IP at L1; *Pooled* $R^2_{L1|IP-1} = 0.50$ declined to *Pooled* $R^2_{L1|IP-3} = 0.35$. Meanwhile, amounts of explained variance at L3 were far less prone to time effects (*Pooled* $R^2_{L3|IP-1} = 0.86$; *Pooled* $R^2_{L3|IP-3} = 0.76$). Only Xu and Nichols (2010) studied temporal declines in IP's predictive power at L2: proportions of explained variance remained at a high level of 70% across two subsequent years. Of note, deteriorations in R^2 seem to be generally more prevalent in elementary than secondary school, especially at L3. To our knowledge, multilevel studies focusing on cross-time validity decay of CP and Gf are lacking to date.

The Present Study

Strong RTs unite cost-efficiency and sophisticated methodology to ensure appropriate design sensitivity. Given that well-selected covariates substantially raise statistical power and precision, evaluation researchers need reliable evidence that substantiates covariate choices by quantifying unique, relative, and incremental yields of the target outcome's most important predictors. We aim to significantly expand the available guidance for IRTs and CRTs on student achievement through a

[8] Note that SC is assumed to be time-invariant (e.g., migration background does change across the lifespan).

comprehensive compilation of reliable single- and multilevel design parameters that were meta-analyzed and applied to simulate precision.[9]

First, both IRTs and CRTs are in their own right cornerstones of evidence-based education. Both designs are frequently implemented (Connolly et al., 2018). However, single-level design parameters on student achievement have not yet been systematically compiled. Indeed, our quantitative research review may be considered a first major step towards this endeavor. Moreover, extant multilevel design parameters remain mostly restricted to two hierarchical levels. To address these gaps, we cover RTs of three different designs: IRTs (with students assumed to be independently sampled), 2L-CRTs (with students nested within schools), and 3L-CRTs (with students nested within classrooms nested within schools).

Second, researchers rely on knowledge about the potential sensitivity-raising effects of specific covariate types, combinations, and time lags. The above research review pointed out that the latest IP is most likely the best among the covariates. Yet, sometimes the inclusion of IP is not feasible, such as when there are multiple outcome domains (e.g., Lortie-Forgues & Inglis, 2019) while testing time is limited, when the outcome changes after the RT has started (e.g., due to political decisions; Bloom et al., 2007, p. 32), when the outcome is subject to strong developmental dynamics and/or presupposes intensive instruction (e.g., reading skills during elementary school), or when individual pretest differences are unlikely to be observed ahead of the intervention (e.g., integral calculus prior to its introduction; Shadish et al., 2002, p. 118). In such situations, CP, Gf, or SC may be meaningful alternatives to IP. However, only a few multilevel studies provide information on the impacts of CP and SC, and none on the impacts of Gf. Beyond that, the combination of IP with CP, Gf, and/or SC may further boost design sensitivity. Past multilevel studies solely assessed incremental validity of SC over and above IP. Further, RTs often span multiple years (e.g., Rickles et al., 2018), especially when long-term intervention effects are of interest. Although the explanatory power of IP, CP, and Gf may be susceptible to temporal decay, prior multilevel studies addressed rather short pre-posttest time lags of 1–3 years to test validity degradation in IP, but not in CP or Gf. To address these gaps, we systematically vary and combine IP, CP, and Gf with 1- to 7-year-lagged data, as well as SC within 11 different covariate sets (in addition to a set 0 without any covariates).

Third, contemporary educational standards refer to a plethora of skills in various domains (National Research Council, 2011; Organisation for Economic Co-operation and Development [OECD], 2018), as do educational RTs (e.g., Morrison, 2020). Past works on multilevel design parameters dealt with a limited number of outcome domains, namely mathematics, science, and reading. To address this gap, we investigate a wide array of eight commonly targeted outcomes from STEM[10] and verbal domains.

[9] This study used, inter alia, the same data as Stallasch et al. (2021), who also reported a small part of the results presented here, namely the two- and three-level results for covariate sets 0, 1, 4, 7, and 9 (see Table 2). However, all single-level results, the multilevel results for the remaining sets, and all meta-analytic integrations are presented for the first time here.

[10] STEM is commonly used to subsume domains of science/technology/engineering/mathematics.

Fourth, educational RTs are conducted all around the globe (Connolly et al., 2018), but existing collections of multilevel design parameters primarily stem from US samples. Estimates for countries whose school system characteristics markedly deviate from those of the United States, such as an (often much) earlier onset of ability-based school-type-tracking as is the case in Germany, are scarce. To address this gap, we capitalize on longitudinal large-scale assessment data from six German probability samples that represent the total student population in elementary (grades 1–4), lower secondary (grades 5–10), and upper secondary school (grades 11–12), as well as the student populations in lower and upper secondary school belonging to the academic and non-academic track.[11]

Finally, many past educational large-scale RTs lacked design sensitivity (Lortie-Forgues & Inglis, 2019). It is therefore essential to reliably judge how the varying covariates types, combinations, and time lags actually affect precision (given the typical desired 80% power). To this end, power analyses contextualizing the respective R^2 values within predefined designs are indispensable: as becomes clear from Eqs. (2), (6), and (7), the MDES is shaped by the interplay of several quantities beyond power and R^2, such as sample size and allocation, and in the multilevel case also values of ρ. Furthermore, since empirical design parameters are tainted with sampling error that may (dramatically) distort power analysis outcomes, proper allowance of uncertainty is best practice (e.g., Jacob et al., 2010; Turner et al., 2004). We consequently ran precision simulations that concede ρ and R^2 uncertainties via a Bayesian rationale to calculate plausible MDES ranges for IRTs and CRTs.

Part I: Two-Stage Individual Participant Data Meta-Analysis—Estimating and Integrating Design Parameters

Method

We briefly sketch the applied methods here (see OSM C for details). We used R 4.2.2 (R Core Team, 2022); package versions are noted in the R scripts.

Large-Scale Assessment Data

Systematic Search To identify German large-scale assessment datasets suitable for analyzing covariate impacts on design sensitivity in RTs on student achievement, we carried out a systematic search in three electronic data repositories (see also Brunner, Stallasch, et al., 2023). Datasets had to meet the following eligibility criteria: (a) representativeness for the German student population, (b) longitudinal design, and (c) assessment of student achievement via standardized tests. We found three

[11] The German secondary school system offers various school types. We differentiate the academic track (most demanding school type: "Gymnasium," up to grade 12) from the non-academic track (subsuming: vocational ["Hauptschule"], intermediate ["Realschule"], and multitrack ["Schule mit mehreren Bildungsgängen"] schools, up to grades 9 or 10; comprehensive school ["Gesamtschule"], up to grades 9, 10, or 12).

large-scale assessments providing data of six independent national probability samples.

National Educational Panel Study (NEPS) NEPS (Blossfeld & Roßbach, 2019) has been tracking multiple cohorts' educational trajectories throughout their lifespan from 2010 to today. We used the data[12] of students from three NEPS starting cohorts: 4-year-olds (in kindergarten) tested through grade 4 (NSC2; NEPS Network, 2020); grade 5 students tested through grade 12 (NSC3; NEPS Network, 2019a); and grade 9 students tested through grade 12 (NSC4; NEPS Network, 2019b). Achievement tests were administered every 1–3 years.

Programme for International Student Assessment (PISA) The PISA cycles 2003 and 2012 were extended as national longitudinal follow-ups in grades 9–10 in Germany (PISA-Konsortium Deutschland, 2006; Reiss et al., 2017). We used the data[13] from PISA-I-Plus 2003, 2004 (PP03; Prenzel et al., 2013), which focuses on students' mathematics and science achievement development and PISA-Plus 2012–2013 (PP12; Reiss et al., 2019), which additionally incorporates a follow-up assessment of reading achievement.

Assessment of Student Achievements in German and English as a Foreign Language (DESI) DESI (DESI-Konsortium, 2008) studied students' verbal achievement during grade 9. We used the DESI data[11] (Klieme, 2012) on verbal skills in German.

Sampling Process and Sample Selection Except for NSC2, all samples were drawn applying a multistage (i.e., multilevel) sampling process where schools were first randomly drawn, followed by at least two intact classrooms per school (Aßmann et al., 2011; Beck et al., 2008; Heine et al., 2017; Prenzel et al. 2006). NSC2 involved sampling kindergarten children and students of the schools that those children entered to ensure representativeness for children entering elementary school (Aßmann et al., 2011).

When studying covariate types and combinations, we drew on the full spectrum of samples. When studying covariate time lags, we drew only on NSC2 and NSC3, as these samples provided longitudinal achievement data across at least three measurement points. As listed in Table 1, we analyzed data from a total of $N = 68{,}502$ students, where sample sizes ranged within 1868 (NSC3, grade 12) $\leq N \leq$ 10,543 (DESI, grade 9), with median cluster sizes of $4 \leq n_{L2} \leq 25$, $14 \leq n_{L3} \leq 50$, and $2 \leq J_{L3} \leq 3$. Note that in grades 11–12, information at L2 did not exist because, in German upper secondary school, the affiliation of students to intact classrooms is usually replaced by a course grouping system catering to students' ability level in a certain school subject (e.g., basic vs. advanced courses).

[12] Provided by the Research Data Center (FDZ) at the Leibniz Institute for Educational Trajectories (LIfBi).
[13] Provided by the FDZ at the Institute for Educational Quality Improvement (IQB).

Table 1 Numbers of students N, classrooms J, and schools K, and median numbers of students per classroom n_{L2}, students per school n_{L3}, and classrooms per school J_{L3}

Grade	Sample	N	J	K	n_{L2}	n_{L3}	J_{L3}
Elementary school							
1	NSC2	6731	1020	374	6	16	2
2	NSC2	6319	986	362	6	15	2
3	NSC2	5554	888	354	6	14	2
4	NSC2	5418	1026	349	4	14	3
Lower secondary school							
7	NSC3	6314	619	268	10	24	2
9	NSC3	4659	631	240	6	20	2
9	DESI	10,543	427	219	25	50	2
10	PP03	6020	275	152	23	42	2
10	PP12	4494	252	134	19	37	2
Upper secondary school							
11	NSC3	2054	n/a	107	n/a	19	n/a
11	NSC4	4565	n/a	175	n/a	26	n/a
12	NSC3	1868	n/a	105	n/a	17	n/a
12	NSC4	3963	n/a	168	n/a	23	n/a
Total		68,502	6124	3007			

Sample sizes refer to the total student population. See Table C10 in OSM C for sample sizes broken down by school track. n/a indicates that information at L2 was not available as students in grades 11 and 12 are not grouped into intact classrooms, but are rather grouped into courses specific to the subject taught

Measures

Achievement Outcomes We analyzed outcomes in three STEM domains, namely mathematics, science, and information and communication technology (ICT), as well as in five verbal domains in German, namely reading, grammar, spelling, vocabulary, and writing.

Covariates We examined four covariate categories: IP, CP, Gf, and SC. We employed reading as CP for STEM outcomes and mathematics as CP for verbal outcomes. Gf was assessed in terms of figural reasoning. IP, CP, and Gf were available with a 1- to 7-year time lag to the outcome, where the smallest pre-posttest gap ranged from 1 to 4 years. SC comprised 4 variables, namely students' gender (0 = male, 1 = female) and migration background (0 = no, 1 = yes) as well as two indicators of socioeconomic status: (1) parents' highest educational attainment was assessed by the greatest number of years of schooling completed (range 9–18) in all studies except the DESI, where the highest school leaving certificate was used; and (2) parents' highest International Socio-Economic Index of Occupational Status (HISEI; Ganzeboom & Treiman, 1996; range: 11–89).

Missing Data

Virtually all measures used in this study contained some missing values. The percent of missings across the datasets varied from 11% (PP03, grade 10) to 42% (NSC2, grade 1).The greatest missing rates occurred in pretests measured in the first two waves of NSC2, as only a small share of kindergarten children continued participating in NEPS after entering elementary school. We performed (groupwise) multilevel multiple imputation and generated 50 multiply-imputed datasets for each sample and grade using the mice (van Buuren & Groothuis-Oudshoorn, 2011) and miceadds (Robitzsch et al., 2021) packages.

Procedure

We applied a two-stage approach to meta-analysis of individual participant data (Brunner, Keller, et al., 2023; see also Brunner, Stallasch, et al., 2023). We estimated and meta-analyzed design parameters for three RT designs, namely single- (individual students), two- (students within schools), and three-level designs (students within classrooms within schools), as well as for three target populations, namely the total, academic track, and non-academic track student populations. Notably, in upper secondary school, only single- and two-level designs were considered due to the lack of L2 information.

Stage 1: Single- and Multilevel Modeling—Estimating Design Parameters We performed single- and multilevel modeling to empirically estimate ρ and R^2. As shown in Table 2, we systematically in- and excluded 1- to 7-year-lagged IP, CP, and Gf, as well as SC within a total of 12 covariate sets, with the number of covariates Q per set ranging between $0 \leq Q \leq 7$. This resulted in up to 363 distinct models per design and population.

Model Fitting For all outcomes, we fitted two model classes separately for each imputation. The first model class consisted of unconditional models without any covariates (set 0). Specifically, for single-level designs, we obtained σ^2_T by taking the outcomes' variances. For multilevel designs, we obtained σ^2_{L1}, σ^2_{L2}, and σ^2_{L3} by specifying two- and three-level random-intercept-only models. The second model class consisted of conditional models with varying covariate types (sets 1–4), combinations (sets 5–8), and time lags (sets 9–11). Specifically, for single-level designs, we obtained $\sigma^2_{T|C_T}$ by specifying single-level regression models. For multilevel designs, we obtained $\sigma^2_{L1|C_{L1}}$, $\sigma^2_{L2|C_{L2}}$, and $\sigma^2_{L3|C_{L3}}$ by specifying two- and three-level random-intercept models. Note that all covariates were assessed at L1. In two-level models, we entered school averages at L3. In three-level models, we entered classroom averages at L2 and school averages at L3. In single-level models, we centered all covariates around their respective total population's means whereas in multilevel models, we applied group-mean centering: L1 covariates were centered around their respective school/classroom means in two-/three-level models and L2 covariate means were centered around their respective school means in three-level models. Single-level modeling was performed using the stats package implemented in base R.

Table 2 Covariate sets analyzed in the present study with numbers of covariates Q, ρ/R^2 effect sizes G, and samples H by design

		Types: bandwidth-fidelity				Combinations: incremental validity				Time lags: validity degradation							
Set	0	1	2	3	4	5	6	7	8	Set							
IP	○	●	○	○	○	●	●	●	●	9	❶	❷	❸	❹	❺	❻	❼
CP	○	○	●	○	○	●	○	○	●	10	❶	❷	❸	❹	❺	❻	❼
Gf	○	○	○	●	○	○	●	○	●	11	❶	❷	❸	❹	n/a	❻	❼
SC	○	○	○	○	●	○	○	●	●								
Q	0	1	1	1	4	2	2	5	7		1	1	1	1	1	1	1
Single-/two-level designs																	
G	34	34	31	31	34	31	31	34	31	9	2	15	6	7	3	1	2
										10	6	14	6	8	3	1	3
										11	5	8	5	7	n/a	1	3
H	6	6	6	6	6	6	6	6	6	9	1	2	2	2	1	1	1
										10	1	2	2	2	1	1	1
										11	1	2	2	2	n/a	1	1
Three-level designs																	
G	26	26	23	23	26	23	23	26	23	9	2	14	3	7	0	0	0
										10	6	13	3	7	0	0	0
										11	5	7	2	7	n/a	0	0
H	5	5	5	5	5	5	5	5	5	9	1	2	2	2	0	0	0
										10	1	2	1	2	0	0	0
										11	1	2	1	2	n/a	0	0

A filled/empty dot marks the in-/exclusion of a covariate, where a numbered dot specifies the pre-posttest time lag in years. Set 0 yielded ρ effect sizes, sets 1–11 yielded R^2 effect sizes. Set 1/2/3 involved the most recently assessed IP/CP/Gf (i.e., with the smallest possible time lag to the outcome, ranging between 1 and 3 years for IP and CP, and between 1 and 4 years for Gf). n/a indicates that the respective covariate was not available. See Table C18 in OSM C for covariate sets broken down by domain area. IP = Domain-identical pretest. CP = Cross-domain pretest (reading for STEM outcomes, mathematics for verbal outcomes). Gf = Fluid intelligence pretest. SC = Sociodemographic characteristics (gender, migration background, socioeconomic status)

Multilevel modeling was performed using the lme4 package (Bates et al., 2015) applying restricted maximum likelihood (REML) estimation.

Calculating Design Parameters and Standard Errors We calculated ρ and R^2 by inserting the variance (component) estimates from the model fits into Eqs. (3)–(5) and (8)–(10). *SE*s of ρ were computed with the formulas for the large sample variances in unbalanced (i.e., with unequal cluster sizes) two-level designs derived in Donner and Koval (1980, Eq. 3) and three-level designs in Hedges et al. (2012, Eqs. 7–9). The latter involves the sampling variances of σ^2_{L2} and σ^2_{L3}, which we obtained by applying the "cases bootstrap" from the lmeresampler package (Loy & Korobova, 2023). We drew 1000 samples (Huang, 2018, p. 303; Schomaker & Heumann, 2018). *SE*s of R^2 were computed with the formula for the large sample variances given in Hedges and Hedberg (2013, p. 451).

Pooling ρ and R^2 with corresponding *SE*s were pooled across the 50 imputations. We used the mitml package (Grund et al., 2021) that employs Rubin's (1987) rules to take into account within- and between-imputation variance.

Stage 2: Meta-Analysis—Integrating Design Parameters We performed meta-analysis to integrate ρ and R^2 for covariate types and combinations, and meta-regression with outcome-covariate time lag as moderator to integrate R^2 for covariate time lags (both across domains and samples, but within hierarchical and grade levels, designs, and populations).[14]

Model Fitting Using the metafor package (Viechtbauer, 2010), we fitted two meta-analytic/meta-regression model classes, conditional on the number of R^2 effect sizes G per covariate set: either (multivariate) fixed-effect models if $G < 10$ or (multivariate multilevel) random-effects models via REML if $G \geq 10$ (see Langan et al., 2019, p. 95). Both methods yield an average (true) effect size *Pooled* R^2, with *SE*(*Pooled* R^2). However, the "real" (i.e., not due to sampling error) heterogeneity among true R^2 values within samples, $\tau^2_{\text{Effect sizes}}$, and between samples, τ^2_{Samples}, can solely be captured by random-effects models (Borenstein et al., 2021, pp. 61–80). We deployed two weighting schemes, conditional on the number of samples H per covariate set: If $H > 1$, we addressed within-sample dependencies among R^2 effect sizes (Hedges, 2019) by multivariate (multilevel) meta-analyses and imputed working variance–covariance matrices using the clubSandwich package (Pustejovsky, 2022). We assumed a within-sample intercorrelation of $r = 0.90$ as a reasonable upper-bound guess (see Brunner, Stallasch, et al., 2023). If $H = 1$, we drew on the sampling variances of R^2 in terms of the standard meta-analytic inverse-variance weighting.

Depicting Heterogeneity With random-effects modeling, we calculated—in addition to the 95% confidence interval (95% CI)—the 95% prediction interval (95% PI). The 95% PI provides a plausible range of R^2; it quantifies the total dispersion (sampling variance plus $\tau^2_{\text{Effect sizes}}$, and if applicable, plus τ^2_{Samples}) of R^2 around *Pooled* R^2 and defines the range in which an R^2 estimated based on data of a new sample randomly drawn from a population of samples will likely (i.e., in 95% of cases) fall (Borenstein et al., 2021, pp. 119–126; Riley et al., 2011). We also calculated (multilevel) I^2 (Higgins & Thompson, 2002), the ratio of real heterogeneity to the total variation across observed R^2 values (Borenstein et al., 2017).

Gauging Sensitivity and Model Convergence For the imputed working variance–covariance matrices, we ran sensitivity analyses over $r \in \{0.00, 0.05, \ldots, 0.95\}$ (Hedges, 2019) to preclude a misspecification of R^2 dependencies. With random-effects modeling, we profiled log-likelihoods of τ^2 values to evaluate their identifiability (see Viechtbauer, 2022).[15]

[14] We concentrate on R^2 as the focus of this study, but all analysis steps described below also applied to ρ.

[15] In all analyses, we did not use sampling weights. After excluding students who did not meet the inclusion criteria for our analyses (see section "Sample Selection" in OSM C), the weights would no longer have been representative for the total German student population.

Key Results

We present major patterns in meta-analytic single- and multilevel (i.e., three-level in grades 1–10 and two-level in grades 11–12) design parameters for the total student population, illustrated in Figs. 3 and 4 (which we refer to in this section, unless otherwise stated; see OSM C for result plots of two-level designs in grades 1–10 and school tracks, and OSM E/F for the full compilation of the empirical/meta-analyzed design parameters).

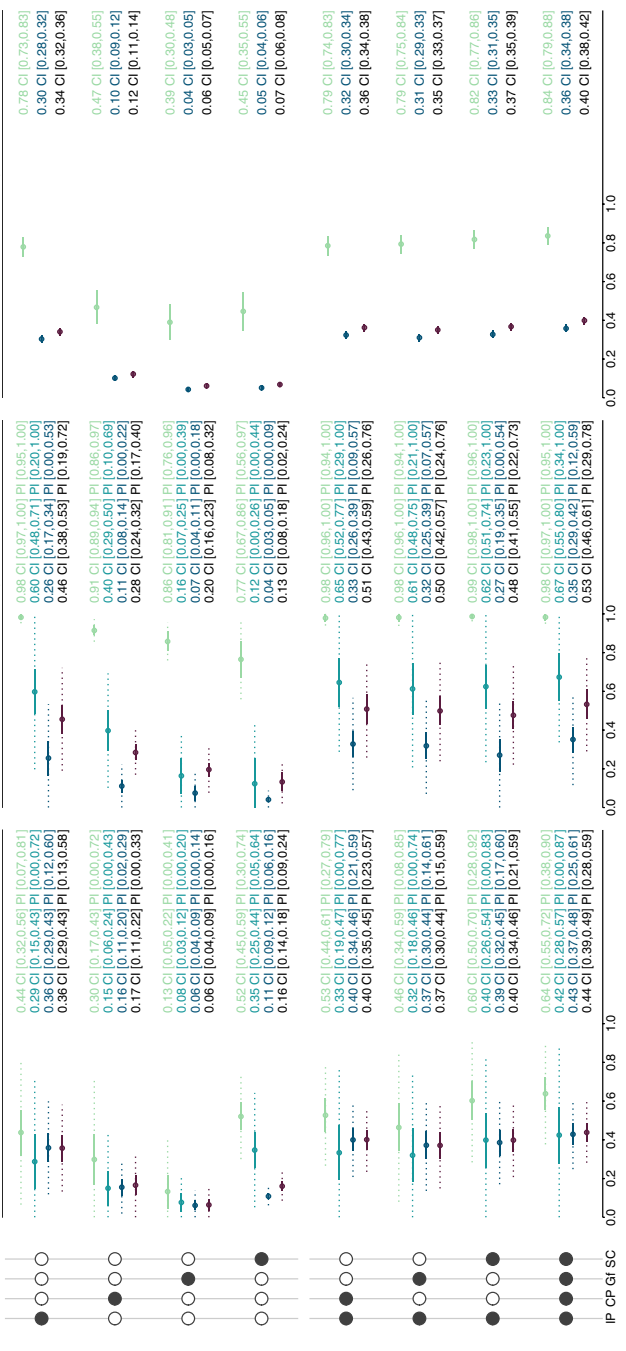
Covariate Types: Bandwidth-Fidelity

Single-Level Perspective IP was consistently the most powerful among all covariate types: IP explained over one third of achievement differences between individual students in elementary and upper secondary school, and almost one half in lower secondary school. Here, the remaining cognitive covariates were also valuable predictors, with *Pooled* $R^2_{T|CP} = 0.28$ and *Pooled* $R^2_{T|Gf} = 0.20$. In elementary and upper secondary school, CP and in particular Gf contributed comparably little to the prediction. In contrast, SC performed best as predictors in elementary school (*Pooled* $R^2_{T|SC} = 0.16$). We registered substantial R^2_T heterogeneities, with the broadest 95% PIs for IP and the narrowest for SC. For example, in elementary school, the 95% PIs were [0.13, 0.58] and [0.09, 0.24], respectively (Table F1).

Multilevel Perspective IP was generally of paramount relevance when predicting student achievement. From grade 5 on, IP was the strongest of all covariate types and showed exceptional prognostic properties at L3, where 98/78% of variance was explained in lower/upper secondary school. Across the entire school career, however, IP turned out to be a weaker predictor at both L1 and L2. CP as well as Gf were very useful to explain differences between schools, in particular in lower secondary school (*Pooled* $R^2_{L3|CP} = 0.91$; *Pooled* $R^2_{L3|Gf} = 0.86$), but less so between classrooms and students—in all grade levels. Gf was moreover consistently the weakest covariate both in elementary and upper secondary school, irrespective of the hierarchical level. Although SC were the poorest predictors in lower secondary school, they still explained over three quarters of between-school variance. Notably, with first–fourth graders, SC outweighed IP at both L2 and L3 (*Pooled* $R^2_{L2|SC} = 0.35$; *Pooled* $R^2_{L3|SC} = 0.52$). Degrees of heterogeneity in multilevel R^2 were often considerable, depending not only on the covariate type but also on the grade and hierarchical level. For instance, the 95% PI for $R^2_{L3|IP}$ was very wide in elementary school with [0.07, 0.81], but considerably narrower in secondary school with [0.95, 1.00] (Tables F1 and F2).

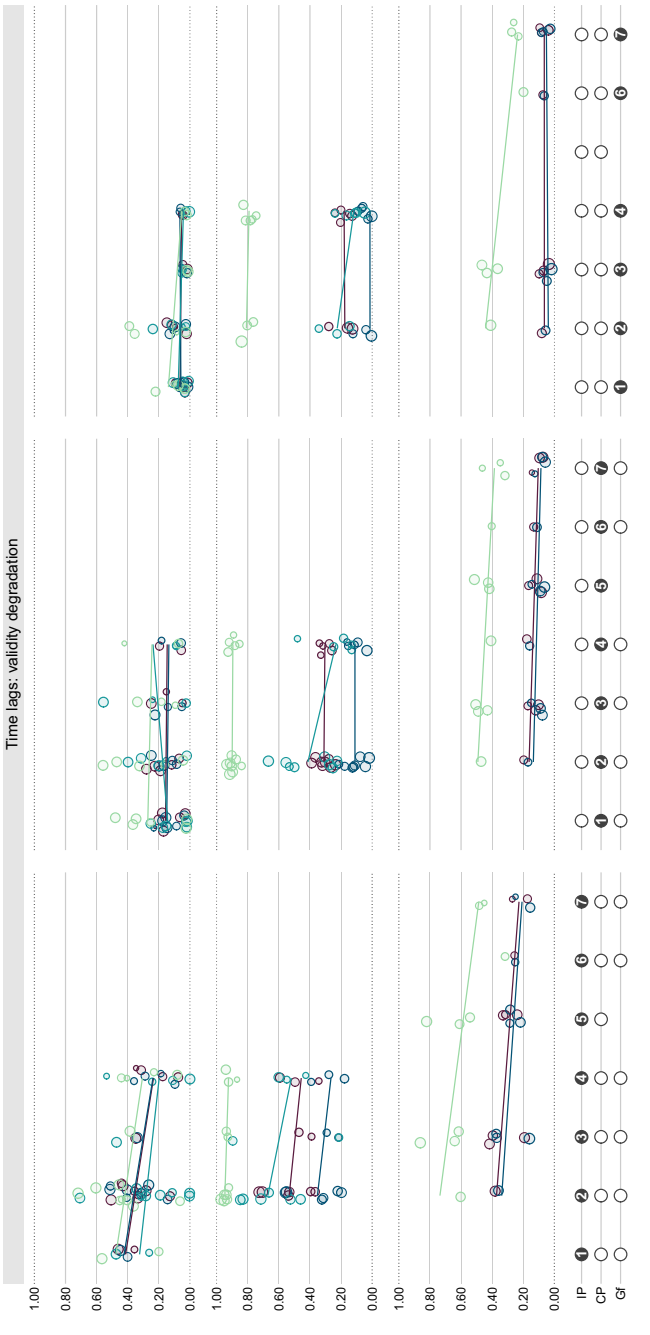
Covariate Combinations: Incremental Validity

Single-Level Perspective In all grade levels, CP explained additional variance in student achievement over and above IP. On average, incremental gains were largest in lower secondary school (+5%) and smallest in upper secondary school (+2%). When



Note. Multivariate fixed-effect (upper secondary school) and (multivariate multilevel) random-effects (elementary and lower secondary school) meta-analysis; dots show *Pooled* R^2; solid/dotted lines represent 95% CIs/PIs. On the y-axis, a filled/empty dot marks the in-/exclusion of a covariate. IP = Domain-identical pretest. CP = Cross-domain pretest (reading for STEM outcomes, mathematics for verbal outcomes). GI = Fluid intelligence pretest. SC = Sociodemographic characteristics (gender, migration background, socioeconomic status).

Fig. 3 Meta-analytic integrations of single- and multilevel R^2 in student achievement: covariate types and covariate combinations



Note. Fixed-effect (set 11 [Gf–lag] throughout secondary school) and random-effects (remaining) meta-regression with time lag as moderator; bubbles show observed R^2 sized by weight; line slopes map b_{lag}. On the x-axis, a filled/empty dot marks the in-/exclusion of a covariate, where a numbered dot specifies the pre-posttest time lag in years. IP = Domain-identical pretest. CP = Cross-domain pretest (reading for STEM outcomes, mathematics for verbal outcomes). Gf = Fluid intelligence pretest. SC = Sociodemographic characteristics (gender, migration background, socioeconomic status).

Fig. 4 Meta-analytic integrations of single- and multilevel R^2 in student achievement: covariate time lags

controlling for IP, benefits through Gf were noteworthy only in lower secondary school ($\Delta Pooled\ R^2_{T|+Gf} = +0.04$). In contrast, SC particularly contributed to the prediction in elementary/upper secondary school, with about +4%/+3%. Joint effects through the full battery of covariates were always largest, with $+0.06 \leq \Delta Pooled\ R^2_{T|CP+Gf+SC} \leq +0.08$. We found signal heterogeneities in all R^2_T. For example, the 95% PI for $R^2_{T|IP+Gf}$ was [0.15, 0.59] (Table F1).

Multilevel Perspective At the school level, CP clearly added to the prediction of achievement differences beyond IP in elementary ($\Delta Pooled\ R^2_{L3|+CP} = +0.09$), but not in secondary ($\Delta Pooled\ R^2_{L3|+CP} \leq +0.01$) school. At L1/L2, CP provided some additional explanatory power over the entire school career; benefits were largest in lower secondary school (+7%/+5%). In general, we found Gf to be a rather poor additional covariate across grade and hierarchical levels. As an exception, contributions in $Pooled\ R^2_{L1}$ over and above IP amounted to +6% in lower secondary school. SC was of notable incremental relevance in elementary school, especially at L3, adding on average +16% of explained variance. In higher grades, additional gains were often negligible though (note, however, that $Pooled\ R^2_{L3|+SC}$ reached 0.99 in lower secondary school). Except for L3 in lower secondary school, the complete set of covariates consistently outweighed all other combinations—at L1, average gains were the highest in lower secondary, and at L2 and L3 in elementary school, with $\Delta Pooled\ R^2_{|+CP+Gf+SC} = +0.09/+0.13/+0.20$ at L1/L2/L3. Multilevel R^2 heterogeneities largely mirrored those of IP, and were substantive (except R^2_{L3} in lower secondary school). For example, the 95% PI of $R^2_{L2|IP+CP}$ was [0.21, 1.00] in grades 5–10 (Table F2).

Covariate Time Lags: Validity Degradation

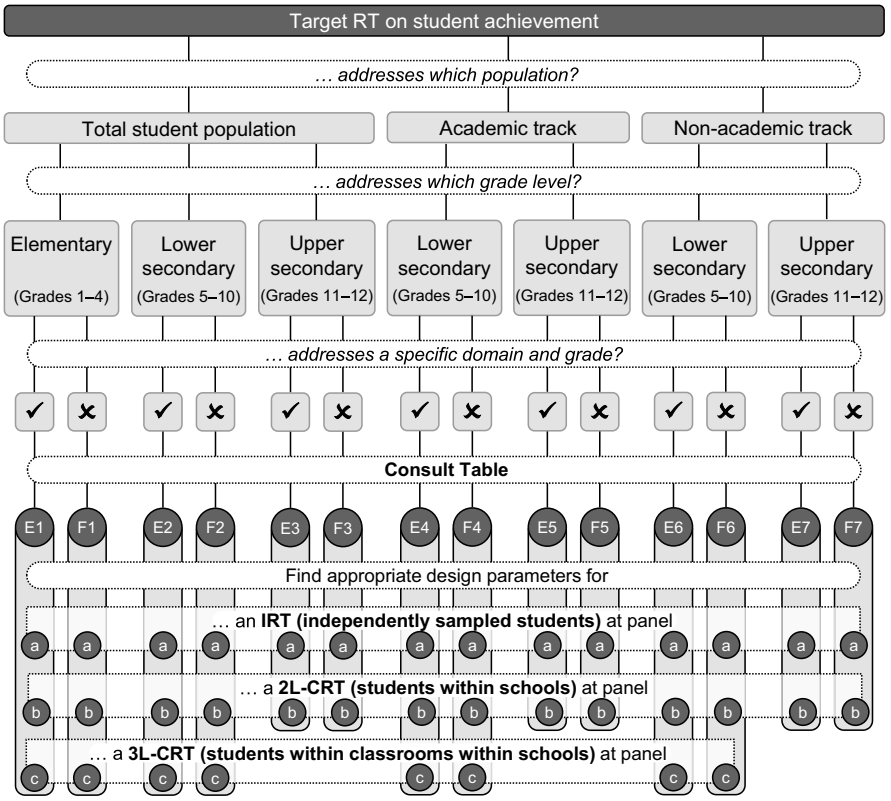
Single-Level Perspective In all grade levels, the predictive power of IP clearly diminished with growing pre-posttest time lags. Validity degradation was most prevalent in elementary school: the meta-regression coefficient $b_{lag} = -0.06$ shows that with each additional year between IP and outcome, $R^2_{T|IP}$ is predicted to decrease by −6%. In lower/upper secondary school, temporal declines in the proportions of explained variance were also noticeable ($b_{lag} = -0.04/-0.03$). In contrast, CP emerged to be far less prone to cross-time decay: until grade 10, prognostic properties remained stable both in elementary and secondary school. In upper secondary school, predicted amounts of explained variance slightly reduced with −1%. Gf turned out to be an extraordinarily time-robust predictor throughout the entire school career.

Multilevel Perspective Validity degradation in $R^2_{|IP}$ was almost always substantial, except for lower secondary school at L3. Here, we recorded remarkable temporal stabilities in the amounts of explained variance ($b_{lag} = -0.01$). In all other cases, the explanatory power of IP is likely to drop about −3% up to −8% per year. CP appeared to be a relatively time-robust covariate; however, decrements in prognostic capacity hinged on both the grade and hierarchical level: in elementary school, solely

predicted $R^2_{L3|CP}$ slightly declined ($b_{lag} = -0.01$); in lower secondary school, only predicted proportions of explained L2 variance dropped, but this strikingly ($b_{lag} = -0.09$); and in upper secondary school, predicted $R^2_{L1|CP}$ ($b_{lag} = -0.01$) and $R^2_{L3|CP}$ ($b_{lag} = -0.02$) showed small reductions over time. While Gf consistently emerged highly time-stable at L1, the predicted validity decay at both L2 and L3 was practically significant in all grade levels ($-1\% \leq b_{lag} \leq -5\%$).

Application

Researchers designing RTs may profit from the flow chart in Fig. 5. It facilitates the choice of single- and multilevel design parameters that are optimally tailored to the specific application context. To showcase the estimates' use in study planning, we developed manifold scenarios to determine the (a) sample size and (b) statistical power of IRTs and CRTs via power analysis. We present one in the following (see OSM C for the remaining).



Note. OSM E is an interactive Excel workbook that contains Tables E1–E7 listing empirically estimated single- and multilevel design parameters. OSM F is an interactive Excel workbook that contains Tables F1–F7 listing meta-analytically integrated single- and multilevel design parameters.

Fig. 5 Flow chart to choose design parameters from our compilation in OSM E and F

An Illustrative Scenario

A research team has programmed an app. It functions as a multidisciplinary digital learning environment which can be used throughout lower secondary school in Germany.

Single-Level Perspective As a first step, the researchers aim to test the general efficacy of the underlying didactic approach. They plan a small-scale pilot IRT involving exclusively mathematical topics from grade 7. A standardized treatment effect of $d = 0.15$ is considered meaningful, representing around one half of the expected annual growth in mathematics for grades 6–7 in the German student population (Brunner, Stallasch et al., 2023, Table 1). The team's objective, therefore, is to sample enough seventh graders to detect MDES = 0.15 at $\alpha = 0.05$ (two-tailed) with $1-\beta = 0.80$, where $P_T = 0.50$. The minimum required sample size (MRSS) to achieve this in an unconditional IRT design is $N = 1397$ students. Striving for parsimony and being aware of the potential virtue of covariate adjustment, the researchers plan to statistically control for IP. Before power analysis, they consult our flow chart (Fig. 5): since the IRT addresses the total population in lower secondary school and a specific grade and domain analyzed in our study, the team is guided to Table E2 (panel a) that lists the suitable empirically estimated single-level design parameters. Inserting $R^2_{T|IP} = 0.53$, the researchers find that the MRSS more than halves to $N = 654$ when adjusting for IP. They then think about optimizing the design by additionally including either a reading CP or SC, where $R^2_{T|IP+CP} = 0.56$ and $R^2_{T|IP+SC} = 0.55$. The MRSS further reduces to $N = 630$ when combining IP with SC and to $N = 622$ when combining IP with CP. They decide to administer both a mathematics and reading test. The team wants to account for uncertainty in $R^2_{T|IP+CP}$. To this end, they determine the 95% CI by means of $SE\ (R^2_{T|IP+CP}) = 0.01$: the lower bound is calculated as $0.56 - 1.96 \times 0.01 = 0.54$ and the upper bound as $0.56 + 1.96 \times 0.01 = 0.57$, which leads to an MRSS range of $649 \geq N \geq 596$. Consequently, when opting for a conservative approach and sampling $N = 649$ students, it is fairly certain that the IRT will be sensitive to uncover a (truly existing) treatment effect of $d = 0.15$ with IP and CP as covariates.

Multilevel Perspective As a second step, the researchers aim to scrutinize the effectiveness of the full app in students' usual school routine. They plan a large-scale 3L-CRT involving the complete spectrum of domains for grades 5–10. $d = 0.11$ is considered reasonable, approximating half of the average academic year-to-year growth observed across lower secondary school in Germany (Brunner, Stallasch, et al., 2023, Table 1). Due to logistical reasons, the total sample is restricted to a maximum of $K = 400$ schools, with $n_{L2} = 20$ and $J_{L3} = 3$. The team's primary concern, thus, is to achieve sufficient power (i.e., $1-\beta \geq 0.80$) to detect MDES = 0.11 at $\alpha = 0.05$ (two-tailed), where $P_{L3} = 0.50$. Since the 3L-CRT addresses the total population in lower secondary school but neither a specific grade nor domain, our flow chart (Fig. 5) directs them to Table F2 (panel c) that lists the suitable meta-analytically integrated three-level design parameters. Entering *Pooled* ρ values at L2/L3 of 0.05/0.35 into power analysis, the researchers learn that an unconditional 3L-CRT

clearly undercuts the desired power rate ($1-\beta=0.43$). They wonder which covariates to use: given the limited testing time, assessing multiple IPs is not a viable option. Instead, controlling for either Gf or SC seems most feasible, with *Pooled* R^2 values at L1/L2/L3 of 0.07/0.16/0.86 for Gf and 0.04/0.12/0.77 for SC. Controlling for both Gf ($1-\beta=0.98$) and SC ($1-\beta=0.92$) leads to adequate power. However, when incorporating total design parameter heterogeneities (i.e., sampling error plus true variation) and adopting a (very) conservative approach by using the upper bounds of 95% PIs of $\rho_{L2}=0.07$ and $\rho_{L3}=0.50$ and the lower bounds of the 95% PIs of $R^2_{L1|Gf}=0.00$, $R^2_{L2|Gf}=0.00$, $R^2_{L3|Gf}=0.76$, $R^2_{L1|SC}=0.00$, $R^2_{L2|SC}=0.00$, and $R^2_{L3|SC}=0.56$, only Gf ($1-\beta=0.81$) likely guarantees enough power, as opposed to SC ($1-\beta=0.59$). The team decides to collect students' Gf scores. Finally, the researchers wish to evaluate the long-term effects of the app. Thus, a possible follow-up 3L-CRT of the same sample should still demonstrate adequate power. The suitable design parameters are *Pooled* $\rho_{L2}=0.04$, *Pooled* $\rho_{L3}=0.38$, and predicted values of $R^2_{L1|Gf-2}=0.01$, $R^2_{L2|Gf-2}=0.22$, $R^2_{L3|Gf-2}=0.81$, as well as $R^2_{L1|Gf-4}=0.01$, $R^2_{L2|Gf-4}=0.12$, $R^2_{L3|Gf-4}=0.79$. Assuming no attrition over time, the team calculates $1-\beta=0.95/0.94$ for a 2-/4-year-lagged Gf. Consequently, even when reevaluating the app's impact 4 years later, the 3L-CRT with Gf as a covariate will likely be adequately powered.

Part II: Precision Simulations—Assessing Design Sensitivity via the MDES

Method

We briefly sketch the applied methods here (see OSM D for details). We used R 4.2.2 (R Core Team, 2022); package versions are noted in the R scripts.

Procedure

We adopted a hybrid Bayesian-classical approach to power analysis (Spiegelhalter et al., 2004; see also Pek & Park, 2019). To this end, we took advantage of the (joint) empirical distribution of single- and multilevel design parameters estimated in stage 1 of Part I to simulate MDES distributions for small, medium, and large IRTs and CRTs.

Simulation Conditions We established typical sample sizes of educational RTs by drawing on data of Lortie-Forgues and Inglis' (2019)[16] review. We computed normative distributions (i.e., percentiles P) across K and categorized $P10(K)=14$ as small, $P50(K)=46$ as medium, and $P90(K)=100$ as large, where $n_{L3}=46$. Sample sizes at L2 were not available; we assumed $J_{L3}=2$, resulting in $n_{L2}=23$. It followed

[16] We thank the authors for providing this data.

that $N = 644/2116/4600$ for small/medium/large RTs.[17] We assumed $\alpha = 0.05$ (two-tailed), $1-\beta = 0.80$, and $P_T = P_{L3} = 0.50$.

Expressing Uncertainty in Design Parameters Random noise in ρ and R^2 can be incorporated into power analysis in a number of ways. One method is to enter the bounds of their (meta-analytic) 95% CIs/PIs, as we illustrated above in Part I (section "Application"). Here, we apply a hybrid technique that implicitly models uncertainty by following a Bayesian notion to treat ρ and R^2 along with their *SE*s as (informative) prior distributions, which are used to perform Monte Carlo simulations within the frequentist framework (see e.g., Moerbeek & Teerenstra, 2016, pp. 211–213). Specifically, for each set of connected design parameters (e.g., in three-level designs, ρ_{L2}, ρ_{L3}, R^2_{L1}, R^2_{L2}, and R^2_{L3} for a certain outcome are interrelated), we specified a multivariate normal distribution. The mean vector was represented by the point estimates of ρ and R^2, the variances by their squared *SE*s, and the covariances were derived assuming an intercorrelation of $r = 0.90$, as a conservative upper-bound guess of dependencies. Using the SimDesign package (Chalmers & Adkins, 2020), we then generated 100 draws from each multivariate design parameter prior distribution.

Calculating the MDES For each draw, we computed the MDES based on Eqs. (2), (6), and (7) employing the PowerUpR package (Bulus et al., 2021).

Gauging Sensitivity For the variance–covariance matrices defined for the multivariate normal distributions, we ran sensitivity analyses over $r \in \{0.00, 0.05, \ldots, 0.95\}$ to preclude a missspecification of ρ and R^2 dependencies.

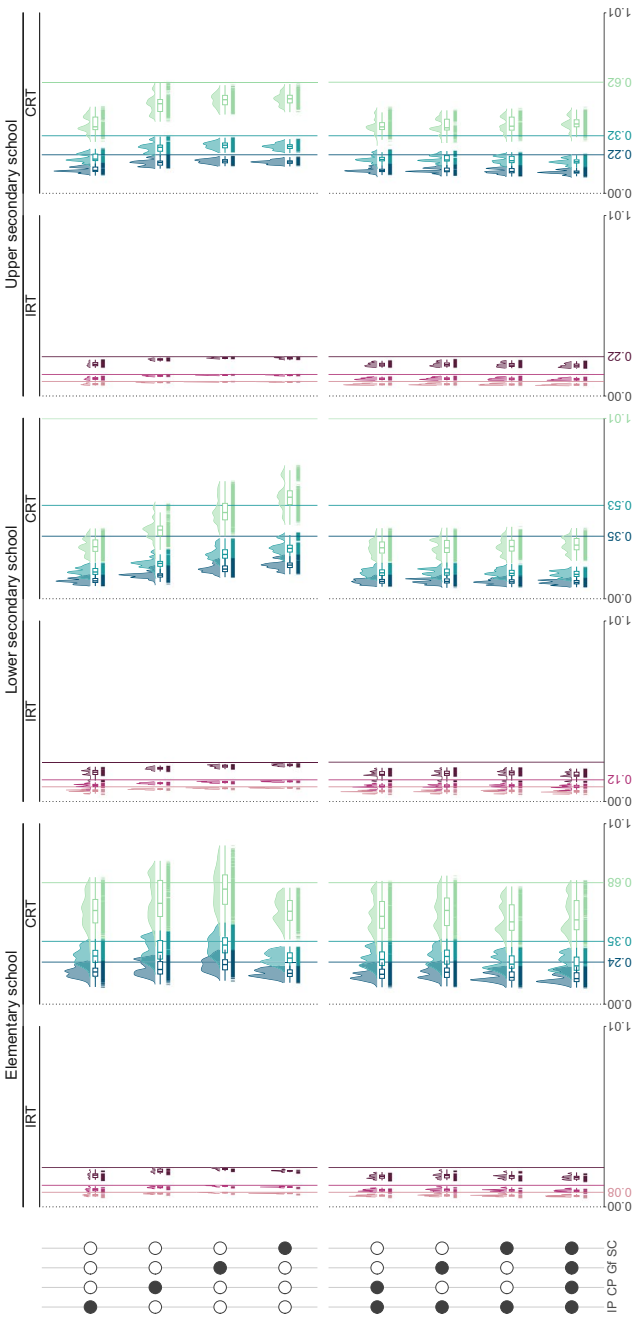
Key Results

We present major patterns in MDES distributions for small, medium, and large IRTs and CRTs (i.e., 3L-CRTs in grades 1–10 and 2L-CRTs in grades 11–12) for the total student population, illustrated in Figs. 6 and 7 (which we refer to in this section; see OSM D for result plots of school tracks and 2L-CRTs in grades 1–10, and OSM G for the full data table of simulated design parameters along with their MDES statistics). Generally, in all simulation conditions, we observed substantive variation in the MDES—between and within outcomes. Further, MDES distributions for small RTs tended to be more sensitive to design parameter uncertainties, and therefore appeared more broadly dispersed than those for large RTs.

Covariate Types: Bandwidth-Fidelity

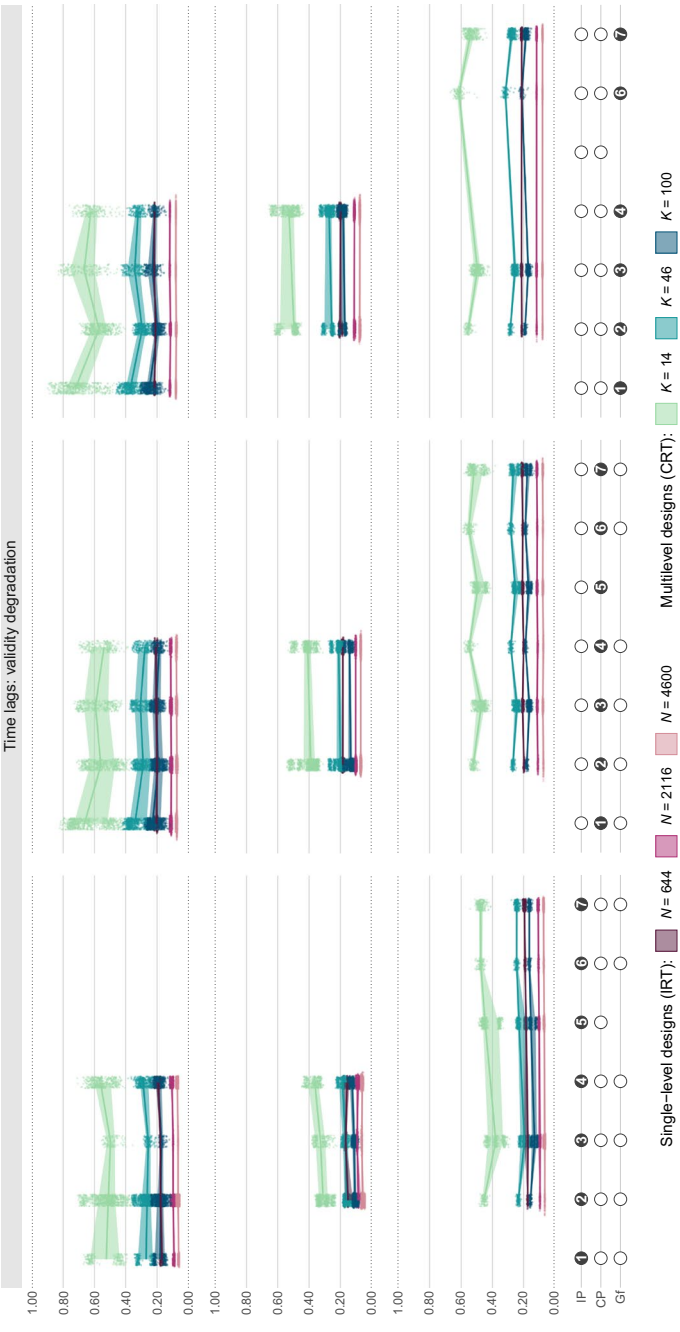
Single-Level Perspective In a medium IRT, $MDES_{IRT}$ was 0.12 without covariates. Precision was then moderately improved by the covariate types; the most by IP, with

[17] Note that Lortie-Forgues and Inglis (2019) reviewed large-scale RTs; thus, we refer to small, medium, and large IRTs and CRTs for interventions whose general effectiveness has already been empirically proven (e.g., via small-scale studies under well-controlled conditions in the lab) and which are now scaled up.



Note. Vertical lines show the unconditional (i.e., unadjusted) $MDES_{IRT}$ (by definition, only varying by sample size) and *Mdn*($MDES_{CRT}$). In multilevel designs, $n_{L2} = 23$ and $J_{L3} = 2$ for 3L-CRTs (elementary and lower secondary school), $n_{L3} = 46$ for 2L-CRTs (upper secondary school). On the y-axis, a filled/empty dot marks the in-/exclusion of a covariate. IP = Domain-identical pretest. CP = Cross-domain pretest (reading for STEM outcomes, mathematics for verbal outcomes). GI = Fluid intelligence pretest. SC = Sociodemographic characteristics (gender, migration background, socioeconomic status).

Fig. 6 MDES distributions for small, medium, and large IRTs and CRTs: covariate types and covariate combinations



Note. Lines connect *Mdn*(MDES) values of consecutive time lags; ribbons depict interquartile ranges; for small/medium/large RTs, unconditional *Mdn*($MDES_{1L\text{-}CRT}$) = 0.68/0.35/0.24 (elementary school) and 1.05/0.54/0.36 (lower secondary school), *Mdn*($MDES_{2L\text{-}CRT}$) = 0.62/0.32/0.21 (upper secondary school). In multilevel designs, n_{L2} = 23 and J_{L3} = 2 for 3L-CRTs (elementary and lower secondary school), n_{L3} = 46 for 2L-CRTs (upper secondary school). On the x-axis, a filled/empty dot marks the in-/exclusion of a covariate, where a numbered dot specifies the pre-posttest time lag in years. IP = Domain-identical pretest. CP = Cross-domain pretest (reading for STEM outcomes, mathematics for verbal outcomes). Gf = Fluid intelligence pretest. SC = Sociodemographic characteristics (gender, migration background, socioeconomic status).

Fig. 7 MDES distributions for small, medium, and large IRTs and CRTs: covariate time lags

median $MDES_{IRT|IP}$ equaling 0.10 in elementary and upper secondary school and 0.09 in lower secondary school. Notably, the percentage MDES reduction for a certain covariate type remained constant across IRT sizes. Since precision is a positive function of sample size, absolute MDES gains were stronger in small IRTs than in large IRTs; furthermore, covariate adjustment reached a point of diminishing returns when sample size increased. For instance, in elementary school, SC somewhat raised precision in an IRT with $N = 644$ ($MDES_{IRT} = 0.22$ vs. $Mdn(MDES_{IRT|SC}) = 0.20$) but not with $N = 4600$ ($MDES_{IRT} = Mdn(MDES_{IRT|SC}) = 0.08$).

Multilevel Perspective In a medium CRT, median $MDES_{CRT}$ was 0.35/0.53/0.32 without covariates in elementary/lower secondary/upper secondary school. In lower secondary school, all covariate types strongly boosted median precision, first and foremost IP ($MDES_{3L\text{-}CRT|IP} = 0.15$), followed by CP, Gf, and SC (in this sequence). Likewise, in upper secondary school, IP markedly reduced the $MDES_{2L\text{-}CRT}$ to around 0.19, around twice as much as the remaining covariates. In elementary school, particularly SC evoked reasonable average precision improvements ($MDES_{3L\text{-}CRT|SC} = 0.26$), while Gf performed the poorest ($MDES_{3L\text{-}CRT|Gf} = 0.33$). Proportionally, the impact of the covariates strengthened somewhat with CRT size: for example, in elementary school, SC reduced the MDES to about 24% in small CRTs ($Mdn(MDES_{3L\text{-}CRT}) = 0.68$ vs. $Mdn(MDES_{3L\text{-}CRT|SC}) = 0.52$) and to 27% in large CRTs ($Mdn(MDES_{3L\text{-}CRT}) = 0.24$ vs. $Mdn(MDES_{3L\text{-}CRT|SC}) = 0.17$). Meanwhile, as with the IRTs, absolute MDES reductions were still (far) more pronounced with $K = 14$ than $K = 100$.

Covariate Combinations: Incremental Validity

Single-Level Perspective In a medium IRT, the additional inclusion of CP over and above IP led to notable MDES drops, but only in elementary/lower secondary school ($Mdn(MDES_{IRT|IP+CP}) = 0.09/0.08$). In these grade levels, no other combination resulted in further tweaks. In upper secondary school, only the complete covariate battery resulted in genuine precision benefits; the median $MDES_{IRT|IP+CP+Gf+SC}$ averaged 0.09.

Multilevel Perspective In a medium CRT targeted at first–fourth graders, adding SC to IP raised precision the most ($Mdn(MDES_{3L\text{-}CRT|IP+SC}) = 0.22$), with no further gains through the full covariate array. From grade 5 on, we did not detect any enhancements in the MDES by pairing IP with CP or Gf. Similarly, the addition of SC, alone or with CP and Gf, returned only miniscule additional MDES declines.

Covariate Time Lags: Validity Degradation

Single-Level Perspective In a medium IRT, precision was slightly affected by temporal validity decrement in IP: we observed the maximum decrement in elementary school, with $\Delta MDES_{IRT|IP}$ equaling +0.02 (from the shortest to the longest time lag). For CP, precision diminished only in upper secondary school, but

starting lately after 7 years ($\Delta Mdn(\text{MDES}_{\text{IRT|CP}-7}) = +0.01$). Of note, precision was more prone to validity deterioration in IP and CP in small rather than large IRTs (e.g., $Mdn(\text{MDES}_{\text{IRT|IP}-2}) = 0.17/0.07$ and $Mdn(\text{MDES}_{\text{IRT|IP}-7}) = 0.19/0.07$ with $N = 644/4600$ in upper secondary school). By contrast, $\text{MDES}_{\text{IRT|Gf}}$ consistently remained highly stable.

Multilevel Perspective In a medium CRT, median MDES_{CRT} was 0.35/0.54/0.32 without covariates in elementary/lower secondary/upper secondary school. The MDES somewhat fluctuated with growing pre-posttest time lags: when subtracting median values for the longest from the shortest time gaps, $\Delta\text{MDES}_{\text{CRT|IP}} = +0.02/+0.03/+0.01$, $\Delta\text{MDES}_{\text{CRT|CP}} = -0.06/+0.01/\pm 0.00$, and $\Delta\text{MDES}_{\text{CRT|Gf}} = -0.05/+0.02/\pm 0.00$. As for IRTs, cross-time precision decay appeared more pronounced in small rather than large CRTs (e.g., $Mdn(\text{MDES}_{\text{2L-CRT|IP}-2}) = 0.45/0.16$ and $Mdn(\text{MDES}_{\text{2L-CRT|IP}-7}) = 0.48/0.16$ for $K = 14/100$ in upper secondary school).

Discussion

Worldwide, the prevalence of educational RTs has been growing sharply (Connolly et al., 2018; Raudenbush & Schwartz, 2020). Reliable knowledge on the effectiveness of programs and innovations to bolster student learning—the foundation of evidence-based policies and practices in education (Hedges, 2018)—requires both well-designed IRTs and CRTs that are sensitive to detect true intervention effects. Highly prognostic covariates are key elements of strong designs; yet, choosing them can be challenging and involves both theoretical and empirical considerations. Our study sought to expand substantive guidance to support informed covariate selection and power analysis for IRTs and CRTs on student achievement: inspired by three psychometric heuristics (the bandwidth-fidelity dilemma, incremental validity concept, and validity degradation principle) and using longitudinal large-scale assessments from Germany, we analyzed unique, relative, and incremental covariate impacts on design sensitivity. Part I covered a wealth of (meta-analytically integrated) single- and multilevel design parameters, and Part II covered a simulation study generating plausible MDES distributions for educational RTs.

Expanding the Range of Designs

We scrutinized covariates in IRTs as well as 2L- and 3L-CRTs. In doing so, our study is unique by covering a large array of the experimental designs implemented to determine the effectiveness of educational interventions (Connolly et al., 2018; Spybrook et al., 2016).

The first central message from our analyses is as follows: *In IRTs, effects on design sensitivity through the covariates largely confirmed the psychometric heuristics; in CRTs, usually all of the covariates noticeably boosted design sensitivity, even long-term.* From a single-level perspective, the higher the fidelity, the lower the bandwidth, and the shorter the pre-posttest time lag of a covariate, the better

the variance explanation between individual students, and the greater the returns in design sensitivity. Thus, the psychometric heuristics are indeed useful to inform covariate choices in IRTs. From a multilevel perspective, however, relations are not always as straightforward. Fortunately, researchers have much more flexibility when choosing covariates for CRTs: all covariates under investigation, regardless of their degree of bandwidth/fidelity and time gap to the outcome, markedly raised design sensitivity. This holds especially true throughout secondary school, where large proportions of between-school differences could be captured by any covariate. This phenomenon, in which aggregated measures tend to correlate much more strongly than their individual-level equivalents, has been described by scholars before (e.g., Bloom et al., 2007; Härnqvist et al., 1994; Robinson, 1950; Snijders & Bosker, 2012).

Expanding the Range of Covariate Types, Combinations, and Time Lags

Previous studies on covariate effects on design sensitivity have systematically analyzed 1- to 3-year-lagged IP, the latest CP, as well as SC; the latter have been examined both uniquely and beyond IP. We added Gf to the spectrum of covariate types, combined IP with CP or Gf as well as with CP plus Gf plus SC, and covered long pre-posttest time lags of up to 7 years. In doing so, we involve the most relevant precursors of students' learning trajectories (e.g., M. C. Wang et al., 1993) and respond to the needs arising from the features of RTs implemented in education (e.g., Connolly et al., 2018; Lortie-Forgues & Inglis, 2019).

The second central message from our analyses is as follows: *Using the latest IP as the only single covariate demonstrated outstanding capacities to improve design sensitivity in both IRTs and CRTs.* IP clearly outweighed all remaining covariate types, although its prognostic property was indeed often affected by temporal deterioration. This pattern of results replicated the pattern that we identified in our meta-analytic research review. However, as noted above, there may be scenarios that necessitate the switch to CP, Gf, or SC, even when assessed long before the target outcome, or that justify their additional inclusion. On a side note, the present values of $R^2_{|CP/Gf/SC}$ may also serve as lower bound estimates when pre-posttest content alignment is less than perfect (Bloom et al., 2007, p. 41). The effectiveness of CP, Gf, and SC to tweak design sensitivity depended on several factors, first and foremost the grade level. Controlling for CP or Gf was a reasonable (alternative) strategy for RTs implemented in lower secondary school. Of importance, Gf appeared to be an exceptionally time-stable predictor, even across numerous years and irrespective of the design. Thus, the idea that Gf may serve as a robust covariate in RTs spanning several years—supported by existing single-level evidence—was generalized to multilevel settings in the present study for the first time. SC, in contrast, performed well as covariates particularly in elementary school, and occasionally also in upper secondary school. Incremental returns of CP, Gf, and/or SC over and above IP were often negligible, largely consonant with previous studies. As an exception, additionally taking into account SC in CRTs with first–fourth graders seems to be a relatively safe option to boost design

sensitivity. Consequently, researchers should always take into account the cost-effectiveness of covariates beyond IP, with regard to the specific application context.

Expanding the Range of Outcome Domains

The bulk of available resources of design parameters to guide covariate choices focus on core domains, namely mathematics and science as STEM outcomes, and reading as a verbal outcome. We further complemented the STEM outcomes by ICT and the verbal outcomes by grammar, spelling, vocabulary, and writing. In doing so, we acknowledge that RTs often seek to enhance skills in domains beyond the core domains (Lortie-Forgues & Inglis, 2019; Morrison, 2020).

The third central message from our analyses is as follows: *Impacts of the covariates on design sensitivity varied widely between achievement outcomes.* For almost all covariates, we observed large heterogeneities in the amounts of explained variance across domains (and, if applicable, samples). Heterogeneity was mostly due to true variation at the level of effect sizes. This observation coincides with the findings of past studies (see also Brunner et al., 2018; Stallasch et al., 2021). Likewise, our simulations emphasize that MDES distributions were considerably dispersed; benefits in precision also strongly hinged on the outcome. Hence, researchers should always strive for an ideal fit between design parameters and the intervention's target outcome. Yet, circumstances may limit this endeavor, such as the unavailability of suitable estimates for a specific domain. Here, our meta-analytic results may inform researchers of possible design parameter ranges and can be used in power analysis to determine expected lower and upper bounds of sample sizes, power rates, or MDES values.

Expanding the Range of National Scopes

Most evidence on sensitivity-enhancing covariate effects is restricted to the United States. We accumulated design parameters drawing on longitudinal large-scale assessment data from six German samples covering the entire school career (i.e., grades 1–12) of the total student population, as well as the student populations in the academic and non-academic tracks. In doing so, we meet the demands of a vast number of RTs that are conducted in countries where the school system more closely resembles the German system (e.g., with respect to the onset of school type tracking; Connolly et al., 2018).

The fourth central message from our analyses is as follows: *The covariates' capabilities to raise design sensitivity cannot be universally generalized across national education contexts.* We found notable differences in multilevel design parameters based on data from German vs. US samples. With the former, explained variances at L3 often appeared more pronounced throughout secondary school, and vice versa at L2. This might be due to the fact that in the tracked German secondary school system, ρ_{L3} tend to be larger and ρ_{L2} tend to be smaller than previously reported in the United States (see Stallasch et al., 2021). Similar patterns in multilevel design parameters by country have also been documented in cross-national works (Brunner et al.,

2018; Kelcey et al., 2016). It is therefore of utmost importance that researchers rely on variance estimates that best depict the characteristics of the interventions' target population.

Essentials of Covariate Adjustment in RTs on Student Achievement at a Glance

Our analyses imply the following general recommendations on pre-treatment covariate inclusion in IRTs and CRTs on student achievement in the German (and similar) school context.

1. A pretest should substantively match the RT's target outcome as closely as possible. In particular, when the outcome is well-aligned with the content of the intervention (e.g., high curricular validity or sensitivity to depict instructional effects), a pretest in the outcome domain may be favorable over one in another domain.
2. A pretest should have high fidelity/low bandwidth rather than low fidelity/high bandwidth. Thus, a domain-specific pretest may be preferable to a domain-general one.
3. A pretest in fluid intelligence may be considered in—especially long-term—RTs in lower secondary school (grades 5–10).
4. Sociodemographic measures may be used in elementary school RTs (grades 1–4).
5. If a pretest in the outcome domain is available, precision gains through additional covariates are often negligible, except for point 4.
6. In IRTs, a pretest in the outcome domain should be granted priority, despite its potential temporal validity degradation. In CRTs—especially in secondary school (grades 5–12)—cost issues should be brought to the fore, as any covariate may be beneficial.
7. Uncertainty in the design parameters should be taken into account, for example via (meta-analytic) 95% CIs/PIs or simulations based on empirical prior distributions.

In addition, we urge researchers planning RTs to keep the following factors in mind:

8. Measurement error in the covariate(s) and/or outcome typically attenuates single- and multilevel R^2 (Cohen et al., 2003; Raudenbush & Bryk, 2002).[18] Reliable measures are expedient for power and precision in RTs (Maxwell et al., 1991). To handle reliability issues, one may (a) add items when newly developing measures, which should reduce measurement error, (b) use a plausible values approach when estimating test scores (Blackwell et al., 2017), or (c) apply latent variable models when analyzing treatment effects in IRTs (Bollen, 2002; Mayer et al., 2016) and CRTs (Lüdtke et al., 2008; Raudenbush & Bryk, 2002), which partials out measurement error. If none of these options is feasible, adjusting for (even error-prone) covariates is—as a rule—still advisable to raise design sensitivity in RTs (Maxwell et al., 2017, p. 481).

[18] In general, measurement error in the outcome tends to attenuate any standardized effect size (Baugh, 2002), such as the standardized treatment effect by increasing the variances of the outcome within experimental groups (Hedges, 1981). However, this bias is most often practically negligible (Maxwell et al., 2017, p. 166).

9. In CRTs, aggregated L1 covariates that demonstrate large ρ values at the implementation level of the intervention are advantageous. When developing new measures for the CRT, using pilot data of ρ estimates for each item to construct multi-item scales that optimize between-group differentiation may be worthwhile (Bliese et al., 2019).
10. Participant attrition during RT implementation hampers the prognostic properties of covariates (Rickles et al., 2018). Hence, planning RTs as conservatively as possible given available financial and personnel resources may be reasonable.

Limitations

Our work has several shortcomings. First, this study's results may generalize well to RT target populations, measures, and design characteristics mimicking those addressed here. More precisely: (a) The ideal application context is the German school system; yet, our design parameters may still serve as valuable benchmarks to plan RTs in countries such as Austria, Czech Republic, Hungary, Slovak Republic, and Turkey with similar performance-based school tracking (Salchegger, 2016). Relatedly, we drew on rather heterogeneous samples. Higher homogeneity (e.g., with convenience samples, as is typical in educational RTs; Tipton & Olsen, 2018) may lead to smaller ρ and R^2 due to range restrictions (Miciak et al., 2016). Further, since we chose not to apply sampling weights, our findings are quasi-representative, and the estimates as well as their *SE*s—albeit presumably only slightly (see e.g., Wenger et al., 2018)—less accurate compared to weighted ones. (b) The present design parameters optimally match measures resembling those used in NEPS, PISA, or DESI. Caution is warranted when designing RTs relying on (substantively) divergent outcome or covariate measures. Importantly, in view of the observed temporal validity degradation, somewhat larger/smaller R^2 are expected for shorter/longer pre-posttest time intervals. (c) The MDES formulae assume homoscedasticity (i.e., TG and CG sharing a common outcome variance). For our simulation conditions with (fully) balanced designs, statistical tests for the treatment effect resting on this assumption were shown robust, even under heteroscedasticity (Blanca et al., 2018; Korendijk et al., 2008). In unbalanced designs, however, the MDES values would probably be too optimistic (see e.g., Gail et al., 1996).

Second, our selection of covariates was theoretically and empirically oriented. Yet, as noteworthy amounts of variance remained unexplained for many outcomes, other individual- or group-level attributes might also function as profitable covariates. For example, socioemotional characteristics—representing important curricular targets (OECD, 2015)—such as domain-specific motivation (Levy et al., 2023), self-concept (Wu et al., 2021), or anxiety (Aldrup et al., 2020) have been shown to predict student achievement over and above IP. Of note, the fully documented R code and our R package multides (Stallasch, 2024) enable researchers to produce R^2 values for additional covariate sets specifically relevant for their prospective RT, drawing on (publicly) available datasets.

Third, we used reasoning ability as assessed by standard figural matrices as our measure of fluid intelligence. Fluid intelligence, however, is a multifaceted construct that encompasses—besides reasoning as one integral component—various further abilities such as perception speed, accuracy, and problem solving (e.g., Baltes et al.,

1999; Cattell, 1987; see also Brunner et al., 2014). Therefore, $R^2_{|Gf}$ values should be interpreted as lower-bound estimates and would possibly have been stronger with a broader spectrum of subtests.

Fourth, virtually no measure in the social sciences is free from measurement error. Ours are no exception: reliabilities of the test scores were $0.51 \leq r_{tt} \leq 0.96$ ($Mdn(r_{tt}) = 0.77$; Table C11). These are fairly typical for applied (experimental) research, where $r_{tt} \geq 0.70$ is desirable, but even smaller values may suffice (Schmitt, 1996). Hence, although fallible measures may lower R^2, our estimates may generalize well to realistic planning scenarios.

Conclusion

Inspired by psychometric heuristics and capitalizing on rich data from several German longitudinal large-scale assessments, we substantively expanded the body of knowledge on covariate impacts to improve design sensitivity in IRTs and CRTs on student achievement. Our study bundles an extensive compilation of (meta-analytic) single- and multilevel design parameters with a precision simulation study implicitly incorporating uncertainty adopting a Bayesian rationale. Our work is enriched by illustrative and empirically supported application guidance and comprehensive OSMs. We hope that these resources support evaluation researchers in making wise covariate selections when planning educational experiments to gather sound evidence on what works to advance student learning.

Acknowledgements While working on her dissertation, Sophie E. Stallasch was a predoctoral fellow at the International Max Planck Research School on the Life Course (LIFE, www.imprs-life.mpg.de; participating institutions: Max Planck Institute for Human Development, Berlin, Germany; Freie Universität Berlin, Germany; Humboldt University of Berlin, Germany; University of Michigan, Ann Arbor, MI, USA; University of Virginia, Charlottesville, VA, USA; University of Zurich, Switzerland).

We would like to thank Elizabeth Parks-Stamm for her editorial assistance with this paper.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant 392108331.

Data Availability This article uses data from: (a) The National Educational Panel Study (NEPS; Blossfeld & Roßbach, 2019): Starting Cohort 2—Kindergarten (NEPS Network, 2020, https://doi.org/10.5157/NEPS:SC2:8.0.1), Starting Cohort 3–5th Grade (NEPS Network, 2019a, https://doi.org/10.5157/NEPS:SC3:9.0.0), and Starting Cohort 4–9th Grade (NEPS Network, 2019b, https://doi.org/10.5157/NEPS:SC4:10.0.0). NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) in cooperation with a nationwide network. NEPS datasets were provided by the Research Data Center (FDZ) at the LIfBi. (b) The Programme for International Student Assessment (PISA): PISA-International Plus 2003, 2004 (PISA-I-Plus 2003, 2004; Prenzel et al., 2013, https://doi.org/10.5159/IQB_PISA_I_Plus_v1) and PISA-Plus 2012–2013 (Reiss et al., 2019, https://doi.org/10.5159/IQB_PISA_Plus_2012-13_v2). PISA datasets were provided by the FDZ at the Institute for Educational Quality Improvement (IQB). (c) The Assessment of Student Achievements in German and English as a Foreign Language (DESI; Klieme, 2012, https://doi.org/10.5159/IQB_DESI_v1). The DESI dataset was provided by the FDZ at the IQB. Instructions on data access can be found in this study's OSF repository at https://osf.io/nhx4w.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences* (2nd ed.). Routledge.

Aldrup, K., Klusmann, U., & Lüdtke, O. (2020). Reciprocal associations between students' mathematics anxiety and achievement: Can teacher sensitivity make a difference? *Journal of Educational Psychology, 112*(4), 735–750. https://doi.org/10.1037/edu0000398

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.).

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., & Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. *Zeitschrift Für Erziehungswissenschaft, 14*(S2), 51–65. https://doi.org/10.1007/s11618-011-0181-8

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. Holt, Rinehart and Winston.

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology, 50*(1), 471–507. https://doi.org/10.1146/annurev.psych.50.1.471

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement, 62*(2), 254–263. https://doi.org/10.1177/0013164402062002004

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review, 4*(3), 165–176. https://doi.org/10.1016/j.edurev.2009.04.002

Bausell, R. B., & Li, Y.-F. (2002). Power analysis for experimental research: A practical guide for the biological, medical and social sciences. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511541933

Beck, B., Bundt, S., & Gomolka, J. (2008). Ziele und Anlage der DESI-Studie [Objectives and design of the DESI study]. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 11–25). Beltz.

Blackwell, M., Honaker, J., & King, G. (2017). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research, 46*(3), 303–341. https://doi.org/10.1177/0049124115585360

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods, 50*(3), 937–962. https://doi.org/10.3758/s13428-017-0918-2

Bliese, P. D., Maltarich, M. A., Hendricks, J. L., Hofmann, D. A., & Adler, A. B. (2019). Improving the measurement of group-level constructs by optimizing between-group differentiation. *Journal of Applied Psychology, 104*(2), 293–302. https://doi.org/10.1037/apl0000349

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review, 19*(5), 547–556. https://doi.org/10.1177/0193841X9501900504

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments. Evolving analytic approaches* (pp. 115–172). Russell Sage Foundation.

Bloom, H. S. (2006). *The core analytics of randomized experiments for social research*. MDRC Working Papers on Research Methodology. http://www.mdrc.org/sites/default/files/full_533.pdf

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30–59. https://doi.org/10.3102/0162373707299550

Bloom, H. S., Zhu, P., Jacob, R., Raudenbush, S. W., Martinez, A., & Lin, F. (2008). *Empirical issues in the design of group-randomized studies to measure the effects of interventions for children*. MDRC Working Papers on Research Methodology. https://www.mdrc.org/sites/default/files/full_85.pdf

Blossfeld, H. P., & Roßbach, H. G. (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd ed.). Springer VS.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*(1), 605–634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). Wiley.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*(1), 371–399. https://doi.org/10.1146/annurev.psych.53.100901.135233

Brod, G. (2021). Toward an understanding of when prior knowledge helps or hinders learning. *Npj Science of Learning, 6*(1), 24. https://doi.org/10.1038/s41539-021-00103-w

Brunner, M., Lang, F. R., & Lüdtke, O. (2014). *Erfassung der fluiden kognitiven Leistungfähigkeit über die Lebensspanne im Rahmen der National Educational Panel Study: Expertise [Measuring fluid intelligence across the lifespan in NEPS: Expert report] (NEPS Working Paper No. 42)*. Leibniz-Institut für Bildungsverläufe. https://www.neps-data.de/Portals/0/Working%20Papers/WP_XLII.pdf

Brunner, M., Keller, L., Stallasch, S. E., Kretschmann, J., Hasl, A., Preckel, F., Lüdtke, O., & Hedges, L. V. (2023). Meta-analyzing individual participant data from studies with complex survey designs: A tutorial on using the two-stage approach for data from educational large-scale assessments. *Research Synthesis Methods, 14*(1), 5–35. https://doi.org/10.1002/jrsm.1584

Brunner, M., Keller, U., Wenger, M., Fischbach, A., & Lüdtke, O. (2018). Between-school variation in students' achievement, motivation, affect, and learning strategies: Results from 81 countries for planning group-randomized trials in education. *Journal of Research on Educational Effectiveness, 11*(3), 452–478. https://doi.org/10.1080/19345747.2017.1375584

Brunner, M., Stallasch, S. E., & Lüdtke, O. (2023). Empirical benchmarks to interpret intervention effects on student achievement in elementary and secondary school: Meta-analytic results from Germany. *Journal of Research on Educational Effectiveness, 17*(1), 119–157. https://doi.org/10.1080/19345747.2023.2175753

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2021). *PowerUpR: Power analysis tools for multilevel randomized experiments. R package version 1.1.0* [Computer software]. https://CRAN.R-project.org/package=PowerUpR

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297–312. https://doi.org/10.1037/h0040950

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. North-Holland.

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology, 16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Chu, F. W., vanMarle, K., Rouder, J., & Geary, D. C. (2018). Children's early understanding of number predicts their later problem-solving sophistication in addition. *Journal of Experimental Child Psychology, 169*, 73–92. https://doi.org/10.1016/j.jecp.2017.12.010

Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research. Advance Online Publication.* https://doi.org/10.1177/00491241221099552

Ciolino, J. D., Palac, H. L., Yang, A., Vaca, M., & Belli, H. M. (2019). Ideal vs. real: A systematic review on handling covariates in randomized controlled trials. *BMC Medical Research Methodology, 19*(1), 1–11. https://doi.org/10.1186/s12874-019-0787-8

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs*. John Wiley & Sons.

Coens, C., Pe, M., Dueck, A. C., Sloan, J., Basch, E., Calvert, M., Campbell, A., Cleeland, C., Cocks, K., Collette, L., Devlin, N., Dorme, L., Flechtner, H.-H., Gotay, C., Griebsch, I., Groenvold, M., King, M., Kluetz, P. G., Koller, M., … Bottomley, A. (2020). International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: Recommendations of the SISAQOL Consortium. *The Lancet Oncology*, *21*(2), e83–e96. https://doi.org/10.1016/S1470-2045(19)30790-9

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

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

Cole, R., Haimson, J., Perez-Johnson, I., & May, H. (2011). *Variability in pretest-posttest correlation coefficients by student achievement level* (NCEE Reference Report 2011–4033). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/pubs/20114033/pdf/20114033.pdf

Committee for Proprietary Medicinal Products. (2004). Points to consider on adjustment for baseline covariates. *Statistics in Medicine, 23*(5), 701–709. https://doi.org/10.1002/sim.1647

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016. *Educational Research, 60*(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Cook, T. D. (2005). Emergent principles for the design, implementation, and analysis of cluster-based experiments in social science. *The ANNALS of the American Academy of Political and Social Science, 599*(1), 176–198. https://doi.org/10.1177/0002716205275738

Cox, D. R., & McCullagh, P. (1982). Some aspects of analysis of covariance. *Biometrics, 38*(3), 541–561.

Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions*. University of Illinois.

DESI-Konsortium. (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie [Teaching and acquisition of competencies in German and English: Results from the DESI study]*. Beltz.

Dochy, F. J. R. C., Segers, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research, 69*(2), 145–186. https://doi.org/10.3102/00346543069002145

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24–67. https://doi.org/10.1080/19345747.2012.673143

Donner, A., & Koval, J. J. (1980). The large sample variance of an intraclass correlation. *Biometrika, 67*(3), 719–722. https://doi.org/10.1093/biomet/67.3.719

Erbeli, F., Shi, Q., Campbell, A. R., Hart, S. A., & Woltering, S. (2021). Developmental dynamics between reading and math in elementary school. *Developmental Science, 24*(1), e13004. https://doi.org/10.1111/desc.13004

European Medicines Agency. (1998). *Statistical principles for clinical trials. ICH harmonised tripartite guideline*. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf

European Medicines Agency. (2015). *Guideline on adjustment for baseline covariates in clinical trials*. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). Oliver & Boyd.

Food and Drug Administration. (2021). *Adjusting for covariates in randomized clinical trials for drugs and biological products. Guidance for industry*. https://www.fda.gov/media/148910/download

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., & Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine, 15*(11), 1069–1092. https://doi.org/10.1002/(SICI)1097-0258(19960615)15:11%3c1069::AID-SIM220%3e3.0.CO;2-Q

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social Science Research, 25*(3), 201–239. https://doi.org/10.1006/ssre.1996.0010

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of a randomized controlled trial. *American Educational Research Journal, 52*(3), 516–546. https://doi.org/10.3102/0002831214565787

Ghiselli, E. E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology, 40*(1), 1–4. https://doi.org/10.1037/h0040429

Grund, S., Robitzsch, A., & Lüdtke, O. (2021). *mitml: Tools for multiple imputation in multilevel modeling. R package version 0.4–3* [Computer software]. https://CRAN.R-project.org/package=mitml

Haertel, G. D., Walberg, H. J., & Weinstein, T. (1983). Psychological models of educational performance: A theoretical synthesis of constructs. *Review of Educational Research, 53*(1), 75–91. https://doi.org/10.3102/00346543053001075

Härnqvist, K., Gustafsson, J.-E., Muthén, B. O., & Nelson, G. (1994). Hierarchical models of ability at individual and class levels. *Intelligence, 18*(2), 165–187. https://doi.org/10.1016/0160-2896(94)90026-4

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment, 15*(4), 456–466. https://doi.org/10.1037/1040-3590.15.4.456

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research*. National Center for Special Education Research. https://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf

Hedges, L. V. (2019). Stochastically dependent effect sizes. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (3rd ed., pp. 245–280). Russell Sage Foundation.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.2307/1164588

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness, 11*(1), 1–21. https://doi.org/10.1080/19345747.2017.1375583

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review, 37*(6), 445–489. https://doi.org/10.1177/0193841X14529126

Hedges, L. V., Hedberg, E. C., & Kuyper, A. M. (2012). The variance of intraclass correlations in three- and four-level models. *Educational and Psychological Measurement, 72*(6), 893–909. https://doi.org/10.1177/0013164412445193

Heine, J.-H., Nagy, G., Meinck, S., Zühlke, O., & Mang, J. (2017). Empirische Grundlage, Stichprobenausfall und Adjustierung im PISA-Längsschnitt 2012–2013 [Empirical basis, sample attrition, and adjustment in PISA 2012–2013]. *Zeitschrift Für Erziehungswissenschaft, 20*(S2), 287–306. https://doi.org/10.1007/s11618-017-0756-0

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity–bandwidth trade-off. *Journal of Organizational Behavior, 17*(6), 627–637. https://doi.org/10.1002/(SICI)1099-1379(199611)17:6%3c627::AID-JOB2828%3e3.0.CO;2-F

Huang, F. L. (2018). Using cluster bootstrapping to analyze nested data with a few clusters. *Educational and Psychological Measurement, 78*(2), 297–318. https://doi.org/10.1177/0013164416678980

Hulin, C. L., Henry, R. A., & Noon, S. L. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships. *Psychological Bulletin, 107*(3), 328–340. https://doi.org/10.1037/0033-2909.107.3.328

Humphreys, L. G. (1960). Investigations of the simplex. *Psychometrika, 25*(4), 313–323. https://doi.org/10.1007/BF02289750

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*(4), 446–455. https://doi.org/10.1037/1040-3590.15.4.446

Jacob, R. T., Zhu, P., & Bloom, H. S. (2010). New empirical evidence for the design of group randomized trials in education. *Journal of Research on Educational Effectiveness, 3*(2), 157–198. https://doi.org/10.1080/19345741003592428

Jensen, A. R. (1993). Psychometric g and achievement. In B. R. Gifford (Ed.), *Policy perspectives on educational testing* (pp. 117–227). Springer Netherlands. https://doi.org/10.1007/978-94-011-2226-9_4

Kahan, B. C., Jairath, V., Doré, C. J., & Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Trials, 15*(1), 139. https://doi.org/10.1186/1745-6215-15-139

Keil, C. T., & Cortina, J. M. (2001). Degradation of validity over time: A test and extension of Ackerman's model. *Psychological Bulletin, 127*(5), 673–697. https://doi.org/10.1037/0033-2909.127.5.673

Kelcey, B., Shen, Z., & Spybrook, J. (2016). Intraclass correlation coefficients for designing cluster-randomized trials in Sub-Saharan Africa education. *Evaluation Review, 40*(6), 500–525. https://doi.org/10.1177/0193841X16660246

Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, J., & Soffer Goldstein, D. (2013). Estimating the effect of web-based homework. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education* (pp. 824–827). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39112-5_122

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Klieme, E. (2012). *Deutsch-Englisch-Schülerleistungen-International (DESI) [Assessment of Student Achievements in German and English as a Foreign Language (DESI)]* (Version 1) [Data set]. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.5159/IQB_DESI_v1

Konstantopoulos, S. (2012). The impact of covariates on statistical power in cluster randomized designs: Which level matters more? *Multivariate Behavioral Research, 47*(3), 392–420. https://doi.org/10.1080/00273171.2012.673898

Korendijk, E. J. H., Maas, C. J. M., Moerbeek, M., & Van Der Heijden, P. G. M. (2008). The influence of misspecification of the heteroscedasticity on multilevel regression parameter and standard error estimates. *Methodology, 4*(2), 67–72. https://doi.org/10.1027/1614-2241.4.2.67

Langan, D., Higgins, J. P. T., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., & Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods, 10*(1), 83–98. https://doi.org/10.1002/jrsm.1316

Levy, J., Brunner, M., Keller, U., & Fischbach, A. (2023). How sensitive are the evaluations of a school's effectiveness to the selection of covariates in the applied value-added model? *Educational Assessment, Evaluation and Accountability, 35*(1), 129–164. https://doi.org/10.1007/s11092-022-09386-y

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, *7*(1), 295–318. https://doi.org/10.1214/12-AOAS583

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. SAGE Publications.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher, 48*(3), 158–166. https://doi.org/10.3102/0013189X19832850

Loy, A., & Korobova, J. (2023). Bootstrapping clustered data in R using lmeresampler. *The R Journal, 14*(4), 103–120. https://doi.org/10.32614/RJ-2023-015

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*(3), 203–229. https://doi.org/10.1037/a0012869.supp

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Routledge. https://doi.org/10.4324/9781315642956

Maxwell, S. E., Cole, D. A., Arvey, R. D., & Salas, E. (1991). A comparison of methods for increasing power in randomized between-subjects designs. *Psychological Bulletin, 110*(2), 328–337. https://doi.org/10.1037/0033-2909.110.2.328

Mayer, A., Dietzfelbinger, L., Rosseel, Y., & Steyer, R. (2016). The EffectLiteR approach for analyzing average and conditional effects. *Multivariate Behavioral Research, 51*(2–3), 374–391. https://doi.org/10.1080/00273171.2016.1151334

McCoach, D. B., Yu, H., Gottfried, A. W., & Gottfried, A. E. (2017). Developing talents: A longitudinal examination of intellectual ability and academic achievement. *High Ability Studies, 28*(1), 7–28. https://doi.org/10.1080/13598139.2017.1298996

Miciak, J., Taylor, W. P., Stuebing, K. K., Fletcher, J. M., & Vaughn, S. (2016). Designing intervention studies: Selected populations, range restrictions, and statistical power. *Journal of Research on Educational Effectiveness, 9*(4), 556–569. https://doi.org/10.1080/19345747.2015.1086916

Moerbeek, M., & Teerenstra, S. (2016). *Power analysis of trials with multilevel data*. CRC Press.

Moerbeek, M. (2006). Power and money in cluster randomized trials: When is it worth measuring a covariate? *Statistics in Medicine, 25*(15), 2607–2617. https://doi.org/10.1002/sim.2297

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science, 62*(3), 760–775. https://doi.org/10.1111/ajps.12357

Morrison, K. (2020). *Taming randomized controlled trials in education: Exploring key claims, issues and debates*. Routledge. https://doi.org/10.4324/9781003042112

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford University Press.

National Research Council. (2011). *Assessing 21st century skills: Summary of a workshop*. National Academies Press.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77–101. https://doi.org/10.1037/0003-066X.51.2.77

NEPS Network. (2019b). *National Educational Panel Study, scientific use file of starting cohort Grade 9* [Data set]. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC4:10.0.0

NEPS Network. (2019a). *National Educational Panel Study, scientific use file of starting cohort Grade 5* [Data set]. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC3:9.0.0

NEPS Network. (2020). *National Educational Panel Study, scientific use file of starting cohort Kindergarten* [Data set]. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC2:8.0.1

Organisation for Economic Co-operation and Development. (2018). *The future of education and skills*. OECD Publishing. https://www.oecd.org/education/2030-project/about/documents/E2030%20Position%20Paper%20(05.04.2018).pdf

Organisation for Economic Co-operation and Development. (2015). *Skills for social progress: The power of social and emotional skills*. OECD Publishing. https://doi.org/10.1787/9789264226159-en

Pek, J., & Park, J. (2019). Complexities in power analysis: Quantifying uncertainties with a Bayesian-classical hybrid approach. *Psychological Methods, 24*(5), 590–605. https://doi.org/10.1037/met0000208

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin, 146*(7), 595–634. https://doi.org/10.1037/bul0000231

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practiceand problems. *Statistics in Medicine, 21*(19), 2917–2930. https://doi.org/10.1002/sim.1296

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology, 34*(4), 383–392. https://doi.org/10.1037/0022-0167.34.4.383

Prenzel, M., Carstensen, C. H., Schöps, K., & Maurischat, C. (2006). Die Anlage des Längsschnitts bei PISA 2003 [The longitudinal design of PISA 2003]. In PISA-Konsortium Deutschland (Ed.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (pp. 29–62). Waxmann.

PISA-Konsortium Deutschland (Ed.). (2006). *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres [PISA 2003. Investigating competence development throughout one school year]*. Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rost, J., & Schiefele, U. (2013). *Programme for International Student Assessment—International Plus 2003, 2004 (PISA-I-Plus 2003, 2004)* (Version 1) [Data set]. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.5159/IQB_PISA_I_Plus_v1

Pustejovsky, J. E. (2022). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. R package version 0.5.8* [Computer software]. https://CRAN.R-project.org/package=clubSandwich

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raab, G. M., Day, S., & Sales, J. (2000). How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials, 21*(4), 330–342. https://doi.org/10.1016/S0197-2456(00)00061-1

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). SAGE Publications.

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

Raudenbush, S. W., Martínez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*(1), 5–29. https://doi.org/10.3102/0162373707299460

Raudenbush, S. W., & Schwartz, D. (2020). Randomized experiments in education, with implications for multilevel causal inference. *Annual Review of Statistics and Its Application, 7*(1), 177–208. https://doi.org/10.1146/annurev-statistics-031219-041205

Reeve, C. L., & Bonaccio, S. (2011). On the myth and the reality of the temporal validity degradation of general mental ability test scores. *Intelligence, 39*(5), 255–272.

Reiss, K., Klieme, E., Köller, O., & Stanat, P. (2017). *PISA Plus 2012 – 2013. Kompetenzentwicklung im Verlauf eines Schuljahres [PISA Plus 2012 – 2013. Competence development throughout one school year]*. Springer VS.

Reiss, K., Heine, J.-H., Klieme, E., Köller, O., & Stanat, P. (2019). *Programme for International Student Assessment—Plus 2012–2013 (PISA Plus 2012–2013)* (Version 2) [Data set]. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.5159/IQB_PISA_Plus_2012-13_v2

Rickles, J., Zeiser, K., & West, B. (2018). Accounting for student attrition in power calculations: Benchmarks and guidance. *Journal of Research on Educational Effectiveness, 11*(4), 622–644. https://doi.org/10.1080/19345747.2018.1502384

Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ, 342*, d549. https://doi.org/10.1136/bmj.d549

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351. https://doi.org/10.2307/2087176

Robitzsch, A., Grund, S., & Henke, T. (2021). *Miceadds: Some additional multiple imputation functions, especially for "mice". R package version 3.11–6* [Computer software]. https://CRAN.R-project.org/package=miceadds

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Salchegger, S. (2016). Selective school systems and academic self-concept: How explicit and implicit school-level tracking relate to the big-fish-little-pond effect across cultures. *Journal of Educational Psychology, 108*(3), 405–423. https://doi.org/10.1037/edu0000063

Salgado, J. F. (2017). Bandwidth-fidelity dilemma. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–4). Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_1280-1

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62–87. https://doi.org/10.3102/1076998607302714

Schomaker, M., & Heumann, C. (2018). Bootstrap inference when using multiple imputation. *Statistics in Medicine, 37*, 2252–2266. https://doi.org/10.1002/sim.7654

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement, 23*(1), 153–158. https://doi.org/10.1177/001316446302300113

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.

Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2022). Quantifying "promising trials bias" in randomized controlled trials in education. *Journal of Research on Educational Effectiveness*, 1–18. https://doi.org/10.1080/19345747.2022.2090470

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist, 55*(1), 21–31. https://doi.org/10.1080/00461520.2019.1611432

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE Publications.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health care evaluation*. John Wiley & Sons.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research & Method in Education, 39*(3), 255–267. https://doi.org/10.1080/1743727X.2016.1150454

Stallasch, S. E. (2024). *multides: R tools for the MULTI-DES project. R package version 1.0.0* [Computer software]. https://github.com/sophiestallasch/multides

Stallasch, S. E., Lüdtke, O., Artelt, C., & Brunner, M. (2021). Multilevel design parameters to plan cluster-randomized intervention studies on student achievement in elementary and secondary school. *Journal of Research on Educational Effectiveness, 14*(1), 172–206. https://doi.org/10.1080/19345747.2020.1823539

Stanat, P., & Chistensen, G. (2006). *Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003*. OECD Publishing.

Steinmayr, R., Meißner, A., Weidinger, A. F., & Wirthwein, L. (2014). Academic achievement. In *Oxford Bibliographies in Education*. Oxford University Press. https://doi.org/10.1093/obo/9780199756810-0108

Stern, E. (2009). The development of mathematical competencies: Sources of individual differences and their developmental trajectories. In M. Bullock & W. Schneider (Eds.), *Human development from early childhood to early adulthood: Findings from a 20 year longitudinal study* (pp. 221–236). Psychology Press.

Tafti, A., & Shmueli, G. (2020). Beyond overall treatment effects: Leveraging covariates in randomized experiments guided by causal structure. *Information Systems Research, 31*(4), 1183–1199. https://doi.org/10.1287/isre.2020.0938

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher, 47*(8), 516–524. https://doi.org/10.3102/0013189X18781522

Träff, U., Olsson, L., Skagerlund, K., & Östergren, R. (2020). Kindergarten domain-specific and domain-general cognitive precursors of hierarchical mathematical development: A longitudinal study. *Journal of Educational Psychology, 112*(1), 93–109. https://doi.org/10.1037/edu0000369

Turner, R. M., Prevost, A. T. & Thompson, S. G. (2004). Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Statistics in Medicine, 23*(8), 1195–1214. https://doi.org/10.1002/sim.1721

Ünal, Z. E., Greene, N. R., Lin, X., & Geary, D. C. (2023). What is the source of the correlation between reading and mathematics achievement? Two Meta-Analytic Studies. *Educational Psychology Review, 35*(1), 4. https://doi.org/10.1007/s10648-023-09717-5

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Viechtbauer, W. (2022). *Analysis examples: Konstantopoulos (2011)*. The Metafor Package. A Meta-Analysis Package for R. https://www.metafor-project.org/doku.php/analyses:konstantopoulos2011

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Wan, F. (2021). Statistical analysis of two arm randomized pre-post designs with one post-treatment measurement. *BMC Medical Research Methodology, 21*(1), 150. https://doi.org/10.1186/s12874-021-01323-9

Wang, J. (2020). Covariate adjustment for randomized controlled trials revisited. *Pharmaceutical Statistics, 19*(3), 255–261. https://doi.org/10.1002/pst.1988

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*(3), 249–294. https://doi.org/10.3102/00346543063003249

Wenger, M., Lüdtke, O., & Brunner, M. (2018). Übereinstimmung, Variabilität und Reliabilität von Schülerurteilen zur Unterrichtsqualität auf Schulebene: Ergebnisse aus 81 Ländern [Interrater agreement, variability and reliability of student ratings of instructional quality at the school-level. Results from 81 countries]. *Zeitschrift für Erziehungswissenschaft, 21*(5), 929–950. https://doi.org/10.1007/s11618-018-0813-3

Whitehurst, G. J. (2012). The value of experiments in education. *Education Finance and Policy, 7*(2), 107–123. https://doi.org/10.1162/EDFP_a_00058

Wilkinson, L., Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Winne, P. H., & Nesbit, J. C. (2010). The psychology of academic achievement. *Annual Review of Psychology, 61*(1), 653–678. https://doi.org/10.1146/annurev.psych.093008.100348

Woolfolk, A. (2020). *Educational psychology* (14th ed.). Pearson Education Canada.

Wright, N., Ivers, N., Eldridge, S., Taljaard, M., & Bremner, S. (2015). A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. *Journal of Clinical Epidemiology, 68*(6), 603–609. https://doi.org/10.1016/j.jclinepi.2014.12.006

Wu, H., Guo, Y., Yang, Y., Zhao, L., & Guo, C. (2021). A meta-analysis of the longitudinal relationship between academic self-concept and academic achievement. *Educational Psychology Review, 33*(4), 1749–1778. https://doi.org/10.1007/s10648-021-09600-1

Xu, Z., & Nichols, A. (2010). *New estimates of design parameters for clustered randomization studies. Findings from North Carolina and Florida*. National Center for Analysis of Longitudinal Data in Education. https://files.eric.ed.gov/fulltext/ED510553.pdf

Yang, S., Starks, M. A., Hernandez, A. F., Turner, E. L., Califf, R. M., O'Connor, C. M., Mentz, R. J., & Roy Choudhury, K. (2020). Impact of baseline covariate imbalance on bias in treatment effect estimation in cluster randomized trials: Race as an example. *Contemporary Clinical Trials, 88*, 105775. https://doi.org/10.1016/j.cct.2019.04.016

Zhang, Q., Spybrook, J., Kelcey, B., & Dong, N. (2023). Foundational methods: Power analysis. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International encyclopedia of education* (4th ed., pp. 784–791). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10088-0

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

Authors and Affiliations

Sophie E. Stallasch[1] · Oliver Lüdtke[2,3] · Cordula Artelt[4,5] · Larry V. Hedges[6] · Martin Brunner[1]

✉ Sophie E. Stallasch
stallasch@uni-potsdam.de

[1] Department of Educational Sciences, Faculty of Human Sciences, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

[2] Leibniz Institute for Science and Mathematics Education, Kiel, Germany

[3] Centre for International Student Assessment, Munich, Germany

[4] Leibniz Institute for Educational Trajectories, Bamberg, Germany

[5] Faculty of Human Sciences and Education, University of Bamberg, Bamberg, Germany

[6] Department of Statistics, Northwestern University, Evanston, IL, USA