Optimizing Power Analysis for Randomized Experiments: Design Parameters for Student Achievement

Dissertation

Zur Erlangung des akademischen Grades

Doktorin der Philosophie (Dr. phil.)

im Fach Psychologie

eingereicht bei der

Humanwissenschaftlichen Fakultät der Universität Potsdam

vorgelegt von

Sophie Elise Stallasch, M.A.

Potsdam, Dezember 2023

Erstgutachter:
Prof. Dr. Martin Brunner, Universität Potsdam

Zweitgutachter:
Prof. Dr. Manuel Völkle, Humboldt-Universität zu Berlin

Tag der Disputation:
19.02.2024

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-62939
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-629396

To Fabian.

CONTENTS

Acknowledgement

During the creation of this doctoral thesis my life has changed irreversibly; both for the better and for the worst. Without the plethora of wonderful people who stood by my side, the finalisation of this thesis would not have been possible.

First and foremost, I would like to thank Martin Brunner. With your impressive expertise, passion for the project, immense support and trust in my abilities, and never-ending positivity you saved this thesis more times than I want to admit. You are the best supervisor I could have asked for: You motivated me in times of crisis, reigned me in when I had too many ideas and offered invaluable guidance for developing my skills.

Oliver Lüdtke as my second supervisor as well as Cordula Artelt provided immeasurably important input, expert knowledge, and methodological and substantive advice, which was integral for grounding the present work on a solid basis. Many thanks to you.

Special thanks goes to Larry Hedges. You rekindled and nurtured a flaming passion for the subject matter and inspired me for thinking beyond this huge project. Thank you for believing in me and supporting the plans coming after this thesis.

Together with the four people mentioned above we form the team of MULTI-DES—the project underlying this thesis. I thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for supporting MULTI-DES under Grant 392108331.

The second assessment on this thesis will be provided by Manuel Völkle. I cannot thank you enough for your interest, expertise, and your time. My dissertation committee is completed by further brilliant people: Julia Kretschmann, Miriam Vock, and Florian Weck. I am very grateful for your support and your interest!

The International Max Planck Research School on the Life Course (IMPRS LIFE) fellow programme created opportunities for valuable exchanges with leading scholars and peers from (educational) psychology as well as other strains of science, which has greatly broadened my horizon. Thank you to Ulman Lindenberger, the Berlin Max Planck Institute for Human Development and all the other excellent LIFE members for this unique chance.

Without the genius ideas and constructive criticism of my amazing colleagues at the University of Potsdam this thesis would be empty. So I would like to thank Andrea, Anta, Gesine, Julia, Lena, Sarah, and all our “HiWis” for your help and friendship both in times of need and happiness.

Without friends who endured putting up with endless stories about statistics and the implementation of new ideas, the whole process would have been more difficult: Thank you, Arum, Max, Marie, and Björn for being there for me in a demanding period of my life.

Finally, Fabian: Words cannot express my deep gratitude for your endless love and support; all the sacrifices you have made, your patience, your unbreakable optimism, and your unconditional faith in me. Not least, thank you for being the (honestly!) most talented hobby cook and bread baker, the most beloved life manager when I was fully immersed in this work, and the dearest companion on all those epic and vitalizing cycling tours of sheer countless kilometers and elevations. From the bottom of my heart, I thank you for removing all the small and large obstacles on this path—without you, I would never have gotten this far.

Summary

Randomized trials (RTs) are promising methodological tools to inform evidence-based reform to enhance schooling. Establishing a robust knowledge base on how to promote student achievement requires sensitive RT designs demonstrating sufficient statistical power and precision to draw conclusive and correct inferences on the effectiveness of educational programs and innovations. Proper power analysis is therefore an integral component of any informative RT on student achievement. This venture critically hinges on the availability of reasonable input variance design parameters (and their inherent uncertainties) that optimally reflect the realities around the prospective RT—precisely, its target population and outcome, possibly applied covariates, the concrete design as well as the planned analysis. However, existing compilations in this vein show far-reaching shortcomings.

The overarching endeavor of the present doctoral thesis was to substantively expand available resources devoted to tweak the planning of RTs evaluating educational interventions. At the core of this thesis is a systematic analysis of design parameters for student achievement, generating reliable and versatile compendia and developing thorough guidance to support apt power analysis to design strong RTs. To this end, the thesis at hand bundles two complementary studies which capitalize on rich data of several national probability samples from major German longitudinal large-scale assessments.

Study I applied two- and three-level latent (covariate) modeling to analyze design parameters for a wide spectrum of mathematical-scientific, verbal, and domain-general achievement outcomes. Three vital covariate sets were covered comprising (a) pretests, (b) sociodemographic characteristics, and (c) their combination. The accumulated estimates were additionally summarized in terms of normative distributions.

Study II specified (manifest) single-, two-, and three-level models and referred to influential psychometric heuristics to analyze design parameters and develop concise selection guidelines for covariate (a) types of varying bandwidth-fidelity (domain-identical, cross-domain, fluid intelligence pretests; sociodemographic characteristics), (b) combinations quantifying incremental validities, and (c) time lags of 1- to 7-year lagged pretests scrutinizing validity degradation. The estimates for various mathematical-scientific and verbal achievement outcomes were meta-analytically integrated and employed in precision simulations.

In doing so, Studies I and II addressed essential gaps identified in previous repertoires in six major dimensions: Taken together, this thesis accumulated novel design parameters and deliberate guidance for RT power analysis (1) tailored to four German student (sub)populations

across the entire school career from Grade 1 to 12, (2) matched to 21 achievement (sub)domains, (3) adjusted for 11 covariate sets enriched by empirically supported guidelines, (4) adapted to six RT designs, (5) suitable for latent and manifest analysis models, (6) which were cataloged along with quantifications of their associated uncertainties. These resources are complemented by a plethora of illustrative application examples to gently direct psychological and educational researchers through pivotal steps in the process of RT design.

The striking heterogeneity of the design parameter estimates across all these dimensions constitutes the overall, joint key result of Studies I and II. Hence, this work convincingly reinforces calls for a close match between design parameters and the specific peculiarities of the target RT's research context.

All in all, the present doctoral thesis offers a—so far unique—nuanced and extensive toolkit to optimize power analysis for sound RTs on student achievement in the German (and similar) school context. It is of utmost importance that research does not tire to spawn robust evidence on what actually works to improve schooling. With this in mind, I hope that the emerging compendia and guidance contribute to the quality and rigor of our randomized experiments in psychology and education.

Preface

> From the outset, it is important to note that a power analysis is only as good as the formulae and parameter estimates that are used and the analyst who uses them. Power analysis should be done by someone who understands the formulae and why they are structured as they are. This allows the analyst to tailor the formulae to the particular needs of the study. In addition, investigators should take care that the parameter estimates accurately reflect the expected state of affairs in the population to be studied. Without good estimates, power analysis is only guesswork.
>
> David M. Murray (1998, pp. 349-350)

The ultimate goal of any randomized experiment in educational psychology is to generate unbiased and valid—in one word: useable—knowledge on how to shape students' developmental trajectories (Hedges, 2018). Without doubt, successful education is a cornerstone of personal, societal, and economic prosperity. Efficient education systems therefore build upon reliable evidence on what works to enhance student achievement. An individual's academic performance not only essentially determines their educational and occupational career and contributes to their well-being, health, and longevity (P. Peng & Kievit, 2020; Spinath, 2012; Steinmayr et al., 2014), but also empowers them to participate in social life and democratic processes (Organisation for Economic Co-operation and Development, 2018). Student achievement is a highly complex, multifaceted construct which is influenced by a myriad of factors (see e.g., Steinmayr et al., 2014; Wang et al., 1993; Winne & Nesbit, 2010). It is therefore of both individual as well as macrosocial interest that research does not tire to raise both the quantity and quality of the empirical information on how to optimally foster student achievement. This information serves as the basis of political decisions and practices in teaching and learning (Whitehurst, 2003). Randomized trials (RTs) function as invaluable methodological tools for this endeavor, as they facilitate causal claims on the actual impacts of educational interventions by assigning units (e.g., individual students or whole schools) by chance to experimental conditions (Institute of Education Sciences & National Science Foundation, 2013; Mosteller & Boruch, 2002; Slavin, 2002; Spybrook, Shi, et al., 2016). Overall, it is quite a recent trend though (spanning around the last 15 years), that RTs literally proliferated in the field of education, however, with large differences between nations (Connolly et al., 2018).

A strong RT is well-designed, well-implemented, and well-analyzed (Spybrook, 2013). Contributing to the former is at the core of this dissertation. Over the last 25 years, educational and psychological research have made great leaps forward in the methodological foundations of RT design. Optimizing power analysis for RTs through innovative statistical methods (e.g.,

Bloom, 2005; Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997) and software (e.g., Borenstein et al., 2012; Dong & Maynard, 2013; Raudenbush et al., 2011) is a key feature of these advancements; and the accumulation of (covariate-adjusted) variance estimates for student achievement outcomes (e.g., Bloom et al., 2007; Hedges & Hedberg, 2007; Westine et al., 2013) has been an integral element of this (Spybrook, 2013). Reasonable assumptions on such so-called *design parameters* are indispensable to plan sensitive RTs that demonstrate sufficient statistical power to detect the hypothesized effect of an educational intervention with a high level of statistical precision (i.e., with a low standard error). As the quote above by Murray (1998) suggests, the accuracy of the design parameters thereby strongly hinges on the idiosyncrasies of the target RT (in terms of, e.g., population, outcome and covariates, or experimental design and analysis model). Note that throughout this thesis, I refer to design parameters as the variance estimands (and their estimates) defining a certain RT design, specifically the intraclass correlation coefficient (ICC) ρ, and the amount of explained variance R^2.[1] Note further that I use *design sensitivity* as a conceptual umbrella term embracing both statistical power and statistical precision (Hedges & Hedberg, 2013).[2]

Overarching Objective of the Present Doctoral Thesis

In the present doctoral thesis, I strive to substantially expand available resources to perform power analysis when planning RTs on student achievement. More precisely, this dissertation accumulates extensive compendia of reliable and versatile (meta-analytically integrated) design parameter estimates consistent with the German (and similar) school context and manifold competence domains across the entire school career (Grade 1 to 12), along with thorough (theoretically derived and empirically established) guidance, for instance, on covariate selection to boost statistical power and precision. The emerging compilation is supposed to assist and guide evaluation researchers in education and psychology to design strong RTs of various designs and with different analysis models for treatment effects, which are conducted with a view to produce useable knowledge on what works to promote student achievement.

---

[1] In doing so, I adopt a narrow meaning of the term "design parameters." Note that in the (methodological) literature, the notion is used with some ambiguity. If used rather inclusively, design parameters refer to all estimands and quantities that define a certain design, including variance parameters, but also the effect size, the noncentrality parameter, the sample size, and so on (see e.g., Spybrook et al., 2014). Note that this thesis does not provide empirical estimates of the treatment effect heterogeneity design parameters additionally required in power analysis when planning multisite RT designs (see Section 1.4.4).

[2] In the (methodological) literature, the use of the notions "(design) sensitivity", "(statistical) power", and "(statistical) precision" is ambiguous; sometimes "power" is simply used in a broader sense including all these concepts (Cumming, 2014).

Structure of the Present Doctoral Thesis

Chapter 1 introduces the substantive and methodological key concepts underlying this work. First, I delineate the rationale motivating the need for design parameters on student achievement, which in essence arose from the growing importance of RTs to inform evidence-based educational practices and policies (Section 1.1). Next is the definition of student achievement (Section 1.2) and a presentation of the RT designs for which the present design parameters are relevant (Section 1.3). Subsequently, I introduce and elaborate on the foundations of power analysis, with a special emphasis on the peculiarities and challenges associated with multilevel RT designs, also encompassing the formal definition of the design parameters at the various hierarchical levels (Section 1.4). Then, I briefly summarize the current state of knowledge on design parameters for student achievement, from both an international as well as a German perspective, in particular identifying vital research gaps (Section 1.5). Finally, I present the objectives of the present thesis alongside the identified research gaps (Section 1.6). The two following chapters constitute the empirical part of the present dissertation. Study I in Chapter 2 analyzes multilevel design parameters for student achievement in various domains across elementary and secondary school, taking into account vital covariates. Study II in Chapter 3 focuses on the improvement of design sensitivity in single- and multilevel RTs by adding design parameters for a large spectrum of diverse covariate sets, which is enriched through concrete guidelines on covariate choice. Chapter 4 provides an overarching, albeit nuanced discussion of the present thesis. Again alongside the identified research gaps, I first summarize the key results derived from the two studies, situating them in the literature, and deriving important implications for the design of RTs on student achievement (Section 4.1). I then put the spotlight on a selection of remaining core challenges in the planning of RTs seeking to inform evidence-based education (Section 4.2). After outlining strengths and limitations of the present work and pointing out some—in my opinion appealing—directions for upcoming works and extensions (Section 4.3), this doctoral thesis closes with some final concluding remarks (Section 4.4).

1

INTRODUCTION AND BACKGROUND

1.1 The Experimental Renaissance of Modern Educational Psychology, Or: Why Do We Need Design Parameters?

Evidence-based policies and practices are key to successful education—and declared goal of governmental authorities all around the globe (Dekker & Meeter, 2022; Hedges & Schauer, 2018; Organisation for Economic Co-operation and Development [OECD], 2007; Pellegrini & Vivanet, 2021; Slavin et al., 2021), as in Germany (Bundesministerium für Bildung und Forschung [BMBF], 2018; Kultusministerkonferenz [KMK], 2016). Since the advent of the millennium, educational policymakers and practitioners increasingly prioritize empirically supported knowledge over traditional, rather hermeneutic notions when investing in and adopting educational innovations, products, and services (Hedges, 2018; Slavin, 2002). In particular, stakeholders request valid answers on how to improve students' achievement: Since academic performance not only fundamentally shapes every student's personal life but also a whole nation's wealth (Steinmayr et al., 2014), its promotion is among the central concerns of evidence-based education (OECD, 2007; Slavin, 2020; see Section 1.2 for the definition of student achievement adopted in this dissertation).

Most apparently in the United States, the shift towards evidence-based education along with its initial legalizations (the U.S. No Child Left Behind Act in 2001 and the Education Sciences Reform Act in 2002) both established whole infrastructures of funding, bundling, and dissemination of rigorous research (e.g., the U.S. Institute of Education Sciences [IES] along with the What Works Clearinghouse [WWC]) as well as creating demands for high-quality studies (Hedges & Schauer, 2018; Whitehurst, 2003). These movements—radiating beyond the United States—initiated a new era of modern educational psychology, culminating in what Raudenbush and Schwartz (2020, p. 177) titled a "methodological renaissance" with a strong emphasis on randomized (field) experiments (because of which I redefine it as *experimental* renaissance; Boruch, 2003; Slavin, 2020; Whitehurst, 2012).

A randomized trial (RT) is a study under controlled conditions, where units (e.g., individual students or entire schools) are allocated by chance (like by tossing a coin[3]) to receive some deliberate intervention (i.e., a treatment) or not, in order to test its effect (Shadish et al., 2002). Widely appreciated as the most unbiased *and* efficient study design to address causality (Shadish et al., 2002, pp. 247–248), RTs have become indispensable tools to rigorously evaluate

[3] Although coin toss is indeed still sporadically used (Bruce et al., 2022), randomization is rather done by computer-based algorithms nowadays (e.g. random number generator). Interestingly though, the randomness of coin toss has itself been experimentally studied (e.g., Clark & Westerberg, 2009; Gelman & Nolan, 2002).

educational interventions (IES & National Science Foundation [NSF], 2013; Mosteller & Boruch, 2002; Slavin, 2002; Spybrook, Shi, et al., 2016).

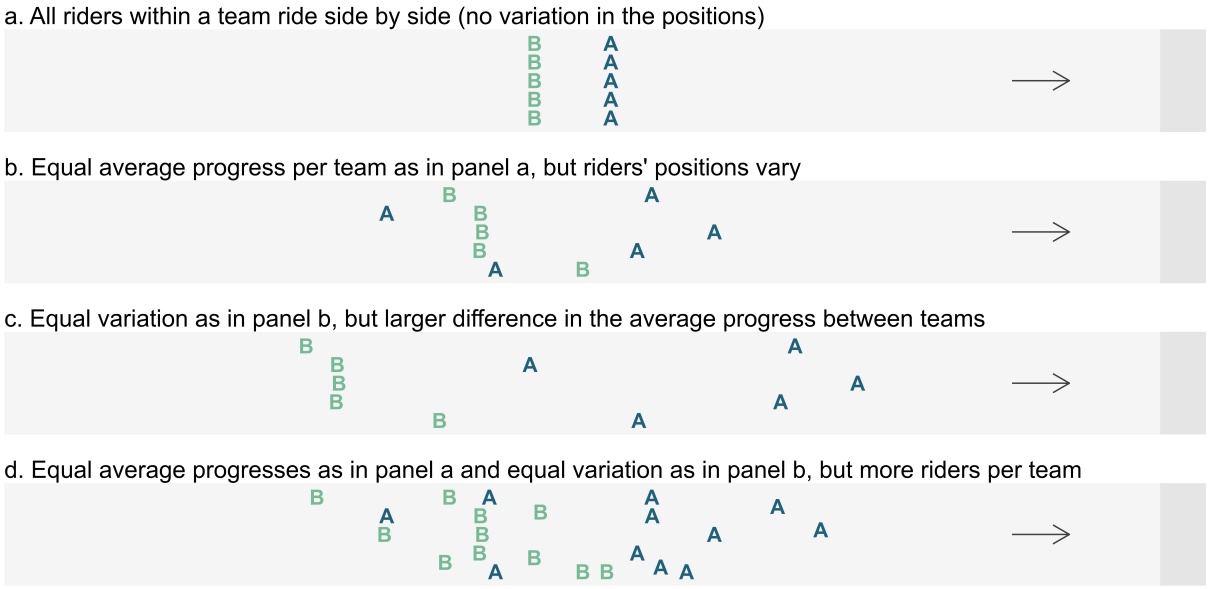
The notion 'renaissance' captures it fairly well: The RT design has by now traversed a (far more than) centennial, fascinating history in which psychology as a field truly plays a pioneering role (here, the famous weight experiments by Charles S. Peirce and his student Joseph Jastrow in the late eighteen-hundreds come to mind,[4] see e.g., Dehue, 1997, 2001; Jamison, 2019; Stigler, 1992, for some excellent digests on the origin of RTs, and their roots in psychology and education). Sir Ronald A. Fisher (1925, 1935) later refined the basic principles of RT design, in agriculture though, but they became popularized far beyond. Not least the fact that since then quite some classics of experimental research methodology appeared in psychological journals (D. T. Campbell, 1957, 1969; Raudenbush, 1997; Rubin, 1974) witnesses that RTs hold a deep-seated key position in (educational) psychology. Notwithstanding, it is a rather new development that educational RTs enormously flourished (see e.g., Connolly, 2017; Gorard et al., 2017; Morrison, 2020; Mosteller & Boruch, 2002), not even two decades old: More than three quarters of the over 1,000 educational RTs reviewed by Connolly et al. (2018) that were internationally carried out between 1980 and 2016 do not predate 2007.

A crucial prerequisite of any RT to allow for conclusive and correct claims on an intervention's impact is, that it is designed to be sensitive to the treatment effect (Hedberg, 2018; Lipsey, 1990). Design sensitivity embraces both statistical power and statistical precision (see also Hedges & Hedberg, 2013). But what does this mean? I basically refer to the sensitivity of an RT design as its capacity (in terms of probability; i.e., statistical power) to detect a *real* contrast between the experimental groups on the studied outcome at a given level of statistical significance with a low standard error (i.e., statistical precision; see Section 1.4.3 for more detail). Of importance, this capacity is heavily determined by the degree of outcome variation in the experimental groups, relative to their contrast (Lipsey, 1990): Usually, low variance barely affects the ability to precisely observe a treatment effect, while large variance, or much random noise fades the signal of any treatment effect, unless this contrast is very pronounced, resulting in limited power and precision (Raudenbush et al., 2007).

[4] Peirce and Jastrow (1885) adapted a self-experiment by Gustav T. Fechner (a pioneer in psychophysics) in which they had to decide which of two concealed and indistinguishable weights was heavier. Aiming for unbiased results, for example by which hand was used or which weight was picked up first, they created a special deck of cards that determined the conditions of each trial. Although anything but perfect, this is gauged the first experiment that applied random assignment—located in psychology (Hacking, 1988; Jamison, 2019; Stigler, 1992).

The pivotal influence of relative variation—or, put differently, the signal-to-noise ratio in group contrasts for design sensitivity can be illustrated by means of the following example (see Figure 1): Imagine a road bike race. We try to judge which of the two Teams A or B is closest to the finish line, on average, as a group. If all riders within Team A ride side by side, and the same happens in Team B (see Figure 1a), we can immediately observe that Team A is leading. The position of the riders within the teams does not vary, which facilitates to precisely assess the contrast. But if the riders spread over the route (see Figure 1b), it becomes much harder to identify the faster team. Yet, the average progresses are exactly the same as in Figure 1a: Team A is leading. The variation in the riders' position within the teams induces noise, blurring the contrast. However, holding the variation in the positions of the single riders constant, if the distance between the two teams grows (see Figure 1c), or the number of riders per team increases (see Figure 1d), the advantage of Team A becomes readily apparent again (see Lipsey, 1990, for a similar illustration).

Figure 1. *Illustration of the Signal-to-Noise Ratio in Detecting a Group Contrast*



Note. Adapted from "Design Sensitivity. Statistical Power for Experimental Research." (Figure 1.1, p. 15) by M. W. Lipsey, 1990, SAGE Publications. Copyright 1990 by SAGE Publications. Reprinted and adapted with permission.

Extrapolated to applied experimental research—where the (mean) contrast between experimental groups is of interest—this illustration implies that both larger treatment effects and larger sample sizes raise design sensitivity, whereas larger outcome variance causes the exact opposite. Unfortunately, various factors induce variation in applied RTs (e.g., sampling, measurement and estimation error; Lipsey, 1990). And, in educational psychology, the dilemma is even exacerbated by the nature of the population to be studied: As in many other scientific

fields, the population of interest is not uniform, but hierarchically structured; individuals are nested within groups—a statistician would say "clusters." For instance, in health care, patients are nested within clinics; in economics, employees are nested within companies; in politics, delegates are nested within parties; and finally in education, students are typically nested within classrooms and schools. Importantly, these groupings do not arise by chance. Rather, the individuals within these clusters do have some kind of connection, they share cluster-specific commonalities, norms, and standards (Kreft & de Leeuw, 1998; Murray, 1998). In the institutionalized school system, this connection is quite obvious: A school represents the (physical) environment for learning and teaching that has its idiosyncratic characteristics (e.g., school climate, teachers' professional level, student composition), and students within these schools co-influence each other. The same principle naturally extends to classrooms within schools. As a result, variation distends among students (within classrooms and schools), classrooms (within schools), as well as schools (Raudenbush & Schwartz, 2020).

Undoubtedly, (small-scale) single-level RTs that randomly sample and assign individual students (irrespective of the membership to a classroom or school; e.g., Harks et al., 2014; Loosli et al., 2012) are fundamental pieces contributing to the generation of knowledge on ways to improve student achievement. Specifically, they offer necessary insights into how to conceive and possibly customize deliberate interventions, and facilitate probing their efficacy (Hedges, 2022; Roland & Torgerson, 1998). At the same time, to test the effectiveness and scalability of these interventions, it is essential to implement RTs at larger scales in ecologically valid settings (D. T. Campbell, 1957; Moerbeek & Teerenstra, 2016; e.g., Gersten et al., 2015). Therefore, more complex RTs that mimic the hierarchical structure (in the sampling, design, as well as analysis stage) have become the methodological instruments of choice when probing educational interventions in the school context (Hedges & Rhoads, 2010a; Spybrook et al., 2020). Such (large-scale) multilevel RTs often address the inherent nesting of the target population by randomly assigning the treatment either to entire student clusters (e.g., classrooms or schools) or to students *within* these clusters (e.g., classrooms or schools serving as sites; see Section 1.3.2 for a presentation of the various multilevel RT designs covered in this thesis).

The major drawback associated with any multilevel RT is an (often dramatically) reduced design sensitivity, as contrasted with a single-level RT (e.g., Bloom et al., 2007; Hedges & Rhoads, 2010a; Lipsey & Hurley, 2009; Raudenbush et al., 2007; Schochet, 2008): The described exposure to a joint environment in conjunction with the mutual influences cause the achievement scores of students within the same classroom and school to resemble each other

(Donner & Klar, 2000). From a statistical point of view, this means that the errors associated with the students within a classroom or school are likely to be correlated (Kreft, 1993; Schochet, 2008). This correlation between individuals within groups, or the other way around, the amount of between-cluster differences, is typically expressed by the intraclass correlation (ICC) ρ (where $0 \leq \rho \leq 1$; ρ is formally defined for different hierarchical levels in Section 1.4.4). In a multilevel RT, the ICC inflates the sampling variances of the outcome (means) in the experimental groups,[5] and thus, compromises the precision of the treatment effect estimate (i.e., bloats its standard error). It follows that, given $\rho > 0$ (which holds virtually always true; Murray, 1998, p. 8), the statistical power of the test in a multilevel RT will always be attenuated as compared to a single-level RT of the exact same total sample size (Hedges & Rhoads, 2010a).[6]

Crucially, one of the techniques proven beneficial to thwart the detrimental consequences of inflated variances is covariate adjustment. In fact, statistically controlling for covariates can (often substantially) raise design sensitivity—notably, regardless of the RT design (e.g., Bloom et al., 2007; Kahan et al., 2014; Maxwell et al., 2017; Porter & Raudenbush, 1987; Raudenbush, 1997; Raudenbush et al., 2007, 2007). As schematically visualized in Figure 2 for a multilevel RT with two hierarchical levels and randomization occurring at the cluster level (e.g., students within schools, and schools are randomly allocated to experimental conditions), baseline covariates that are correlated with (i.e., explain variance in) the outcome reduce error variance in the outcome (Porter & Raudenbush, 1987). This mechanism improves the signal-to-noise ratio of the treatment effect estimate (Raudenbush et al., 2007), which boosts power and precision (unless the sample size is very small; see Konstantopoulos, 2012; Liu, 2011; Moerbeek & Teerenstra, 2016). The proportion of explained variance by one or more covariates is indexed by R^2 (where $0 \leq R^2 \leq 1$; R^2 is defined for different hierarchical levels in Section 1.4.4).

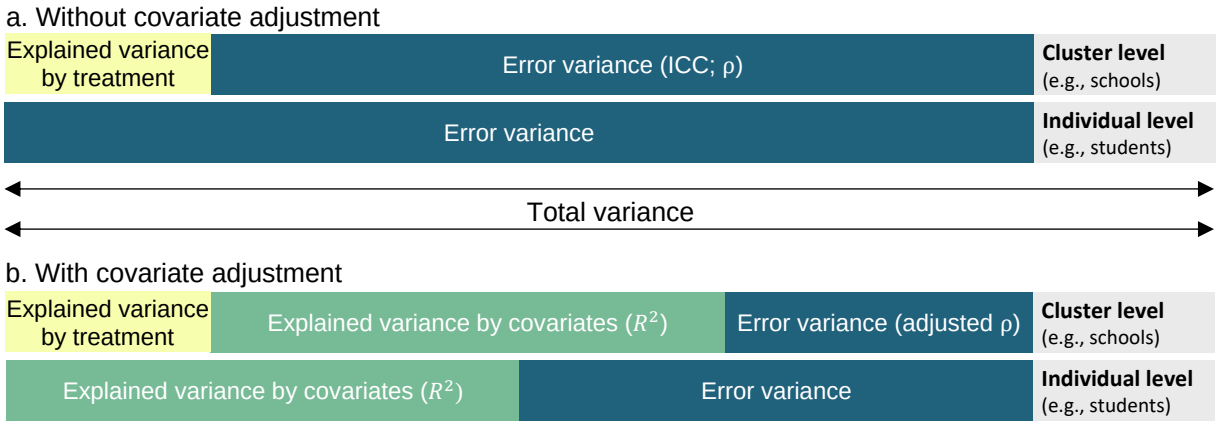
---

[5] For continuous (or binary) outcomes, this property is reflected in the so-called variance inflation factor (VIF; Donner et al., 1981; also referred to as the "design effect" in sampling theory; Kish, 1965). The VIF is conceived as the ratio of the variance in a clustered random sample (see also Section 1.3.2) to the variance in a simple random sample of the same total sample size (irrespective of any randomization). See Appendix A for details.

[6] Major foundations of this understanding have been laid as early as almost one century ago, which, interestingly, root in educational psychology (see Hedges & Schauer, 2018): Everett F. Lindquist (1940), one of the pioneers of modern psychometrics and large-scale assessment, spread the idea of biased conventional single-level significance tests when using data actually obtained from multilevel RTs. Soon after, Walsh (1947) demonstrated how ignoring the hierarchical variance structure in such data may overrate precision, and how growing cluster differences aggravates bias. Finally, in his seminal paper, Cornfield (1978) was the first who formalized the analytical aspects posed by cluster randomization. He is also the originator of the (under experimental methodologists) well-known credo: "Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception." (Cornfield, 1978, pp. 101–102)

In total, whereas between-cluster differences (with $\rho > 0$) generally hamper design sensitivity in multilevel RTs, predictive covariates (with $R^2 > 0$) tend to raise design sensitivity in both single- and multilevel RTs. However, which are the concrete implications of growing ICCs and explained variances for well-designed (i.e., sufficiently powered and precise) RTs on student achievement?

Figure 2. *Variance Decomposition in a Multilevel RT with Cluster Randomization*

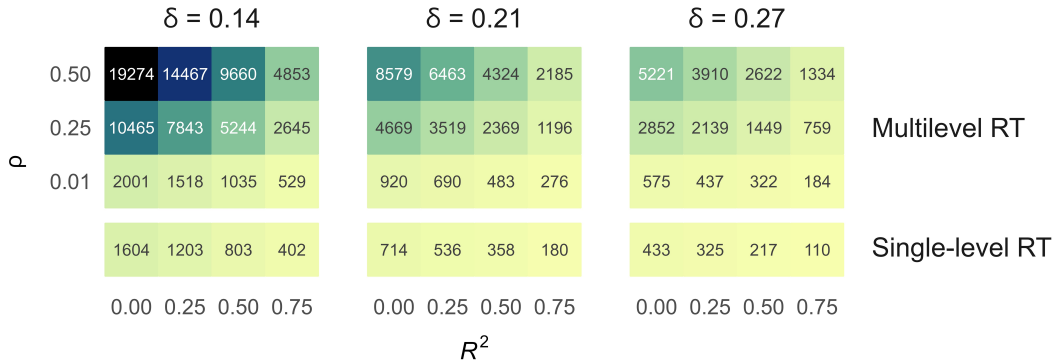


Note. Variance component partitioning when the treatment explains 20% of the total variance at the cluster level, and (a set of) individual- and cluster-level covariates explain 50% of the total variance at either hierarchical level.

Consider, for example, a multilevel RT of the form depicted in Figure 2 in which K schools are randomly allocated to experimental groups, with (a fixed number of) $n_k = 23$ students nested within each school (representing the typical school size in large-scale educational RTs; see Lortie-Forgues & Inglis, 2019). Figure 3 portrays the total sample size $N = Kn_k$ that is required for such a multilevel RT to be sensitive to the (standardized) treatment effects that define students' typical (range of) annual academic growth in German lower secondary school (meta-analytic average and 95% confidence interval [CI]; Brunner, Stallasch, et al., 2023, Table 1), in juxtaposition to a conventional single-level RT (i.e., ignoring clustering) as placed at the bottom row. First, the sample size requirements are quickly augmented with growing homogeneity among students within schools, especially without covariate adjustment ($R^2 = 0$): For instance, to find an effect of $\delta = .14\ SD$ (lower bound of the 95% CI) in some achievement test, almost 400 more students are required for a multilevel RT ($N = 2{,}001$, $K = 87$) compared to a single-level RT ($N = 1{,}604$) when only 1% of the total variance in the studied outcome can be attributed to between-school achievement differences; and when $\rho = .50$, the required sample size for a multilevel RT strikingly inflates to $N = 19{,}274$ ($K = 838$). Notably, such large ICCs are by no means uncommon in German lower secondary school (see Brunner et al., 2018; Knigge & Köller, 2010). The same pattern of results is

observed for $\delta = .21$ (average) as well as $\delta = .27$ (upper bound of the 95% CI), although it becomes evident again (remember the road bike race scenario) that treatment effects of larger magnitudes are detectable with fewer students (everything else held constant).[7] Second, the adjustment for strong (i.e., highly prognostic) covariates substantively diminish sample size requirements: When including one student-level and one school-level covariate that explain large amounts of variance of $R^2 = .75$ at either level, the necessary sample sizes can be virtually quartered, irrespective of the RT design and the degree of clustering (e.g., to detect $\delta = .21$, $N = 8{,}579$ students would have to be sampled without covariates, but only $N = 2{,}185$ with covariates).

Figure 3. *Total Sample Size Required to Detect Treatment Effects in a Single- vs. Multilevel RT as a Function of Design Parameters*



Note. Power analysis (see Section 1.4 for details) for a single-level RT vs. multilevel RT (students within schools; schools form the unit of randomization) under complete balance (i.e., samples are randomly assigned to the treatment and control condition in equal shares; constant school size of $n_k = 23$ students) to detect a statistically significant standardized treatment effect of $\delta = .14/.21/.27$ at $\alpha = .05$ (two-tailed) in a two-sample independent t-test with 80% statistical power. Designs with covariate adjustment (i.e., $R^2 > 0$) include one covariate at the student level and one covariate at the school level which explain the same amount of variance at either level. The hypothesized effect sizes represent the typical range (i.e., 95% confidence interval) and the average of the annual growth in students' achievement in German lower secondary school, meta-analytically averaged across Grades 5 to 10 (Brunner, Stallasch, et al., 2023, Table 1). The typical school size was calculated as the average number of students within schools across the educational large-scale RTs reviewed by Lortie-Forgues and Inglis (2019).

To sum up: Strong (i.e., well-designed) RTs maximize the chances for conclusive and correct causal inferences on the effectiveness of educational interventions, programs, and innovations, while maintaining cost-efficiency. Crucially, only if RTs are adequately powered and precise (in short: sensitive), they can serve as indispensable methodological tools to generate valid knowledge informing evidence-based policies and practices in education. Therefore, sound power analysis is an integral component in the planning of any RT on student achievement for which substantive guidance is needed—irrespective of its particular design.

[7] Of importance, in this concrete example of $n_k = 23$ students per school, it holds true across hypothesized effect sizes, that the increase in the required sample size is almost tenfold when moving from $\rho = .01$ to $\rho = .50$. However, the exact percentage increase depends on n_k (as also implied by the VIF, see Appendix A).

Yet, with multilevel designs, by contrast with single-level designs, this is a more sophisticated venture additionally requiring educated assumptions on the degree of clustering in the outcome's variance structure. Thus, researchers rely on reliable values of the design parameters ρ and (in both single- and multilevel RTs) R^2. As numerous scholars stressed for decades, these estimates are sharply context-specific and should therefore be optimally tailored to the RT's target population, achievement outcome, and possibly applied covariate set, as well as its concrete design and planned analysis model to guarantee strong designs (e.g., Bloom et al., 2007; Brunner et al., 2018; M. Campbell et al., 2000; Cohen, 1988; Donner & Klar, 2000; Hedges & Hedberg, 2007; Lipsey et al., 2012; Moerbeek & Teerenstra, 2016; Murray, 1998; Schochet, 2008; Spybrook, 2013; Zhang et al., 2023).

The remainder of this chapter is organized as follows. Section 1.2 defines student achievement as a core target outcome of educational RTs. Section 1.3 presents the RT designs covered in this thesis. Section 1.4 introduces and elaborates on power analysis, putting the spotlight on the peculiarities of multilevel RTs, and formulizes the level-specific design parameters. Section 1.5 summarizes the current research state on design parameters for student achievement, from an international and German perspective, identifying vital research gaps. Section 1.6 details the objectives of the present thesis alongside these identified research gaps.

1.2 Student Achievement as Target Outcome of Randomized Experiments

The enhancement of student achievement ranks among the pivotal endeavors of evidence-based education (OECD, 2007; Slavin, 2020), as it has far-reaching ramifications—in individual as well as macrosocial respect (Steinmayr et al., 2014). Hence, it might not come as a surprise that more than one third of the educational RTs reviewed by Connolly et al. (2018) explicitly targeted student achievement, in various domains. In two other reviews propounded by Spybrook and Raudenbush (2009) and Spybrook et al. (2016) that cover RTs funded by the IES in the years 2002 to 2004 and 2011 to 2013, the dominance of achievement-related over -unrelated RTs was even more pronounced (55% and 73%, respectively).

Student achievement is a multifaceted construct (Steinmayr et al., 2014). It results from long-term and cumulative mechanisms of knowledge acquisition determined by a myriad of factors (Baumert et al., 2009; Winne & Nesbit, 2010). Typically, an achievement outcome in a certain domain (e.g., mathematics) indicates the degree of goal attainment in learning tasks or

activities in that very same domain (e.g., the number of correctly solved calculations), and is often assessed via school grades or standardized achievement tests (Steinmayr et al., 2014). Since school grades are given by teachers with reference to a particular classroom or school framework, standardized tests represent more objective, and thus, comparable measures—across students, classrooms, or schools (Borghans et al., 2016; Brookhart, 2015). Contemporary conceptualizations of domain-specific achievement, as underlying, for instance, the standardized tests of large-scale assessment studies reflect the multidimensionality of domain-specific achievement by uniting both content dimensions (e.g., mathematical quantity, space and shape) and cognitive process dimensions (e.g., mathematical modeling, problem solving; Neumann et al., 2013; OECD, 2013).

1.3 Experimental Designs

RTs facilitate unbiased causal claims on the impact of educational programs and innovations (IES & NSF, 2013; Mosteller & Boruch, 2002; Slavin, 2002; Spybrook et al., 2016), given some basic assumptions (Shadish et al., 2002). The core feature of an RT as opposed to any non-experimental study is, as the name implies, randomization. Randomization means to select units to receive a treatment or not, completely by chance. In such a two-arm or two-group study[8], the first resulting group forms the treatment group TG and the second the control group CG (often doing "business as usual"). This tactic, in principle, eliminates any systematic differences between experimental groups that might exist prior to the treatment. This way, factors both known and unknown to be related to the treatment are offset and differences in the outcome under investigation are likely exclusively due to the treatment (M. J. Campbell & Walters, 2014; Shadish et al., 2002). For instance, randomizing students to receive a French language training or not allows a "fair comparison" (Boruch, 2003, p. 107) of the outcomes in a respective posttest between the TG and CG, irrespective of whether single students spent a year abroad in France or show higher linguistic affinity than others. Obviously, this would only work perfectly with infinitely large samples; therefore, a certain likelihood of imbalance between the experimental groups always remains (M. J. Campbell & Walters, 2014).

Power analysis for RTs hinges on the target experimental design. Many different experimental designs have been developed (see e.g., Kirk, 2013, for a thorough presentation).

[8] The delineations and discussions in this thesis are limited to two-arm RT designs with one single treatment group TG and one single control group CG.

The design parameters accumulated in the present doctoral thesis inform the RT designs most frequently implemented in educational research (see Connolly et al., 2018; Hedges & Rhoads, 2010a; Spybrook, Shi, et al., 2016; Spybrook & Raudenbush, 2009). Figure 4 illustrates their respective randomization schemes and data structures; the Appendix B provides formulations of the corresponding unconditional (i.e., without covariates) and conditional (i.e., with covariates) statistical models. These designs may be roughly classified with respect to two dimensions: (a) the sampling process (simple single-level vs. complex multilevel), and (b) the randomization unit (individual vs. cluster vs. subcluster).

The implications for both the design and the analysis of an RT that follow from the underlying *sampling process* can hardly be overstated (Hedges & Rhoads, 2010a): The sampling technique fundamentally determines the assumptions on the stochastic (in)dependence of the students, and thus, the complexity of the data structure and the properties of the derived statistics.

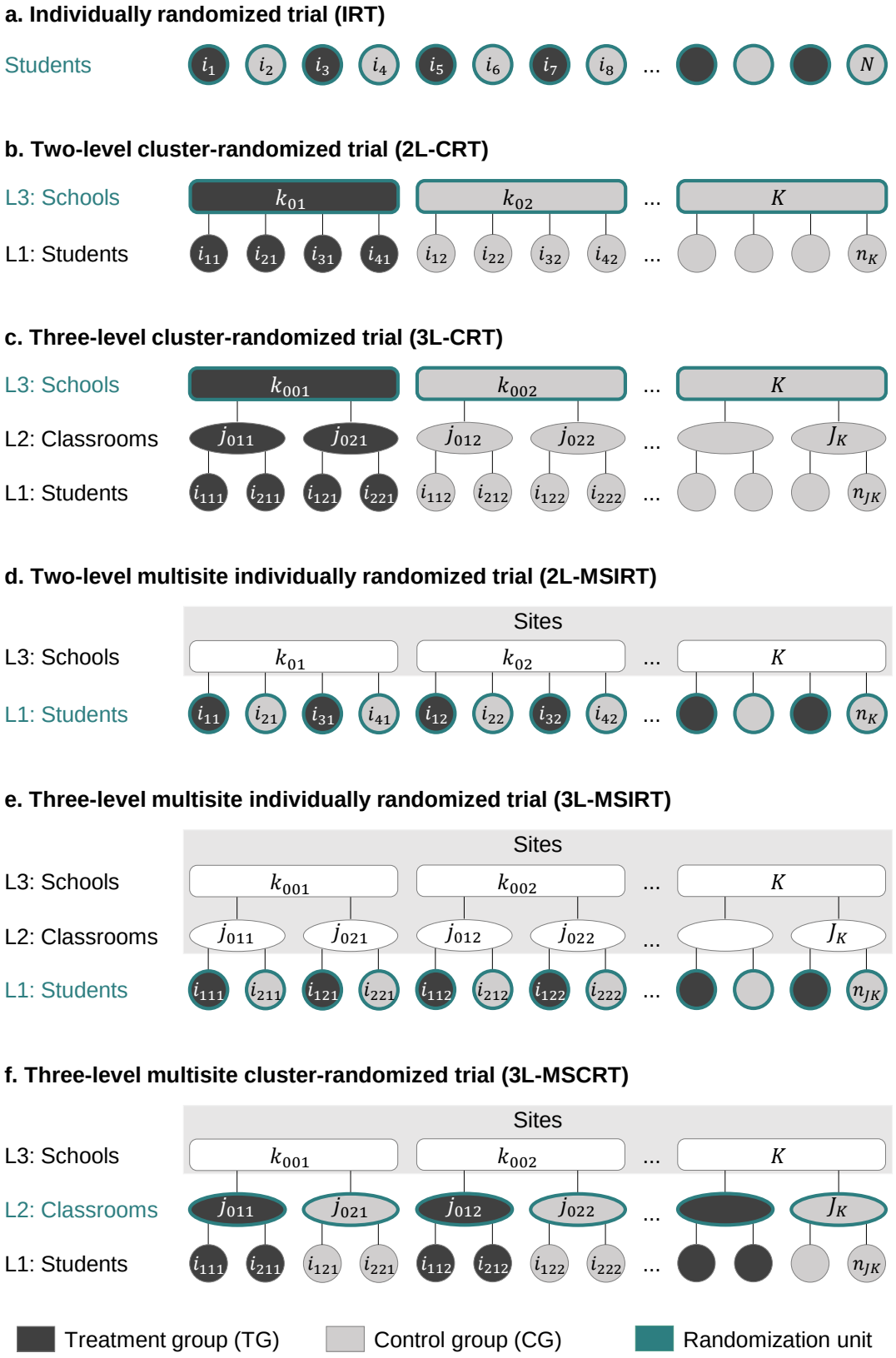
1.3.1 Single-Level Designs

Under simple random sampling, the students are selected independently of one another, at an entirely individual basis. In statistical terms, this means that all observations are assumed to be stochastically independent.

Individually Randomized Trial

From a simple random sampling procedure, the most basic design emerges: an *individually randomized trial* (IRT; e.g., Bloom, 2006; Dong & Maynard, 2013; see Equations (B1)–(B2) in Appendix B). An IRT is also referred to as "single-level" (e.g., Zhang et al., 2023), "completely" (e.g., Hedges & Rhoads, 2010b), or "simple" (e.g., Moerbeek & Teerenstra, 2016) randomized design. In an IRT as portrayed in Figure 4a, individual students are sampled independently of each other and are randomly assigned to the experimental groups, regardless of the classroom or school they attend; likewise, the experimental and control protocols are delivered at an individual basis (Lohr et al., 2014). An IRT design may be used in a number of scenarios. Examples are predominated by studies conducted in laboratory-similar settings under well-controlled conditions by well-trained staff (e.g., Harks et al., 2014; Karbach et al., 2017; Karbach & Kray, 2009; Loosli et al., 2012), but also include studies conducted in informal learning settings (e.g., Biggart et al., 2013) or with students from only one single classroom or school (e.g., Kelly et al., 2013).

Figure 4. *Experimental Designs Covered in the Present Doctoral Thesis*



Note. The design parameters estimated in the present doctoral thesis are appropriate for power analysis to plan the six RT designs shown. Figure 4a: $i \in \{1, 2, \ldots, N\}$ students are independently randomized at an individual basis (i.e., irrespective of classroom or school membership). Figures 4b and 4d: $i \in \{1, 2, \ldots, n_k\}$ students at Level (L) 1 are nested within $k \in \{1, 2, \ldots, K\}$ schools at L3. Figures 4c, 4e, and 4f: $i \in \{1, 2, \ldots, n_{jk}\}$ students at L1 are nested within $j \in \{1, 2, \ldots, J_k\}$ classrooms at L2 which are nested within $k \in \{1, 2, \ldots, K\}$ schools at L3.

Note, however, that IRTs represent usually small-scale experiments, and are in fact most often based on convenience samples rather than fully random samples from a population, raising concerns about their limited generalizability (Hedges & Rhoads, 2010b; Stuart et al., 2011; Tipton & Olsen, 2018; see also Section 4.2.2). Nonetheless, IRTs are integral cornerstones in educational evaluation research as they not only allow to develop, calibrate, and modulate innovative programs and measures but may also offer unique possibilities to draw causal conclusions on their general efficacy (e.g., for answering research questions such as: "Does the underlying didactic approach of this literacy training would improve students' reading comprehension under ideal conditions?"; see Hedges, 2022; Roland & Torgerson, 1998; Stuart et al., 2011).

1.3.2 Multilevel Designs

The great majority of RT designs in education involves some sort of complex cluster and/or multistage random sampling (Connolly et al., 2018; Hedges & Rhoads, 2010b), where a pool of (intact) clusters instead of individual subjects is selected from a pre-defined population (see e.g., S. K. Thompson, 2012). In many educational experiments, for instance, whole schools are sampled. The complete student body in each school may then be collectively assigned to either the TG or the CG (representing a cluster sample; e.g., Stullich et al., 2007). Alternatively, from this pool of schools, students may be sampled in a second step (representing a two-stage sample; e.g., Corrin et al., 2015), or again, (intact) clusters—typically classrooms—may be sampled (representing a two-stage cluster sample; e.g., Cook et al., 2000), possibly followed by the sampling of students in a third step (representing a three-stage cluster sample; e.g., Itzek-Greulich et al., 2017). In either case, observations of students can no longer be regarded as stochastically independent. Rather, such sampling techniques lead to complex hierarchical data structures whose (statistical) properties often deviate drastically from those of simple random samples (Hedges & Rhoads, 2010b). As IRTs, RTs incorporating multilevel samples lead to unbiased inferences, as long as design and analysis adequately account for the clustering (Raudenbush, 1997).

Notably, hierarchically nested samples do not exclusively result from cluster or multistage sampling procedures. Even when students are independently sampled (and individually assigned to experimental conditions at random), many educational treatments are either delivered in group settings or operate at the group level by definition (Bloom, 2005; Boruch & Foley, 2000; Cook, 2005), still inducing a clustered variance structure (Lipsey & Hurley, 2009). Such kind of RTs are often labeled individually randomized group treatment

trials (IRGT; e.g., Moerbeek & Teerenstra, 2016) where randomization determines the cluster membership of individuals (Moerbeek & Teerenstra, 2016). One of the most prominent examples is the project STAR (Finn & Achilles, 1990), also known as the Tennessee class size experiment, where kindergarten children (as well as teachers) were independently and randomly enrolled into small or large classrooms installed within participating elementary schools. The treatment itself (i.e., teaching within small vs. large classrooms) then occurred at the cluster level. Similarly, in the evaluation of U.S. charter schools (Gleason et al., 2010), students were allocated at the individual level to either a reform (TG) or another (CG) school by means of lotteries.

This thesis considers hierarchically clustered (i.e., multilevel) designs with, in total, two and three hierarchical levels (L). In the two-level designs, $i \in \{1, 2, \dots, n_k\}$ students at L1 are nested within $k \in \{1, 2, \dots, K\}$ schools at L3. In the three-level designs, $i \in \{1, 2, \dots, n_{jk}\}$ students at L1 are nested within $j \in \{1, 2, \dots, J_k\}$ classrooms at L2 which are, in turn, nested within $k \in \{1, 2, \dots, K\}$ schools at L3. Multilevel RTs may be further distinguished with regard to the level or *randomization unit*. Specifically, whether treatment allocation occurs at the top hierarchical level or within the top hierarchical level has important consequences for—not exclusively, but especially—the design of the RT. In the former case, schools (i.e., the top-level units) are nested within experimental conditions, which characterizes a cluster-randomized trial (CRT); in the latter case, experimental conditions are crossed with the random effects of the schools within which randomization occurs, constituting a multisite randomized trial (MSRT). Sample sizes at either hierarchical level held constant, it can be shown that CRTs demonstrate less efficiency than MSRTs (due to a larger variance in CRTs as compared to MSRTs; Moerbeek & Teerenstra, 2011). Hence, design sensitivity tends to be smaller for CRTs as opposed to MSRTs; yet, power calculations for MSRTs require a larger set of input parameters.

Cluster-Randomized Trial

A CRT (e.g., Raudenbush, 1997), which is also referred to as "group-randomized" (e.g., Murray, 1998) or "place-based" (e.g., Bloom, 2005) design, assigns intact clusters of individuals to experimental conditions. Importantly, in a straight CRT, treatment allocation always occurs at the top hierarchical level (Hedges & Rhoads, 2010a). Especially in the United States, CRTs have a long—in fact over 100 year-old (see Hedges & Schauer, 2018)—tradition and are nowadays deeply seated tools in educational research. In fact, Connolly et al. (2018) found that 58% of educational RTs represent CRTs. This prevalence is motivated by several advantages encapsulated in this design (see Bloom, 2005, for a comprehensive list of

appropriate application scenarios): CRT designs are especially useful for interventions that operate best at the group level or that do this by definition, such as whole school reforms (Boruch & Foley, 2000; Cook, 2005). Moreover, in some situations, it may be unfeasible or unethical to favor individual over cluster assignment, for example, when an intact group of students has to be spatially and physically separated (Bloom, 2005). Finally, the most frequent rationale for a CRT is the prevention of contamination or spillover effects, reflecting undesired interdependencies among several students and/or among the outcomes of a single student (Bloom, 2005; Hemming & Taljaard, 2023). Yet, if not carefully designed, CRTs can be susceptible to different kinds of methodological bias (Hahn et al., 2005; Hemming & Taljaard, 2023), for instance, selection bias when recruitment is not blinded and/or students within classrooms or schools are selected post-hoc to randomization (Brierley et al., 2012; F. Li et al., 2022).

In a *two-level cluster-randomized trial* (2L-CRT; see Equations (B3)–(B8) in Appendix B) as shown in Figure 4b, students at L1 are nested within schools at L3, and entire schools receive the treatment. Inserting the classroom level, while keeping randomization at the level of schools, so that students at L1 are nested within classrooms at L2, which are, in turn nested within schools at L3, renders this design into a *three-level cluster randomized trial* (3L-CRT; see Equations (B9)–(B16) in Appendix B) as depicted in Figure 4c. Spybrook and Raudenbush (2009) as well as the follow-up review of Spybrook et al. (2016) revealed that both designs are employed with equal frequency.

Multisite Individually Randomized Trial

In recent years, multisite individually randomized trials (MSIRT; e.g., Raudenbush & Liu, 2000), which have also been called “blocked” designs (e.g., Konstantopoulos, 2008a) have gained popularity. An MSIRT randomly delivers the treatment to individuals within clusters, forming the sites or blocks. Put differently, in an MSIRT, one and the same experiment is replicated in several superordinate clusters, making it more feasible (Liu, 2014). Such designs offer appealing opportunities to evaluate educational interventions beyond what is possible via CRTs: In addition to the average treatment effect, they allow studying the extent and the determinants (e.g., mediator and moderator variables) of cross-site heterogeneity in treatment effects (Bloom & Spybrook, 2017; Raudenbush & Bloom, 2015; Weiss et al., 2014). Moreover, similar individuals can be matched within sites (therewith a main rationale behind this design is named) which should reduce between-site variance; and this between-site variance does not

affect the overall heterogeneity in the average treatment effect, so that power and precision may substantially be raised (Hedges & Rhoads, 2010a; Spybrook & Raudenbush, 2009).

A *two-level multisite individually randomized trial* (2L-MSIRT; see Equations (B17)–(B26) in Appendix B) as illustrated in Figure 4d randomly assigns students at L1 within schools at L3 to experimental conditions. For instance, Dynarski et al. (2004) conducted a 2L-MSIRT in which randomly selected students within each school attended an after-school program at an 21st Century Community Learning Center. Such a design may be extended by adding the classroom level, leading to a *three-level multisite individually randomized trial* (3L-MSIRT; see Equations (B27)–(B41) in Appendix B) as mapped in Figure 4e, wherein the treatment is delivered to students at L1 within classrooms at L2 within schools at L3. Thus, classrooms and schools form (nested) sites (Raudenbush & Schwartz, 2020). Such a design was implemented, for instance, by Torkildsen et al. (2022). In their study, randomization occurred within four large school blocks, and students within classrooms were individually allocated to either a morphological (TG) or mathematical (CG) training. Along with the quantification of cross-site treatment effect variability, Weiss et al. (2017) provide a comprehensive list of such MSIRTs with various hierarchical levels.

Multisite Cluster-Randomized Trial

A multisite cluster-randomized trial (MSCRT) couples cluster randomization and blocking: Figure 4f shows a *three-level multisite cluster-randomized trial* (3L-MSCRT; e.g., Dong et al., 2023, see Equations (B42)–(B53) in Appendix B) in which entire classrooms within school sites are randomly assigned to experimental conditions. In Spybrook and Raudenbush (2009), 50% of the reviewed RTs involving cluster randomization represented 3L-MSCRTs; in their follow-up examination, even 55% were of such a design (Spybrook, Shi, et al., 2016). A 3L-MSCRT was used, for instance, to evaluate the Open Court Reading curriculum in the United States (Borman et al., 2008). As noted by Weiss et al. (2017), such RTs, however, typically suffer from low site-level precision because the number of classrooms within a school is limited.

1.4 Power Analysis for Randomized Experiments

Statistical power is definitely part of the bedrock of statistics: Around the 1930s, Jerzy Neyman and Egon S. Pearson (1928, 1933) espoused the idea of stating an alternative hypothesis of *an*

effect (thought of as addition to Sir Ronald Fisher's null hypothesis of *no* effect), dichotomizing correct vs. incorrect regions in the space of possible statistical results (Pernet, 2016). Basically, their approach facilitates to differentiate between Type I and Type II errors in statistical decision making, and statistical power is the probability of not committing an error of the second type. At that time, however, the concept of statistical power was not at all warmly embraced by statisticians, above all not by the influential Fisher (1955) who denied that it is even possible (or helpful) to determine power (Descôteaux, 2007; Liu, 2014; Sedlmeier & Gigerenzer, 1989).

The consolidation of power analysis in psychology is widely credited to Jacob Cohen (1969, 1988), whose ground-breaking introductory book "Statistical Power Analysis for the Behavioral Sciences", as well as numerous further works (e.g., Cohen, 1962, 1973, 1992, 1994), has been sculpting the methodological landscape in the field and beyond until today (Perugini et al., 2018).

1.4.1 Preliminaries: Type I and Type II Error in Statistical Decision Making

The estimand of interest in an RT on student achievement is usually the average treatment effect on an achievement outcome Y. The estimator of the average treatment effect is the difference between the means in Y observed for the TG and the CG, that is $\bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}}$ (Bloom, 2006). For instance, one may ask: "Is the new curriculum, on average, effective to improve students' reading achievement?" The average treatment effect would then be computed as the mean reading score of the TG minus the mean reading score of the CG.

When making decisions on the (in)effectiveness of such an educational intervention in the framework of conventional null hypothesis significance testing (NHST)[9], 2×2 conclusion scenarios are possible (see Table 1; e.g., Cumming & Calin-Jageman, 2017, p. 148; Lipsey & Hurley, 2009). First, a true treatment effect either does exist or it does not exist in the population. The former corresponds to the alternative hypothesis H_1 (i.e., $\bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}} \neq 0$; e.g., the curriculum improves reading achievement); the latter corresponds to the null hypothesis H_0 (i.e., $\bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}} = 0$; e.g., the curriculum does not affect reading achievement).[10] Second, in

[9] For the sake of clearness, here, "conventional" refers to the classical NHST approach to statistical inference as typically used in applied (experimental) research, which in fact mixes the approaches introduced by Fisher (1925) on the one hand and Neyman and Pearson (1928, 1933; see Pernet, 2016; Sedlmeier & Gigerenzer, 1989). As Sedlmeier and Gigerenzer (1989) pointedly note, this hybrid approach would actually have been validated neither by Fihser nor the Neyman-Pearson team.

[10] There are also other possibilities to state these hypotheses. For instance, the H_1 may be formulated as a one-sided hypothesis of only a positive treatment effect (i.e., $\bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}} > 0$), instead of the stated two-sided hypothesis ($\bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}} \neq 0$) that implies that the treatment effect may be either positive or negative. Crucially, H_0 and H_1 cannot overlap though.

both cases, the statistical test for the treatment effect either is statistically significant or it is not statistically significant based on the sample data.

Table 1. *Types of Error and Statistical Power in Null Hypothesis Significance Testing*

Statistical conclusion	Population	
	Treatment effect exists H_0 is false	No treatment effect H_0 is true
Significant treatment effect ($p < \alpha$), reject H_0	Correct conclusion Probability: $1 - \beta$ (power)	Type I error Probability: α
No significant treatment effect ($p \geq \alpha$), fail to reject H_0	Type II error Probability: β	Correct conclusion Probability: $1 - \alpha$

In these four scenarios, two types of errors can lead to invalid inferences: (a) The test yields a statistically significant result while there is actually no treatment effect in the population. Rejecting the H_0 when it is in fact true is referred to as Type I error, with (in the long run) probability α. This is called a false positive. α is the criterion for H_0 rejection (often set to $\alpha = .05$; see Section 1.4.3). (b) The test yields a statistically nonsignificant result while there is actually a treatment effect in the population. Failure to reject the H_0 when it is in fact false is referred to as Type II error, with (in the long run) probability β (often set to $\beta = .20$; see Section 1.4.3). This is called a false negative. Note that the probabilities α and β are conditional on H_0: If the H_0 is true, then an erroneous statistical conclusion has probability α; if the H_0 is false, then a statistical conclusion error occurs with probability β (Lipsey & Hurley, 2009).

The remaining two scenarios represent correct statistical conclusions. Most relevant for the present thesis, the probability of correctly rejecting the H_0 when it is in fact false, is $1 - \beta$, and denotes statistical power (often set to $1 - \beta = .80$; see Section 1.4.3). In other words: Statistical power is the likelihood of not committing a Type II error, or detecting an effect, if it truly exists, which is significant at a stated α level.

1.4.2 The Why of A Priori Power Analysis

Power analysis is an indispensable step in RT planning.[11] Disclosing the assumptions and key results of power analyses—sample size, power, and precision—ranks among the "basic

[11] Generally, power analysis is considered a venture in study design (i.e., a priori), not in study analysis (i.e., post-hoc). Post-hoc power analysis have long been and still are subject to controversial debates (see e.g., Dziak et

expectations" of the reporting standards for quantitative studies (American Psychological Association, 2020, pp. 79, 83–84; see also Wilkinson & Task Force on Statistical Inference, 1999). This also complies with the Consolidated Standards of Reporting Trials 2010 Statement for social and psychological interventions (CONSORT-SPI 2018) urging researchers to make sample size justification in RTs explicit by reporting power analysis (Grant et al., 2018; Montgomery et al., 2018). Not surprisingly then that third-party funding nowadays presupposes in general full transparency on power analysis and sample size determination (e.g., Education Endowment Foundation, 2022; German Research Foundation, 2022; Institute of Education Sciences, 2023). Power analysis basically showcases the prospective capacity of a design to distinguish real from chance differences (Schochet, 2008). Given that educational RTs are typically resource-intensive (Wozny et al., 2018), "it would be almost unthinkable to embark on a large-scale study without conducting a power analysis." (Hedges & Rhoads, 2010b, p. 436)

The overarching purpose of an a priori power analysis for RTs on student achievement is twofold: designing studies that are likely to produce (a) conclusive results and (b) correct results on the (in)effectiveness of an educational intervention (see Hedberg, 2018). Obtaining *conclusive* results (i.e., being informative with regard to a particular inferential goal; Lakens, 2022; Lortie-Forgues & Inglis, 2019) is important to guarantee considerate handling of scarce monetary and human resources (Bausell & Li, 2002; Halpern et al., 2002; Lenth, 2001). Importantly, both under- and overpowered RTs may waste these resources: the former by probably overlooking a meaningful effect (i.e., Type II error) or by being incapable to prove the (true) absence of an effect; the latter by needlessly overusing funds as well as the time and commitment of investigators, school principals, teachers, students and so on (Ahn et al., 2020). Striving for *correct* results tells its own tale. It was repeatedly shown that an underpowered RT may inflate or even invert the estimate of the population effect (Gelman & Carlin, 2014; Ioannidis, 2005, 2008; Sims et al., 2022). Such erroneous findings not only rule out reproducibility (Open Science Collaboration, 2015) but also—if experimental education

al., 2020; Lakens, 2022; Onwuegbuzie & Leech, 2004; Quach et al., 2022), whose coverage is, however, beyond the scope of the present thesis. Shortly recapped, unless for the sake of research reviews scrutinizing achieved power rates in previous studies, the meaningfulness of post-hoc power calculations seems weak at best in most cases. In essence, they provide very little new, but potentially flawed information (Aberson, 2019, p. 15; Dziak et al., 2020; Levine & Ensom, 2001). The level of retrospectively observed power of an underpowered RT failing to reject the null hypothesis H_0 implies nothing more and nothing less than that the design lacked power to detect a for this design too small effect. Fisher (1938, p. 17) somewhat ironically get to the heart of this problem by stating that "to consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of." Furthermore, an underpowered RT that only "slightly fail significance" (within quotation marks as this wording is to be criticized itself) would demonstrate higher post-hoc power than an underpowered RT that not even approach significance (Aberson, 2019; Dziak et al., 2020). As an alternative to post-hoc power, one may better investigate the confidence interval of the effect size (Levine & Ensom, 2001).

research claims to inform evidence-based policy and practices—could lead to (fundamentally) wrong political decisions in the education system.

Sufficient design sensitivity is therefore a paramount indicator of the rigor and quality of RTs. All the more alarming is that quite a few RTs in education lack power and precision (Lortie-Forgues & Inglis, 2019; Spybrook, Shi, et al., 2016; Spybrook & Raudenbush, 2009).[12]

1.4.3 Fundamentals: Types of Power Analysis and Core Factors

Power analysis typically has one of three outputs (see e.g., Bausell & Li, 2002; Hedberg, 2018): (a) the required sample size, or (b) the statistical power, or (c) the statistical precision of the treatment effect estimate.[13] These three quantities are interrelated concepts, meaning that computing one of them requires assumptions on the two remaining; and on the α level (i.e., the Type I error rate) as well as the type of the test (often a t-test, but sometimes also F-test, Mann-Whitney U test, etc.). By convention, α = .05 in a two-tailed test (Cohen, 1988), although it should be noted that, theoretically, one could just as well determine the α criterion as a function of sample size, power, and effect size (Cohen, 1988; Murphy & Myors, 2004). Practically however, this strategy is rarely adopted: moving from α = .05 to α = .01, for example, dramatically reduces power with almost no substantial benefit with regard to the protection against a Type I error (Murphy & Myors, 2004).

Sample Size

Researchers planning RTs on student achievement frequently conduct power analysis to find the required sample size for a given design which facilitates detecting a hypothesized effect size that can be expected from a specific educational intervention, or that is deemed worthwhile, in terms of practical or political relevance (Kraft, 2020; Lipsey et al., 2012). The sensitivity of an RT design to uncover a (true) treatment effect is an increasing function of the sample size, although this relation is not linear. As a rule, however, larger samples are associated with lower sampling error, improving the signal of a potential effect (Lipsey, 1990). Yet, in educational settings, the total sample size is often sharply restricted (not only due to constraints on the resources provided by a funding agency or a governmental authority, but also due to the

[12] This is in no way symptomatic for educational experimentation science in particular. Older and more recent reviews coincide in baring that low power and precision is a frequent issue in psychology in general (e.g., Bakker et al., 2012; Cohen, 1962; Fraley & Vazire, 2014; Rossi, 1995; Sedlmeier & Gigerenzer, 1989). Maxwell (2004) propounded a thorough investigation of possible reasons.

[13] It should be mentioned that there co-exist counter-projects to this rather broad, more generic notion of "power analysis" which unites multiple concepts pertinent to RT design. Schönbrodt and Wagenmakers (2018), for instance, defend the term "design analysis" (notably, within the Bayesian framework).

institutional frame of the school system). It may therefore also be helpful to assess statistical power or precision, given a certain sample size.

Statistical Power

Statistical Power, $1 - \beta$, is defined as the likelihood of a statistical test to detect a treatment effect if it exists in the population (importantly: in the long run; Cohen, 1988). In the social and behavioral sciences, it is common to strive for an at least 80% chance of detecting an effect, that is $1 - \beta = .80$ (Cohen, 1988).[14] The achieved level of statistical power is influenced by various design characteristics, however, the core determinants include sample size, hypothesized effect size, and alpha level/statistical test (e.g., Aberson, 2019).[15]

Statistical Precision

The precision of the treatment effect estimate basically refers to its standard error (Hedges, 2022). Generally, smaller standard errors mean higher precision. There are several ways to conceptualize and assess the target precision in power calculations. Some have proposed to set a desired width of the CI of the treatment effect (see e.g., Ahn et al., 2020, for a coverage of precision analysis; see e.g., Maxwell et al., 2008; Pornprasertmanit & Schneider, 2014, for the accuracy in parameter estimates [AIPE] approach), or, the other way around, defining a region of practical equivalence (ROPE) within the Bayesian framework (e.g., Kruschke, 2018). When precision is the outcome of power analysis, it seems intuitive to think about it as the minimum detectable effect size ($MDES$; Bloom, 1995, 2005) which may then be juxtaposed against an effect size of educational importance (Kraft, 2020; Lipsey et al., 2012). The $MDES$ quantifies the smallest possible standardized effect size that reaches statistical significance in a given design (Bloom, Zhu, et al., 2008).

Figure 5 visualizes that the $MDES$ is conceived as a multiple of the standardized standard error of the treatment effect (Bloom, 2005), and as such, represents a metric of precision. On the left, the t-distribution under the H_0 of no treatment effect is shown, and on the right, the t-distribution under the H_1 of a positive[16] standardized treatment effect of size

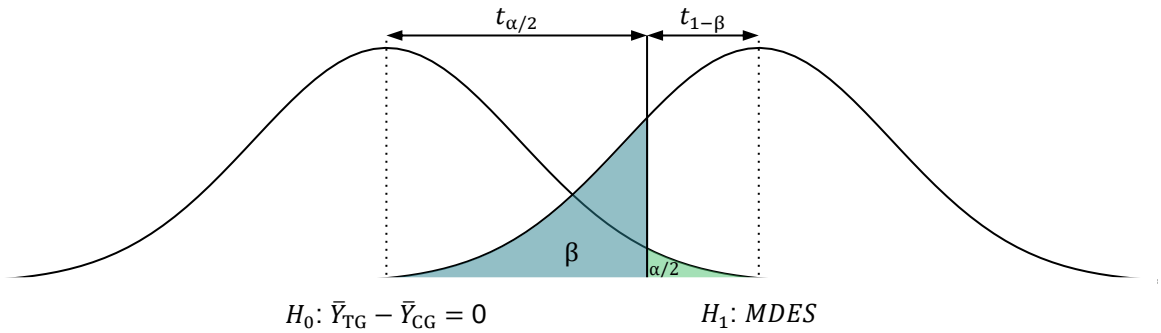
[14] At the same time, most researchers are well aware of the fact that the 80% power benchmark lacks formal justification (e.g., Lakens, 2022). Lipsey and Hurley (2009), for instance, suggest to carefully weighing costs against risks with regard to the specific research context when deciding on the relative seriousness of potential Type I and Type II errors—missing an effect of an actually effective and promising intervention may mean a deprivation of beneficial learning and teaching conditions for students and teachers.

[15] Further factors include the variance of the target outcome measure, the reliability of outcome and covariate measures, and the experimental error (Lipsey, 1990, pp. 14–15).

[16] For simplicity, a positive treatment effect is assumed. Nevertheless, while the treatment effect, of course, may also be negative, the $MDES$ can not.

$MDES$. To have a probability of $1 - \beta$ to be statistically significant at $\alpha/2$ in a two-tailed test, this effect must be larger by $t_{1-\beta}$ than the critical *t*-value of the H_1 and larger by $t_{\alpha/2} + t_{1-\beta}$ than the H_0 (Bloom, 2005). Therefore, the smallest distance in standardized standard error (t-statistic) units between H_0 and H_1 equals the $MDES$ (Dong & Maynard, 2013).

Figure 5. *Minimum Detectable Effect Size Multiplier for a Two-Tailed Test*



Note. Multiplier for a two-tailed test: $M_{df} = t_{\alpha/2} + t_{1-\beta}$, with df degrees of freedom. Adapted from "The Core Analytics of Randomized Experiments for Social Research" (Figure 1, p. 22) by H. S. Bloom, 2006, MDRC Working Papers on Research Methodology, MDRC. Copyright 2006 by MDRC. Reprinted and adapted with permission.

The generalized form of an approximate $MDES$ is:

$$MDES = M_{df} \frac{SE(\bar{Y}_{TG} - \bar{Y}_{CG})}{\sigma_T} \quad (1)$$

$SE(\bar{Y}_{TG} - \bar{Y}_{CG})$ is the standard error of the treatment effect, and σ_T is the (pooled) total student population's standard deviation. The term $SE(\bar{Y}_{TG} - \bar{Y}_{CG})/\sigma_T$ can be rewritten as a function of sample size, proportional division of the sample into the TG and the CG, and (multilevel) design parameters. This reformulation is specific to the particular design of the RT, where for a two-tailed test, $M_{df} = t_{\alpha/2} + t_{1-\beta}$ with df degrees of freedom. df increases with larger samples, but decreases by the number of included covariates. For $\alpha = .05$ and $1 - \beta = .80$, M_{df} approaches 2.8 in a two-tailed test when $df \geq 20$. Since in multilevel designs, the effective or operational sample size undercuts the total sample size (given $\rho > 0$; Bloom, 2006; Hedges & Rhoads, 2010a; Lipsey & Hurley, 2009), the threat of loss of dfs is more pertinent to multilevel than single-level designs (Bloom, 2005).

Notably, Bloom (1995) originally introduced the minimum detectable effect (MDE) in a natural (i.e., unstandardized) metric. In general, standardized effect sizes are preferable in most scenarios of power analysis (Lipsey, 1990) as researchers evade the estimation of the population variance of an RT's target outcome. Nonetheless, in doing so researchers should be aware that (a) a large standardized effect size could either point to a large unstandardized effect

or little variation (Liu, 2014), and (b) the meaning of effect sizes are strongly tied to the specific research context (target population, outcome, etc.; e.g., Brunner, Stallasch, et al., 2023; C. J. Hill et al., 2008). For instance, our own meta-analytic study, wherein we estimated versatile effect size benchmarks for student achievement in the German school context, testifies considerable variation in each benchmark type among subpopulations, age groups, and outcomes (Brunner, Stallasch, et al., 2023). Hence, standardized effect sizes should be well rationalized within the RT's actual contextual conditions (Pek & Flora, 2018).

1.4.4 Power Analysis for Multilevel Designs

Unlike in single-level designs, power analysis in multilevel designs not only circuits around the properties of the statistical test, effect size, and sample size but also involves concretizing the structure of clustering inherent in the sample. Here, two additional factors conspire to render power analysis for multilevel designs more complex (Hedges & Rhoads, 2010a; Konstantopoulos, 2009): (a) the sample allocation among hierarchical levels and (b) the variance design parameters at the various hierarchical levels.

Sample Allocation Among Hierarchical Levels

When planning CRTs and MSRTs, the level-specific sample sizes have to be configured. This means, depending on the desired output of a power analysis, the researcher has to specify the number of schools at L3; within schools, the number of classrooms at L2; and within classrooms, the number of students at L1. Holding the total sample size constant, different sample allocations across the hierarchical levels will likely yield different degrees of design sensitivity. Importantly, this also implies that—in stark contrast to single-level designs, when everything else is equal—augmenting the total sample size does not necessarily result in enhanced power and precision (Hedges & Hedberg, 2013).

Multilevel Design Parameters

For any CRT or MSRT[17] design (and as a consequence, any CRT or MSRT analysis), the variance components of the random effects at each hierarchical level *must* be taken into account

[17] When planning MSRTs, researchers will additionally rely on reasonable assumptions on the treatment effect heterogeneity (i.e., the variation of the treatment effect between the sites). Reliable estimates of these parameters can only be obtained from experimental studies themselves, but not on the basis of observational (large-scale assessment) studies. Extensive resources of cross-site heterogeneity parameters are still scarce (for an exception see Weiss et al., 2017) as these heavily depend on the actual availability of MSRTs, which are largely lacking so far in Germany.

to allow for conclusive and correct inferences on an intervention's impact. The internal homogeneity of students' outcomes within classrooms and schools, which is usually quantified through the ICC, often noticeably limits statistical power and precision in multilevel RTs (e.g., Raudenbush, 1997; Schochet, 2008). Hence, researchers *should* carefully consider covariate adjustment to improve design sensitivity, which requires assumptions on the amounts of explained variance at the various hierarchical levels (e.g., Bloom et al., 2007; Raudenbush, 1997; Raudenbush et al., 2007). Of importance, this strategy has been proven effective across RT designs, meaning that statistically controlling for covariates also is a beneficial strategy to raise design sensitivity in single-level IRTs (e.g., Kahan et al., 2014; Maxwell et al., 2017; Porter & Raudenbush, 1987).

Next, I formally define the design parameters ρ and R^2, decomposed each hierarchical level to plan three- and two-level RTs. Note that I also include the definition of the single-level R^2 (i.e., not decomposed) to be used in power analysis for IRTs. The discussion assumes a continuous achievement outcome Y with constant, unconditional total variance σ^2_{T}, and common within-cluster variances as well as infinite populations at either hierarchical level a (see Snijders & Bosker, 2012). In the multilevel scenarios, the sources of σ^2_{T} are typically identified through multilevel regression modeling (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). In the single-level scenario, conventional OLS modeling may be applied (Cohen et al., 2003). All underlying statistical models can be found in the Online Supplemental Materials (OSMs) A for Studies I and II.

Intraclass Correlation Coefficients: Between-Classroom and Between-School Achievement Differences. The achievement outcomes of students within the same classroom or school show the tendency to intercorrelate (i.e., they are stochastically dependent; Kreft, 1993). In other words: there is some redundancy in the scores of students who belong to the same classroom or school *cluster* (Scherbaum & Pesner, 2019). The degree of redundancy in *Y* due to cluster membership—or equivalently, the extent of variation between clusters—is measured via the intraclass correlation coefficient (ICC), denoted by ρ.[18]

In a *three-level design* (students at L1 within classrooms at L2 within schools at L3), σ^2_{T} can be decomposed into the between-student-within-classroom variance located at L1, σ^2_{L1}, the between-classroom-within-school variance located at L2, σ^2_{L2}, and the between-school variance

[18] Put differently, ρ equals the Pearson correlation between any two achievement scores of students within one and the same classroom or school cluster. This implies that, as soon as $\rho > 0$, clusters can no longer be regarded "*interchangeable* [emphasis added] with regard to the experimental endpoint." (Donner & Klar, 2000, p. 2)

located at L3, σ^2_{L3}. Thus, with three levels of nesting, σ^2_T is the sum of the variance components at L1, L2, and L3: $\sigma^2_T = \sigma^2_{L1} + \sigma^2_{L2} + \sigma^2_{L3}$.

The ICC at L2 expresses the ratio of the variance at the classroom level to the total variance and can accordingly be interpreted as the proportion of the total variance in Y that can be attributed to between-classroom achievement differences:

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

The ICC at L3 expresses the ratio of the variance at the school level to the total variance and is therefore the proportion of the total variance in Y that can be attributed to between-school differences:

$$\rho_{L3} = \frac{\sigma^2_{L3}}{\sigma^2_T} \tag{3}$$

Having determined ρ_{L2} and ρ_{L3}, the remainder proportion of σ^2_T can be attributed to students within classrooms in schools, which is simply their complement: $1 - \rho_{L2} - \rho_{L3}$ (Hedges & Hedberg, 2013).

In a *two-level design* (students at L1 within schools at L3, i.e., skipping L2), σ^2_{L1} becomes the between-student-within-school variance at L1, and σ^2_T becomes the sum of the variance components at L1 and L3 (one could think about this as if σ^2_{L2} was set to zero). Thus, with two levels of nesting, $\sigma^2_T = \sigma^2_{L1} + \sigma^2_{L3}$; the ICC at L3 is then computed with Equation (2).

Either ICC is, in theory, defined for $0 \leq \rho \leq 1$, where $\rho = 0$ signifies that the only source of variation in Y are achievement differences between individual students (i.e., there are no between-classroom or between-school differences). Inversely, $\rho = 1$ signifies that students within the same classroom or school show identical achievement, so that classroom or school affiliation accounts for all the variation in Y.[19] Expressions for the large-sample variances of the ICCs are given in Appendix C.

Squared Multiple Correlation Coefficients: Explained Variances by Covariates. Well-selected covariates can substantially boost statistical power and precision in all RT designs (e.g., Bloom et al., 2007; Kahan et al., 2014; Konstantopoulos, 2012; Maxwell et al., 2017; Porter & Raudenbush, 1987; Raudenbush et al., 2007). Crucially, well-selected basically means highly predictive to the outcome. The mechanism behind this idea is as follows. When

[19] As Eldridge et al. (2009) note, $\rho < 0$ is in fact possible (e.g., when using analysis of variance or generalized estimating equations, but not, by definition, with multilevel or mixed effects modeling), however, this is implausible in most scenarios (Donner & Klar, 2000, pp. 10–11). A negative ICC is likely due to estimation error, but also occurs when the population ICC is actually negative with naturally finite cluster sizes (e.g., a family has always a finite number of members; Eldridge et al., 2009). A word of caution: In either case, power analysis should *never* draw on a value of $\rho < 0$.

covariates predict achievement differences in Y, they reduce the unexplained total variance σ_T^2. This implies an attenuation of error variance which results in an improved signal-to-noise ratio of the treatment effect estimate (Raudenbush et al., 2007). Or equivalently, the reduction of σ_T^2 reduces the standard error of the treatment effect estimate, and thus, the $MDES$ (Bloom, 2006). This gain in precision translates into a gain in statistical power. Notably, given the basic principles of randomization, covariates do neither bias the treatment effect estimator nor do they change its estimate at all (in magnitude or direction)—covariates exclusively affect the treatment effect estimate's standard error (Borenstein & Hedges, 2019; Maxwell et al., 2017, p. 471; Porter & Raudenbush, 1987). The amount to which covariates can account for achievement differences in Y is quantified by the squared multiple correlation coefficient, denoted by R^2.[20]

In multilevel designs, covariates can act at either hierarchical level, although it is not necessary to specify covariates at all levels. Cluster-level covariates may be represented by directly observable and non-decomposable "global" or "integral" measures (e.g., classroom size, school budget) or "contextual" or "analytical" measures aggregated from the individual student level (e.g., mean prior knowlegde, share of female students; Lüdtke et al., 2008, pp. 203–204). In general, group-mean centering of within-cluster covariates is recommended to guarantee that the covariates contribute to the variance explanation only at those hierarchical level at which they are introduced (Konstantopoulos, 2012; Raudenbush & Bryk, 2002).

In a *three-level design* (students at L1 within classrooms at L2 within schools at L3) that adjusts for one or more classroom-mean centered covariates C_{L1} at L1, one or more school-mean centered covariates C_{L2} at L2, and one or more covariates C_{L3} at L3, the variance decomposition produces a conditional between-student-within-classroom variance located at L1, $\sigma^2_{L1|C_{L1}}$, a conditional between-classroom-within-school variance located at L2, $\sigma^2_{L2|C_{L2}}$, and a between-school variance located at L3, $\sigma^2_{L3|C_{L3}}$.

The explained variance at L1 by covariates C_{L1} is:

$$R^2_{L1} = \frac{\sigma^2_{L1} - \sigma^2_{L1|C_{L1}}}{\sigma^2_{L1}} \tag{4}$$

The explained variance at L2 by covariates C_{L2} is:

$$R^2_{L2} = \frac{\sigma^2_{L2} - \sigma^2_{L2|C_{L2}}}{\sigma^2_{L2}} \tag{5}$$

[20] Elsewhere, the impacts of covariates C have also been indexed in the form of an covariate-adjusted ICC, $\rho_{|C}$ (Eldridge et al., 2009; Hedges & Hedberg, 2007), simply expressing the complement of R^2: $R^2 = 1 - \rho_{|C}$.

The explained variance at L3 by covariates C_{L3} is:

$$R^2_{\mathrm{L3}} = \frac{\sigma^2_{\mathrm{L3}} - \sigma^2_{\mathrm{L3}|C_{\mathrm{L3}}}}{\sigma^2_{\mathrm{L3}}} \tag{6}$$

In a *two-level design* (students at L1 within schools at L3) that controls for one or more school-mean centered covariates C_{L1} at L1 and one or more covariates C_{L3} at L3, $\sigma^2_{\mathrm{L1}|C_{\mathrm{L1}}}$ becomes the conditional between-student-within-school variance at L1. The explained variances at L1 and L3 are computed as in Equations (4) and (6), respectively.

In a *single-level design* (with students assumed to be independently sampled) with one or more (possibly grand-mean centered) covariates C_{T}, $\sigma^2_{\mathrm{T}|C_{\mathrm{T}}}$ quantifies the conditional total variance, among all individual students.

The total explained variance by covariates C_{T} is:

$$R^2_{\mathrm{T}} = \frac{\sigma^2_{\mathrm{T}} - \sigma^2_{\mathrm{T}|C_{\mathrm{T}}}}{\sigma^2_{\mathrm{T}}}, \tag{7}$$

Generally, $0 \leq R^2 \leq 1$, so that covariates can account for 0% to 100% of achievement variance.[21] Expressions for the large-sample variances of the R^2 values are provided in Appendix C.

Implications for Statistical Power

To make the influences of the cluster structure precise, Figure 6 illustrates statistical power as a function of (a) level-specific sample allocation and (b) multilevel design parameters for a 2L-CRT (students at L1 within schools at L3). At the y-axis, the estimated power is shown for detecting a standardized treatment effect size of $\delta = .15$ at $\alpha = .05$ by a two-tailed t-test when the design is completely balanced (i.e., schools were randomly split fifty-fifty into one TG and one CG, $K_{\mathrm{TG}} = K_{\mathrm{CG}}$; and the number of students per school does not vary, $n_k = n_{k'}$).[22]

The power analyses shown in Figure 6a assume $\rho_{\mathrm{L3}} = .20$ and no covariate adjustment ($R^2_{\mathrm{L1}} = R^2_{\mathrm{L3}} = .00$). Three main conclusions follow: (1) Power increases when the numbers of students per school n_k and the number of schools K increase, whereby (2) K exerts greater impact on power than n_k, and (3) there is a point of diminishing returns in power for n_k (Konstantopoulos, 2008b, provides the formal derivation why this is the case). For instance,

---

[21] Under the above definitions, negative R^2 may occur (Snijders & Bosker, 2012). In the multilevel case, $R^2 < 0$ can be caused by estimation error when an unconditional variance component (i.e., σ^2_{L1}, σ^2_{L2}, or σ^2_{L3}) approximates zero (Jacob et al., 2010). Another possible reason might be that the formulae for the multilevel R^2 values ignore the interplay between the unconditional variance components (LaHuis et al., 2019; Rights & Sterba, 2018).

[22] I chose an effect size of $\delta = .15$ as a realistic example because it depicts the average annual academic growth in reading across Grades 5 to 10 in German lower secondary school (Brunner, Stallasch, et al., 2023, Table 1).

with $K = 200$ schools in total (i.e., $K = 100$ per experimental condition) and $n_k = 20$ students per school, power equals $1 - \beta = .58$. When quintupling the number of students per school to $n_k = 100$, while holding the total sample size of $Kn_k = 4{,}000$ constant (i.e., decreasing the number of schools to $K = 40$ schools), power substantially drops to $1 - \beta = .17$. By instead doubling the number of schools to $K = 400$ schools (i.e., $n_k = 10$), power makes a great leap to $1 - \beta = .81$. Notably, this sums up to a relative power gain of 40% as compared to the starting design with $K = 200$ and $n_k = 20$. Finally, when additionally doubling the number of students per school to $n_k = 20$, so that the total sample size is also doubled to $Kn_k = 8{,}000$, power equals $1 - \beta = .86$, which yields a 48% power increase as opposed to the starting design (i.e., doubling the total sample size only yields 8% additional power).[23] Therefore, everything else being equal, the number of schools is the main driver for power. Note that K influences power through df, but even more strongly through the noncentrality parameter defining the cumulative distribution function of the noncentral t-distribution (see Equation (C65) in the OSM C for Study II).

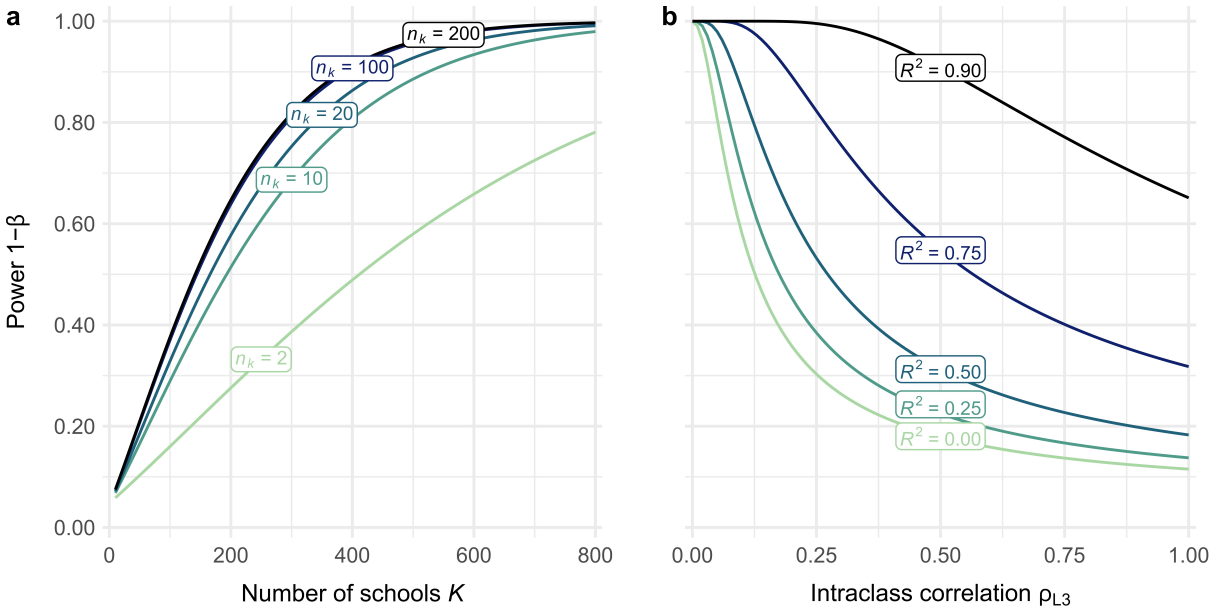
Figure 6b portrays power as a function of ρ_{L3}, R^2_{L1} and R^2_{L3} design parameters, where the sample size is fixed to a total of $K = 100$ and $n_k = 50$ ($Kn_k = 5{,}000$). Note that the designs with $R^2_{L1} \geq .25$ and $R^2_{L3} \geq .25$ are all based on one single covariate at both L1 and L3. In contrast, the design with $R^2_{L1} = R^2_{L3} = .00$ does not involve any covariate, at neither level. The three main conclusions from Figure 6b are: (1) Power decreases when ρ_{L3} increases, (2) power increases when R^2_{L1} and R^2_{L3} increase, (3) the impacts of both ρ_{L3} as well as R^2_{L1} and R^2_{L3} on power level off at a certain point, albeit in diametric directions. For instance, holding $R^2_{L1} = R^2_{L3} = .00$ constant, the relative drop in power when $\rho_{L3} = .05$ ($1 - \beta = .81$) augments to $\rho_{L3} = .10$ ($1 - \beta = .58$) is −28%, while the drop for $\rho_{L3} = .45$ ($1 - \beta = .19$) heightening to $\rho_{L3} = .50$ ($1 - \beta = .18$) is only −5%. The other way around, holding ρ_{L3} constant, gains in power increase with higher values of R^2_{L1} and R^2_{L3}.

Finally, a simple comparison between a completely balanced 2L-CRT as just discussed and an IRT (where students are assumed to be stochastically independent) further highlights the utmost importance of taking into account the cluster structure when planning multilevel educational experiments: When assuming $\rho_{L3} = .20$, $n_k = 50$, and foregoing any covariate adjustment, in total $K = 304$ schools are required to achieve 80% power in a two-tailed t-test

[23] Note that in this example, the ratio of the (higher) costs associated with sampling additional schools to the (lower) costs associated with sampling additional students within a school is neglected for illustrative purposes. In real RT settings, however, cost efficiency plays a major role for a successful research design (e.g., Konstantopoulos, 2009).

to uncover $\delta = .15$ at $\alpha = .05$. This translates to a total sample size requirement of $Kn_k = 15{,}150$ students. An equivalent IRT (ignoring clustering) would instead only need $N = 1{,}397$ students in total.

Figure 6. *Statistical Power as a Function of (a) Sample Allocation and (b) Design Parameters*



Note. Power analysis for a two-level cluster-randomized trial (2L-CRT; students at L1 within schools at L3) under complete balance ($K_{TG} = K_{CG}$ and $n_k = n_{k'}$ for $i \in \{1, 2, \ldots, n_k\}$ students nested within $k \in \{1, 2, \ldots, K\}$ schools randomly assigned to the treatment group TG and the control group CG) to detect a standardized effect size of $\delta = .15$ at $\alpha = .05$ (two-tailed) in a two-sample independent t-test. Figure 6a: Power for different sample sizes at L1 and L3 when $\rho_{L3} = .20$ and $R^2_{L1} = R^2_{L3} = .00$ are fixed. Figure 6b: Power for different values of ρ_{L3} and $R^2_{L1} = R^2_{L3}$ when $K = 100$ and $n_k = 50$ are fixed.

Note that the general pattern of relationships between statistical power on the one hand and (a) sample allocation and (b) design parameters on the other also holds for 2L-MSRT designs (see Raudenbush & Liu, 2000), as well as—albeit associated with greater complexity—3L-CRT (see Konstantopoulos, 2008b), 3L-MSRT, and 3L-MSCRT (see Konstantopoulos, 2008a) designs: (a) Everything else being equal, the sample size at the top hierarchical school level shapes power (far) more than those at lower levels. In MSRTs and MSCRTs, however, the influences of the sample sizes at lower levels is somewhat stronger by contrast with their influences in CRTs (Konstantopoulos, 2008a; Raudenbush & Liu, 2000).[24] (b) Everything else held constant, ρ is a decreasing and R^2 an increasing function of power, across the various hierarchical levels. In MSRTs and MSCRTs, these relations are additionally modulated by the

[24] In particular, in a 3L-MSCRTs, df is not only a function of the number of schools (and, if applicable, the number of the covariates included) but also depends on the number of classrooms, as shown in Equation (8) in Chapter 2.

treatment effect heterogeneity parameters capturing the variability in intervention effects across sites.

Formulae to compute the minimum required sample size, statistical power, as well as $MDES$ for the various designs can be found in (the OSMs for) Studies I and II (see also, e.g., Dong & Maynard, 2013, for expressions on sample size and the $MDES$; Liu, 2014, for expressions on power).

1.4.5 Incorporating Uncertainties and Heterogeneities

The accuracy of power analysis relies on the goodness of their input parameters. Unfortunately, each output quantity obtained from power analysis (i.e., sample size, power, precision/$MDES$) is, by definition, only locally optimal (Moerbeek & Teerenstra, 2016, p. 203; see also Du & Wang, 2016 for a related discussion of the local optimization problem in effect sizes specified in power analysis). What does this mean? Power analysis heavily depends, inter alia, on the (point) estimates of ρ and R^2, while, in fact, every single RT has its own set of true design parameter estimands. It is simply impossible to foresee the exact magnitudes of these estimates before the data have been collected; thus, they remain unknown.[25] Putting it in a nutshell in the words of Hedberg (2018, p. 99): "power analysis is all about assumptions." Hence, their a priori output estimates (being it sample size, power, or precision) will inevitably deviate from the actual—true—state of affairs. No matter how elaborated and justified a researcher's guesses on ρ and R^2 are.

Basically, design parameters are subject to two major sources of variation: (a) sampling error, and (b) true heterogeneity. First, values of ρ and R^2 to be entered in power analysis are usually empirically derived (Bloom, Zhu, et al., 2008), either from former observational or experimental studies, pilot studies, or (large-scale) sample surveys (Hedges et al., 2012; Turner et al., 2004). Consequently, as based on finite samples, they are subject to random noise and therefore (sometimes fairly) imprecise (Eldridge & Kerry, 2012; Hedges et al., 2012; Jacob et al., 2010; Turner et al., 2004, 2005). Second, it is reasonable to assume that there is also true heterogeneity in ρ and R^2, systematically varying by populations, outcomes and so forth (e.g., Brunner et al., 2018; Spybrook, Westine, et al., 2016; Zhang et al., 2023). Here, random-effects meta-analysis (e.g., Borenstein et al., 2021) is an excellent methodological tool to disentangle these two sources of variation in the design parameters, which has been applied to summarize

[25] Although the tests in power analyses actually assume that the input parameters, such as the effect size but also the design parameters, represent known quantities (Konstantopoulos, 2011).

design parameters estimated based on individual participant data (e.g., Hedberg & Hedges, 2014; see Brunner et al., 2022, for a gentle introduction to individual participant data meta-analysis based on large-scale assessments).

Ignoring uncertainty in the empirically derived input parameters may severely distort power analysis (Liu, 2014; Perugini et al., 2014). For instance, Liu (2014, pp. 51–52) showed that the probability of nominal power (i.e., based on sample variance) overestimating actual power (i.e., based on population variance) exceeds 50%, and increases with decreasing sample size. It was therefore repeatedly highlighted that best-practice power analysis properly incorporate uncertainties in ρ and R^2 values (Bausell & Li, 2002; Donner & Klar, 2000; Hedges et al., 2012; Jacob et al., 2010; Liu, 2014; Moerbeek & Teerenstra, 2016; Turner et al., 2004, 2005). Several strategies have been proposed to conduct sensitivity analyses when determining sample size, power, or precision in RT planning.

Using Confidence and Prediction Intervals

One prominent option to explicitly address ρ and R^2 uncertainties in power analysis is to draw on the (95%) CIs constructed from their nominal standard errors or on the meta-analytic prediction intervals (PIs) to define a plausible range of values for a certain design parameter. The meta-analytic 95% PI provides a plausible range of ρ and/or R^2; it quantifies the total dispersion (sampling variance plus true heterogeneity) around the meta-analytic average of ρ and/or R^2 (e.g., Borenstein et al., 2021). The bounds of the CI or PI can be used to estimate ranges of required sample sizes, or power and precision rates within which the actual target quantities will likely lie (Bausell & Li, 2002; Donner & Klar, 2000; Jacob et al., 2010; Liu, 2014; Perugini et al., 2014). This has also been labeled the "safeguard" approach (Perugini et al., 2014). However, such strategies have been criticized for being overly conservative (Turner et al., 2004; Williamson et al., 2023). Moreover, such techniques treat each value within the 95% CI or PI as equally likely, which might not be justifiable in most scenarios (Turner et al., 2004; Williamson et al., 2023). Turner et al. (2005, p. 114) lamented that "it seems insufficient to simply provide a series of "what-if" scenarios conditioning on various ICC values: a measure of their respective plausibility appears necessary."

Running Simulations

Another, more elegant solution is to simulate power analysis outcomes based on Monte Carlo methods. Especially promising are approaches that make use of (empirically informed) prior distributions in the spirit of Bayesian statistics to implicitly take into account uncertainty in ρ

and R^2 (Moerbeek & Teerenstra, 2016; Pek & Park, 2019; Spiegelhalter et al., 2004; Turner et al., 2004, 2005; Williamson et al., 2023). For instance, Spiegelhalter et al. (2004) introduced a hybrid "Bayesian-classical" approach to power analysis (see also Pek & Park, 2019). Quite a few variations and extensions of the concept have been developed (see Kunzmann et al., 2021, for a review); yet, the gist uniting all of them is: while the design stage of an RT contains Bayesian elements, the analysis stage is assumed to be purely classical. This strategy has been frequently applied in clinical research (e.g., Brus et al., 2022; Moerbeek & Teerenstra, 2011; O'Hagan & Stevens, 2002; Sarkodie et al., 2023; Turner et al., 2004, 2005). Note that the idea of placing priors on ρ and R^2 can be easily extended to the remaining input parameters in power formulas (depending on the desired output such as effect sizes; see e.g., Du & Wang, 2016; Pek & Park, 2019)—a major appeal of Bayesian-based power analysis simulations.

1.5 How Much Is (Not) Known on Design Parameters for Student Achievement? A Brief Outline of Previous Research

Apt design parameters are the fundament of expedient power analysis to plan sound RTs. Researchers who aim to evaluate interventions to foster student achievement therefore need reliable estimates and standard errors of ρ and R^2 optimally customized to the planned study's specifications (e.g., Brunner et al., 2018; M. Campbell et al., 2000; Cohen, 1988; Murray, 1998; Spybrook, 2013; Zhang et al., 2023). Over the past two decades—encouraged by the turn towards evidence-based policies and practices in education and its strong focus on RTs—vast repertoires of ICCs and explained variances have been built.

This section recapitulates where these collections are already well-equipped; and also where not so well. Specifically, I pinpoint five major dimensions in which existing resources on empirical guidance for designing RTs on student achievement show several important research gaps: (1) target populations, (2) target outcome domains, (3) target covariates, (4) target experimental designs, (5) target analysis models, and (6) quantifications of uncertainties and heterogeneities. Thereby, Figure 7 visualizes the extent of these shortcomings by mapping available studies that accumulated design parameters explicitly devoted to inform power analysis for RTs on student achievement in Grades 1 to 12. In particular, Figure 7 lists scopes and features, as well as summarizes estimates of ρ_{L2} and ρ_{L3}, as well as R^2_T, R^2_{L1}, R^2_{L2}, or R^2_{L3} for several covariate sets. These depict the unique, incremental, and relative impacts of those

factors that have previously been theoretically and empirically identified as core predictors of student achievement (detailed below).

Note that Studies I and II in the present doctoral thesis enclose research reviews which thoroughly elaborate on the current body of knowledge on design parameters for student achievement. Specifically, the research review included in Study I offers a detailed examination and visualization of previously reported ρ and R^2 values appropriate for two-level (students within schools) and three-level (students within classrooms within schools) RT designs. The research review included in Study II complements this picture via a meta-analytic integration of previous R^2 values for the various covariate sets for single-level (students assumed to be independently sampled), two-level, and three-level RT designs.

1.5.1 Current Research Gaps

Target Populations—National Scopes and Grade Levels

School systems markedly differ from each other in vital characteristics (OECD, 2010), and so do students in different grades (i.e., cohorts/age groups) with regard to their proficiency levels and taught curricula. Hence, design parameters should accurately mirror the RT's target population (e.g., Lipsey et al., 2012).

As Figure 7 shows, many studies that accumulated design parameters on student achievement to inform RT power analysis stem from the United States. These works drew either on national probability samples (Hedberg et al., 2004; Hedges & Hedberg, 2007; Konstantopoulos, 2009), state-wide assessments in one single state (Brandon et al., 2013; Jacob et al., 2010; Konstantopoulos, 2009; Westine et al., 2013) or multiple states (Cole et al., 2011; Hedges & Hedberg, 2013; Spybrook, Westine, et al., 2016; Xu & Nichols, 2010; Zhu et al., 2012), or individual (experimental) studies in multiple districts (Bloom et al., 2007; Hedberg et al., 2004; Jacob et al., 2010; Schochet, 2008; Zhu et al., 2012), in one city (Bloom et al., 1999; Gargani & Cook, 2005, as cited in Schochet, 2008), or multiple cities (Schochet, 2008). Notably, the majority of these studies provide ρ and R^2 values for several, albeit mostly selected grades in both elementary[26] as well as secondary school.

The few studies going beyond the U.S. school context listed in Figure 7 were carried out at an international level, as cross-country research. These publications capitalized on large-scale assessments for 81 nations and economies from the 2000, 2003, 2006, 2009, and 2012 cycles of the Programme for International Student Assessment (PISA) study (Brunner et al.,

[26] In the United States, elementary school covers Grades 1 to 6.

2018), up to 84 nations and economies from the 1995, 1999, 2003, and 2007 cycles of the Trends in International Mathematics and Science Study (TIMSS) plus the 2001 and 2006 cycles of the Progress in International Reading Literacy Study (PIRLS; Zopluoglu, 2012), and 15 sub-Saharan African countries of the third cycle of the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ III) study (Kelcey et al., 2016). Therefore, the sheer volume of resulting ICCs and explained variances is stunning (e.g., Zopluoglu, 2012, spawned 646 distinct ρ_{L3} values). Notably, these compendia are based on representative, yet cross-sectional data of national probability samples. Although the international studies offer estimates of ρ and R^2 to design RTs in many different school systems (including the German one), they remain limited to Grade 4 or 6[27] in elementary and/or Grade 8 or 9 in secondary school. Nevertheless, consistently testifying substantial variation in ρ and R^2 values between countries, all of the mentioned studies of international scope provide strong evidence that the design parameters documented for the United States may not generalize well across national contexts.

Target Outcome Achievement Domains

Student achievement is multifarious, and may be measured in several more broader or narrower defined (sub)domains (Brunner, Preckel, et al., 2023; Steinmayr et al., 2014). Accordingly, contemporary educational curricula reach far beyond the typical core domains, such as mathematics, science, and reading (OECD, 2018). Importantly, estimates of ρ and R^2 to be entered in power analysis should align with the target outcome domain (e.g., Westine et al., 2013).

However, as shown in Figure 7, the grand majority of past studies focus on outcomes in only one or two of these core achievement domains, namely mathematics and/or reading (Bloom et al., 1999, 2007; Brandon et al., 2013; Cole et al., 2011; Gargani & Cook, 2005; Hedges & Hedberg, 2007, 2013; Jacob et al., 2010; Kelcey et al., 2016; Konstantopoulos, 2009; Schochet, 2008). Some investigations broadened this spectrum by adding one or more science-related subdomains such as biology, physics, and chemistry (Brunner et al., 2018; Westine et al., 2013; Xu & Nichols, 2010; Zhu et al., 2012; Zopluoglu, 2012). Yet, other important achievement domains, for instance specific verbal skills or domain-general cognitive abilities are still severely underrepresented, although this line of research suggests that design parameters may differ substantially depending on the achievement domain.

[27] In sub-Saharan African countries, elementary school covers Grades 1 to 6.

Figure 7. *Overview on Previous Studies on Design Parameters for Student Achievement*

Note. The color code corresponds to the median ρ or (Δ)R^2 value. The number in a bubble counts the achievement (sub)domains analyzed. Outer triangles map the R^2 value (absolute) for a combination, inner triangles map the ΔR^2 value (increment) for a covariate over and above a domain-identical pretest. 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. LA = ρ and/or R^2 values are also suitable for a latent variable modeling target analysis of the treatment effect. *SE* = standard errors of ρ and/or R^2 were reported. INT = International.

Target Covariates

Highly predictive covariates of student achievement are powerful means to raise power and precision; in both single- as well as multilevel RT designs (e.g., Bloom et al., 2007; Kahan et al., 2014; Konstantopoulos, 2012; Porter & Raudenbush, 1987). Numerous scholars and agencies have repeatedly emphasized that decisions on which covariates to include should be both empirically as well as theoretically justified, ideally in combination with preregistration (Committee for Proprietary Medicinal Products, 2004; Cook, 2005; European Medicines Agency [EMA], 1998, 2015; Maxwell et al., 2017; Moerbeek & Teerenstra, 2016; Murray, 1998; Raab et al., 2000; U.S. Food and Drug Administration, 2021). As a multifaceted construct, domain-specific student achievement may be shaped by various factors (e.g., Steinmayr et al., 2014; Winne & Nesbit, 2010). In educational psychology, prominent models of school learning (Haertel et al., 1983; Wang et al., 1993) coincide with existing empirical evidence in highlighting the relevance of the following determinants: domain-identical pretests (e.g., previous mathematics skills predict future mathematics skills; Dochy et al., 1999), cross-domain pretests (e.g., previous reading skills predict future mathematics skills; Baumert et al., 2009), fluid intelligence pretests (Cattell, 1987), as well as sociodemographic characteristics (e.g., gender, migration background, socioeconomic status; Bradley & Corwyn, 2002; see Chapter 3 for a more detailed rationalization of these covariates as well as their hypothesized unique, incremental, and relative impacts).

Previous studies on design parameters have scrutinized precision-enhancing impacts of covariates from three angles. (a) Unique effects of key *covariate types*, namely domain-identical pretests and/or sociodemographics were frequently quantified, and occasionally also of cross-domain pretests. Yet, the impact of fluid intelligence has been neglected so far. (b) The largest part of the studies that estimated R^2 values for both domain-identical pretests and sociodemographics also investigated their joint or incremental effect. However, further *covariate combinations* have been ignored. (c) Few studies providing R^2 values for domain-identical pretests scrutinized the influence of *covariate time lags*. The temporal decay in their predictive power has only been modeled for baseline measures lagged three years at maximum. Of importance, this choice of covariates was largely not (or, at least not explicitly) motivated by substantive theory but was rather data-driven instead. Relatedly, to the best of my knowledge, former works forego any theoretically founded derivation of hypotheses about the unique, relative, or incremental effectiveness of the covariates to explain variation in student achievement.

Briefly summarized, this strand of research indicates that a domain-identical pretest is the most powerful covariate for student achievement, explaining impressive amounts of variance, both in total as well as decomposed at the various hierarchical levels. Moreover, its explanatory power has been proven to only slowly decline with growing pre-posttest time lags. Meanwhile, sociodemographics turned out to be useful covariates at L2 and especially L3 but not L1, and are in general of little incremental value when combined with a domain-identical pretest.

Target Experimental Designs

Basically, the concise sampling strategy (i.e., simple single-level vs. complex two- or three-level), and thus, the structure of the data used to estimate the design parameters has to match the planned experimental design to obtain valid power calculations.[28] In applied experimental educational and psychological research, both single-level IRTs as well as two-, but first and foremost three-level CRTs and MSRTs represent the most commonly implemented designs (Connolly et al., 2018; Spybrook, Shi, et al., 2016; Spybrook & Raudenbush, 2009).

To date, only one single investigation offers explained variances by covariates explicitly compiled to inform the planning of IRTs (students treated as independently sampled; Cole et al., 2011). Apart from this, R_{T}^2 values are widely scattered across single empirical studies and have not yet been systematically integrated. In stark contrast, Figure 7 illustrates that the bulk of design parameter studies are multilevel in nature, and consequently, produced ICCs and explained variances relevant to design CRTs and MSRTs. All of these works—and most of them exclusively—quantified between-school achievement differences; therefore much is already known on the typical magnitudes of ρ_{L3} for CRT and MSRT designs with two hierarchical levels (students at L1 within schools at L3). At an international level, these unconditional variance components varied broadly, but appeared on average larger than in the United States.

These are undoubtedly relevant pieces of information to plan 2L-CRTs or 2L-MSIRTs. At the same time, reliable values of ρ_{L2} *and* ρ_{L3} that are required to plan 3L-CRTs, 3L-MSIRTs, or 3L-MSCRTs (with students at L1 within classrooms at L2 within schools at L3) are still scarce. Of note, the few studies that decomposed the total variance in student achievement into the shares that can be attributed to differences between students, classrooms, and schools all apply to the U.S. school context (Jacob et al., 2010; Konstantopoulos, 2009; Xu & Nichols,

[28] When aiming to plan MSRTs, the unit where randomization occurs (i.e., students or classrooms) further determines the choice of the treatment effect heterogeneity parameters which are then additionally needed.

2010; Zhu et al., 2012). Respective evidence suggests that in secondary school, between-classroom differences substantially outweighed between-school differences whereas it was the other way around in elementary school.

Target Analysis Models

In applied effectiveness research, a test for the treatment effect may generally be performed drawing on various statistical analysis methods (see e.g., M. J. Campbell & Walters, 2014; Hayes & Moulton, 2017; Maxwell et al., 2017, for comprehensive treatises). Importantly, the review by Blanca et al. (2018) indicates that most tests of comparison in RTs reported in psychological articles draw on manifest variables (e.g., via ANOVA, ANCOVA, or "conventional" regression modeling) rather than latent variables (via structural equation modeling). Concerning CRTs and MSRTs, the review by Luo et al. (2021, Table 7) substantiates this picture: Only around 11% of multilevel analysis that have been carried out in educational and psychological research during the last decade used M*plus* (Muthén & Muthén, 2017; which was, as far as I am aware, the only software implementing multilevel latent variable modeling for a long time). In spite of this, techniques for group comparison within a latent variable framework have been developed—for single-level (Bollen, 2002; Mayer et al., 2016) as well as multilevel RTs (Lüdtke et al., 2008; Raudenbush & Bryk, 2002)—that offer the advantageous possibility to partial out measurement error in the (outcome and covariate) measures. It was repeatedly recommended that design parameters entered into power analysis should mirror the planned analysis procedure (Ahn et al., 2020; Kleinman & Huang, 2017; Schochet, 2008), certainly not least because latently modeled R^2 are expected to be larger than their manifest counterparts, due to their higher reliabilities (Cohen et al., 2003, pp. 119–124; Raudenbush & Bryk, 2002, p. 346).

Yet, as Figure 7 indicates, design parameters have typically been generated using fallible manifest variables. So far only the international study by Brunner et al. (2018) embedded the estimation of ρ and R^2 values into a general latent variable modeling framework by using M*plus*. Specifically, as the applied multilevel latent covariate models (Lüdtke et al., 2008) involve latently aggregated cluster means of L1 covariates that correct for measurement error, resulting R^2_{L2} and R^2_{L3} should be more pronounced than they would have been without this (default) option (i.e., based on manifestly aggregated cluster means). However, nothing is known so far whether, and if yes, how much design parameters may differ by the employed statistical model to analyze the treatment effect, posing problems when researchers intend to use some kind of structural equation model in the analysis stage of RTs (Schochet, 2008).

Quantifications of Uncertainties and Heterogeneities

The outputs from power analysis are inextricably linked to their input parameters (referred to as local optimization; Du & Wang, 2016; Moerbeek & Teerenstra, 2016, p. 203). Already small deviations between a priori assumed design parameters and retrospectively observed (true) values may result in misleading sample size, power, or $MDES$ calculations. However, ρ and R^2 as empirical estimates are subject to uncertainty due to sampling error, and meta-analytic aggregations of ρ and R^2 (to be used, e.g., in the absence of more specific estimates that would optimally fit the target RT, or when there are multiple competing more specific estimates) additionally contain true heterogeneity (apart from sampling error). A robust RT design, thus, should take into account the statistical uncertainty, and if applicable, also the true heterogeneity associated with the design parameters, either explicitly (e.g., using CIs/PIs; Liu, 2014) or implicitly (e.g., running simulations that involve empirical distributions; Moerbeek & Teerenstra, 2016).

Unfortunately, Figure 7 discloses that fewer than half of the extant design parameter compilations provide respective standard errors (Hedges & Hedberg, 2007, 2013; Jacob et al., 2010; Kelcey et al., 2016; Spybrook, Westine, et al., 2016; Westine et al., 2013). Notably, with the exception of Hedges and Hedberg (2013), all standard errors pertain to ρ, but not to R^2 values. This is astonishing as the results of Hedges and Hedberg (2013) imply that the sampling uncertainties associated with the values of R^2_{L3} are typically (much) larger than those of ρ_{L3} (and also of R^2_{L1}). Meanwhile, meta-analytic estimates informing on the degree of true variation among two-level ICCs (but neither three-level ICCs nor explained variances at either hierarchical level) have been, as far as I am aware, so far only provided by Hedberg and Hedges (2014; not shown in Figure 7).

1.5.2 A Closer Look at German Research

As stated above, design parameters to plan RTs on student achievement need to optimally fit the target population in terms of the national context and the attended grade. Figure 7 does not list any exclusive research from Germany though. Rather, the international studies of Brunner et al. (2018) and Zopluoglu (2012) encompass reliable estimates of ρ_{L3}, and Brunner et al. (2018) additionally calculated R^2_{L1}, and R^2_{L3} for sociodemographic characteristics. Drawing on nationally representative cross-sectional data from PISA, TIMSS, and PIRLS, these collections are suitable to design sound RTs which test interventions targeted at the general (i.e., total) student population in Grade 4, 8, or 9. Nevertheless, other grades or certain subpopulations that

arise from the special characteristics of the German school system were not covered, neither were quite a few of the central covariates.

Characteristics of the German School System

The school system in Germany has some distinctive peculiarities that create special requirements for the design parameters. Germany constitutes a federal republic, where each of the 16 federal states takes the primary responsibility for legislation and administration of schooling (Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany, 2021). In 14 federal states, elementary school ("Grundschule") comprises Grades 1 to 4, but in two federal states (Berlin and Brandenburg), elementary school comprises Grades 1 to 6. In secondary school, all school systems are characterized by extensive school type tracking, as is the case for many other school systems (Reichelt et al., 2019; Salchegger, 2016). As a rule, five school types are distinguished that address students with different achievement levels: the academic track school ("Gymnasium"; up to Grade 12 or 13), vocational school ("Hauptschule"; up to Grade 9 or 10), intermediate school ("Realschule"; up to Grade 10), multitrack school ("Schulen mit mehreren Bildungsgängen"; up to Grade 9, 10, 12, or 13), and comprehensive school ("Gesamtschule"; up to Grade 12 or 13). The more demanding academic track school is offered across all federal states, but the remaining less demanding school types—which I subsume under the umbrella term "non-academic track"—partly differ by federal state.

Note that throughout this thesis, I differentiate between three grade levels: for Grades 1 to 4, I refer to elementary school; for Grades 5 to 10, I refer to lower secondary school; for Grades 11 to 12, I refer to upper secondary school. Importantly, in upper secondary school, students are typically not taught into intact classrooms, but are rather enrolled into courses differentiated by the aspiration level chosen for a certain subject (e.g., basic vs. advanced mathematics courses). However, all upper secondary school students are taught in the core domains (i.e., mathematics, German as first language, and a science-related subject).

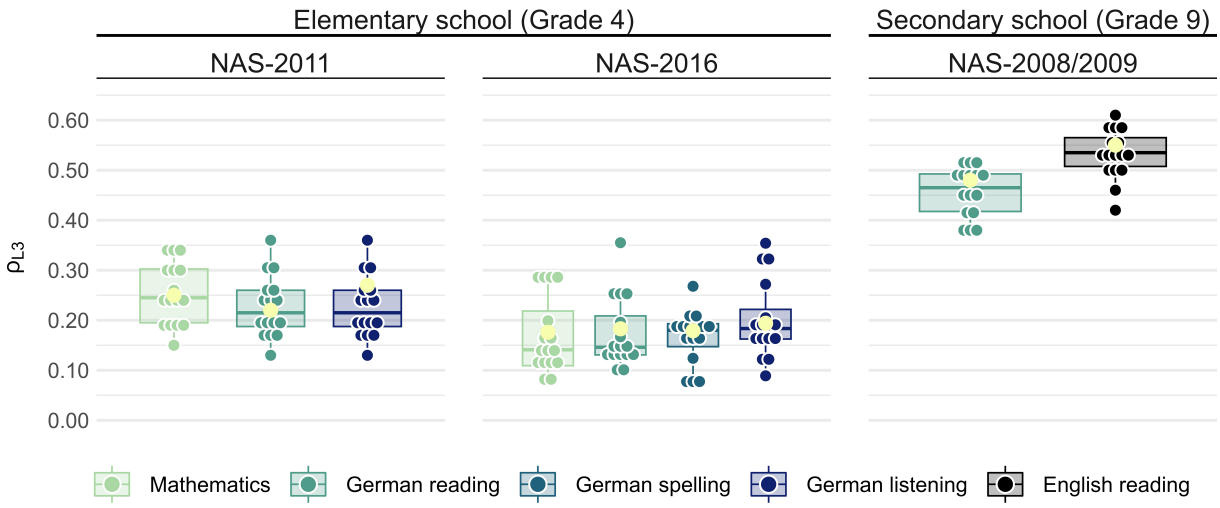
Design Parameters Reflecting the German School Context

To date, design parameters appropriate for the design of RTs implemented in the German school system have not yet been systematically cataloged. Apart from the mentioned exceptions of Brunner et al. (2018) and Zopluoglu (2012), the respective body of knowledge is fairly scattered. Empirical estimates of ρ and R^2 were—if at all—reported as by-product alongside

with research on educational effectiveness and social inequalities (e.g., Baumert et al., 2000, 2003; Knigge & Köller, 2010; Lehmann & Lenkeit, 2008).

The so far largest collection comprises ICCs at L3 based on data from the 2008/2009, 2011, and 2016 cycles of the German National Assessment Study. Figure 8 shows the distributions of ρ_{L3} values for mathematics, several verbal skills in German (as first language), and reading in English (as foreign language) by federal state, and for the total German student population (i.e., analyzed as a whole, highlighted yellow). Figure 8 accentuates three aspects: Achievement differences between German schools (a) appear significantly larger in secondary than in elementary school, (b) markedly deviate from those registered for U.S. schools, and (c) vary considerably across federal states.

Figure 8. *Between-School Achievement Differences for the German School Context*



Note. Dots show ρ_{L3} values as estimated for the various federal states in Germany. ρ_{L3} as estimated for the total German student population is depicted in yellow. NAS = National Assessment Study (NAS-2008/2009: "Ländervergleich 2008/2009", NAS-2011/2016: "IQB-Bildungstrend"). Data has been retrieved from Knigge and Köller (2010) for NAS-2008/2009, Böhme and Weirich for NAS-2011 German reading and listening, Haag and Roppelt for NAS-2011 mathematics, Wittig and Weirich for NAS-2016 spelling, reading, listening, Hagg and Kohrt for NAS-2016 mathematics. The NAS is carried out by the Institute for Educational Quality Improvement (IQB).

1.6 The Present Doctoral Thesis

The virtue of power analyses for RTs to be designed to evaluate interventions on student achievement hinge on reliable input design parameter estimates. Specifically, ρ and R^2 values should properly reflect the particular study context (e.g., Zhang et al., 2023) as defined by the RT's target population, achievement outcome domain, possibly applied covariates,

experimental design, and analysis model, as well as they should be accompanied by quantifications of their associated uncertainties (e.g., Hedges et al., 2012). However, as the many gaps in Figure 7 immediately unveil: currently available collections in this vein suffer from several crucial shortcomings. And despite the fact that educational stakeholders in Germany increasingly prioritize rigorous RTs to generate useable knowledge about interventions that make students succeed, the necessary evidence footing to design such studies falls even further behind this international knowledge base.

The overarching objective of the present doctoral thesis is to analyze versatile compendia of (meta-analytically integrated) design parameters in order to build comprehensive, reliable resources and thorough guidance to optimize power analysis for the design of RTs on student achievement in the German (and similar) school context. To this end, I conducted two comprehensive studies directly addressing the gaps identified in previous research.

First, the current knowledge base on design parameters for *target populations* outside the United States is meager. Even if such estimates have been propounded, they are restricted to some single grades. In particular, a systematic compilation of ρ and R^2 values across the entire school career that map the distinctive characteristics of the German school system is still lacking. Studies I and II used rich, representative data from three German longitudinal large-scale assessments (National Educational Panel Study [NEPS], PISA, Assessment of Student Achievements in German and English as a Foreign Language [DESI]) to generate design parameters for students in Grades 1 to 12. Study I drew on five samples (starting cohorts 2, 3, 4 from NEPS; the follow-up of the 2003 PISA cycle; DESI). Study II capitalized on six samples (samples from Study I plus the follow-up of the 2012 PISA cycle) as identified via a systematic search whose results were meta-analyzed within grade levels to support the design of RTs targeting multiple grades. Both studies covered several student (sub)populations to reflect the German school context: the total population as well as the subpopulations in the academic and non-academic track In Study I, ICCs and explained variances were additionally adjusted for mean-level achievement differences between the various school types in German secondary education.

Second, the coverage of *target outcome domains* by past design parameters is insufficient in that it remains limited to the core subjects of mathematics, science, and reading. Studies I and II significantly broaden this spectrum. Study I accumulates ρ and R^2 estimates for, in total, 21 different subdomains (core domains, multifarious verbal skills in German and English, domain-general skills such as information and communication technology or basic cognitive functions). Study II meta-analytically summarized design parameters across, in total,

eight STEM[29] and German verbal skills to inform power analysis for RTs addressing multiple domains.

Third, available guidance on the selection of *target covariates* is poor. Not only that many critical covariate types, combinations, and time lags have been excluded so far. Also, the choice of studied covariate sets as well as their hypothesized unique, incremental, and relative return fully lack theoretical justification. Study I lays a foundation by revisiting three previously scrutinized covariate sets which in fact include factors that are—according to influential models of school learning (Haertel et al., 1983; Wang et al., 1993)—among the most important predictors for student achievement: pretest (domain-identical or proxy scores) and/or sociodemographics. Study II referred to three psychometric heuristics (bandwidth-fidelity; Cronbach & Gleser, 1957; incremental validity; Sechrest, 1963; validity degradation; Ghiselli, 1956; Humphreys, 1960) to scrutinize precision-enhancing impacts of 11 distinct covariate sets of varying types (domain-identical, cross-domain, fluid intelligence pretests, sociodemographics), their combinations (domain-identical pretests plus each of the remaining and all together), and time lags (1- to 7-year lagged domain-identical, cross-domain, fluid intelligence pretests). To evaluate covariate impacts on precision, Study II encloses simulations which followed a hybrid Bayesian-classical approach to power analysis. The results of Study II were used to develop empirically supported guidelines on covariate adjustment.

Fourth, existing collections of design parameters neglect several *target experimental designs* beyond RTs with two hierarchical levels (students at L1 within schools at L3). Studies I and II compiles reliable estimates appropriate to plan six different RT designs (see Figure 4) with up to three hierarchical levels (students at L1 within classrooms at L2 within schools at L3): IRTs (students assumed to be independently sampled), 2L- and 3L-CRTs, 2L- and 3L-MSIRTs, as well as 3L-MSCRTs. Specifically, Study I applied two- and three-level modeling to estimate ρ_{L2} and ρ_{L3} as well as R^2_{L1}, R^2_{L2}, and R^2_{L3}. Study II specified single-, two-, and three-level models to estimate ρ_{L2} and ρ_{L3} as well as R^2_{T}, R^2_{L1}, R^2_{L2}, and R^2_{L3}.

Fifth, virtually all previous repertoires of design parameters are suitable for manifest *target analysis model* for the test of the treatment effect and disregard potential applications within a latent variable modeling framework (Lüdtke et al., 2008; Mayer et al., 2016). Acknowledging the coexistence of both options and offering the chance to juxtapose the performances of the techniques to estimate ρ and R^2 (Schochet, 2008), Study I applied latent (covariate) modeling in M*plus* when estimating the multilevel variance components, while Study II relies on a manifest estimation approach using R.

[29] The term STEM is commonly used to subsume science, technology, engineering, and mathematics.

Sixth, most past studies failed to report statistical *uncertainties and heterogeneities* associated with the empirically estimated and meta-analyzed design parameters. Studies I and II consistently documented all ρ and R^2 values along with their corresponding standard errors and/or 95% CIs. Study II quantified meta-analytic heterogeneities among ρ and R^2 and registered 95% PIs. Studies I and II were complemented by diverse illustrative application scenarios which guide through the process of RT planning when incorporating uncertainties and heterogeneities associated with the design parameters by using their 95% CIs/PIs (explicit uncertainty handling). In Study II, the simulation study showcases power analysis involving priors based on the joint empirical distributions of ρ and R^2 (implicit handling of uncertainty).

All in all, tackling crucial gaps of extant resources in six major dimensions, this dissertation strives to support educational researchers and psychologists in designing strong RTs on student achievement. The emerging resources couple—so far unique—nuanced design parameter compendia and guidance for power analysis.

References

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences* (2nd Edition). Routledge, Taylor & Francis Group.

Ahn, C., Heo, M., & Zhang, S. (2020). *Sample size calculations for clustered and longitudinal outcomes in clinical research* (First issued in paperback). CRC Press.

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

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). *The rules of the game called* psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Baumert, J., Köller, O., Lehrke, M., & Brockmann, J. (2000). Anlage und Durchführung der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie zur Sekundarstufe II (TIMSS/III)—Technische Grundlagen [Design and implementation of the third trends in international mathematics and science study (TIMSS/III)—Technical information]. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Vol. 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichschulzeit* (pp. 31–84). Leske+Budrich.

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

Baumert, J., Trautwein, U., Artelt, C., Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten—Institutionelle Bedingungen des Lehrens und Lernens [School contexts—Institutional conditions for teaching and learning]. In Deutsches PISA-Konsortium, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 261–331). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-97590-4_11

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

Biggart, A., Kerr, K., O'Hare, L., & Connolly, P. (2013). A randomised control trial evaluation of a literacy after-school programme for struggling beginning readers. *International Journal of Educational Research*, *62*, 129–140. https://doi.org/10.1016/j.ijer.2013.07.005

Blanca, M. J., Alarcón, R., & Bono, R. (2018). Current practices in data analysis procedures in psychology: What has changed? *Frontiers in Psychology*, *9*, 2558. https://doi.org/10.3389/fpsyg.2018.02558

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., Bos, J. M., & Lee, S.-W. (1999). Using cluster random assignment to measure program impacts. Statistical implications for the evaluation of education programs. *Evaluation Review*, *23*(4), 445–469. https://doi.org/10.1177/0193841X9902300405

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., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness*, *10*(4), 877–902. https://doi.org/10.1080/19345747.2016.1271069

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://files.eric.ed.gov/fulltext/ED502531.pdf

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. (2019). Effect sizes for meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine, *Handbook of research synthesis and meta-analysis* (3rd ed., pp. 207–243). Russell Sage Foundation.

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

Borenstein, M., Hedges, L. V., & Rothstein, H. (2012). *CRT-Power – Power analysis for cluster-randomized and multi-site studies* [Computer software]. Biostat.

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences*, *113*(47), 13354–13359. https://doi.org/10.1073/pnas.1601135113

Borman, G. D., Dowling, N. M., & Schneck, C. (2008). A multisite cluster randomized field trial of open court reading. *Educational Evaluation and Policy Analysis*, *30*(4), 389–407. https://doi.org/10.3102/0162373708326283

Boruch, R. F. (2003). Randomized Field Trials in Education. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International Handbook of Educational Evaluation* (pp. 107–124). Springer Netherlands. https://doi.org/10.1007/978-94-010-0309-4_9

Boruch, R. F., & Foley, E. (2000). The honestly experimental society: Sites and other entities as the units of allocation and analysis in randomized trials. In L. Bickman (Ed.), *Validity and social experimentation: Donald campbell's legacy* (pp. 193–239). Sage Publications.

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

Brandon, P. R., Harrison, G. M., & Lawton, B. E. (2013). SAS code for calculating intraclass correlation coefficients and effect size benchmarks for site-randomized education experiments. *American Journal of Evaluation*, *34*(1), 85–90. https://doi.org/10.1177/1098214012466453

Brierley, G., Brabyn, S., Torgerson, D., & Watson, J. (2012). Bias in recruitment to cluster randomized trials: A review of recent publications. *Journal of Evaluation in Clinical Practice*, *18*(4), 878–886. https://doi.org/10.1111/j.1365-2753.2011.01700.x

Brookhart, S. M. (2015). Graded achievement, tested achievement, and validity. *Educational Assessment*, *20*(4), 268–296. https://doi.org/10.1080/10627197.2015.1093928

Bruce, C. L., Juszczak, E., Ogollah, R., Partlett, C., & Montgomery, A. (2022). A systematic review of randomisation method use in RCTs and association of trial design

characteristics with method selection. *BMC Medical Research Methodology*, *22*(1), 314. https://doi.org/10.1186/s12874-022-01786-4

Brunner, M., Keller, L., Stallasch, S. E., Kretschmann, J., Hasl, A., Preckel, F., Lüdtke, O., & Hedges, L. V. (2022). 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*, jrsm.1584. 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., Preckel, F., Götz, T., Lüdtke, O., & Keller, L. K. (2023). *The relationship between math anxiety and math achievement: New perspectives from combining individual participant data and aggregated data in a meta-analysis* [Unpublished manuscript].

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*, 1–39. https://doi.org/10.1080/19345747.2023.2175753

Brus, D. J., Kempen, B., Rossiter, D., Balwinder-Singh, & McDonald, A. J. (2022). Bayesian approach for sample size determination, illustrated with Soil Health Card data of Andhra Pradesh (India). *Geoderma*, *405*, 115396. https://doi.org/10.1016/j.geoderma.2021.115396

Bundesministerium für Bildung und Forschung (Ed.). (2018). *Rahmenprogramm empirische Bildungsforschung [Framework program educational research]*. https://www.empirische-bildungsforschung-bmbf.de/img/Rahmenprogramm%20empirische%20Bildungsforschung_barrierefrei_NEU(1).pdf

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

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, *24*(4), 409–429. https://doi.org/10.1037/h0027982

Campbell, M., Grimshaw, J., Steen, N., & Changing Professional Practice in Europe Group (EU BIOMED II Concerted Action). (2000). Sample Size Calculations for Cluster Randomised Trials. *Journal of Health Services Research & Policy*, *5*(1), 12–16. https://doi.org/10.1177/135581960000500105

Campbell, M. J., & Walters, S. J. (2014). *How to design, analyse and report cluster randomised trials in medicine and health related research*. John Wiley & Sons.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. North-Holland ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co.

Clark, M. P. A., & Westerberg, B. D. (2009). How random is the toss of a coin? *Canadian Medical Association Journal*, *181*(12), E306–E308. https://doi.org/10.1503/cmaj.091733

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). L. Erlbaum Associates.

Cohen, J. (1973). Brief notes: Statistical power analysis and research results. *American Educational Research Journal*, *10*(3), 225–229. https://doi.org/10.3102/00028312010003225

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997–1003. https://doi.org/10/bm96wk

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (Eds.). (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. (2017). *Using randomised controlled trials in education*. SAGE PUBLICATIONS.

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

Cook, T. D., Murphy, R. F., & Hunt, H. D. (2000). Comer's school development program in Chicago: A theory-based evaluation. *American Educational Research Journal*, *37*(2), 535–597. https://doi.org/10.3102/00028312037002535

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology*, *108*(2), 100–102. https://doi.org/10.1093/oxfordjournals.aje.a112592

Corrin, W., Parise, L. M., Cerna, O., Haider, Z., & Somers, M.-A. (2015). Case management for students at risk of dropping out: Implementation and interim impact findings from the communities in schools evaluation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2609366

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

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge Taylor & Francis Group.

Dehue, T. (1997). Deception, efficiency, and random groups: Psychology and the gradual origination of the random group design. *Isis, 88*(4), 653–673. https://doi.org/10.1086/383850

Dehue, T. (2001). Establishing the experimenting society: The historical origin of social experimentation according to the randomized controlled design. *The American Journal of Psychology*, *114*(2), 283. https://doi.org/10.2307/1423518

Dekker, I., & Meeter, M. (2022). Evidence-based education: Objections and future directions. *Frontiers in Education, 7*, 941410. https://doi.org/10.3389/feduc.2022.941410

Descôteaux, J. (2007). Statistical power: An historical introduction. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 28–34. https://doi.org/10.20982/tqmp.03.2.p028

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., Kelcey, B., & Spybrook, J. (2023). Experimental design and power for moderation in multisite cluster randomized trials. *The Journal of Experimental Education*, 1–17. https://doi.org/10.1080/00220973.2023.2226934

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., Birkett, N., & Buck, C. (1981). Randomization by cluster. Sample size requirements and analysis. *American Journal of Epidemiology*, *114*(6), 906–914. https://doi.org/10.1093/oxfordjournals.aje.a113261

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

Du, H., & Wang, L. (2016). A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate Behavioral Research*, *51*(5), 589–605. https://doi.org/10.1080/00273171.2016.1191324

Dziak, J. J., Dierker, L. C., & Abar, B. (2020). The interpretation of statistical power after the data have been gathered. *Current Psychology*, *39*(3), 870–877. https://doi.org/10.1007/s12144-018-0018-1

Education Endowment Foundation. (2022). *Statistical analysis guidance for EEF evaluations*. https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1698086955

Eldridge, S., & Kerry, S. (2012). *A practical guide to cluster randomised trials in health services research: Eldridge/A practical guide to cluster randomised trials in health services research*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119966241

Eldridge, S., Ukoumunne, O. C., & Carlin, J. B. (2009). The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions. *International Statistical Review*, *77*(3), 378–394. https://doi.org/10/c2qvn8

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

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, *27*(3), 557–577. https://doi.org/10.3102/00028312027003557

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17*(1), 69–78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

Fisher, R. A. (1938). Presidential address. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, *4*(1).

Fraley, R. C., & Vazire, S. (2014). The N-Pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, *9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Gargani, J., & Cook, T. D. (2005). *How many schools? Limits of the conventional wisdom about sample size requirements for cluster randomized trials* [Working paper].

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

Gelman, A., & Nolan, D. (2002). You can load a die, but you can't bias a coin. *The American Statistician*, *56*(4), 308–311. https://doi.org/10.1198/000313002605

German Research Foundation (Ed.). (2022). *Proposal preparation instructions. Project proposals*. https://www.dfg.de/formulare/54_01/54_01_en.pdf

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

Gleason, P., Clark, M., Tuttle, C. C., Dwoyer, E., & Silverberg, M. (2010). *The evaluation of charter school impacts: Final report* (NCEE 2010-4029). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education: The promises, opportunities and problems and of trials in education*. Routledge.

Grant, S., Mayo-Wilson, E., Montgomery, P., Macdonald, G., Michie, S., Hopewell, S., Moher, D., & on behalf of the CONSORT-SPI Group. (2018). CONSORT-SPI 2018 explanation and elaboration: Guidance for reporting social and psychological intervention trials. *Trials*, *19*(1), 406. https://doi.org/10.1186/s13063-018-2735-z

Hacking, I. (1988). Telepathy: Origins of randomization in experimental design. *Isis*, *79*(3), 427–451. https://doi.org/10.1086/354775

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

Hahn, S., Puffer, S., Torgerson, D. J., & Watson, J. (2005). Methodological bias in cluster randomised trials. *BMC Medical Research Methodology*, *5*(1), 10. https://doi.org/10.1186/1471-2288-5-10

Halpern, S. D., Karlawish, J. H. T., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *JAMA*, *288*(3), 358. https://doi.org/10.1001/jama.288.3.358

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: The role of feedback's perceived usefulness. *Educational Psychology*, *34*(3), 269–290. https://doi.org/10.1080/01443410.2013.785384

Hayes, R. J., & Moulton, L. H. (2017). *Cluster randomised trials* (Second edition). CRC Press.

Hedberg, E. C. (2018). *Introduction to power analysis: Two-group studies*. SAGE Publications, Inc. https://doi.org/10.4135/9781506343105

Hedberg, E. C., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics results from a meta-analysis of district-specific values. *Evaluation Review*, *38*(6), 546–582. https://doi.org/10.1177/0193841X14554212

Hedberg, E. C., Santana, R., & Hedges, L. V. (2004). *The variance structure of academic achievement in America* [Working paper].

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. (2022, June 9). *The design and analysis of randomized field experiments in education and the social sciences* [Workshop].

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

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

Hedges, L. V., & Rhoads, C. (2010a). *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., & Rhoads, C. H. (2010b). Statistical power analysis. In *International encyclopedia of education*. Elsevier. http://www.sciencedirect.com/science/referenceworks/9780080448947

Hedges, L. V., & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research*, *60*(3), 265–275. https://doi.org/10.1080/00131881.2018.1493350

Hemming, K., & Taljaard, M. (2023). Key considerations for designing, conducting and analysing a cluster randomized trial. *International Journal of Epidemiology*, dyad064. https://doi.org/10.1093/ije/dyad064

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

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

Institute of Education Sciences. (2023). *Education research grants program. Request for applications.* (ALN: 84.305A). https://ies.ed.gov/funding/pdf/2021_84305A.pdf

Institute of Education Sciences, & National Science Foundation. (2013). *Common guidelines for education research and development*. https://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Itzek-Greulich, H., Flunger, B., Vollmer, C., Nagengast, B., Rehm, M., & Trautwein, U. (2017). Effectiveness of lab-work learning environments in and out of school: A cluster randomized study. *Contemporary Educational Psychology*, *48*, 98–115. https://doi.org/10/gf38kj

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

Jamison, J. C. (2019). The entry of randomized assignment into the social sciences. *Journal of Causal Inference*, *7*(1), 20170025. https://doi.org/10.1515/jci-2017-0025

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

Karbach, J., Könen, T., & Spengler, M. (2017). Who benefits the most? Individual differences in the transfer of executive control training across the lifespan. *Journal of Cognitive Enhancement*, *1*(4), 394–405. https://doi.org/10.1007/s41465-017-0054-z

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental Science*, *12*(6), 978–990. https://doi.org/10.1111/j.1467-7687.2009.00846.x

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

Kirk, R. (2013). *Experimental design: Procedures for the behavioral sciences*. SAGE Publications, Inc. https://doi.org/10.4135/9781483384733

Kish, L. (1965). *Survey sampling*. Wiley.

Kleinman, K., & Huang, S. S. (2017). Calculating power by bootstrap, with an application to cluster-randomized trials. *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *4*(1), 32. https://doi.org/10.13063/2327-9214.1202

Knigge, M., & Köller, O. (2010). Effekte der sozialen Zusammensetzung der Schülerschaft [Impact of the social classroom composition of schools]. In O. Köller, M. Knigge, & B. Tesch, *Sprachliche Kompetenzen im Ländervergleich [Verbal competencies in the National Assessment Study]* (pp. 227–244). Waxmann.

Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, *1*(4), 265–288. https://doi.org/10.1080/19345740802328216

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, *1*(1), 66–88. https://doi.org/10.1080/19345740701692522

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation Review*, *33*(4), 335–357. https://doi.org/10.1177/0193841X09337991

Konstantopoulos, S. (2011). A more powerful test in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, *4*(4), 354–369. https://doi.org/10.1080/19345747.2010.519824

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

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kreft, I. G. G. (1993). Using multilevel analysis to assess school effectiveness: A study of dutch secondary schools. *Sociology of Education*, *66*(2), 104. https://doi.org/10.2307/2112796

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. SAGE Publications.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kultusministerkonferenz. (2016). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Wolters Kluwer. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

Kunzmann, K., Grayling, M. J., Lee, K. M., Robertson, D. S., Rufibach, K., & Wason, J. M. S. (2021). A Review of Bayesian Perspectives on Sample Size Derivation for Confirmatory Trials. *The American Statistician*, *75*(4), 424–432. https://doi.org/10.1080/00031305.2021.1901782

LaHuis, D. M., Blackmore, C. E., & Bryant-Lees, K. B. (2019). Explained variance measures for multilevel models. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis.* (pp. 353–364). American Psychological Association. https://doi.org/10.1037/0000115-016

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lehmann, R., & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien [ELEMENT: Study of reading and mathematics literacy. Development from grades 4 to 6 in Berlin. Final research report on the 2003, 2004, and 2005 assessments at primary schools and undergraduate academic tracks in Berlin]*. Humboldt-Universität zu Berlin. https://www.researchgate.net/profile/Jenny_Lenkeit/publication/273380369_ELEMENT_Erhebung_zum_Lese-_und_Mathematik-verstandnis_-_Entwicklungen_in_den_Jahrgangsstufen_4_bis_6_in_Berlin_Abschlussbericht_uber_die_Untersuchungen_2003_2004_und_2005_an_Berliner_Grundschulen_und_/links/553f61600cf23e796bfb38c2.pdf?origin=publication_detail

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*(3), 187–193. https://doi.org/10.1198/000313001317098149

Levine, M., & Ensom, M. H. H. (2001). Post hoc power analysis: An idea whose time has passed? *Pharmacotherapy: Official Journal of the American College of Clinical Pharmacy*, *21*(4), 405–409. https://doi.org/10.1592/phco.21.5.405.34503

Li, F., Tian, Z., Bobb, J., Papadogeorgou, G., & Li, F. (2022). Clarifying selection bias in cluster randomized trials. *Clinical Trials*, *19*(1), 33–41. https://doi.org/10.1177/17407745211056875

Lindquist, E. F. (1940). *Statistical analysis in educational research*. Houghton Mifflin.

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

Lipsey, M. W., & Hurley, S. (2009). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. Rog, *The SAGE Handbook of Applied Social Research Methods* (pp. 44–76). SAGE Publications, Inc. https://doi.org/10.4135/9781483348858.n2

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research. http://eric.ed.gov/?id=ED537446

Liu, X. S. (2011). The effect of a covariate on standard error and confidence interval width. *Communications in Statistics - Theory and Methods*, *40*(3), 449–456. https://doi.org/10.1080/03610920903391337

Liu, X. S. (2014). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. Taylor&Francis. http://site.ebrary.com/id/10801501

Lohr, S., Schochet, P. Z., & Sanders, E. (2014). *Partially nested randomized controlled trials in education research: A guide to design and analysis. NCER 2014-2000.* National Center for Education Research. https://ies.ed.gov/ncer/pubs/20142000/pdf/20142000.pdf

Loosli, S. V., Buschkuehl, M., Perrig, W. J., & Jaeggi, S. M. (2012). Working memory training improves reading processes in typically developing children. *Child Neuropsychology*, *18*(1), 62–78. https://doi.org/10.1080/09297049.2011.575772

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

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

Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting practice in multilevel modeling: A revisit after 10 years. *Review of Educational Research*, *91*(3), 311–355. https://doi.org/10.3102/0034654321991229

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective* (Third edition). Routledge.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

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

Moerbeek, M., & Teerenstra, S. (2011). Optimal design in multilevel experiments. In *Handbook of Advanced Multilevel Analysis*. Routledge. https://doi.org/10.4324/9780203848852.ch14

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

Montgomery, P., Grant, S., Mayo-Wilson, E., Macdonald, G., Michie, S., Hopewell, S., Moher, D., & on behalf of the CONSORT-SPI Group. (2018). Reporting randomised trials of social and psychological interventions: The CONSORT-SPI 2018 Extension. *Trials*, *19*(1), 407. https://doi.org/10.1186/s13063-018-2733-1

Morrison, K. (2020). *Taming randomized controlled trials in education: Exploring key claims, issues and debates* (1st ed.). Routledge. https://doi.org/10.4324/9781003042112

Mosteller, F., & Boruch, R. F. (Eds.). (2002). *Evidence matters: Randomized trials in education research*. Brookings Institution Press.

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed). L. Erlbaum Associates, Publishers.

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

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide*. Muthén & Muthén.

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing of mathematical competence over the lifespan. *Journal for Educational Research Online*, *5*, 80–109.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, *20A*(1/2), 175. https://doi.org/10.2307/2331945

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, *231*(694–706), 289–337. https://doi.org/10.1098/rsta.1933.0009

OECD. (2010). *PISA 2009 results: What makes a school successful?: Resources, policies and practices (Volume IV)*. OECD. https://doi.org/10.1787/9789264091559-en

O'Hagan, A., & Stevens, J. W. (2002). Bayesian methods for design and analysis of cost-effectiveness trials in the evaluation of health care technologies. *Statistical Methods in Medical Research*, *11*(6), 469–490. https://doi.org/10.1191/0962280202sm305ra

Onwuegbuzie, A. J., & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, *3*(4), 201–230. https://doi.org/10.1207/s15328031us0304_1

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Organisation for Economic Co-operation and Development. (2007). *Evidence in education: Linking research and policy*. OECD Publishing. https://doi.org/10.1787/9789264033672-en

Organisation for Economic Co-operation and Development (Ed.). (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD.

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

Peirce, C. S., & Jastrow, J. (1885). On small differences of sensation. *Memoirs of the National Academy of Sciences for 1884*, *3*, 75–83.

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208–225. https://doi.org/10.1037/met0000126

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

Pellegrini, M., & Vivanet, G. (2021). Evidence-based policies in education: Initiatives and challenges in Europe. *ECNU Review of Education*, *4*(1), 25–45. https://doi.org/10.1177/2096531120924670

Peng, P., & Kievit, R. A. (2020). The development of academic achievement and cognitive abilities: A bidirectional perspective. *Child Development Perspectives*, *14*(1), 15–20. https://doi.org/10.1111/cdep.12352

Pernet, C. (2016). Null hypothesis significance testing: A short tutorial. *F1000Research*, *4*, 621. https://doi.org/10.12688/f1000research.6963.3

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. https://doi.org/10.1177/1745691614528519

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1), 20. https://doi.org/10.5334/irsp.181

Pornprasertmanit, S., & Schneider, W. J. (2014). Accuracy in parameter estimation in cluster randomized designs. *Psychological Methods*, *19*(3), 356–379. https://doi.org/10/f6hb77

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

Quach, N. E., Yang, K., Chen, R., Tu, J., Xu, M., Tu, X. M., & Zhang, X. (2022). Post-hoc power analysis: A conceptually valid approach for power based on observed study data. *General Psychiatry*, *35*(4), e100764. https://doi.org/10.1136/gpsych-2022-100764

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. (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., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475–499. https://doi.org/10.1177/1098214015600515

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

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. https://doi.org/10.1037/1082-989X.5.2.199

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

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., Martínez, A., Bloom, H. S., & Hill, C. J. (2011). *Optimal Design Plus Empirical Evidence* (3.0) [Computer software]. https://wtgrantfoundation.org/optimal-design-with-empirical-information-od

Reichelt, M., Collischon, M., & Eberl, A. (2019). School tracking and its role in social reproduction: Reinforcing educational inheritance and the direct effects of social origin. *The British Journal of Sociology*, *70*(4), 1323–1348. https://doi.org/10.1111/1468-4446.12655

Rights, J. D., & Sterba, S. K. (2018). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*. https://doi.org/10.1037/met0000184

Roland, M., & Torgerson, D. J. (1998). Understanding controlled trials: What are pragmatic trials? *BMJ*, *316*(7127), 285–285. https://doi.org/10.1136/bmj.316.7127.285

Rossi, J. S. (1995). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*(5), 646–656. https://doi.org/10/ftzfqf

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

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

Sarkodie, S. K., Wason, J. M., & Grayling, M. J. (2023). A hybrid approach to comparing parallel-group and stepped-wedge cluster-randomized trials with a continuous primary outcome when there is uncertainty in the intra-cluster correlation. *Clinical Trials*, *20*(1), 59–70. https://doi.org/10.1177/17407745221123507

Scherbaum, C. A., & Pesner, E. (2019). Power analysis for multilevel research. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis.* (pp. 329–352). American Psychological Association. https://doi.org/10.1037/0000115-015

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

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

Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (Ed.). (2021). *The education system in the federal republic of Germany 2018/2019. A description of the responsibilities, structures and developments in education policy for the exchange of information in Europe*. KMK. https://www.kmk.org/fileadmin/Dateien/pdf/Eurydice/Bildungswesen-engl-pdfs/dossier_en_ebook.pdf

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309–316. https://doi.org/10.1037/0033-2909.105.2.309

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. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15–21. https://doi.org/10.3102/0013189X031007015

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

Slavin, R. E., Cheung, A. C. K., & Zhuang, T. (2021). How could evidence-based reform advance education? *ECNU Review of Education, 4*(1), 7–24. https://doi.org/10.1177/2096531120976060

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

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

Spinath, B. (2012). Academic achievement. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior* (2nd ed., pp. 1–8). Academic Press.

Spybrook, J. (2013). Introduction to special issue on design parameters for cluster randomized trials in education. *Evaluation Review, 37*(6), 435–444. https://doi.org/10.1177/0193841X14527758

Spybrook, J., Hedges, L., & Borenstein, M. (2014). Understanding statistical power in cluster randomized trials: Challenges posed by differences in notation and terminology. *Journal of Research on Educational Effectiveness, 7*(4), 384–406. https://doi.org/10.1080/19345747.2013.848963

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the institute of education sciences. *Educational Evaluation and Policy Analysis, 31*(3), 298–318. https://doi.org/10.3102/0162373709339524

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

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *2*(1), 15. https://doi.org/10.1177/2332858415625975

Spybrook, J., Zhang, Q., Kelcey, B., & Dong, N. (2020). Learning from cluster randomized trials in education: An assessment of the capacity of studies to determine what works , for whom , and under what conditions. *Educational Evaluation and Policy Analysis*, *42*(3), 354–374. https://doi.org/10.3102/0162373720929018

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

Stigler, S. M. (1992). A historical view of statistical concepts in psychology and educational research. *American Journal of Education*, *101*(1), 60–70. https://doi.org/10.1086/444032

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Stullich, S., Eisner, E., McCrary, J., & Policy and Program Studies Service. (2007). *National assessment of Title I. Final report: Volume I: Implementation*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/pdf/20084012_rev.pdf

Thompson, S. K. (2012). *Sampling* (3. ed). Wiley.

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/gd32gj

Torkildsen, J. V. K., Bratlie, S. S., Kristensen, J. K., Gustafsson, J.-E., Lyster, S.-A. H., Snow, C., Hulme, C., Mononen, R.-M., Næss, K.-A. B., López-Pedersen, A., Wie, O. B., & Hagtvet, B. (2022). App-based morphological training produces lasting effects on word knowledge in primary school children: A randomized controlled trial. *Journal of Educational Psychology*, *114*(4), 833–854. https://doi.org/10.1037/edu0000688

Turner, R. M., Thompson, S. G., & Spiegelhalter, D. J. (2005). Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials*, *2*(2), 108–118. https://doi.org/10.1191/1740774505cn072oa

Turner, R. M., Toby Prevost, A., & 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

U.S. 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

Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. *The Annals of Mathematical Statistics*, *18*(1), 88–96. https://doi.org/10.1214/aoms/1177730495

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

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808. https://doi.org/10.1002/pam.21760

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites?

Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876. https://doi.org/10.1080/19345747.2017.1300719

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*(6), 490–519. https://doi.org/10.1177/0193841X14531584

Whitehurst, G. J. (2003). *The institute of education sciences: New wine, new bottles* (2003 Annual Meeting Presidential Invited Session, p. 15). American Educational Research Association. https://files.eric.ed.gov/fulltext/ED478983.pdf

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

Williamson, S. F., Tishkovskaya, S. V., & Wilson, K. J. (2023). *Hybrid sample size calculations for cluster randomised trials using assurance* (arXiv:2308.11278). arXiv. http://arxiv.org/abs/2308.11278

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

Wozny, N., Balser, C., & Ives, D. (2018). Low-cost randomized controlled trials in education. *AEA Papers and Proceedings*, *108*, 307–311. https://doi.org/10.1257/pandp.20181054

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

Zhang, Q., Spybrook, J., Kelcey, B., & Dong, N. (2023). Foundational methods: Power analysis. In *International Encyclopedia of Education(Fourth Edition)* (pp. 784–791). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10088-0

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2012). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Educational Evaluation and Policy Analysis*, *34*(1), 45–68. https://doi.org/10.3102/0162373711423786

Zopluoglu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(1), 233–270.

2

STUDY I

Multilevel Design Parameters to Plan Cluster-Randomized Intervention Studies on Student Achievement in Elementary and Secondary School

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*, pp. 172-206. https://doi.org/10.1080/19345747.2020.1823539

This article has been posted as a preprint on EdArXiv.org. https://doi.org/10.35542/osf.io/f3p7q

Abstract

To plan cluster-randomized trials with sufficient statistical power to detect intervention effects on student achievement, researchers need multilevel design parameters, including measures of between-classroom and between-school differences and the amounts of variance explained by covariates at the student, classroom, and school level. Previous research has mostly been conducted in the United States, focused on two-level designs, and limited to core achievement domains (i.e., mathematics, science, reading). Using representative data of students attending grades 1 to 12 from three German longitudinal large-scale assessments ($3{,}963 \leq N \leq 14{,}640$), we used three- and two-level latent (covariate) models to provide design parameters and corresponding standard errors for a broad array of domain-specific (e.g., mathematics, science, verbal skills) and domain-general (e.g., basic cognitive functions) achievement outcomes. Three covariate sets were applied comprising (a) pretest scores, (b) sociodemographic characteristics, and (c) their combination. Design parameters varied considerably as a function of the hierarchical level, achievement outcome, and grade level. Our findings demonstrate the need to strive for an optimal fit between design parameters and target research context. We illustrate the application of design parameters in power analyses.

Keywords: explained variance, intraclass correlation, large-scale assessment, multilevel latent (covariate) models, power analysis

Multilevel Design Parameters to Plan Cluster-Randomized Intervention Studies on Student Achievement in Elementary and Secondary School

Educational research strongly moved towards evidence-based policies and practices at the outset of the 21st century, when educational stakeholders around the world increasingly demanded sound evidence of what actually works to foster student achievement (Kultusministerkonferenz, 2015; Organisation for Economic Co-operation and Development [OECD], 2007; Slavin, 2002). Formal education is usually organized within intact classrooms and schools. Further, various interventions operate by definition at the group level, such as teaching methods, curricular programs, or school reforms (Bloom, 2005; Boruch & Foley, 2000; Cook, 2005). A fundamental question of evidence-based education is therefore whether results on the effectiveness of interventions tested in small-scale laboratory experiments can be replicated when implementing these interventions, for instance, in the regular school day by teachers at the classroom or school level (see e.g., Gersten et al., 2015). An efficient way for educational researchers to address this concern is to conduct large-scale experiments where entire classrooms or schools rather than individual students are randomly assigned to the treatment or control condition. Studies of this type are known as cluster-randomized trials (CRTs; Donner & Klar, 2000; Raudenbush, 1997), place-based trials (Bloom, 2005), or group-randomized trials (Murray, 1998). CRTs can provide unbiased causal inferences about the impacts of interventions in the field at larger scales, and thus generate reliable knowledge to inform evidence-based educational policies and practices (Institute of Education Sciences & National Science Foundation, 2013; Slavin, 2002; Spybrook, Shi, et al., 2016).

Given their scale, CRTs are by nature very expensive. Hence, when planning such trials educational researchers should make every effort to ensure that their study design will allow for valid causal conclusions (Shadish et al., 2002). In this respect, a power analysis is an essential step in the planning phase of any CRT (American Educational Research Association, 2006, p. 37; American Psychological Association, 2019, pp. 83-84). However, power analysis for CRTs is particularly challenging as it requires reasonable assumptions on design parameters that take into account the multilevel (i.e., nested) structure of the outcome data. The reviews on CRTs in educational research (Spybrook & Raudenbush, 2009; Spybrook, Shi, et al., 2016) indicated that most studies (between 82 and 90%) had at least three hierarchical levels (e.g., students nested within classrooms, and classrooms nested within schools), with treatment allocation at either the classroom or school level. Thus, most educational researchers

conducting CRTs need multilevel design parameters that inform about the proportions of variance located at the student, classroom, and school level, as well as the respective amounts of variance that can be explained by vital covariates (e.g., pretest scores or sociodemographic characteristics) at these levels. Crucially, leading scholars strongly recommend using empirically established estimates of design parameters that match the target population, the target hierarchical level, and the target outcome measure rather than conventional benchmarks with unclear ties to the research context under investigation (Bloom et al., 2008; Brunner et al., 2017; Lipsey et al., 2012). To date most knowledge on design parameters is based on U.S. samples, only pertains to two-level designs (i.e., students within schools), and is limited to mathematics, science, and reading achievement (cf. Spybrook, 2013; Spybrook & Kelcey, 2016). Hence, the overarching goal of this article is to substantially expand the empirical body of knowledge on design parameters for CRTs in these three major dimensions. Our study is the first to compile (normative distributions of) design parameters with standard errors that are relevant to (I) the German school context or similar school systems, (II) three- as well as two-level designs, and (III) a broad variety of achievement domains.

Statistical Framework

Researchers need several multilevel design parameters to perform power analyses for CRTs aimed at enhancing student achievement based on three-level designs (Bloom et al., 2008; Hedges & Rhoads, 2010; Konstantopoulos, 2008a), where students at level one (L1) are nested within classrooms at level two (L2) which, in turn, are nested within schools at level three (L3):[30] (a) Intraclass correlations ρ quantifying the proportions of total variance in students' achievement that can be attributed to achievement differences between classrooms within schools (ρ_{L2}) and between schools (ρ_{L3}), as well as (b) the amounts of variance in students' achievement that can be explained by covariates, typically measured as squared multiple correlations R^2, at the student (R^2_{L1}), classroom (R^2_{L2}), and school level (R^2_{L3}).

The intraclass correlation at L2 is given by

$$\rho_{L2} = \frac{\sigma^2_{L2}}{\sigma^2_{T}}, \quad (1)$$

[30] Equivalent specifications for two-level designs (where students at L1 are nested within schools at L3) are recorded in the Supplemental Online Material A.

and at L3 by

$$\rho_{L3} = \frac{\sigma^2_{L3}}{\sigma^2_T}, \tag{2}$$

where $\sigma^2_T = \sigma^2_{L1} + \sigma^2_{L2} + \sigma^2_{L3}$ represents the total variance in students' achievement across all individual students, with σ^2_{L1}, σ^2_{L2}, and σ^2_{L3} denoting the variances between students within classrooms in schools, between classrooms within schools, and between schools, respectively. $\rho = 0$ implies that there are no between-classroom or between-school achievement differences, but rather that the total variance in students' achievement is located at L1. $\rho = 1$ means, inversely, that students within a classroom do not differ in their achievement, but rather that the total variance in students' achievement is located at L2 and L3.

A major challenge when designing a CRT is to ensure adequate precision (i.e., small standard errors) for any estimated intervention effects. It is well-documented that vital covariates (e.g., pretest scores or sociodemographic characteristics) may significantly raise the precision of randomized experiments (e.g., Bloom et al., 2007; Hedges & Hedberg, 2007a, 2013; Konstantopoulos, 2012; Raudenbush, 1997; Raudenbush et al., 2007). Covariates remove noise in the variance of an outcome measure (i.e., reduce σ^2_T), which improves the signal of the intervention effect (Raudenbush et al., 2007, p. 18). Although not necessary for validity, covariates can operate in CRTs at various hierarchical levels. When covariates explain a substantial proportion of variance in an outcome (in particular at higher levels), they are an efficient way to improve statistical power and precision, and thus reduce the required sample sizes and therefore the cost of CRTs (Bloom et al., 2007; Konstantopoulos, 2012; Raudenbush, 1997).

The explained variance at L1 is computed as

$$R^2_{L1} = \frac{\sigma^2_{L1} - \sigma^2_{L1|C_{L1}}}{\sigma^2_{L1}}, \tag{3}$$

at L2 as

$$R^2_{L2} = \frac{\sigma^2_{L2} - \sigma^2_{L2|C_{L2}}}{\sigma^2_{L2}}, \tag{4}$$

and at L3 as

$$R^2_{L3} = \frac{\sigma^2_{L3} - \sigma^2_{L3|C_{L3}}}{\sigma^2_{L3}}. \tag{5}$$

Here, $\sigma^2_{L1|C_{L1}}$, $\sigma^2_{L2|C_{L2}}$, and $\sigma^2_{L3|C_{L3}}$ are the covariate-adjusted within-classroom variance at L1, within-school variance at L2, and between-school variance at L3, respectively. C_{L1}, C_{L2} and C_{L3} denote a set of covariates introduced at L1, L2, and L3, respectively. Typically, multilevel modeling is applied to estimate the variance components σ^2_{L1}, σ^2_{L2}, and σ^2_{L3}, as well as the

covariate-adjusted variance components $\sigma^2_{\mathrm{L1}|C_{\mathrm{L1}}}$, $\sigma^2_{\mathrm{L2}|C_{\mathrm{L2}}}$, and $\sigma^2_{\mathrm{L3}|C_{\mathrm{L3}}}$ (for further details, see Supplemental Online Material A). Of note, C_{L2} and C_{L3} may include covariates assessed at L1 which are aggregated to L2 and/or L3 (e.g., the classroom and school mean of a pretest) as well as covariates assessed only at L2 (e.g., class size) or L3 (e.g., school size). Note that aggregated L1 covariates should be entered as group-mean centered variables in the multilevel models. Doing so ensures that the covariates explain variance only at the level at which they are specified (Konstantopoulos, 2008a, 2012). Consequently, the R^2 values (that may vary between 0 and 1) quantify the proportions of the variances observed at each level that can be explained by a certain set of covariates at the corresponding level.

The values for the design parameters ρ and R^2 at each level are entered into power calculations to determine the number of students, classrooms, and schools that are needed to achieve a certain minimum detectable effect size ($MDES$; Bloom, 1995). The $MDES$ can be described as the smallest true intervention effect that a study design could detect with confidence (Jacob et al., 2010) and thus is a measure of the precision of a CRT (Bloom, 2005). In formal terms, the $MDES$ is defined as the smallest possible standardized intervention effect that can be detected in a study of a certain sample size with, by convention, a power of $1 - \beta = .80$ and a significance level of $\alpha = .05$ in a two-tailed test (Bloom et al., 2008). Since the $MDES$ is standardized with respect to the total student-level standard deviation in the outcome, it can be conceived as a standardized effect size measure. For instance, an $MDES$ of 0.25 implies 80% power to detect an intervention effect on the outcome measure of one quarter of the total student-level standard deviation (Bloom et al., 2007).

The size of the $MDES$ depends on the type of CRT. Assuming no covariates and equal variances for the treatment and control group, the $MDES$ of a three-level CRT with treatment assignment at L3 is calculated as follows (see Bloom et al., 2008, Equation 2):

$$MDES = M_{df}\sqrt{\frac{\rho_{\mathrm{L3}}}{P(1-P)K} + \frac{\rho_{\mathrm{L2}}}{P(1-P)KJ} + \frac{(1-\rho_{\mathrm{L3}}-\rho_{\mathrm{L2}})}{P(1-P)KJn}}, \quad (6)$$

where n is the harmonic mean number of students per classroom, J is the harmonic mean number of classrooms per school, and K is the total number of schools. The multiplier M_{df} is a function of the t-distributions specific to α and $1 - \beta$ for the applied test procedure (i.e., one- or two-tailed) with $df = K - 2$ degrees of freedom (for details, see Bloom, 2005, pp. 158–160). For example, when 20 or more schools are randomly assigned to both the treatment and the control condition (i.e., $K \geq 40$), M_{df} equals approximately 2.8 (Bloom et al., 2008). Finally, P represents the proportion of schools assigned to the treatment group. From Equation (6) it becomes clear that the $MDES$ increases with growing values of ρ.

Adding covariates yields an adjusted $MDES$ (see Bloom et al., 2008, Equation 3):

$$MDES_{\text{adj}} = M_{df}\sqrt{\frac{\rho_{\text{L3}}(1-R^2_{\text{L3}})}{P(1-P)K} + \frac{\rho_{\text{L2}}(1-R^2_{\text{L2}})}{P(1-P)KJ} + \frac{(1-\rho_{\text{L3}}-\rho_{\text{L2}})(1-R^2_{\text{L1}})}{P(1-P)KJn}}, \quad (7)$$

with $df = K - g^*_{\text{L3}} - 2$ degrees of freedom where g^*_{L3} is the number of L3 covariates. Given that ρ_{L2} and ρ_{L3} are fixed values, adding covariates (especially at higher levels), as shown in Equation (7), leads to a lower $MDES$, or in other words, a higher precision of the CRT.

The formula for the adjusted $MDES$ of a three-level *multisite* or *blocked* cluster randomized trial (MSCRT; e.g., Konstantopoulos, 2008b; Raudenbush & Liu, 2000), where treatment assignment occurs at L2 subclusters (e.g., classrooms) within L3 clusters (serving as sites or blocks; e.g., schools), is given in Dong and Maynard (2013, pp. 53–55):

$$MDES_{\text{MSCRT}_{\text{adj}}} = M_{df}\sqrt{\frac{\tau^2_{\delta_{\text{L3}}}\rho_{\text{L3}}\,(1-R^2_{\delta_{\text{L3}}})}{K} + \frac{\rho_{\text{L2}}(1-R^2_{\text{L2}})}{P(1-P)KJ} + \frac{(1-\rho_{\text{L3}}-\rho_{\text{L2}})(1-R^2_{\text{L1}})}{P(1-P)KJN}}, \quad (8)$$

where $\tau^2_{\delta_{\text{L3}}} = \sigma^2_{\delta_{\text{L3}}} / \sigma^2_{\text{L3}}$ is the effect size variability at L3 (i.e., the heterogeneity of the intervention effect δ across schools) with $\sigma^2_{\delta_{\text{L3}}}$ denoting the between-school variance in δ. Further, $R^2_{\delta_{\text{L3}}}$ is defined as the proportion of $\tau^2_{\delta_{\text{L3}}}$ that can be explained by covariates at L3: $R^2_{\delta_{\text{L3}}} = (\tau^2_{\delta_{\text{L3}}} - \tau^2_{\delta_{\text{L3}|C_{\text{L3}}}}) / \tau^2_{\delta_{\text{L3}}}$, where $\tau^2_{\delta_{\text{L3}|C_{\text{L3}}}}$ is the covariate-adjusted effect size variability at L3. If δ is considered to be constant across schools (as represented by a fixed effect), $\tau^2_{\delta_{\text{L3}}}$ and ρ_{L3} equal zero and thus, the first term within the square root (i.e., $\tau^2_{\delta_{\text{L3}}}\rho_{\text{L3}}\,(1 - R^2_{\delta_{\text{L3}}}) / K$) vanishes and is dropped from Equation (8). In this fixed effect scenario, df becomes $K(J - 2) - g^*_{\text{L2}}$, where g^*_{L2} is the number of L2 covariates. If δ is considered to vary across schools (as represented by a random effect), df is $K - g^*_{\text{L3}} - 1$. As in the computation of the unadjusted $MDES$ for CRTs, the values for g^* and R^2 equal zero (and are therefore dropped from Equation (8)) when no covariates are used.

Previous Empirical Research on Multilevel Design Parameters

A critical question that any educational researcher faces when performing power analyses is which values of ρ and R^2 at each hierarchical level should be entered in the equations presented above. Unfortunately, many applied researchers (still) draw on conventional guidelines: For example, they interpret values of $\rho = .01$ as "small", $\rho = .10$ as "medium", and $\rho = .25$ as "large" (LeBreton & Senter, 2008, p. 838). These guidelines, though, were proposed as "operational definitions", with the strong recommendation to use better estimates whenever possible – "better" means that they should match the target population, hierarchical level, and

outcome measure of the study (e.g., Cohen, 1988, pp. 12–13 and 534; Lipsey et al., 2012, p. 4). Thus, what do we know about design parameters at the various levels for student achievement?

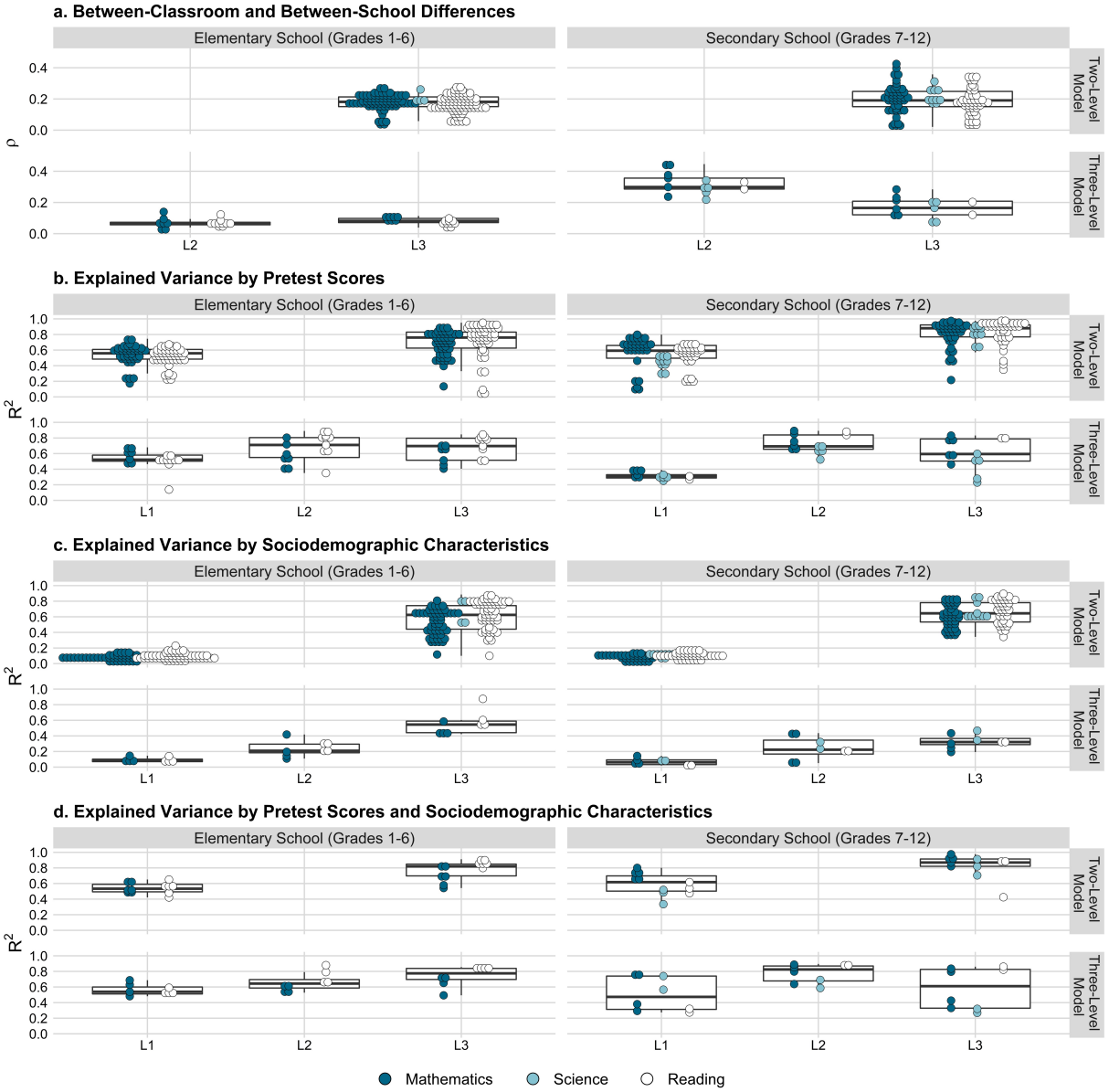
International Research

First, in the United States, the body of knowledge on design parameters has substantially expanded in recent years (cf. Spybrook, 2013; Spybrook & Kelcey, 2016), especially for the core achievement domains mathematics, science, and reading. Figure 1 summarizes design parameters based on U.S. samples as reported in previous research.

Second, it is evident from Figure 1 that most studies in the United States have cataloged design parameters that are relevant for planning two-level CRTs (i.e., students within schools; see upper panels in Figure 1a to d). Despite expected variation across samples, domains, and grade levels, this line of research indicates that the variance attributable to between-school achievement differences in the United States only occasionally exceeds a value of $\rho_{L3} = .25$ (see Figure 1a).

In contrast, few studies have compiled variance components for three-level designs (i.e., students within classrooms within schools; see lower panels in Figure 1a-d). Figure 1a reveals that intraclass correlations at L2 vary by grade level. For instance, in the study by Zhu and colleagues (2012), values of ρ_{L2} were usually smaller than .14 (with $\rho_{L3} \leq .10$) in both mathematics and reading in elementary school. In secondary school, however, Zhu et al. (2012) reported between-classroom differences within a range of $.29 \leq \rho_{L2} \leq .38$ in tests related to mathematics and science (with $.07 \leq \rho_{L3} \leq .17$). The authors argue that this increase in ρ_{L2} probably reflects a more extensive student tracking within secondary schools than within elementary schools (Zhu et al., 2012, p. 53).

Figure 1. *Results from Previous Research on Multilevel Design Parameters for Student Achievement in Elementary and Secondary School in the United States: (a) Between-Classroom (ρ_{L2}) and Between-School Differences (ρ_{L3}), and Explained Variances by (b) Pretest Scores, (c) Sociodemographic Characteristics, and (d) Pretest Scores and Sociodemographic Characteristics at the Student (R^2_{L1}), Classroom (R^2_{L2}), and School Level (R^2_{L3})*



Note. Boxplots show distributions across all domains. The distributions in mathematics/science/reading are based on 341/12/370 values for elementary school (grades 1-6) and 266/93/223 values for secondary school (grades 7-12). The underlying data table can be obtained from the OSF (https://osf.io/2w8nt). In the upper panels of Figures 1a to 1d, design parameters obtained from two-level models (students at L1 within schools at L3) are shown as reported in the following studies: Bloom et al. (1999) reported ρ_{L3} for elementary schools in 1 city. Bloom et al. (2007) reported ρ_{L3}, R^2_{L1} and R^2_{L3} for pretests and sociodemographics for elementary and secondary schools in 5 districts. Brandon et al. (2013) reported upper bounds of the means of ρ_{L3} across several years for elementary and secondary schools in one state. Hedberg et al. (2004) reported ρ_{L3}, and R^2_{L3} for sociodemographics for elementary schools in 120 districts and for secondary schools on a nationwide basis (values are retrieved from Schochet, 2008). Hedges and Hedberg (2007a) reported ρ_{L3}, R^2_{L1} and R^2_{L3} for pretests, sociodemographics, and their combination for elementary and secondary schools on a nationwide basis (across districts and states).

(Figure continues)

Figure 1. *(continued)*

Note. Hedges and Hedberg (2013) reported ρ_{L3}, R^2_{L1} and R^2_{L3} for pretests and sociodemographics for elementary and secondary schools in 11 states (with between-district variance pooled into between-school variance within states). Schochet (2008) reported ρ_{L3} for elementary schools based on 3 studies conducted in 6 cities, 12 districts, and 7 states, respectively. Spybrook, Westine, et al. (2016) reported means of ρ_{L3}, R^2_{L1} and R^2_{L3} across several years for pretests and sociodemographics for elementary and secondary schools in 3 states. Westine et al. (2013) reported means of ρ_{L3}, R^2_{L1} and R^2_{L3} across 5 years for pretests, sociodemographics, and their combination for elementary and secondary schools in 1 state. In the lower panels of Figures 1a to 1d, design parameters obtained from three-level models (students at L1 within classrooms at L2 within schools at L3) are shown as reported in the following studies: Jacob et al. (2010) reported ρ_{L2}, ρ_{L3}, R^2_{L1}, R^2_{L2} and R^2_{L3} for pretests, sociodemographics and their combination for elementary schools in 6 districts. Xu and Nichols (2010) reported ρ_{L2}, ρ_{L3}, R^2_{L1}, R^2_{L2} and R^2_{L3} for pretests, sociodemographics, and their combination for elementary and secondary schools in 2 states. Zhu et al. (2012) reported ρ_{L2}, ρ_{L3}, R^2_{L1}, R^2_{L2} and R^2_{L3} for pretests for elementary and secondary schools on a nationwide basis.

Third, a small number of studies outside the United States have investigated intraclass correlations focusing on between-school achievement differences. The study by Kelcey et al. (2016) drew on representative samples of grade 6 students in 15 sub-Saharan African countries. Their results showed that between-school differences in mathematics and reading varied widely across countries ($.08 \leq \rho_{L3} \leq .60$). Zopluoglu (2012) reanalyzed data from several cycles of the Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) and found that ρ_{L3} varied considerably across countries in mathematics, science and reading. For example, in the year 2007 cycle of TIMSS the average intraclass correlation at L3 in mathematics were $\rho_{L3} = .27$ across 44 countries ($SD = .14$, $.07 \leq \rho_{L3} \leq .62$) in grade 4, and $\rho_{L3} = .31$ across 57 countries ($SD = .14$, $.03 \leq \rho_{L3} \leq .65$) in grade 8. Similar results were found for science and reading. Finally, capitalizing on five cycles of the Programme for International Student Assessment (PISA) with representative data from 15-year-old students from 81 different countries and economies, Brunner and colleagues (2017) found large international variation in between-school achievement differences with median values of ρ_{L3} lying around .40 (ranging from .10 to over .60). In sum, these results from international studies clearly show that design parameters obtained for the United States do not generalize well to the large majority of other countries. For instance, the analyses by Brunner et al. (2017, p. 21) reveal that in about 80% of the countries that participated in PISA, achievement differences at L3 are (much) larger than those typically found for U.S. schools.

Fourth, pretest scores have proven to be highly powerful in explaining variance in students' achievement at all levels (e.g., Bloom et al., 2007; Hedges & Hedberg, 2007a; Westine et al., 2013; Zhu et al., 2012; see Figure 1b). For example, in the study by Zhu and colleagues (2012, p. 66, Table A1), median values for the proportions of variance explained by pretests were $R^2_{L1} = .59$, $R^2_{L2} = .72$, and $R^2_{L3} = .52$.

Fifth, as a rule, sociodemographic characteristics (i.e., as typically represented by a covariate set comprising socioeconomic status, gender, and migration background) explain a smaller proportion of variance in students' achievement at L1, and a larger proportion at L3. As shown in Figure 1c, in the United States (e.g., Bloom et al., 2007; Hedges & Hedberg, 2013; Spybrook, Westine, et al., 2016), values of R^2_{L3} typically lie in the range of .42 to .79. The corresponding average values of R^2_{L1} typically lie around .10. This general pattern of results was also found for sub-Saharan countries in the study by Kelcey et al. (2016) as well as in the analyses of Brunner et al. (2017) for 81 countries participating in PISA. Notably, these international studies also demonstrated that achievement differences adjusted for sociodemographics varied widely across countries. For example, values of R^2_{L3} for reading ranged between .18 and .89 across countries (Brunner et al., 2017).

To the best of our knowledge, only Jacob et al. (2010) and Xu and Nichols (2010) have provided empirical estimates of R^2_{L2} for the application of sociodemographic covariates. Drawing on data from 3rd graders, Jacob and colleagues (2010) reported that sociodemographics explained 42%/20% of the variance located at L2 for mathematics/reading achievement. In the investigation of Xu and Nichols (2010) the proportions of explained variance at L2 varied by state, domain, and grade level: The values for R^2_{L2} in elementary school were between .11 (mathematics; Florida) and .32 (reading; North Carolina), and in secondary school between .05 (mathematics; Florida) and .44 (geometry; North Carolina).

Sixth, drawing on data from the United States for K-12th graders, Hedges and Hedberg (2007a) found that sociodemographics provided (almost) no incremental gain in explaining variance in mathematics and reading at either L1 or L3, once pretests were controlled for at these levels. However, the analyses of Jacob et al. (2010; Table 2) as well as Xu and Nichols (2010, Table NC-7) suggest that sociodemographics may contribute to the prediction over and above pretests, especially at L2 (see Figure 1d).

Research in Germany

To date, design parameters for student achievement in Germany have typically been reported in the context of research on educational effectiveness or social inequalities, mainly as ancillary results. Hence, the knowledge base is scattered and design parameters for Germany have not been systematically summarized. Table 1 provides an overview of intraclass correlations as reported in several key German large-scale studies.

Table 1. *Results from Previous Large-Scale Studies on Student Achievement in Germany: Between-Classroom (ρ_{L2}) and Between-School Differences (ρ_{L3}) by Grade and Domain*

Grade	Mathematics		Science		Verbal Skills in German				English Reading	
					Reading		Listening			
	ρ_{L2}	ρ_{L3}	ρ_{L2}	ρ_{L3}	ρ_{L2}	ρ_{L3}	ρ_{L2}	ρ_{L3}	ρ_{L2}	ρ_{L3}
Elementary School										
1										
2										
3										
4	.02 [a]	*.25* [b]/*.27* [c]/.15 [a]		*.26* [d]	.04 [e]	*.22* [f]/*.24* [g]/.17 [e]		*.27* [h]		
Secondary School										
5										
6										
7	.03 [i]	.45 [i]								
8		*.51* [j]/*.49* [k]		*.41* [l]						
9		*.56* [m]		*.54* [m]		*.58* [m]/*.48* [n]				*.55* [n]
10	.03 [i]	*.47* [o]/.62 [i]		*.44* [p]						
11										
12		*.52* [q]								

Note. Design parameters in italic/normal print are based on national probability samples/representative samples of a certain state. [a] Lehmann & Lenkeit (2008, Table 3.6). [b] Haag & Roppelt (2012, Figure 5.11). [c] Martin et al. (2013, p. 139). [d] Martin et al. (2013, p. 140). [e] Lehmann & Lenkeit (2008, Table 3.3). [f] Böhme & Weirich (2012, Figure 5.3). [g] Martin et al. (2013, p. 138). [h] Böhme & Weirich (2012, Figure 5.4). [i] Baumert et al. (2003, Figure 10.6) where values of ρ_{L3} represent the sum of the variances between schools and between school types. [j] Martin et al. (2000, p. 77). [k] Baumert et al. (2000, p. 68). [l] Martin et al. (2000, p. 76). [m] Brunner et al. (2017, Table S2) with data from 15-year-old students of which about 65% attend grade 9; most remaining students attend grade 10. [n] Knigge & Köller (2010, Table 10.1). [o] Senkbeil (2006, p. 298). [p] Senkbeil (2006, p. 299). [q] Baumert et al. (2000, p. 69).

The following results are noteworthy in Table 1: First, intraclass correlations were only available at L3 for the majority of studies, and these differed markedly between elementary school and secondary school. Elementary school values of ρ_{L3} lay between .15 and .27, whereas secondary school values of ρ_{L3} lay between .41 and .62. In the very few studies where intraclass correlations at L2 were reported, they appeared rather small (with $\rho_{L2} \leq .04$) compared to between-school differences. Finally, the many empty cells in Table 1 demonstrate that the existing empirical research on design parameters for German schools is limited to selected hierarchical levels, achievement domains, and grades.

Second, as in most countries, the amount of variance explained by sociodemographics differs substantively between levels in Germany. For example, in the reanalysis of data from five PISA cycles by Brunner et al. (2017), the average proportion of L1 variance explained by socioeconomic status, gender, and migration background was $R^2_{L1} = .09/.09/.10$ for German students' achievement in mathematics/science/reading. On the other hand, the respective average proportion of explained L3 variance was $R^2_{L3} = .75/.77/.77$. Similar patterns of results

were also found in other studies (Baumert et al., 2003; Knigge & Köller, 2010). Of note, to the best of our knowledge, multilevel models have not yet been used to decompose the variance that can be explained by pretests at L1, L2, and L3 for German schools.

Third, the design parameters reported in Table 1 refer to the general (i.e., total) student population. At Germany's elementary level, there is only a single type of elementary school across all 16 federal states ("Grundschule"; up to grade 4 in most German federal states). However, at the secondary level, Germany's school system is characterized – like many other countries (Salchegger, 2016) – by tracking into different school types that cater to students with different performance levels. Typically, five major school types are distinguished in large-scale studies: The academic track school ("Gymnasium"; up to grade 12 or 13), vocational school ("Hauptschule"; up to grade 9 or 10), intermediate school ("Realschule"; up to grade 10), multitrack school ("Schulen mit mehreren Bildungsgängen"; up to grade 9, 10, 12, or 13), and comprehensive school ("Gesamtschule"; up to grade 12 or 13). Notably, all federal states offer schools in the academic track but they vary with respect to the other school types. In the remainder of this article, we will therefore subsume the latter four school types under the umbrella term "non-academic track" to describe this broad class of schools.

Importantly, when statistically controlling for mean-level differences between school types in secondary education (e.g., by introducing school type as a L3 covariate), ρ_{L3} may decrease markedly. For instance, Baumert and colleagues (2003, p. 270) found that around 47%/45% of the total variance in mathematics/reading achievement of 9th graders were accounted for by differences between school types whereas 7%/12% were attributable to differences between schools of the same type; the remaining 46%/43% were attributable to differences between students within schools. Drawing on data from the German federal state North Rhine-Westphalia, Baumert et al. (2003) also delineated that the amount of variance attributable to school types may increase with higher grades, while the amount of variance attributable to differences between schools of the same type decreases. In summary, these results for the German school system have two implications: first, school types are an important feature of the German school system that explain a substantial proportion of between-school differences in students' achievement and, second, design parameters obtained for certain grades cannot be easily generalized to other grades.

The Present Study

Multilevel design parameters that are tied to the target population, hierarchical level, and outcome measure are indispensable for designing CRTs on student achievement with sufficient statistical power and precision. However, our literature review showed that the corresponding empirical knowledge base is limited in several ways: First, existing compendia of design parameters are based almost exclusively on U.S. samples, whereas the body of knowledge is rather weak for Germany and other countries with similar school systems. Second, most previous research on design parameters focused on two-level structures (i.e., students within schools), but little research has been done using three-level analyses yielding classroom-level estimates in the United States and elsewhere. Third, design parameters are most frequently available for the core achievement domains mathematics, science, and reading. Yet, contemporary educational curricula go far beyond these core domains (National Research Council, 2011; OECD, 2018): They cover a multifaceted skills portfolio including, for instance, verbal skills in foreign languages and domain-general skills such as information and communication technology literacy and problem solving. Although cognitive outcomes of different domains are correlated, their unique characteristics may introduce considerable variation in design parameters and, therefore, in the required samples sizes for CRTs (Westine et al., 2013). Finally, it is important to quantify the statistical uncertainty associated with empirically estimated design parameters due to sampling error (Hedges et al., 2012). To date, standard errors or confidence intervals have rarely been reported for ρ and R^2 at L1 and L3 (Hedges & Hedberg, 2007a, 2007b, 2013; Jacob et al., 2010) and, as far as we are aware, never at L2.

The present study directly addresses these research gaps. Specifically, this is the first study to rigorously investigate design parameters and their standard errors (I) based on rich, large-scale data from German samples spanning the entire school career (grades 1 to 12), (II) for three- as well as two-level designs, and (III) for a very wide array of achievement domains. Following prior work (e.g., Bloom et al., 2007; Hedges & Hedberg, 2007a, 2013; Westine et al., 2013), we use pretest scores and sociodemographic characteristics as covariates at each level to determine the increase in the precision of CRTs when estimating causal effects on student achievement. We analyze three-level design parameters (i.e., ρ_{L2}, ρ_{L3}, R^2_{L1}, R^2_{L2}, R^2_{L3}) and two-level design parameters (i.e., ρ_{L3}, R^2_{L1}, R^2_{L3}) for the general student population. Given that tracking is a key characteristic of the secondary school system in Germany and many other school systems around the world (Salchegger, 2016), we additionally estimate design

parameters both by adjusting them for mean-level differences in achievement between school types as well as separately for the academic and non-academic track. Finally, we illustrate how the present design parameters can be applied in power analysis in the planning phase of CRTs.

Method

Large-Scale Assessment Data

This study drew on several national probability samples from three German longitudinal large-scale assessments: The National Educational Panel Study (NEPS; Blossfeld et al., 2011), the Assessment of Student Achievements in German and English as a Foreign Language (DESI; DESI-Konsortium, 2008), and the longitudinal extension of the year 2003 cycle of the Programme for International Student Assessment (PISA-I-Plus 2003, 2004 [PISA-I+]; PISA-Konsortium Deutschland, 2006). NEPS is an ongoing complex multi-cohort study on the interplay of student achievement, educational processes, and life outcomes across the lifespan. We analyzed data from students attending grades 1 to 12 using the starting cohorts (SC) 2, 3, and 4. DESI investigated the development of first (i.e., German) and foreign language (i.e., English) achievement during grade 9. PISA-I+ focused on the development of mathematics and science achievement from grade 9 to 10 and additionally contains assessments of reading and problem solving in grade 9.

All studies followed a multistage sampling procedure. In NEPS-SC3 and -SC4, as well as in DESI and PISA-I+, two entire classrooms per school were randomly drawn (Aßmann et al., 2011; Beck et al., 2008; Prenzel et al., 2006). For NEPS-SC2, the sample did not consist of intact classrooms but rather was representative of children entering elementary school (Aßmann et al., 2011).

Our analysis sample of NEPS-SC2 included students who took part in the study in grade 1. It was composed of two subsamples: students who started participating as 4-year-old kindergarten children (school year 2010/11; wave 1) and a refreshment sample of 1st graders (2012/13; wave 3), both providing data up to grade 4 (2015/16; wave 6). The analysis sample of NEPS-SC3 comprised students from grade 5 (2010/11; wave 1) up to 9 (2014/15, wave 6) and, again, included two subsamples: 5th graders of wave 1, and a refreshment sample of grade 7 students (2012/13; wave 3). For NEPS-SC4, we analyzed data from students from grade 9 (2010/11; wave 1) up to 12 (2013/14; wave 7). For DESI, we analyzed data of the full student sample at the outset (wave 1) and end (wave 2) of grade 9 in 2003/04. The analysis sample of

PISA-I+ covered students from grade 9 (2002/03; wave 1) up to 10 (2003/04; wave 2). Datasets for each large-scale study and grade consisted of those students who participated in the studies in the respective grade and for whom the exclusion criteria[31] did not apply. Table 2 contains detailed information on sample sizes by grade, large-scale study, and school track.

Table 2. *Number of Students (L1), Classrooms (L2), and Schools (L3), and Median Cluster Sizes by Grade, Large-Scale Study, and School Track*

Grade	Study	Total					Academic Track					Non-Academic Track				
		N			*Mdn*		*N*			*Mdn*		*N*			*Mdn*	
		L1	L2	L3	L1	L2	L1	L2	L3	L1	L2	L1	L2	L3	L1	L2
Elementary School																
1	NEPS-SC2	6,731	1,020	374	6	2	-	-	-	-	-	-	-	-	-	-
2	NEPS-SC2	6,319	986	362	6	2	-	-	-	-	-	-	-	-	-	-
3	NEPS-SC2	5,554	888	354	6	2	-	-	-	-	-	-	-	-	-	-
4	NEPS-SC2	5,419	1,026	349	4	3	-	-	-	-	-	-	-	-	-	-
Secondary School																
5	NEPS-SC3	5,380	452	225	12	2	2,340	155	76	15	2	3,040	297	149	10	2
6	NEPS-SC3	5,026	452	211	11	2	2,287	170	76	13	2	2,739	282	135	10	2
7	NEPS-SC3	6,279	614	266	10	2	2,980	254	105	11	2	3,299	360	161	8	2
9	NEPS-SC3	4,651	627	239	6	2	2,255	271	95	8	2	2,396	356	144	5	2
9	NEPS-SC4	14,640	975	518	15	2	5,098	292	146	18	2	9,542	683	372	14	2
9	DESI	10,543	427	219	25	2	4,308	163	82	27	2	6,235	264	137	24	2
9	PISA-I+	6,020	275	152	23	2	2,664	116	61	23	2	3,356	159	91	22	2
10	NEPS-SC4	10,031	824	402	12	2	3,770	298	118	12	2	6,261	526	284	12	2
10	PISA-I+	6,020	275	152	23	2	2,664	116	61	23	2	3,356	159	91	22	2
11	NEPS-SC4	4,566	n/a	175	26	n/a	4,087	n/a	143	29	n/a	479	n/a	32	14	n/a
12	NEPS-SC4	3,963	n/a	168	23	n/a	3,596	n/a	137	27	n/a	367	n/a	31	12	n/a

Note. Cells containing a dash indicate that tracking into different school types does not occur in elementary school. Cells containing n/a indicate that classroom-level information was not available as 11th and 12th grade students did not attend intact classrooms, but rather the grouping of students varied depending on the subject taught.

The sample sizes varied from $N = 3{,}963$ students from 168 schools (NEPS-SC4, grade 12) and $N = 14{,}640$ students in 975 classrooms in 518 schools (NEPS-SC4, grade 9). Notably, none of samples from the three large-scale studies comprises 8th grade students as achievement tests were not conducted in this grade. Furthermore, in the German school system, the majority of 11th and 12th graders are not grouped in intact classrooms, but rather attend courses that are specific to the subject taught at different ability levels (e.g., basic and advanced courses). Information on classroom affiliation in grades 11 to 12 consequently did not exist.

[31] The exclusion criteria applied for the present analyses are outlined in the Supplemental Online Material A. Table A1 itemizes the number of excluded students. Sensitivity analyses showed no systematic differences in the study measures between students that were included and those that were excluded (see Tables A2 to A4).

Measures

Achievement Outcomes

We examined a broad spectrum of domain-specific and domain-general achievement measures (for a comprehensive overview, see Table A5 in the Supplemental Online Material A). The datasets included data at L1 in various domains: mathematics, science, specific verbal skills in German as a first language (reading comprehension, reading speed, spelling, grammar, vocabulary, writing, argumentation, listening), and specific verbal skills in English as a foreign language (reading comprehension, text reconstruction, language awareness, writing, listening). Likewise, we investigated domain-general areas: declarative metacognition, information and communication technology, problem solving, and basic cognitive functions (perception speed, reasoning).

Assessments were conducted in all grades from 1 to 12 except grade 8. All tests were administered using a paper-and-pencil format. Test scores were provided either as weighted likelihood estimates (WLE; Warm, 1989) that were derived from item-response models, or as sum or mean scores that were computed by the number of correctly solved items.

Pretest Scores

For each outcome measure we used the corresponding previously-collected domain-identical achievement score as predictor, if available. If there were multiple pretests from different years for a certain domain, we selected the pretest with the smallest time lag between pre- and posttest. When studying mathematics, science, and German vocabulary and grammar as outcomes in grade 1, and basic cognitive functions in grade 2, we included the corresponding pretests that were assessed in kindergarten (waves 1 and 2 of NEPS-SC2). If no domain-identical pretest was available, we used predictors that were conceptually related to the target outcome (so-called “proxy” pretests; Shadish et al., 2002, p. 118; see Table A6 in the Supplemental Online Material A). However, some grade-specific achievement outcomes did not have any relevant pretest available.

Sociodemographic Characteristics

We used four sociodemographic characteristics as covariates. Specifically, we used two measures of socioeconomic status, including the highest International Socio-Economic Index of Occupational Status within a family (HISEI; Ganzeboom & Treiman, 1996) and an indicator of the highest educational attainment within the family. The highest educational attainment was

based on the greatest number of years of schooling completed within a family (ranged between 9 and 18) for NEPS and PISA-I+ and the highest school-leaving qualification within a family (with 1 = no qualification up to 5 = "Abitur") for DESI. Indicator variables were used to represent students' gender (0 = male, 1 = female) and migration background (0 = no migration background, 1 = migration background).

Statistical Analyses

Missing Data

Missing data is an unavoidable reality in any large-scale assessment (for missing data statistics, see Tables A7 to A11 in the Supplemental Online Material A). Across all datasets, the total percentage of missing values varied from 6% (NEPS-SC2, grade 3) to 32% (NEPS-SC2, grade 1). The highest missing rates occurred in pretests measured in the first two waves of NEPS-SC2, as only a small share of the kindergarten children continued participating in NEPS after entering elementary school. To deal with missing data we used (groupwise) multilevel multiple imputation and generated 50 multiply imputed datasets for each large-scale study and grade using the mice (van Buuren & Groothuis-Oudshoorn, 2011) and miceadds (Robitzsch et al., 2018) packages (for details, see Supplemental Online Material A).

Multilevel Models

Adapting the approach of Hedges and Hedberg (2007), we estimated four sets of three-level (i.e., students within classrooms within schools) and two-level (i.e., students within schools) multilevel latent (covariate) models (Lüdtke et al., 2008) with random intercepts for each grade and achievement outcome. Notably, all covariates were assessed at L1. The classroom and school means of these covariates were estimated by applying the default options for the latent multilevel modeling framework as implemented in M*plus* 8 (Muthén & Muthén, 2017), and thus were entered as latent group means in the models. Doing so also implies that in the three-level models both L1 covariates and L2 means were "implicitly" centered at the respective classroom and school means (Muthén & Muthén, 2017, pp. 274–275).

Model set 1 was an *intercept-only model* that did not contain any covariates. Model set 2 was a *pretest covariate(s) model* that drew on the respective pretest scores (or proxy pretest scores, if necessary) as predictors at each level. Model set 3 was a *sociodemographic covariates model* that included at each level students' socioeconomic status (i.e., HISEI and the highest educational attainment within the family), gender, and migration background. Model set 4 was

a *pretest and sociodemographic covariates model* that combined the pretest covariate(s) model and the sociodemographic covariates model. All analysis models are specified in Equations (A13) to (A30) in the Supplemental Online Material A.

To estimate design parameters at L1, L2, and L3 for grades 1 to 10, we applied three-level modeling. For grades 11 to 12, we specified two-level models to estimate design parameters at L1 and L3 because German education in grades 11 and 12 is usually not organized within intact classrooms. As noted above, secondary students in Germany are tracked into different school types. We therefore also applied two different adjustments to model sets 1 to 4 to estimate design parameters in secondary education taking tracking into account. First, we adjusted the design parameters for mean-level differences in achievement between school types. To accomplish this, we added dummy-coded indicator variables representing the various school types as covariates at L3 in all multilevel models (see Table A12 in the Supplemental Online Material A). Second, we examined model sets 1 to 4 separately for the subpopulations of students in the academic and non-academic track.

Finally, we ran model sets 1 to 4 as two-level models for grades 1 to 10 for the general student population (both with and without adjusting for mean-level differences between school types), and separately for the academic and non-academic track. This approach allows us to provide design parameters at L1 and L3 that are appropriate for research lacking information at the classroom level.

Estimation of Design Parameters and Standard Errors

The analyses were conducted in three steps. First, model sets 1 to 4 were run separately for each large-scale study, grade, achievement outcome, and for each of the 50 imputed datasets in M*plus* 8 (Muthén & Muthén, 2017) using the maximum likelihood estimator with robust standard errors (MLR) which were computed based on a sandwich estimator.[32] These analyses were run via R (R Core Team, 2018) using the MplusAutomation package (Hallquist & Wiley, 2018).

Second, the calculation of the design parameters and their standard errors was done in R (R Core Team, 2018) using the estimates obtained in the first step: We employed Equations (1) and (2) to calculate ρ_{L2} and ρ_{L3}, respectively, Equations (3), (4), and (5) to calculate R^2_{L1}, R^2_{L2}, and R^2_{L3}, respectively, as well as Equations (A18) and (A22) displayed in the Supplemental Online Material A to calculate school-type-adjusted values of ρ_{L3} and R^2_{L3}, respectively. The

[32] In very few cases, negative R^2 values or estimation problems occurred. Different strategies applied to resolve these estimation issues are described in the Supplemental Online Material A.

standard errors of the ρ values were computed using the formulas for large sample variances in unbalanced three-level designs (i.e., with unequal cluster sizes) presented in Hedges et al. (2012, Equations 7 to 9), and the formula for the large sample variance in unbalanced two-level designs given in Donner and Koval (1980, Equation 3). The standard errors of the R^2 values were calculated drawing on Hedges and Hedberg (2013, p. 451).

Third, the design parameters and standard errors obtained in the second step were pooled across imputations using Rubin's (1987) rules in R (R Core Team, 2018) using the mitml package (Grund et al., 2019) to combine the estimates into a single set of results and to obtain standard errors that take into account within and between imputation variance. Of note, for grade 9, design parameters for the same achievement domain were available from several large-scale studies. We integrated these results in R (R Core Team, 2018) with the metafor package (Viechtbauer, 2010) and applied a meta-analytic fixed effects model to determine the average design parameter estimates across the grade 9 samples (Hedges & Vevea, 1998).[33]

Results

The complete compilation of multilevel design parameters, corresponding standard errors, and normative distributions are available in Tables B1 to B16 in the Supplemental Online Material B on the Open Science Framework (OSF; https://osf.io/2w8nt; see also Figure 4). Table 3 aggregates the results based on three-level (grades 1 to 10) and two-level (grades 11 to 12) models for the general student population (with and without adjustment for mean-level differences between school types), the academic track, and the non-academic track, yielding normative distributions of design parameters. Figure 2 visualizes the results for the general student population as well as the school-type-adjusted results at L3 by grade level and achievement domain.

Design Parameters for the General Student Population

The results obtained for the intercept-only models demonstrated substantial between-school differences in students' achievement across grade levels and domains. As displayed in Figure 2a, values of ρ_{L3} were noticeably smaller in elementary ($Mdn(\rho_{L3}) = .11$) than in secondary

[33] After careful consideration we decided not to use sampling weights in our analyses. As we had to exclude students who did not meet the criteria required for our analyses, applying the weights to the remaining students would have no longer represented the total German student population.

school ($Mdn(\rho_{L3}) = .35$; see also Table 3). Moreover, achievement differences at L3 varied widely between outcome measures and grade levels, and even within grade levels (see Tables 3 and B1): In elementary school, ρ_{L3} ranged from .04 (e.g., basic cognitive functions in reasoning, grade 2) to .22 (German vocabulary, grade 1), and in secondary school from .09 (German reading comprehension, grade 12) to .59 (English language awareness in grammar, grade 9). Compared to between-school differences, between-classroom differences were considerably smaller ranging from $\rho_{L2} = .03$ (e.g., German grammar, grade 1) to $\rho_{L2} = .09$ (basic cognitive functions in perception speed, grade 2) in elementary school ($Mdn(\rho_{L2}) = .05$), and from .01 (declarative metacognition, grade 9) to .13 (e.g., German reading speed, grade 5) in secondary school ($Mdn(\rho_{L2}) = .04$; see Figure 2a, Tables 3 and B1).

The results of the pretest covariate(s) models showed that pretest scores (including proxy pretests) explained substantial amounts of variance in students' achievement at all hierarchical levels with median values of $R^2_{L1} = .21$, $R^2_{L2} = .51$, and $R^2_{L3} = .89$ across elementary and secondary school (see Table B2). Table 3 and Figure 2b reveal that the effectiveness of pretests in reducing variability in students' achievement at L1 were relatively consistent across grade levels with $Mdn(R^2_{L1}) = .24/.20$ in elementary/secondary school. The explanatory power of pretests at L2 and L3, however, depended on the grade level: Pretests explained substantively larger proportions of L2 and L3 variance in secondary school (median values: $R^2_{L2} = .65$, $R^2_{L3} = .96$) than in elementary school (median values: $R^2_{L2} = .27$, $R^2_{L3} = .39$). The corresponding standard errors were exceptionally large in grade 1 (e.g., German grammar: $SE(R^2_{L2}) = .34$, $SE(R^2_{L3}) = .32$; see Table B1). The proportion of explained variance varied considerably across domains for all grade levels ($.01 \leq R^2_{L1} \leq .56$, $.00 \leq R^2_{L2} \leq .95$, $.00 \leq R^2_{L3} \leq 1.00$; see Table B2).

The results of the sociodemographic covariates models indicated that these student characteristics were in general very powerful predictors at L2 and L3 across grade levels (median values: $R^2_{L2} = .55$ and $R^2_{L3} = .85$) but considerably less effective at L1 ($Mdn(R^2_{L1}) = .04$; see Table B2). Again, we found a wide range in the amount of variance explained by sociodemographic characteristics across outcome measures ($.00 \leq R^2_{L1} \leq .14$, $.16 \leq R^2_{L2} \leq .89$, $.14 \leq R^2_{L3} \leq .97$; see Table B2). Broken down by grade levels as mapped in Table 3 and Figure 2c, median values for R^2 at L1/L2 were greater in elementary than secondary school with .10/.61 and .03/.53, respectively. At L3 explained variances were lower in elementary ($Mdn(R^2_{L3}) = .63$) than in secondary school ($Mdn(R^2_{L3}) = .88$) instead.

Table 3. *Normative Distributions of Multilevel Design Parameters for Student Achievement: (a) Between-Classroom (ρ_{L2}) and Between-School Differences (ρ_{L3}), and Explained Variances by (b) Pretest Scores, (c) Sociodemographic Characteristics, and (d) Pretest Scores and Sociodemographic Characteristics at the Student (R^2_{L1}), Classroom (R^2_{L2}), and School Level (R^2_{L3})*

Statistic	a. Model Set 1		b. Model Set 2			c. Model Set 3			d. Model Set 4		
	Intercept-Only Model		Pretest Covariate(s) Model			Sociodemographic Covariates Model			Pretest and Sociodemographic Covariates Model		
	ρ_{L2}	ρ_{L3}	R^2_{L1}	R^2_{L2}	R^2_{L3}	R^2_{L1}	R^2_{L2}	R^2_{L3}	R^2_{L1}	R^2_{L2}	R^2_{L3}
Elementary School (Grades 1-4)											
Minimum	.03	.04	.01	.00	.06	.00	.16	.14	.02	.27	.27
25th percentile	.04	.10	.09	.11	.33	.04	.38	.51	.14	.46	.66
Median	.05	.11	.24	.27	.39	.10	.61	.63	.30	.67	.77
75th percentile	.06	.15	.36	.37	.59	.10	.66	.69	.38	.82	.82
Maximum	.09	.22	.51	.85	.90	.14	.84	.92	.52	.89	.92
Secondary School (Grades 5-12)											
General Student Population											
Minimum	.01	.09	.08	.09	.00	.00	.16	.16	.08	.31	.61
25th percentile	.03	.27	.14	.45	.87	.01	.35	.80	.17	.69	.95
Median	.04	.35	.20	.65	.96	.03	.53	.88	.22	.77	.98
75th percentile	.06	.39	.30	.75	.98	.05	.66	.91	.31	.86	.99
Maximum	.13	.59	.56	.95	1.00	.10	.89	.97	.57	.97	1.00
General Student Population with Adjustment for Mean-Level Differences Between School Types											
Minimum	.02	.03	.08	.06	.01	.00	.15	.24	.08	.27	.51
25th percentile	.04	.08	.14	.41	.72	.02	.34	.54	.17	.70	.84
Median	.06	.10	.21	.63	.81	.03	.46	.70	.22	.77	.90
75th percentile	.09	.12	.30	.75	.91	.05	.64	.80	.31	.86	.97
Maximum	.18	.22	.56	.94	.99	.10	.90	.96	.57	.97	1.00
Academic Track											
Minimum	.01	.01	.07	.05	.01	.00	.07	.27	.08	.19	.64
25th percentile	.04	.04	.11	.51	.52	.02	.44	.46	.15	.72	.82
Median	.05	.06	.20	.61	.68	.03	.64	.66	.22	.85	.89
75th percentile	.09	.09	.30	.82	.84	.05	.75	.81	.32	.92	.95
Maximum	.21	.23	.54	.94	.97	.10	.92	.94	.55	.98	.98
Non-Academic Track											
Minimum	.01	.07	.07	.07	.02	.00	.11	.14	.07	.52	.45
25th percentile	.04	.16	.16	.41	.81	.02	.42	.66	.18	.74	.91
Median	.06	.20	.20	.65	.88	.03	.53	.79	.24	.84	.96
75th percentile	.09	.23	.33	.80	.98	.06	.67	.89	.34	.91	.99
Maximum	.15	.36	.59	.94	1.00	.21	.90	.99	.61	.97	1.00

Note. Statistics were calculated across achievement domains and are based on the estimates obtained from three-level models (students at L1 within classrooms at L2 within schools at L3) for grades 1 to 10 and two-level models (students at L1 within schools at L3) for grades 11 to 12 because 11th and 12th grade students did not attend intact classrooms, but rather the grouping of students varied depending on the subject taught. This means that statistics for estimates at the classroom level (i.e., ρ_{L2}, R^2_{L2}) were calculated for grades 1 to 10 only. Statistics were calculated excluding meta-analytically pooled results of grade 9. The complete collection of normative distributions is available in Tables B2, B4, B6, B8, B10, B12, B14 and B16 in the Supplemental Online Material B on the OSF (https://osf.io/2w8nt).

Figure 2. *Multilevel Design Parameters for Student Achievement for the General Student Population Without and With Adjustment for Mean-Level Differences Between School Types: (a) Between-Classroom (ρ_{L2}) and Between-School Differences (ρ_{L3}), and Explained Variances by (b) Pretest Scores, (c) Sociodemographic Characteristics, and (d) Pretest Scores and Sociodemographic Characteristics at the Student (R^2_{L1}), Classroom (R^2_{L2}), and School Level (R^2_{L3})*

Note. Boxplots show distributions across all achievement domains. For grades 1 to 10, design parameters are based on three-level models (students at L1 within classrooms at L2 within schools at L3). For grades 11 to 12, design parameters are based on two-level models (students at L1 within schools at L3) as 11th and 12th grade students did not attend intact classrooms, but rather the grouping of students varied depending on the subject taught. This means that design parameters at the classroom level (i.e., ρ_{L2}, R^2_{L2}) were estimated for grades 1 to 10 only. In Figure 2a, intraclass correlations ρ were estimated in intercept-only models (model set 1). In Figure 2b, explained variances R^2 by pretests were estimated in pretest covariate(s) models (model set 2). In Figure 2c, explained variances R^2 by sociodemographics were estimated in sociodemographic covariates models (model set 3). In Figure 2d, explained variances R^2 by pretests and sociodemographics were estimated in pretest and sociodemographic covariates models (model set 4). To estimate design parameters that were adjusted for mean-level achievement differences between school types offered in German secondary education (L3 adjusted), dummy-coded indicator variables representing the various school types were added as additional covariates at L3. The complete collection of design parameters is available in Tables B1, B3, B5, B7, B9, B11, B13 and B15 in the Supplemental Online Material B on the OSF (https://osf.io/2w8nt).

As evident from Figure 2d, the results of the pretest and sociodemographic covariates models suggested that pretests and sociodemographics may explain incremental amounts of variance in students' achievement over and above each other (see also Table 3): In secondary school, this was most noticeable at L2, where we observed a significant increase in the median value for R^2 of .12 relative to the pretest covariate(s) models. In elementary school, the respective gains were even larger at both L2 ($\Delta Mdn(R^2_{L2}) = .40$) and L3 ($\Delta Mdn(R^2_{L3}) = .38$). Averaged across grade levels, pretests plus sociodemographics could explain about 23% of the variance at L1, 75% at L2, and 95% at L3 (see Table B2).

Design Parameters with Adjustment for Mean-Level Achievement Differences between School Types

When comparing the design parameters with and without adjustment for mean-level achievement differences between secondary school types, we observed several key results (see Table 3 and Figure 2). First, when adjusting for mean-level differences, intraclass correlations at L2 were slightly larger ($.02 \leq \rho_{L2} \leq .18$, $Mdn(\rho_{L2}) = .06$) whereas intraclass correlations at L3 were considerably smaller ($.03 \leq \rho_{L3} \leq .22$; $Mdn(\rho_{L3}) = .10$). Second, the results obtained for the adjusted pretest covariate(s) models showed that the explanatory power of pretests remain roughly the same at L1 and L2 with median R^2 values of .21 and .63, respectively, but that it was decreased at L3 ($Mdn\,(R^2_{L3}) = .81$). Third, the pattern of results from the adjusted sociodemographic covariates models largely mirrored the results of the unadjusted pretest covariate(s) models. Fourth, in the adjusted pretest and sociodemographic covariates models, median amounts of explained variance remained unchanged at L1/L2 (22%/77%), but were slightly decreased at L3 (90%).

Design Parameters for the Academic and Non-Academic Track

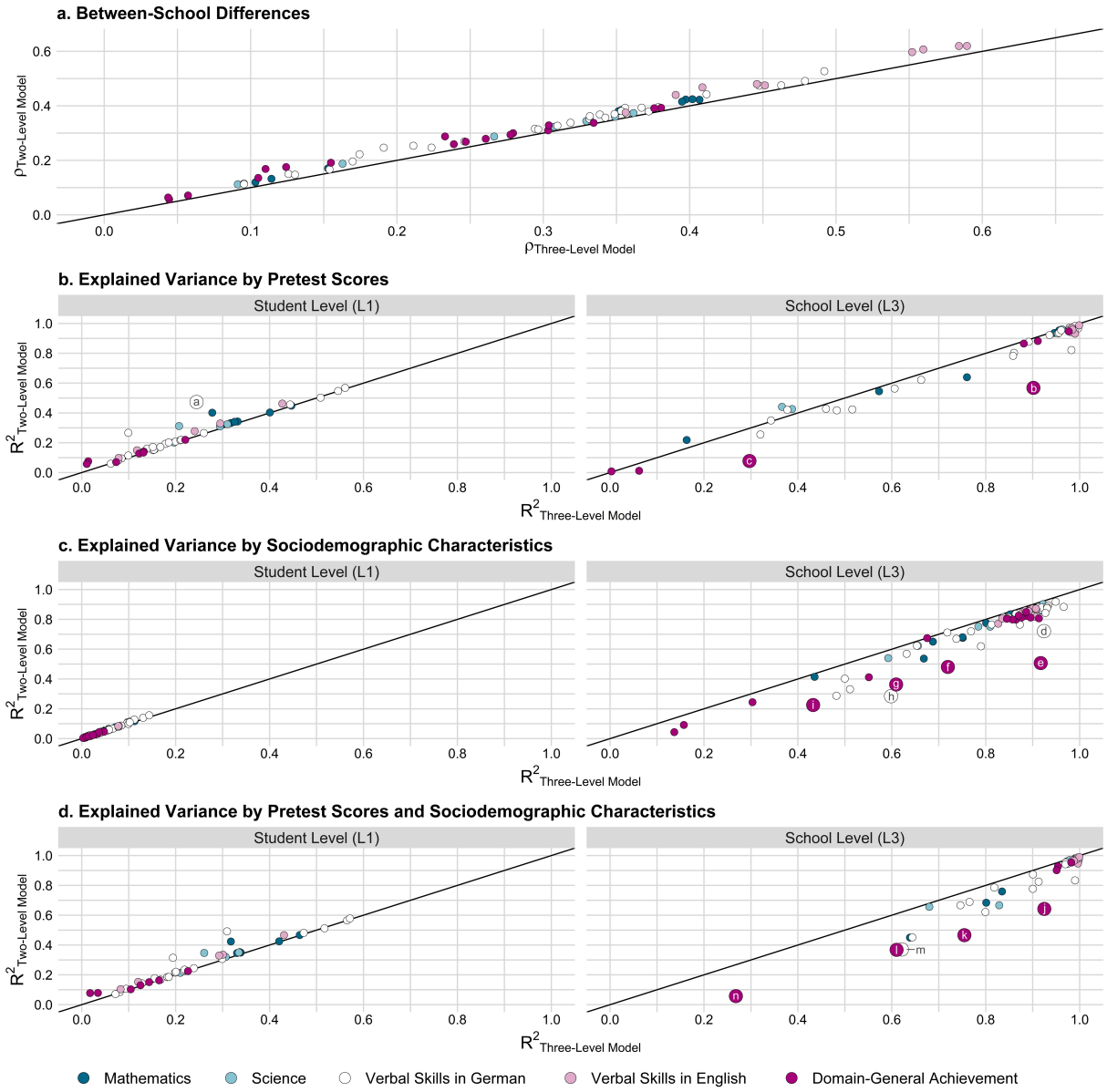
The following major findings emerged from the analyses performed separately for the academic track and the non-academic track (see Table 3). First, the results of the intercept-only models showed that between-classroom differences in students' achievement for the academic ($.01 \leq \rho_{L2} \leq .21$; $Mdn(\rho_{L2}) = .05$) and non-academic track ($.01 \leq \rho_{L2} \leq .15$; $Mdn(\rho_{L2}) = .06$) were very similar. However, median proportions of achievement differences located at L3 were found to be smaller in the academic than non-academic track, with 6% (ranging between $.01 \leq \rho_{L3} \leq .23$) and 20% (ranging between $.07 \leq \rho_{L3} \leq .36$), respectively. Second, the results of the pretest covariate models demonstrated that pretests explained on average about the same

amount of variance at L1/L2 in both tracks (20%/approximately 63%). The amount of variance explained at L3, however, was smaller in the academic track ($Mdn(R^2_{L3}) = .68$) than the non-academic track ($Mdn(R^2_{L3}) = .88$). Third, the pattern of results from the sociodemographic covariates models mirrored those obtained from the pretest covariate(s) models. Fourth, the results obtained from the pretest and sociodemographic covariates models revealed that the amount of incremental variance explained by either the pretests or sociodemographics differed only marginally between the academic and non-academic track at all levels.

Design Parameters for Two-Level versus Three-Level Designs

We additionally studied design parameters and standard errors for student achievement assuming only a two-level structure (i.e., students within schools) for grades 1 to 10 to simulate situations where no classroom-level information is available. Concerning the (unadjusted) results obtained for the general student population, values for ρ_{L3} and R^2_{L1} are highly similar between two- and three-level designs, as clearly seen in Figure 3, indicating that information at L2 barely affects the design parameters. On the other hand, applying two-level instead of three-level models underestimated the values for R^2_{L3} in several cases, sometimes considerably. Similar patterns of results were observed when these comparisons were performed for the adjusted and track-specific design parameters (see Figures A1 to A3 in the Supplemental Online Material A).

Figure 3. *How Much Bias May Result in Design Parameters for Student Achievement for the General Student Population at the Student (L1) and School Level (L3) When the Classroom Level (L2) Is Ignored? Comparison of Corresponding Design Parameters Obtained From Three-Level Models versus Two-Level Models: (a) Between-School Differences (ρ_{L3}), and Variances Explained by (b) Pretest Scores, (c) Sociodemographic Characteristics, and (d) Pretest Scores and Sociodemographic Characteristics at the Student (R^2_{L1}) and School Level (R^2_{L3})*



Note. The graph juxtaposes corresponding design parameters estimated by three-level models (x-coordinate; students at L1 within classrooms at L2 within schools at L3) with design parameters estimated by two-level models (y-coordinate; students at L1 within schools at L3). The black line marks congruence of three- and two-level design parameters. Larger labeled dots exceed a deviation of ± .20 between three- and two-level design parameters. For example, in Figure 3b, left grid ("Student Level (L1)"), the dot labeled with "a" (representing German vocabulary in grade 1) shows that R^2_{L1} was .24 when specifying a three-level pretest covariate model, whereas R^2_{L1} was .47 when specifying a two-level pretest covariate model.

[a] Vocabulary (NEPS-SC2, grade 1). [b] Declarative metacognition (NEPS-SC2, grade 3). [c] Basic cognitive functions: Reasoning (NEPS-SC2, grade 2). [d] Reading speed (DESI, grade 9, wave 2). [e] Declarative metacognition (NEPS-SC2, grade 1). [f] Declarative metacognition (NEPS-SC2, grade 3). [g] Basic cognitive functions: Perception speed (NEPS-SC3, grade 9). [h] Reading speed (NEPS-SC2, grade 2). [i] Basic cognitive functions: Perception speed (NEPS-SC3, grade 5). [j] Declarative metacognition (NEPS-SC2, grade 3). [k] Basic cognitive functions: Reasoning (NEPS-SC2, grade 2). [l] Basic cognitive functions: Perception speed (NEPS-SC3, grade 9). [m] Reading speed (NEPS-SC2, grade 2). [n] Basic cognitive functions: Perception speed (NEPS-SC2, grade 2).

Applications

This section discusses three research scenarios to illustrate how the design parameters and their standard errors that we provided in this paper can be used in power analyses to plan CRTs (and MSCRTs) on student achievement. Figure 4 can help researchers select an appropriate set of design parameters as a function of key characteristics of the planned intervention. For each scenario, we assumed that classrooms or schools would be randomly assigned to the experimental conditions in equal shares (i.e., 50% of the target [sub]clusters obtain the educational treatment, and the remaining 50% represent the control group; $P = .50$). Further, we assume a two-tailed test with a significance level of $\alpha = .05$ and set the desired power at 80% ($1 - \beta = .80$). A constitutive step when planning CRTs is to define a reasonable value for the $MDES$; this decision can take into account political, economic, and programmatic perspectives or a combination thereof (see Bloom, 2006; Brunner et al., 2017; Schochet, 2008, for thorough discussions). We used the package PowerUpR (Bulus et al., 2019) in R (R Core Team, 2018) for the calculations.

Scenario 1: How Many Schools Are Required for a CRT?

Research Team 1 would like to conduct a three-level CRT on the effectiveness of a school-wide intervention to improve 4th graders mathematical achievement. Team 1 plans to sample $J = 3$ classrooms with $n = 20$ students per classroom from every school. The researchers are interested in K, the number of schools necessary to detect a typical intervention effect on student achievement. According to the research synthesis by Hill and colleagues (2008), the mean standardized effect size for intervention effects on student achievement ranges between $.20 \leq \delta \leq .30$ across domains and grade levels. Thus, Team 1 choses a target intervention effect size of $\delta = .25$. After consulting Figure 4, Team 1 chooses Table B1 containing the appropriate estimates of design parameters for their study. According to this table, the intraclass correlations at L2 and L3 for mathematics in grade 4 were $\rho_{L2} = .05$ and $\rho_{L3} = .10$, respectively. As recommended in Hedges et al. (2012) and Jacob et al. (2010), the researchers want to take into account the statistical uncertainty (due to sampling error) associated with these point estimates. Team 1 therefore determines the lower and upper bound estimates for K by computing the 95% confidence interval of ρ_{L2} and ρ_{L3} using their standard errors of $SE(\rho_{L2}) = SE(\rho_{L3}) = .02$ (see Table B1). The lower bound of the 95% confidence interval of ρ_{L2} is thereby computed as $.05 - 1.96 * .02 = .01$ and the upper bound as $.05 + 1.96 * .02 = .09$. Analogously, the 95% confidence interval of ρ_{L3} equals 95% CI [.06, .14]. When using these values for the power

calculations, Team 1 needs $K = 42$ schools for the lower bound estimates, $K = 68$ schools for the point estimates, and $K = 94$ schools for the upper bound estimates of ρ.

In order to improve statistical precision, Team 1 plans to assess pretest scores and to use them as covariates. As listed in Table B1, the explained variances and corresponding standard errors for a mathematics pretest were $R^2_{\mathrm{L1}} = .40$ $(SE = .01)$, $R^2_{\mathrm{L2}} = .35$ $(SE = .04)$, and $R^2_{\mathrm{L3}} = .76$ $(SE = .03)$. These values yield a lower bound estimate for R^2_{L1} of $.40 - 1.96 * .01 = .38$ and an upper bound estimate for R^2_{L1} of $.40 + 1.96 * .01 = .42$. Likewise, the 95% confidence intervals of R^2_{L2} and R^2_{L3} are 95% CI [.27, .43] and 95% CI [.70, .82], respectively. Hence, when including a pretest and using the point estimates of ρ_{L2} and ρ_{L3}, the required number of schools is $K = 28$ for the lower bound estimates, $K = 24$ for the point estimates, and $K = 20$ for the upper bound estimates of the R^2 values.

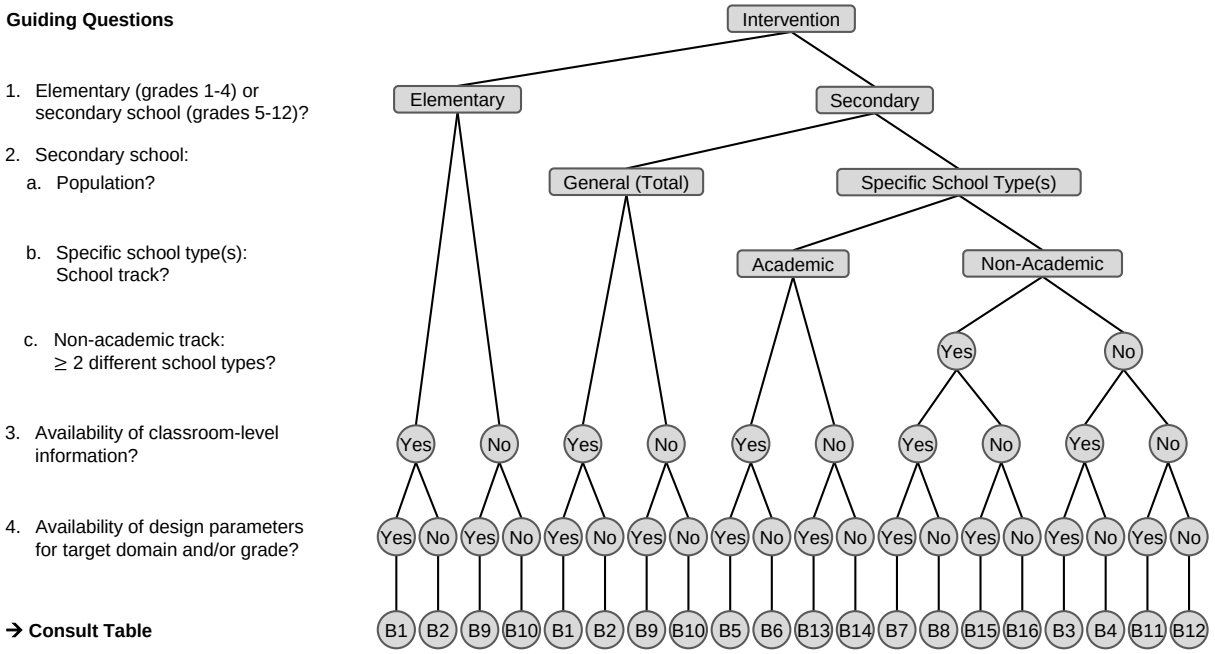
When opting for a conservative approach, Team 1 should use the upper bound estimates of ρ and the lower bound estimates of R^2 at each hierarchical level (i.e., $\rho_{\mathrm{L2}} = .09$, $\rho_{\mathrm{L3}} = .14$, $R^2_{\mathrm{L1}} = .38$; $R^2_{\mathrm{L2}} = .27$; $R^2_{\mathrm{L3}} = .70$), resulting in a required number of schools of $K = 38$. Of note, if Team 1 employed pretests as well as sociodemographic characteristics as covariates, the required number of schools would decrease significantly ($K = 26$, when using the upper bound estimates of ρ and lower bound estimates of R^2 at each level). In conclusion, Team 1 should carefully balance the cost of additionally assessing sociodemographics against the cost of sampling a larger number of schools to achieve an equal level of precision (see Schochet, 2008).

Scenario 2: Which $MDES$ is Attainable for a CRT?

Suppose that research Team 2 plans a three-level CRT to study the impact of an intervention that is intended to affect students' history achievement in comprehensive schools (grades 5 to 12). Due to budgetary constraints (see Spybrook, Shi, et al., 2016), a fixed maximum number of $K = 40$ schools (with $J = 2$ classrooms, and $n = 20$ students each) are at the researchers' disposal. Given these limits, the primary concern of Team 2 is to ensure that the attainable $MDES$ lies within the range of typical intervention effects on student achievement (i.e., $.20 \leq \delta \leq .30$; Hill et al., 2008). Team 2 consults Figure 4 to find the suitable table of design parameters. Since the intervention is targeted at a single, specific school type within the non-academic track, Team 2 uses the design parameters that are adjusted for mean-level differences between school types. Moreover, since design parameters for history are not available, Team 2 consults Table B4 outlining the normative distributions across the various achievement domains to determine small (i.e., 25th percentile [P25]), medium (i.e., median), and large values (i.e., 75th percentile [P75]) of the design parameters. Entering the respective values for the intraclass

correlations (P25: $\rho_{L2} = .04$, $\rho_{L3} = .08$; median: $\rho_{L2} = .06$, $\rho_{L3} = .10$; P75: $\rho_{L2} = .09$, $\rho_{L3} = .12$; see Table B4), Team 2 learns that the attainable $MDES$ is .32/.35/.39 for small/medium/large values of ρ_{L2} and ρ_{L3}. Including both pretests and sociodemographics as covariates (P25: $R^2_{L1} = .17$, $R^2_{L2} = .70$, $R^2_{L3} = .84$; median: $R^2_{L1} = .22$, $R^2_{L2} = .77$, $R^2_{L3} = .90$; P75: $R^2_{L1} = .31$ $R^2_{L2} = .86$, $R^2_{L3} = .97$; see Table B4) and using the 75th percentiles of the values for ρ_{L2} and ρ_{L3} (as more conservative upper bounds), the respective values for the $MDES$ reduce to .20/.18/.14 for small/medium/large values of R^2 at the various levels. Consequently, Team 2 can be quite confident that their CRT design offers sufficient sensitivity to detect a true intervention effect within the desired range when including both pretests and sociodemographics.

Figure 4. *Flow Chart to Guide the Choice of Design Parameters as a Function of Key Characteristics of the Target Intervention*



Note. Tables B1 to B16 can be retrieved from Supplemental Online Material B. A comprehensive overview of the achievement measures analyzed in the present study is given in Table A5 in the Supplemental Online Material A. The Supplemental Online Materials are available on the OSF (https://osf.io/2w8nt).

Scenario 3: How Many Schools Are Required for a MSCRT?

Research Team 3 would like to study the effects of a new teaching method involving learning software developed to enhance grade 9 students' English listening comprehension skills in the academic track. Due to practical constraints (e.g., limited availability of computers in the schools), classrooms within schools (serving as sites or blocks) are randomly assigned to

experimental conditions, making this design a three-level MSCRT. Since most academic track schools have at least four 9th grade classrooms of at least 20 students each, Team 3 plans to have $J = 4$ and $n = 20$. Team 3 considers an intervention effect of $\delta = .10$ policy-relevant (see Bloom, 2006; Bloom et al., 2007; Brunner et al., 2017; Schochet, 2008). Since the goal of Team 3 is to generalize the study findings to the population of German academic track schools beyond those sampled for their MSCRT, they treat the school effects as random (Bloom et al., 2017; Bloom & Spybrook, 2017; Spybrook & Raudenbush, 2009). Recall, this requires a reasonable assumption on the estimate of the cross-site effect size variability $\tau^2_{\delta_{L3}}$. According to Weiss et al. (2017), the values for the standard deviations of standardized intervention effects across schools often range between $.10 \leq \tau_{\delta_{L3}} \leq .25$. Since schools in the academic track form a comparatively homogeneous sample, Team 3 assumes that $\tau_{\delta_{L3}} = .15$. Team 3 consults Figure 4 and chooses Table B5 for the appropriate design parameters. Team 3 draws on the estimates that were meta-analytically pooled across 9th grade samples, with $\rho_{L2} = .19$ and $\rho_{L3} = .07$ (see Table B5). Under these conditions and in the absence of covariates, $K = 198$ schools are necessary to detect an intervention effect of $\delta = .10$, if it exists. In order to raise statistical precision, Team 3 intends to assess vital sociodemographics. The researchers enter the meta-analytically pooled R^2 values at L1 and L2 given for the sociodemographic covariates models in the power calculations ($R^2_{L1} = .01$, $R^2_{L2} = .72$; see Table B5). A particular challenge is to define the amount of variance in $\tau^2_{\delta_{L3}}$ that can be explained by L3 covariates because empirical guidance on values for $R^2_{\delta_{L3}}$ is scarce. According to Schochet et al. (2014) as well as Weiss et al. (2014), site-level covariates may explain a substantial proportion of $\tau^2_{\delta_{L3}}$. Nevertheless, as can be derived from Equation (8), when $\tau^2_{\delta_{L3}}$ and ρ_{L3} are rather small, $R^2_{\delta_{L3}}$ has a negligible effect on statistical power and precision, and thus, on the required number of schools. Opting for a conservative approach, Team 3 assumes that sociodemographics will explain considerably less variability in the intervention effect across schools compared to between-school differences in achievement (i.e., 1/10). Following this rationale, Team 3 estimates $R^2_{\delta_{L3}} = R^2_{L3} * 0.10 = .87 * 0.10 = .09$. Using sociodemographics as covariates at all levels decreases the required number of schools markedly to $K = 89$. Team 3 should therefore sample at least $K = 89$ schools (with $J = 4$ classrooms of $n = 20$ students each) and include vital sociodemographics in their study design in order to uncover a true intervention effect of $\delta = .10$ with confidence.

Discussion

CRTs on the effectiveness of large-scale educational interventions are valuable tools to inform evidence-based educational policies and practices (Institute of Education Sciences & National Science Foundation, 2013; Slavin, 2002; Spybrook, Shi, et al., 2016). When planning CRTs, educational researchers need reliable multilevel design parameters that match the target population, hierarchical level, and outcome domain to derive the number of students, classrooms, and schools needed to ensure sufficient statistical power to detect intervention effects. Capitalizing on data from three German longitudinal large-scale assessments, the present study provides three- and two-level design parameters (and respective standard errors) for student achievement across a very broad array of domains throughout the school career. This research expands the existing body of knowledge in three major dimensions.

(I) Expanding the Knowledge Base of Design Parameters to Germany

The large majority of previous research provided design parameters for the United States We added design parameters based on German samples of 1st to 12th graders to this knowledge base. We observed the following key results:

First, for the general student population, we found substantially larger (unadjusted) between-school differences in achievement than those typically reported for U.S. samples. In our study, the average value of ρ_{L3} lay around .31, whereas in the United States ρ_{L3} does not often exceed .25 (e.g., Bloom et al., 2007; Hedges & Hedberg, 2013; Spybrook, Westine, et al., 2016). This difference between schools in Germany and the United States corroborates the results of international studies pointing to a significant variation of ρ_{L3} across countries (Brunner et al., 2017; Kelcey et al., 2016; Zopluoglu, 2012). Looking at different grade levels, however, yields a more differentiated picture. As mentioned before, the German school system is characterized by an early tracking into different school types that cater to students with different performance levels. In elementary school, the discrepancy between the results from the United States (with $Mdn(\rho_{L3}) = .18$; see Figure 1a) and the present German samples (with $Mdn(\rho_{L3}) = .11$) was therefore considerably smaller than the discrepancy observed for secondary school. When German students were placed into different school types in secondary education, achievement differences at L3 were considerably smaller in the United States ($Mdn(\rho_{L3}) = .19$; see Figure 1a) than in Germany ($Mdn(\rho_{L3}) = .35$). This finding supports previous results from German large-scale studies indicating that values of ρ_{L3} are larger in secondary than in elementary school (see Table 1). However, when adjusting for mean-level

differences between school types or when conducting the analyses separately for schools in the academic or non-academic track, values of ρ_{L3} dropped considerably. This observation is well-aligned with past research showing that school types explain a vast proportion of achievement differences between schools in Germany (Baumert et al., 2003).

Second, we replicated and extended the well-documented finding that covariates are a powerful way to increase statistical power and precision of CRTs in educational research. Specifically, we confirmed the discovery that pretest scores are highly effective in explaining variance, especially at higher levels (Bloom et al., 2007; Hedges & Hedberg, 2007a; Spybrook, Westine, et al., 2016). Overall, pretests explained about 21% of the variance at L1 and 89% of the variance at L3. We also observed substantial variation in the amounts of variance explained by pretests. Very low values of R^2 might be partly due to the application of proxy pretests in some instances. In line with previous research (e.g., Bloom et al., 2007; Hedges & Hedberg, 2013; Westine et al., 2013), we also found that the explanatory power of sociodemographic characteristics was quite strong at L3, but relatively weak at L1: while sociodemographics on average explained 85% of between-school variability in students' achievement, the amount of variance explained at L1 was relatively low with an average of about 4%. Finally, sociodemographics contributed to the prediction of variance over and above pretests (and vice versa) at all levels. In our analyses, the combined covariate set, however, was markedly less effective at L1 than in studies in the United States, but more effective at L3. Divergences in the composition of variance components might explain this observation: achievement differences at L1 are more pronounced in the United States than in Germany, leading to a better signal-to-noise ratio at L1 for U.S. samples, whereas the reverse pattern was found at L3, resulting in a better signal-to-noise ratio in Germany than in the United States (Raudenbush et al., 2007).

(II) Providing Three-Level Design Parameters and Standard Errors for the Student, Classroom, and School Level

Previous research has established a wealth of two-level design parameters (i.e., students within schools). Yet, little was known about classroom-level estimates within schools. Further, the statistical uncertainty associated with the design parameters (particularly those at L2) was rarely reported – although it is a decisive piece of information when conducting power analyses (Hedges et al., 2012; Jacob et al., 2010). To address these gaps, we fitted multilevel latent (covariate) models (Lüdtke et al., 2008) with three levels (i.e., students within classrooms within schools) whenever students were in intact classroom settings (i.e., for grades 1 to 10) and estimated standard errors for all design parameters. We observed the following key results:

First, in line with previous research from Germany (see Table 1), between-classroom differences in students' achievement were substantially smaller in size than between-school differences. In total, values for ρ_{L2} were around .05 and usually smaller than .13. These values appeared relatively stable across grade levels, but varied by domain to a certain degree. Overall, our results suggested markedly lower achievement differences at L2 than in the United States, especially in secondary school (elementary school in the United States: $Mdn(\rho_{L2}) = .07$, $\rho_{L2} \leq .14$; secondary school in the United States: $Mdn(\rho_{L2}) = .30$, $\rho_{L2} \leq .45$; Jacob et al., 2010; Xu & Nichols, 2010; Zhu et al., 2012).

Second, the explanatory power of both pretests and sociodemographics at L2 strongly varied as a function of achievement domain and grade level. Values for R^2_{L2} ranged from .00 to .95 for pretests, from .16 to .89 for sociodemographics, and from .27 to .97 when combining both covariate sets. Sociodemographics consistently contributed incremental amounts of variance to the prediction of students' achievement over and above pretests (and vice versa), in particular at L2. These results align with those presented in Jacob and colleagues (2010). Therefore, depending on the level of treatment assignment, collecting data on sociodemographics in addition to measuring baseline achievement appears to be a sound strategy to improve the precision of CRTs. Notably, the wide range observed for R^2_{L2} and the corresponding standard errors may be attributable to estimation error caused by the very small size of certain variance components at L2 (Jacob et al., 2010, p. 177).

Third, we specified two-level models to assess the degree of bias when omitting information on the classroom-level cluster variance structure. In line with existing research addressing this question (Xu & Nichols, 2010; Zhu et al., 2012) we found negligible deviations between the intraclass correlations as estimated based on three-level versus two-level designs. Some values for R^2_{L3} were markedly higher in three-level than two-level models. As Xu and Nichols (2010, p. 28-29) described, the degree of bias should hinge on the degree of clustering in the outcome at L2: if there is substantial between-classroom variability, the omission of L2 can lead to severely biased design parameters at L1 and/or L3, and thus to erroneous results in power analyses. Our findings suggest that students' achievement varied only to a small degree at L2 for most outcome measures. Thus, the present results suggest that ignoring the classroom-level variance and using two-level instead of three-level design parameters is unlikely to produce biased estimates from power analyses for the German school context, at least regarding intraclass correlations. Nevertheless, we recommend educational researchers to use three-level design parameters for sample size calculations whenever these parameters are available in order to obtain the most accurate results in power analysis for CRTs.

Fourth, capitalizing on data from three large-scale studies allowed us to achieve a satisfactory to high level of precision when estimating design parameters (in terms of small standard errors). A major exception was found in the large standard errors of the estimates for R^2_{L2} and R^2_{L3}, primarily in grade 1, obtained from the pretest covariate(s) models involving pretests assessed in Kindergarten. The high percentage of missing values (over 90%) in these measures induced significant variation across the imputed datasets (i.e., between-imputation variance) resulting in large standard errors. When planning CRTs, we therefore recommend that researchers apply the provided values in their power analyses with caution (e.g., using conservative strategies as illustrated in the applications), or use both pretests and sociodemographics as covariates in grade 1 as we observed much smaller standard errors for design parameters in this case.

(III) Providing Design Parameters for a Very Broad Array of Achievement Domains

The bulk of previously presented design parameters were restricted to mathematics, science, and reading achievement. However, schools aim to foster a considerably broader spectrum of achievement domains. Thus, in addition to the core domains, we also estimated design parameters that have not previously been available, including specific verbal skills in student's first language (i.e., German) and foreign languages (i.e., English), and domain-general measures such as declarative metacognition, information and communication technology, problem solving, and basic cognitive functions. We observed the following key results:

First, the present findings corroborate those from previous research stressing that design parameters do not generalize well across achievement (sub)domains (e.g., Spybrook, Westine, et al., 2016; Westine et al., 2013; Xu & Nichols, 2010). Specifically, median values of between-classroom and between-school differences were typically lower for domain-general achievement ($\rho_{L2} = .04$, $\rho_{L3} = .24$) and science ($\rho_{L2} = .04$, $\rho_{L3} = .29$), and higher for verbal skills in English as foreign language ($\rho_{L2} = .07$, $\rho_{L3} = .45$) than for other domains (mathematics: $\rho_{L2} = .05$, $\rho_{L3} = .35$; verbal skills in German as first language: $\rho_{L2} = .05$, $\rho_{L3} = .33$).

Second, the present study showed that design parameters may even not generalize well across skills of the same domain. For instance, we examined German reading comprehension and German reading speed in grade 5: ρ_{L2} and ρ_{L3} were strikingly different from each other for these outcome measures (reading comprehension: $\rho_{L2} = .04$, and $\rho_{L3} = .32$; reading speed: $\rho_{L2} = .13$ and $\rho_{L3} = .19$).

Taken together, these findings underscore the importance of striving for the best fit between design parameters and target achievement measure when performing power analyses for CRTs because borrowing design parameters that do not match well can yield severely biased sample size requirements (Westine et al., 2013).

Limitations and Outlook

This study has several limitations. First, given the large international variability of design parameters detected in previous studies (e.g., Brunner et al., 2017; Zopluoglu, 2012), our findings are first and foremost applicable to the German school system. Notably, the school systems in Austria, Czech Republic, Hungary, Slovak Republic, and Turkey are also characterized by an early onset of school-level tracking after elementary school as in Germany (Salchegger, 2016). When design parameters are not available, intervention researchers conducting trials in such countries may apply the present design parameters in their power analyses because they are still better guesses than conventional benchmarks.

Second, we did not apply sampling weights. Hence, our results are representative only for those students selected for the present analyses. In general, the present design parameters are likely somewhat less accurate compared to those obtained from analyses using sampling weights. However, differences may be small as indicated in previous studies drawing on international large-scale assessment data (e.g., Wenger et al., 2018).

Third, the present design parameters were derived from national probability samples. Federal states within Germany as well as districts within federal states may vary in their mean achievement levels. The outcome measures analyzed in this paper contain some degree of variance that may be located at those higher levels. Thus, the reported values for between-school differences may be considered upper bound rather than lower bound estimates (see Hedges & Hedberg, 2007a, 2013).

Fourth, the present design parameters focus on student achievement as outcomes. Yet, apart from cognitive achievement, educational curricula worldwide identify a large range of further outcomes as key learning targets in school (World Economic Forum, 2015), such as socio-emotional skills (e.g., skills needed for task performance, to cooperate with others, or to regulate emotions; Organisation for Economic Co-operation and Development, 2017). Future research should therefore also supply design parameters for these skills (see e.g., Brunner et al., 2017).

Fifth, the present design parameters go well with outcome measures that are identical or highly similar to the measures that were used in NEPS, DESI, or PISA-I+. Researchers

should be cautious when relying on the present design parameters for planning CRTs with outcome measures that differ substantially from those used in the present study (see Brunner et al., 2017).

Finally, we provided standard errors for design parameters that quantify the statistical uncertainty associated with these estimates due to sampling error. Importantly, variability in research contexts (e.g., student populations, outcome measures) may further increase statistical uncertainty. When planning CRTs for research designs that are not covered in our study (e.g., for modestly dissimilar student populations and outcome measures), we recommend using our compilation of normative distributions of design parameters (Tables B2, B4, B6, B8, B10, B12, B14, and B16). In general, little is still known about the factors that affect the value of design parameters (e.g., why certain R^2_{L3} values equal 1.00; see Figures 1 and 2). An important next step for future research is therefore to conduct meta-analyses that quantify variability in design parameters across research contexts and examine moderator variables (e.g., outcome domain, onset of school type tracking, time lag between pre- and posttest, reliability of measures) that might explain this variability.

Conclusion and Recommendations

Capitalizing on representative data from three German longitudinal large-scale assessments, our study provides reliable three- and two-level design parameters with standard errors for a broad spectrum of achievement domains across the school career. Design parameters varied considerably as a function of the hierarchical level, achievement outcome, and grade level. Importantly, our analyses show that pretest and sociodemographic covariates improve the precision of educational CRTs at the student, classroom, and school level over and above each other. The present design parameters and their standard errors are therefore fundamental when planning CRTs in the German or similar school systems. Specifically, researchers may benefit from consulting Figure 4 to select the set of design parameters that offers the best fit to the planned educational intervention (e.g., in terms of population, domain, grade level) so CRTs can be adequately powered to generate high-quality evidence of what actually works to foster student achievement in Germany and elsewhere.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*(6), 33–40. https://doi.org/10.3102/0013189X035006033

American Psychological Association (Ed.). (2019). *Publication manual of the American Psychological Association* (7th ed.). American Psychological Association.

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*, 51–65. https://doi.org/10.1007/s11618-011-0181-8

Baumert, J., Köller, O., Lehrke, M., & Brockmann, J. (2000). Anlage und Durchführung der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie zur Sekundarstufe II (TIMSS/III)—Technische Grundlagen [Design and implementation of the Third Trends in International Mathematics and Science Study (TIMSS/III)—Technical information]. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Vol. 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichschulzeit* (pp. 31–84). Leske+Budrich.

Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten—Institutionelle Bedingungen des Lehrens und Lernens [School contexts—Institutional conditions for teaching and learning]. In Deutsches PISA-Konsortium (Ed.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 261–331). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-97590-4_11

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.

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., Bos, J. M., & Lee, S.-W. (1999). Using cluster random assignment to measure program impacts. Statistical implications for the evaluation of education programs. *Evaluation Review, 23*(4), 445–469. https://doi.org/10.1177/0193841X9902300405

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817–842. https://doi.org/10.1080/19345747.2016.1264518

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., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness*, *10*(4), 877–902. https://doi.org/10.1080/19345747.2016.1271069

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://files.eric.ed.gov/fulltext/ED502531.pdf

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Education as a lifelong process: The German National Educational Panel Study (NEPS)*. VS-Verl.

Böhme, K., & Weirich, S. (2012). Der Ländervergleich im Fach Deutsch [National Assessment Study in German]. In P. Stanat, H. A. Pant, K. Böhme, & D. Richter (Eds.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (pp. 103–116). Waxmann.

Boruch, R. F., & Foley, E. (2000). The honestly experimental society: Sites and other entities as the units of allocation and analysis in randomized trials. In L. Bickman (Ed.), *Validity and social experimentation: Donald campbell's legacy* (pp. 193–239). Sage Publications.

Brandon, P. R., Harrison, G. M., & Lawton, B. E. (2013). SAS code for calculating intraclass correlation coefficients and effect size benchmarks for site-randomized education experiments. *American Journal of Evaluation*, *34*(1), 85–90. https://doi.org/10.1177/1098214012466453

Brunner, M., Keller, U., Wenger, M., Fischbach, A., & Lüdtke, O. (2017). 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

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2019). *PowerUpR: Power analysis tools for multilevel randomized experiments. R package version 1.0.4*. https://CRAN.R-project.org/package=PowerUpR

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.
http://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf

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

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

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., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. Wiley & Sons.

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

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

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

Grund, S., Robitzsch, A., & Lüdtke, O. (2019). *Mitml: Tools for multiple imputation in multilevel modeling. R package version 0.3-7*. https://cran.r-project.org/web/packages/mitml/mitml.pdf

Haag, N., & Roppelt, A. (2012). Der Ländervergleich im Fach Mathematik [National Assessment Study in Mathematics]. In P. Stanat, H. A. Pant, K. Böhme, & D. Richter (Eds.), *Kompetenzen von Schülerinnen und Schülern am Ender der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (pp. 117–127). Waxmann. https://www.iqb.hu-berlin.de/bt/LV2011/Bericht

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Hedberg, E. C., Santana, R., & Hedges, L. V. (2004). *The variance structure of academic achievement in America*. Annual meeting of the American Educational Research Association, San Diego, CA.

Hedges, L. V., & Hedberg, E. C. (2007a). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hedges, L. V., & Hedberg, E. C. (2007b). Intraclass correlations for planning group randomized experiments in rural education. *Journal of Research in Rural Education*, *22*(10). http://jrre.vmhost.psu.edu/wp-content/uploads/2014/02/22-10.pdf

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

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., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Institute of Education Sciences, & National Science Foundation. (2013). *Common guidelines for education research and development*. https://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf

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

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

Knigge, M., & Köller, O. (2010). Effekte der sozialen Zusammensetzung der Schülerschaft [Impact of the social classroom composition of schools]. In O. Köller, M. Knigge, & B. Tesch, *Sprachliche Kompetenzen im Ländervergleich* (pp. 227–244). Waxmann.

Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1*(1), 66–88. https://doi.org/10.1080/19345740701692522

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness, 1*(4), 265–288. https://doi.org/10.1080/19345740802328216

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

Kultusministerkonferenz. (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lehmann, R., & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien [ELEMENT: Study of reading and mathematics literacy. Development from grades 4 to 6 in Berlin. Final research report on the 2003, 2004, and 2005 assessments at primary schools and undergraduate academic tracks in Berlin]*. Humboldt-Universität zu Berlin. https://www.researchgate.net/profile/Jenny_Lenkeit/publication/273380369_ELEMENT_Erhebung_zum_Lese-_und_Mathematik-verstandnis_-_Entwicklungen_in_den_Jahrgangsstufen_4_bis_6_in_Berlin_Abschlussbericht_uber_die_Untersuchungen_2003_2004_und_2005_an_Berliner_Grundschulen_und_/links/553f61600cf23e796bfb38c2.pdf?origin=publication_detail

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research. http://eric.ed.gov/?id=ED537446

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

Martin, M. O., Foy, P., Mullis, I. V. S., & O'Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at the fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—Implications for early learning.* (pp. 109–179). TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timsspirls2011/downloads/TP11_Relationship_Report.pdf

Martin, M. O., Mullis, I. V. S., Gregory, K., D., Hoyle, C., & Shen, C. (2000). *Effective schools in science and mathematics: IEA's Third International Mathematics and Science Study*.

International Study Center, Boston College. https://timssandpirls.bc.edu/timss1995i/TIMSSPDF/T95_EffSchool.pdf

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

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

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

Organisation for Economic Co-operation and Development. (2007). *Evidence in education: Linking research and policy*. OECD Publishing. https://doi.org/10.1787/9789264033672-en

Organisation for Economic Co-operation and Development. (2017). *Social and Emotional Skills. Well-being, connectedness and success*. OECD Publishing. http://www.oecd.org/education/school/UPDATED%20Social%20and%20Emotional%20Skills%20-%20Well-being,%20connectedness%20and%20success.pdf%20(website).pdf

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

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., 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.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213. https://doi.org/10.1037/1082-989X.5.2.199

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

Robitzsch, A., Grund, S., & Henke, T. (2018). *miceadds: Some additional multiple imputation functions, especially for mice. R package version 2.15-6*. https://CRAN.R-project.org/package=miceadds

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

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

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

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (p. 50). Institute of Education Sciences (IES). https://ies.ed.gov/ncee/pubs/20144017/pdf/20144017.pdf

Senkbeil, M. (2006). Die Bedeutung schulischer Faktoren für die Kompetenzentwicklung in Mathematik und in den Naturwissenschaften [The relevance of school context factors

for competence development in mathematics and science]. In PISA-Konsortium Deutschland (Ed.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (pp. 277–308). Waxmann.

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

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15–21. https://doi.org/10.3102/0013189X031007015

Spybrook, J. (2013). Introduction to special issue on design parameters for cluster randomized trials in education. *Evaluation Review*, *37*(6), 435–444. https://doi.org/10.1177/0193841X14527758

Spybrook, J., & Kelcey, B. (2016). Introduction to three special issues on design parameter values for planning cluster randomized trials in the social sciences. *Evaluation Review*, *40*(6), 491–499. https://doi.org/10.1177/0193841X16685646

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the institute of education sciences. *Educational Evaluation and Policy Analysis*, *31*(3), 298–318. https://doi.org/10.3102/0162373709339524

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

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *2*(1), 15. https://doi.org/10.1177/2332858415625975

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

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808. https://doi.org/10.1002/pam.21760

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876. https://doi.org/10.1080/19345747.2017.1300719

Wenger, M., Lüdtke, O., & Brunner, M. (2018). Übereinstimmung, Variabilität und Reliabilität von Schülerurteilen zur Unterrichtsqualität auf Schulebene: Ergebnisse aus 81 Ländern. *Zeitschrift für Erziehungswissenschaft*, *21*(5), 929–950. https://doi.org/10.1007/s11618-018-0813-3

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*(6), 490–519. https://doi.org/10.1177/0193841X14531584

World Economic Forum. (2015). *New Vision for Education. Unlocking the Potential of Technology*. http://www3.weforum.org/docs/WEFUSA_NewVisionforEducation_Report2015.pdf

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. http://www.urban.org/sites/default/files/alfresco/publication-pdfs/1001394-New-Estimates-of-Design-Parameters-for-Clustered-Randomization-Studies.PDF

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2012). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Educational Evaluation and Policy Analysis, 34*(1), 45–68. https://doi.org/10.3102/0162373711423786

Zopluoglu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(1), 233–270.

3

STUDY II

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

Stallasch, S. E., Lüdtke, O., Artelt, C., Hedges, L. V. & Brunner, M. (2023). Single- and Multilevel Perspectives on Covariate Selection in Randomized Intervention Studies on Student Achievement [Manuscript submitted for publication]. *Educational Psychology Review*.

This article has been posted as a preprint on PsyArXiv.org.
https://doi.org/10.31234/osf.io/5ajmg

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 representative German samples ($1{,}868 \leq N \leq 10{,}543$) 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; https://tinyurl.com/osf-blinded).

Keywords: covariate selection, design parameters, individual participant data meta-analysis, individually and cluster-randomized trials, power analysis, student achievement

Single- and Multilevel Perspectives on Covariate Selection in Randomized Intervention Studies on 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 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. 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 4,080 students (68 schools, each with 3 classrooms of 20 students) to detect an effect of $d = .25$ on 4th graders' mathematics achievement ($\alpha = .05$, $1 - \beta = .80$). 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).[34] In the

[34] 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).

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 1,440 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.[35] 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, pp. 494–495; Murray, 1998, pp. 137–140). 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).

The overall aim of this two-part study is to offer thorough empirical guidance on covariate selection to optimize design sensitivity in IRTs and CRTs on student achievement. In Part I, we estimate and meta-analytically integrate single- and multilevel design parameters for a broad array of outcomes in Grades 1–12 by capitalizing on large-scale assessment data from multiple German samples. In doing so, we quantify impacts of varying (a) covariate types (i.e., 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 (i.e., 1-7 years), drawing on the psychometric heuristics of bandwidth-fidelity (Cronbach & Gleser, 1957), incremental validity (Sechrest, 1963), and validity degradation (Ghiselli, 1956; Humphreys, 1960). In Part II, we use the empirically estimated design parameters in precision simulations to assess the actual returns of the covariates for the design sensitivity in IRTs and CRTs.

Statistical Underpinnings

Sufficient design sensitivity is a vital methodological quality criterion of rigorous research (American Psychological Association, 2020, pp. 83–84, 86; 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., $1 - \beta = .80$;

[35] Note that covariate adjustment might be worthless or even harmful in certain cases (Aberson, 2019, p. 136; Berk et al., 2013; Liu, 2011; Moerbeek & Teerenstra, 2016, p. 85; Raab & Butcher, 2005). Perks and perils of the method have been intensively and controversially discussed (see e.g., Moerbeek, 2006; J. Wang, 2020).

Cohen, 1988) to detect a true treatment effect.[36] 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):

$$MDES = M_{df} SE(\bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}})/\sigma_{\mathrm{T}} \tag{1}$$

M_{df}, reflects the t-distributions specific to α and $1 - \beta$, 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 $\alpha = .05$ and $1 - \beta = .80$ (Bloom, 2006, p. 5). The term $SE(\bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}})/\sigma_{\mathrm{T}}$ represents the treatment effect's $\bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}}$ 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, $MDES = .25$ 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).

As we show below, $SE(\bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}})/\sigma_{\mathrm{T}}$ is a function of three factors:[37] (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). Equation (1) then transforms to (Bloom, 2006, Equation 14; Dong & Maynard, 2013, p. 45):

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

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 $P_{\mathrm{T}} = .50$ (i.e., a balanced design with 50%/50% are randomly assigned to TG/CG) minimizes the $MDES$. The design parameter R_{T}^2 is

[36] 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.

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

of special interest in this study because it quantifies the amount of the total variance σ^2_{T} in Y that can be explained by covariates C_{T}:

$$R^2_{\mathrm{T}} = \left(\sigma^2_{\mathrm{T}} - \sigma^2_{\mathrm{T}|C_{\mathrm{T}}}\right)/\sigma^2_{\mathrm{T}} \tag{3}$$

$\sigma^2_{\mathrm{T}|C_{\mathrm{T}}}$ 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 within classrooms at L2 which are 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 classrooms or schools. The degree of within-cluster similarity is typically expressed by the multilevel design parameters ρ_{L2} and ρ_{L3} (i.e., the intraclass 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:

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

$$\rho_{\mathrm{L3}} = \sigma^2_{\mathrm{L3}}/\sigma^2_{\mathrm{T}}, \tag{5}$$

For a 2L-CRT, $\sigma^2_{\mathrm{T}} = \sigma^2_{\mathrm{L1}} + \sigma^2_{\mathrm{L3}}$, and for a 3L-CRT, $\sigma^2_{\mathrm{T}} = \sigma^2_{\mathrm{L1}} + \sigma^2_{\mathrm{L2}} + \sigma^2_{\mathrm{L3}}$, 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, Equation (1) then transforms to (Bloom, 2006, Equation 21; Dong & Maynard, 2013, p. 33):

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

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

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

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; $K = K_{\mathrm{TG}} + K_{\mathrm{CG}}$). 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., $P_{\mathrm{L3}} = K_{\mathrm{TG}}/K$)

with P_{L3} = .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:[38]

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

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

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

$\sigma^2_{\mathrm{L1}|\mathrm{C}_{\mathrm{L1}}}$, $\sigma^2_{\mathrm{L2}|\mathrm{C}_{\mathrm{L2}}}$, and $\sigma^2_{\mathrm{L3}|\mathrm{C}_{\mathrm{L3}}}$ signify the conditional between-student, -classroom, and -school variances, respectively. $df = K - Q_{\mathrm{L3}} - 2$, 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 should result in smaller $MDES$ values, and therefore, higher design sensitivity—provided adequate power and covariate-treatment orthogonality, usually holding for large samples (Moerbeek & Teerenstra, 2016, p. 83). Altogether, adjusting for highly prognostic covariates is a powerful way to raise RT design sensitivity.

Theoretical and Empirical Considerations on Covariate Selection

Well-founded decisions on the choice of covariates are key to designing strong RTs. Scholars and agencies agree that these decisions should be based on both substantive theory and empirical results (Committee for Proprietary Medicinal Products, 2004; Cook, 2005; EMA, 1998, 2015; Maxwell et al., 2017, pp. 494–495; Moerbeek & Teerenstra, 2016, pp. 84–87; Murray, 1998, pp. 137–140; Raab et al., 2000; Tafti & Shmueli, 2020; U.S. Food and Drug Administration, 2021; Wright et al., 2015). When the target outcome is student achievement—a multifaceted, complex construct influenced by numerous factors (Haertel et al., 1983; M. C. Wang et al., 1993; Winne & Nesbit, 2010)—several covariates are worth considering. 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

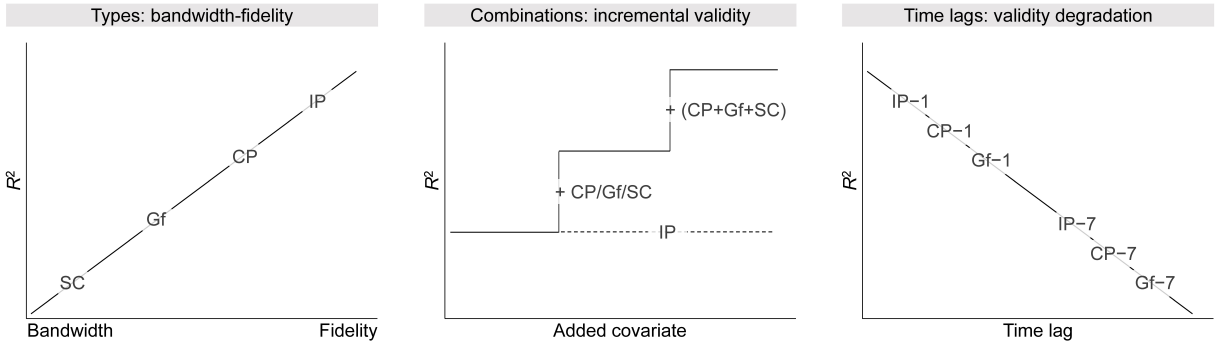
[38] C_{L1} and C_{L2} are group-mean centered (see e.g., Konstantopoulos, 2012). C_{L2} and C_{L3} may be either covariates directly assessed at L2 and L3 or classroom and school means of L1 covariates, respectively.

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 reading skills predicting future mathematics skills), 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, pp. 123–124) 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, p. 32), and often span several years (Connolly et al., 2018; Rickles et al., 2018). Moreover, apart from the fact that RTs should always be designed as parsimoniously as possible, they are usually subject to limited resources. Therefore, 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. We then systematically review previous evidence on the links between standardized achievement tests and the covariate sets relevant to each heuristic. We present meta-analytic integrations of R^2 (see OSM B for methodology and detailed results) for past studies providing estimates based on either (a) single-level methods (i.e., that do not hierarchically decompose the variances between students, classrooms, and schools) which are informative for planning IRTs or (b) multilevel methods to compile multilevel design

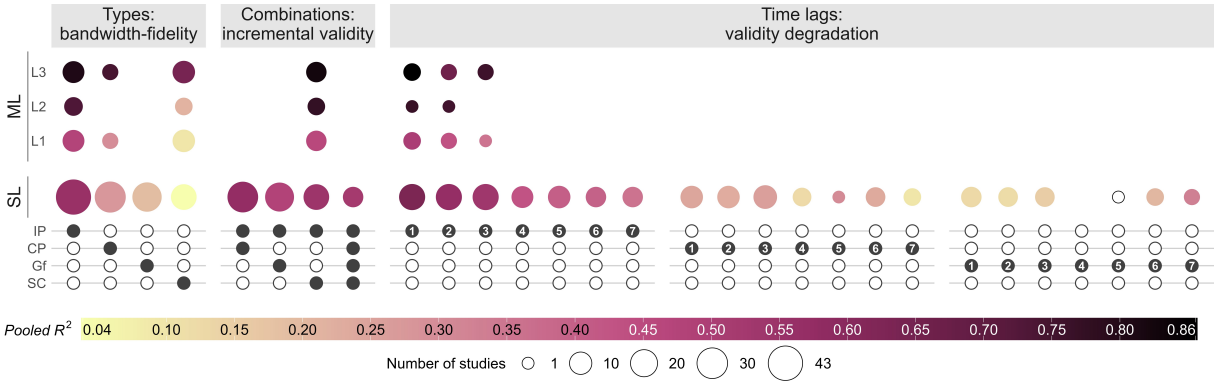
parameters which are informative for planning CRTs. Figure 2 portrays the *Pooled* R^2 values discussed below.

Figure 1. *Schematic Visualization of the Theoretical Predictions Implied by Psychometric Heuristics for Covariate Impacts on* R^2 *in Student Achievement*



Note. IP = Domain-identical pretest. CP = Cross-domain pretest. Gf = Fluid intelligence. 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.

Figure 2. *Previous Research on Covariate Impacts: Meta-Analytically Pooled* R^2 *in Student Achievement for Single- and Multilevel Designs*



Note. Multivariate fixed-effect meta-analysis with correlated effect sizes (with an assumed within-study correlation of $r = .90$). For single-level designs, we reviewed in total $S = 44$ studies, with $H = 53$ independent samples yielding $G = 1{,}633$ 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 $G = 2{,}394$ 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. SL = Single-level designs. ML = Multilevel designs. IP = Domain-identical pretest. CP = Cross-domain pretest. Gf = Fluid intelligence. SC = Sociodemographic characteristics.

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). Following this 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 non-cognitive measures and should be covariates of even broader bandwidth.

Previous Empirical Evidence

Single-Level Perspective. Many studies demonstrated the high predictive power of IP for student achievement, with $Pooled\ R^2_{\mathrm{T|IP}} = .56$. For example, to inform power analyses for IRTs, Cole et al. (2011) calculated the year-to-year pre-posttest correlations for 3rd–8th graders in five U.S. states that translated to around two thirds of the explained variance in both mathematics and verbal skills and remained fairly stable across grades. However, others showed that IP gains in relevance with higher grades (e.g., McCoach et al., 2017). There was substantial between-study variation ($.17 \leq R^2_{\mathrm{T|IP}} \leq .73$) due to variation across grade levels and/or domains and pre-posttest time lags. CP was half as effective as IP ($Pooled\ R^2_{\mathrm{T|CP}} = .28$). Amounts of explained variance by CP varied widely between studies, from 4% for mathematics as predicted by phonological abilities (Passolunghi & Lanfranchi, 2012) to around 49% for associations between mathematics and reading (Bailey et al., 2020). Gf turned out to be a significant predictor, with $Pooled\ R^2_{\mathrm{T|Gf}} = .19$. For instance, using large-scale data from six German secondary school samples, Saß et al. (2021) recorded that Gf explained about one quarter of achievement differences in mathematics and reading. However, the prognostic validity of Gf ranged broadly across studies ($.04 \leq R^2_{\mathrm{T|Gf}} \leq .38$). Finally, SC explained a meaningful but—relative to IP, CP, and Gf—small proportion of variance of about 4%. Again, there was

considerable between-study heterogeneity in $R^2_{\mathrm{T|SC}}$ from .00 (derived from Spencer et al., 2022, for gender as single covariate) and .29 (derived from Li et al., 2022, for a set of gender, migration background, and socioeconomic status). 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 predictor for student achievement, explaining astonishing proportions of variance at the group levels: $Pooled\ R^2_{\mathrm{L2|IP}} = .73$ and $Pooled\ R^2_{\mathrm{L3|IP}} = .81$. At L1, IP explained on average 48% of achievement differences. Nevertheless, there was notable variation between studies ($.23 \leq R^2_{\mathrm{L1|IP}} \leq .58$, $.49 \leq R^2_{\mathrm{L2|IP}} \leq .70$, $.54 \leq R^2_{\mathrm{L3|IP}} \leq .83$). Of note, the prognostic value of IP seems to strengthen throughout the school career, in particular at L3: $R^2_{\mathrm{L3|IP}}$ in mathematics/reading as reported in Hedges and Hedberg (2013) averaged to .69/.75 with 1st–6th graders and to .79/.84 with 7th–11th graders. This trend was replicated in several works (see Stallasch et al., 2021, Figure 1). Despite domain mismatch, CP proved a highly robust predictor, particularly at L3: $Pooled\ R^2_{\mathrm{L3|CP}}$ amounted to .74, whereas $Pooled\ R^2_{\mathrm{L1|CP}}$ was .30. Spybrook, Westine et al. (2016), for example, found that reading explained about 77% of school-level variance in science achievement. Cross-study variations were moderate ($.24 \leq R^2_{\mathrm{L1|CP}} \leq .35$, $.56 \leq R^2_{\mathrm{L3|CP}} \leq .77$). As far as we are aware, the predictive capacity of Gf has not yet been partitioned into its hierarchical variance components. Overall, SC exerted substantial predictive power at L3 with 64% of explained variance, but rather limited predictive properties at L1/L2 with 10%/21%. It is noteworthy that $R^2_{\mathrm{L3|SC}}$ show considerable heterogeneity, specifically between domains and countries: Brunner et al. (2018), for instance, documented $R^2_{\mathrm{L3|SC}} = .01$ for mathematics in Azerbaijan and $R^2_{\mathrm{L3|SC}} = .97$ for reading in Liechtenstein. In summary, the available multilevel evidence fit the assumptions about the differential impacts of covariate types with 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.

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 $Pooled\ R^2_{\mathrm{T|IP+CP}} = .57$. Yet, the maximum incremental returns from CP summed to +13% (Chu et al., 2018). Overall, Gf showed no additional benefits over and above IP ($Pooled\ R^2_{\mathrm{T|IP+Gf}} = .48$). There was, however, notable between-study variation: In Georgiou et al. (2021), for instance, Gf added around +12% to the proportion of explained variance in mathematics and reading. Combining IP and SC did not lead to a general improvement over controlling for IP alone ($Pooled\ R^2_{\mathrm{T|IP+SC}} = .55$), but increments occasionally reached $\Delta R^2_{\mathrm{T|+SC}} = +.08$ (Li et al., 2022). Taken together, the full covariate battery did not raise the amount of explained variance beyond IP ($Pooled\ R^2_{\mathrm{T|IP+CP+Gf+SC}} = .52$). Yet, some studies revealed meaningful increments, peaking at +15% (Chu et al., 2018). Notably, the $Max(\Delta R^2_{\mathrm{T}})$ were consistently found 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 SC: SC 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. Jacob et al.'s (2010) and Stallasch et al.'s (2021) analyses revealed that SC contributed around +13%/+4% and +21%/+13% to the prediction of L2/L3 achievement differences beyond IP, respectively. Of note, additional returns in R^2 at the various hierarchical levels appeared to be more pronounced in elementary than secondary school (Stallasch et al., 2021).

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.[39]

Previous Empirical Evidence

Single-Level Perspective. The vast majority of reviewed studies indicate that $R^2_{\mathrm{T|IP}}$ in student achievement decreases with greater pre-posttest time lags: Values considerably dropped from $Pooled\ R^2_{\mathrm{T|IP-1}} = .63$ to $Pooled\ R^2_{\mathrm{T|IP-7}} = .36$. For example, drawing on large-scale data from U.S. colleges and universities, Dahlke et al. (2018) showed that the prognostic validities of high school students' mathematics and reading IPs clearly deteriorates over time ($R^2_{\mathrm{T|IP-1}} = .62$, $R^2_{\mathrm{T|IP-3}} = .55$). 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: $Pooled\ R^2_{\mathrm{T|CP-1}} = .24$ declined to $Pooled\ R^2_{\mathrm{T|CP-7}} = .10$. Specifically, McCoach et al. (2017) found that correlations between mathematics and reading in Grades 2 through 12 steadily weakened as the time gap grew ($R^2_{\mathrm{T|CP-1}} = .36$, $R^2_{\mathrm{T|CP-7}} = .28$). However, there was significant between-study heterogeneity. In some studies, $R^2_{\mathrm{T|CP}}$ barely diminished (e.g., Erbeli et al., 2021) or even increased with growing time lags (e.g., Träff et al., 2020). The scant available studies on the potential validity degradation of Gf suggest 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. Stern (2009), for instance, demonstrated that Gf was an exceptionally stable predictor of Grade 11 mathematics after 7 years and even beyond ($R^2_{\mathrm{T|Gf-5}} = .17$ and $R^2_{\mathrm{T|Gf-7}} = .16$).

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

Multilevel Perspective. The few existing investigations on multilevel design parameters addressing the temporal validity degradation of covariates substantiated a notable decrement of explanatory power of IP at L1; $Pooled\ R^2_{\mathrm{L1|IP-1}} = .50$ declined to $Pooled\ R^2_{\mathrm{L1|IP-3}} = .35$. Meanwhile, amounts of explained variance at L3 were far less prone to time effects: Pooled across studies, IP−1 accounted for 86% and IP−3 accounted for 76% of achievement differences between schools. Only Xu and Nichols (2010) studied deterioration in the prognostic property of IP at L2. The authors found that explanatory power remained at a high level of 70% across two subsequent years. Of note, declines in R^2 seem to be more prevalent in elementary than secondary school, especially at L3. In Bloom et al. (2007), mean $R^2_{\mathrm{L3|IP-1}}/R^2_{\mathrm{L3|IP-2}}/R^2_{\mathrm{L3|IP-3}}$ was .56/.49/.26 in elementary school, and .83/.79/.77 in secondary school. This finding held true for both mathematics and reading and could largely be replicated by Xu and Nichols (2010). To the best of 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 comprehensive compilation of reliable single- and multilevel design parameters that were meta-analyzed and applied to simulate precision.[40]

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

[40] 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.

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, 2018), as do educational RTs (e.g., Morrison, 2020, pp. 123–124). 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[41] and verbal domains.

Fourth, educational RTs are conducted all around the globe (Connolly et al., 2018), but existing collections of multilevel design parameters primarily stem from U.S. 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 are representative for the total

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

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[42].

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 Equations (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.

The remainder of this paper is structured as follows. Part I covers empirically estimated and meta-analytically integrated design parameters, and demonstrates their use in sample size and power computations. Part II covers the $MDES$ simulation study. Note that this study is accompanied by an extensive OSF repository at https://tinyurl.com/osf-blinded. In addition to all R scripts and brief instructions for data access, 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) simulated design parameters, the latter along with their $MDES$ statistics.

[42] 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).

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 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[43] 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); 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 (Prenzel, Baumert, et al., 2006; Reiss et al., 2017). We used the data[44] 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.

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

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, Carstensen, 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 3 measurement points. As listed in Table 1, we analyzed data from a total of $N = 68{,}502$ students, where sample sizes ranged within 1,868 (NSC3, Grade 12) $\leq N \leq$ 10,543 (DESI, Grade 9), with median cluster sizes of $4 \leq n_{\mathrm{L2}} \leq 25$, $14 \leq n_{\mathrm{L3}} \leq 50$, and $2 \leq J_{\mathrm{L3}} \leq 3$. 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).

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	6,731	1,020	374	6	16	2
2	NSC2	6,319	986	362	6	15	2
3	NSC2	5,554	888	354	6	14	2
4	NSC2	5,418	1,026	349	4	14	3
			Lower secondary school				
7	NSC3	6,314	619	268	10	24	2
9	NSC3	4,659	631	240	6	20	2
9	DESI	10,543	427	219	25	50	2
10	PP03	6,020	275	152	23	42	2
10	PP12	4,494	252	134	19	37	2
			Upper secondary school				
11	NSC3	2,054	n/a	107	n/a	19	n/a
11	NSC4	4,565	n/a	175	n/a	26	n/a
12	NSC3	1,868	n/a	105	n/a	17	n/a
12	NSC4	3,963	n/a	168	n/a	23	n/a
Total		68,502	6,124	3,007			

Note. 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.

Table 2. *Covariate Sets Analyzed in the Present Study with Numbers of Covariates Q, and 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

Note. 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. SC = Sociodemographic characteristics (gender, migration background, socioeconomic status).

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}, $\sigma^2_{L2}2$, 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. 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 Equations (3)–(5) and (8)–(10). SEs 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, Equation 3) and three-level designs in Hedges et al. (2012, Equations 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, 2021). We drew 1,000 samples (Huang, 2018, p. 303; Schomaker & Heumann, 2018). SEs of R^2were computed with the formula for the large sample variances given in Hedges and Hedberg (2013, p. 451).

Pooling. ρ and $R^2$with corresponding SEs 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).[45]

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.,

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

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, 2021). We assumed a within-sample intercorrelation of $r = .90$ 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).

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, as illustrated in Figure 3 (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/upper secondary school ($Pooled\ R^2_{\text{T|IP}} = .36/.34$), and even almost one half in

lower secondary school ($Pooled\ R^2_{\mathrm{T|IP}} = .46$). Despite domain mismatch, CP was a valuable predictor, particularly in lower secondary school, where $Pooled\ R^2_{\mathrm{T|CP}} = .28$. In elementary/upper secondary school, CP was almost half as effective as IP, with $Pooled\ R^2_{\mathrm{T|CP}} = .17/.12$. Gf only served as a useful covariate type with 5th–10th graders: Pooled across domains and samples, Gf contributed 20% to the prediction in lower secondary school, but only 6% in the other grade levels. SC turned out to be meaningful predictors only in younger students: in elementary school, SC performed as well as CP ($Pooled\ R^2_{\mathrm{T|SC}} = .16$), but their explanatory power significantly fell behind all other covariates in lower secondary school ($Pooled\ R^2_{\mathrm{T|SC}} = .13$) and was as weak as Gf in upper secondary school ($Pooled\ R^2_{\mathrm{T|SC}} = .07$).

We registered substantial R^2_{T} heterogeneities, in particular for IP and least for SC: in elementary school, for example, the respective 95% PIs were [.13, .58] and [.09, .24], with $\tau^2_{\text{Effect sizes}} = .0119$ and .0013 (Table F1). Consistently, most of the observed variability was due to true variance rather than random noise ($I^2_{\text{Effect sizes}} \geq 94\%$; Table F1).

Multilevel Perspective. IP was of paramount relevance when predicting student achievement. This holds true for all grade and hierarchical levels. Nevertheless, it is noteworthy that while from Grade 5 on, IP was the strongest among all covariate types and showed exceptional prognostic properties at L3 with $Pooled\ R^2_{\mathrm{L3|IP}} = .98/.78$ in lower/upper secondary school, the respective value lay around .44 in elementary school. Across the entire school career, IP explained less variance at both L1 and L2 ($.26 \leq Pooled\ R^2_{\mathrm{L1|IP}} \leq .36$; $Pooled\ R^2_{\mathrm{L2|IP}} = .29/.60$ in elementary/lower secondary school). CP was a powerful predictor, explaining about 91% of L3 variance in lower secondary school, and still about 30%/47% in elementary/upper secondary school. At lower hierarchical levels, the explanatory power of CP was clearly reduced, with $Pooled\ R^2_{\mathrm{L1|CP}}$ being between .10 and .16 and $Pooled\ R^2_{\mathrm{L2|CP}} = .15/.40$ in elementary/lower secondary school. Gf appeared to be of utmost importance to explaining differences between lower secondary schools ($Pooled\ R^2_{\mathrm{L3|Gf}} = .86$), but less so between classrooms ($Pooled\ R^2_{\mathrm{L2|Gf}} = .16$) and students ($Pooled\ R^2_{\mathrm{L1|Gf}} = .07$). Gf was consistently the weakest covariate type both in elementary ($Pooled\ R^2_{\mathrm{|Gf}} = .06/.08/.13$ at L1/L2/L3) and upper secondary school ($Pooled\ R^2_{\mathrm{|Gf}} = .04/.39$ at L1/L3). Although SC were the poorest predictors in lower secondary school, amounts of explained between-school differences still amounted to around 77% (with 4%/12% at L1/L2). In upper secondary school, the explanatory power of SC at L3 was similar to that of CP ($Pooled\ R^2_{\mathrm{L3|SC}} = .45$, with $Pooled\ R^2_{\mathrm{L1|SC}} = .05$). Notably,

with 1st–4th graders, SC outweighed IP at both L2 and L3, explaining 35% and 52% of variance, respectively (with $Pooled\ R^2_{\text{L1|SC}} = .11$).

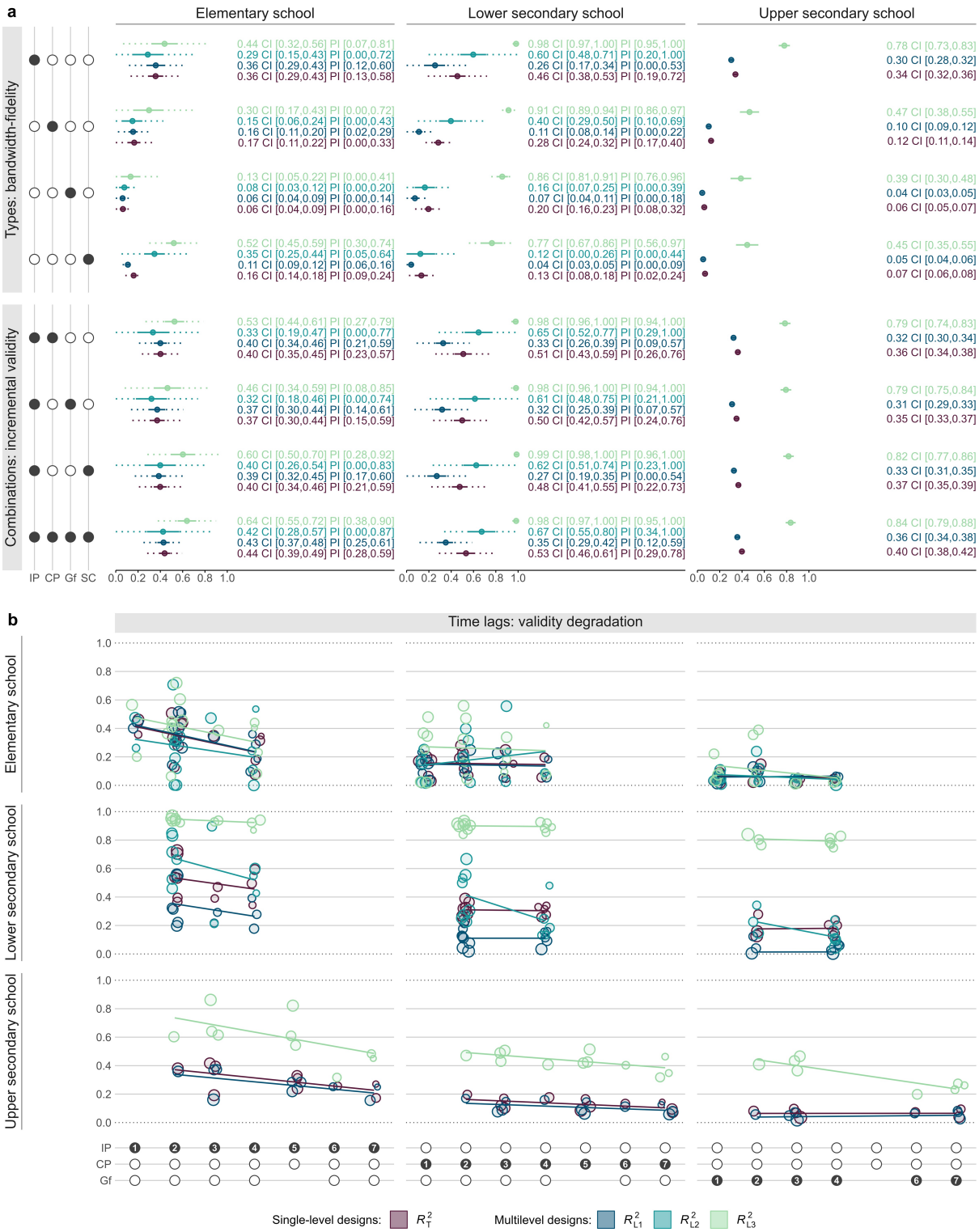
Degrees of heterogeneity in multilevel R^2 were often substantial, depending not only on the covariate type but also on the grade and hierarchical level. Consider, for instance, IP: Predicted $R^2_{\text{L3|IP}}$ ranged widely from .07 to .81 in elementary ($\tau^2_{\text{Effect sizes}} = .0325$; $I^2_{\text{Effect sizes}} = 93\%$; Table F1) but only between .95 and 1.00 in lower secondary school ($\tau^2_{\text{Effect sizes}} = \tau^2_{\text{Samples}} = .0001$; $I^2_{\text{Effect sizes}} = 26\%$, $I^2_{\text{Samples}} = 60\%$; Table F2). At L1/L2, 95% PIs were always sizeable, with [.12, .60]/[.00, .72] ($\tau^2_{\text{Effect sizes}} = .0136/.0430$; $I^2_{\text{Effect sizes}} = 99\%/96\%$; Table F1) in elementary and [.00, .53]/[.20, 1.00] ($\tau^2_{\text{Effect sizes}} = .0140/.0383$, $\tau^2_{\text{Samples}} = .0041/.0000$; $I^2_{\text{Effect sizes}} = 77\%/94\%$, $I^2_{\text{Samples}} = 23\%/0\%$; Table F2) in lower secondary school.

Covariate Combinations: Incremental Validity

Single-Level Perspective. In all grade levels, CP explained additional variance in student achievement over and above IP. Incremental gains were largest in lower secondary school, with around +5% ($Pooled\ R^2_{\text{T|IP+CP}} = .51$), but were also noticeable in elementary school, with around +4% ($Pooled\ R^2_{\text{T|IP+CP}} = .40$). In upper secondary school, however, increments were small, with an average gain of +2% ($Pooled\ R^2_{\text{T|IP+CP}} = .36$). When controlling for IP, Gf further contributed to the prediction in lower secondary school ($Pooled\ R^2_{\text{T|IP+Gf}} = .50$; $\Delta Pooled\ R^2_{\text{T|+Gf}} = +.04$), but in elementary/upper secondary school, benefits were negligible ($Pooled\ R^2_{\text{T|IP+Gf}} = .37/.35$, $\Delta Pooled\ R^2_{\text{T|+Gf}} = +.01/+.01$). In contrast, SC explained more additional variance in elementary/upper secondary school, with about +4%/+3% ($Pooled\ R^2_{\text{T|IP+SC}} = .40/.37$) than in lower secondary school, with about +2% ($Pooled\ R^2_{\text{T|IP+SC}} = .48$). Joint effects through the full battery of covariates were always largest, with around 44%/53%/40% of variance explained in elementary/lower secondary/upper secondary school ($\Delta Pooled\ R^2_{\text{T|+CP+Gf+SC}} = +.08/+.07/+.06$).

We found signal heterogeneities in all R^2_{T}. For example, future $R^2_{\text{T|IP+Gf}}$ will likely fall between .15 and .59 with 1st–4th graders ($\tau^2_{\text{Effect sizes}} = .0117$; $I^2_{\text{Effect sizes}} = 98\%$; Table F1).

Figure 3. *Meta-Analytic Integrations of Single- and Multilevel R^2 in Student Achievement*



Note. Figure 3a: 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. Figure 3b: Fixed-effect (Set 11 [Gf−lag] throughout secondary school) and random-effects (remaining) meta-regression with time lag as moderator; bubbles show observed R^2sized by weight; line slopes map b_{lag}. On the axes, 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 pretests. CP = Cross-domain pretests (reading for STEM outcomes, mathematics for verbal outcomes). Gf = Fluid intelligence pretests. SC = Sociodemographic characteristics (gender, migration background, socioeconomic status).

Multilevel Perspective. At L3, CP added to the prediction of achievement differences beyond IP between elementary schools ($Pooled\ R^2_{\text{L3|IP+CP}} = .53$, $\Delta Pooled\ R^2_{\text{L3|+CP}} = +.09$), but not between lower/upper secondary schools, where increments did not exceed values of around +1% ($Pooled\ R^2_{\text{L3|IP+CP}} = .98/.79$). At both L1 and L2, CP provided some additional explanatory power beyond IP over the entire school career: benefits were largest in lower secondary school with on average +7% at L1 ($Pooled\ R^2_{\text{L1|IP+CP}} = .33$) and +5% at L2 ($Pooled\ R^2_{\text{L2|IP+CP}} = .65$). We found Gf to be a rather poor additional covariate beyond IP across grade and hierarchical levels. In elementary school, average gains in the amounts of explained variance from Gf was the greatest at L2 with +3% ($Pooled\ R^2_{\text{L2|IP+Gf}} = .32$). In lower secondary school, Gf was not useful at the group levels with $\Delta Pooled\ R^2_{\text{|+Gf}} \leq +.01$ at L2 and L3, but added around +6% at L1 ($Pooled\ R^2_{\text{L1|IP+Gf}} = .32$). In upper secondary school, contributions of Gf were negligible, with $\Delta Pooled\ R^2_{\text{|+Gf}} = +.01$ at all hierarchical levels. SC was of notable incremental relevance, especially in elementary school at L1/L2, adding +16%/+11% of explained variance ($Pooled\ R^2_{\text{|SC}} = .39/.40$). For 5th–10th graders, increments were consistently small ($\Delta Pooled\ R^2_{\text{|+SC}} \leq +.02$); nevertheless, here, pooled values of R^2_{L3} maximized, with 99% of explained variance. For 11th–12th graders, SC contributed about +4%/+3% at L1/L3 ($Pooled\ R^2_{\text{|SC}} = .33/.82$). Except for L3 in lower secondary school, the complete set of covariates consistently outweighed all other combinations: together they explained 43%/35%/36% of the variance at L1 ($\Delta Pooled\ R^2_{\text{L1|+CP+Gf+SC}} = +.07/+.09/+.06$), 42%/67% at L2 ($\Delta Pooled\ R^2_{\text{L2|+CP+Gf+SC}} = +.13/+.07$), and 60%/98%/84% at L3 ($\Delta Pooled\ R^2_{\text{L3|+CP+Gf+SC}} = +.20/+.00/+.06$) in elementary/lower secondary/upper secondary school.

Multilevel R^2 heterogeneities largely mirrored those of IP, and were substantive (except R^2_{L3} in lower secondary school). $R^2_{\text{L2|IP+CP}}$, for instance, had a 95% PI [.21, 1.00] in Grades 5–10 ($\tau^2_{\text{Effect sizes}} = .0380$, $\tau^2_{\text{Samples}} = .0002$; $I^2_{\text{Effect sizes}} = 92\%$, $I^2_{\text{Samples}} = 0\%$; Table F2).

Covariate Time Lags: Validity Degradation

Single-Level Perspective. In all grade levels, the predictive power of IP clearly reduced with growing pre-posttest time lags. Validity degradation was most prevalent in elementary school, where predicted $R^2_{\text{T|IP}-1} = .41$ almost halved to $R^2_{\text{T|IP}-4} = .23$. The meta-regression coefficient $b_{\text{lag}} = -.06$ shows that with each additional year between IP and outcome, $R^2_{\text{T|IP}}$ decreases by 6%. In lower/upper secondary school, temporal declines in the proportions of explained

variance were also noticeable, with b_{lag}= −.04/−.03; predicted $R^2_{\text{T|IP}-2}$ = .53/.37 declined to $R^2_{\text{T|IP}-4}$ = .46/$R^2_{\text{T|IP}-7}$ = .23. In contrast, CP emerged to be far less prone to cross-time decay. Until Grade 10, prognostic properties remained stable (b_{lag} = .00) over 4 years both in elementary ($R^2_{\text{T|CP}-1}$ = .16, $R^2_{\text{T|CP}-4}$ = .15) and secondary school ($R^2_{\text{T|CP}-2}$ = .31, $R^2_{\text{T|CP}-4}$ = .30). In upper secondary school, predicted amounts of explained variance slightly reduced from 16% to 10% for a 2- to 7-year-lagged CP (b_{lag} = −.01). Gf turned out to be an extraordinarily time-robust predictor throughout the entire school career, with b_{lag} =.00: in lower/upper secondary school, predicted $R^2_{\text{T|Gf}}$ = .18/.06 remained unchanged across 4/7 years; in elementary school, the drop was infinitesimal ($R^2_{\text{T|Gf}-1}$ = .06, $R^2_{\text{T|Gf}-4}$ = .05).

Multilevel Perspective. Validity degradation in $R^2_{\text{|IP}}$ was substantial for almost all grade and hierarchical levels, except for lower secondary school at L3. Here, we recorded remarkable temporal stabilities in the amounts of explained variance (b_{lag} = −.01): predicted $R^2_{\text{L3|IP}-2}$ = .95 persisted at high levels of $R^2_{\text{L3|IP}-4}$= .92 after 4 years. In all other cases, explanatory power of IP is likely to drop around 3% per year, as documented at L1 in upper secondary school ($R^2_{\text{L1|IP}-2}$ = .34, $R^2_{\text{L1|IP}-7}$ = .21), up to about 8% per year as found at L2 in lower secondary school ($R^2_{\text{L2|IP}-2}$ = .67, $R^2_{\text{L2|IP}-4}$ = .52). CP appeared to be a relatively time-stable covariate; however, decrements in prognostic capacity hinged on both the grade and hierarchical level: in elementary school, solely predicted $R^2_{\text{L3|CP}-1}$ = .27 slightly declined to $R^2_{\text{L3|CP}-4}$ = .24 (b_{lag} = −.01); in lower secondary school, only predicted proportions of explained L2 variance dropped, and this strikingly from 41% to 23% across 2 to 4 years (b_{lag} = −.09); and in upper secondary school, both predicted $R^2_{\text{L1|CP}}$ and $R^2_{\text{L3|CP}}$ showed small reductions over time, with b_{lag} = −.01 ($R^2_{\text{L1|CP}-2}$ = .14, $R^2_{\text{L1|CP}-7}$ = .09) and b_{lag} = −.02 ($R^2_{\text{L3|CP}-2}$ = .49, $R^2_{\text{L3|CP}-7}$ = .39), respectively. In all other cases, we registered any temporal decline in the proportions of explained differences (.00 ≤ b_{lag} ≤ .03). Gf emerged as highly time-stable at L1, losing 0% of explanatory power across time: predicted $R^2_{\text{L1|Gf}}$ stagnated at .06/.01 across 4 years in elementary/lower secondary school, and $R^2_{\text{L1|Gf}-2}$ = .04 even slightly raised to $R^2_{\text{L1|Gf}-7}$ = .05 in upper secondary school. At the same time, validity decay was significant at L2/L3: In 95% of studies with 1st–4th graders, $R^2_{\text{|Gf}-1}$ = .08/.14 will fall to $R^2_{\text{|Gf}-4}$ = .04/.05 (b_{lag} = −.01/−.03). With 5th–10th graders, predicted $R^2_{\text{|Gf}-2}$ = .22/.81 dropped to $R^2_{\text{|Gf}-4}$ = .12/.79 (b_{lag}= −.05/−.01). With 11th–12th graders, a 2-year-lagged Gf is predicted to explain 44% of between-school differences, but a 7-year-lagged Gf only 24% (b_{lag} = −.04).

Application

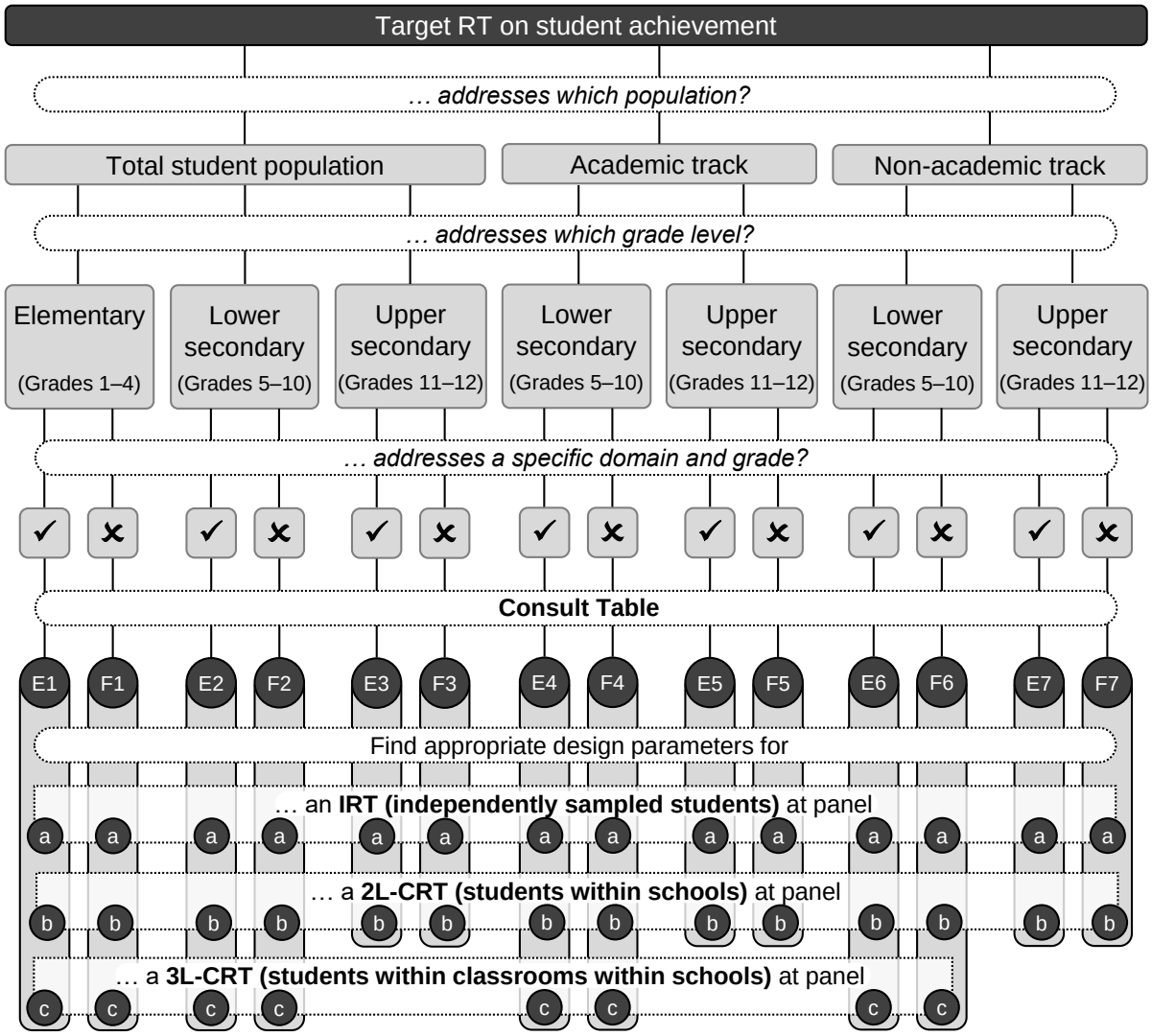
Researchers designing RTs may profit from the flow chart in Figure 4. It facilitates the choice of single- and multilevel design parameters that are optimally tailored to the specific application context. To showcase the estimates' use in study planning, we developed manifold scenarios to determine the (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).

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 $d = .15$ is considered meaningful, representing around one half of the expected annual growth in mathematics for Grade 6–7 in the German student population (Brunner, Stallasch, et al., 2023, Table 1). The team's objective, therefore, is to sample enough 7th graders to detect $MDES = .15$ at $\alpha = .05$ (two-tailed) with $1 - \beta = .80$, where $P_T = .50$. The minimum required sample size ($MRSS$) to achieve this in an unconditional IRT design is $N = 504$ 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 (Figure 4): 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 $R^2_{T|IP} = .53$, the researchers find that the $MRSS$ more than halves to $N = 237$ when adjusting for IP. They then think about optimizing the design by additionally including either a reading CP or SC, where $R^2_{T|IP+CP} = .56$ and $R^2_{T|IP+SC} = .55$. The $MRSS$ further reduces to $N = 228$ when combining IP with SC and to $N = 225$ 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 $SE(R^2_{T|IP+CP}) = .01$: the lower bound is calculated as $.56 - 1.96 * .01 = .54$ and the upper bound as $.56 + 1.96 * .01 = .57$, which leads to an $MRSS$ range of $235 \geq N \geq 216$. Consequently, when opting for a conservative approach and sampling $N = 235$ students, it is fairly certain that the IRT will be sensitive to uncover a (truly existing) treatment effect of $d = .15$ with IP and CP as covariates.

Figure 4. *Flow Chart to Choose Design Parameters from Our Compilation in OSM E and F*



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.

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. $d = .11$ 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 $n_{\mathrm{L2}} = 20$ and $J_{\mathrm{L3}} = 3$. The team's primary concern, thus, is to achieve sufficient power (i.e., $1 - \beta \geq .80$) to detect $MDES = .11$ at $\alpha = .05$ (two-tailed), where $P_{\mathrm{L3}} = .50$. Since the 3L-CRT addresses the total population in lower secondary school but neither a specific grade nor domain, our flow chart (Figure 4) 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 .05/.35 into

power analysis, the researchers learn that an unconditional 3L-CRT clearly undercuts the desired power rate ($1 - \beta = .43$). 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 .07/.16/.86 for Gf and .04/.12/.77 for SC. Controlling for both Gf ($1 - \beta = .98$) and SC ($1 - \beta = .92$) 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 $\rho_{L2} = .07$ and $\rho_{L3} = .50$ and the lower bounds of the 95% PIs of $R^2_{L1|Gf} = .00$, $R^2_{L2|Gf} = .00$, $R^2_{L3|Gf} = .76$, $R^2_{L1|SC} = .00$, $R^2_{L2|SC} = .00$, and $R^2_{L3|SC} = .56$, only Gf ($1 - \beta = .81$) likely guarantees enough power, as opposed to SC ($1 - \beta = .59$). 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 $Pooled\ \rho_{L2} = .04$, $Pooled\ \rho_{L3} = .38$, and predicted values of $R^2_{L1|Gf-2} = .01$, $R^2_{L2|Gf-2} = .22$, $R^2_{L3|Gf-2} = .82$, as well as $R^2_{L1|Gf-4} = .01$, $R^2_{L2|Gf-4} = .12$, $R^2_{L3|Gf-4} = .79$. Assuming no attrition over time, the team calculates $1 - \beta = .95/.94$ 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, pp. 189–202; 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)[46] 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 that $N = 644/2116/4600$ for small/medium/large RTs.[47] We assumed $\alpha = .05$ (two-tailed), $1 - \beta = .80$, and $P_T = P_{L3} = .50$.

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 SEs 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 SEs, and the covariances were derived assuming an intercorrelation of $r = .90$, 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 Equations (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.

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, as illustrated in Figure 5 (which we refer to in this section; see OSM D for result plots of school

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

[47] 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.

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, $MDES_{\text{IRT}} = .12$ (i.e., unconditional). Precision was then moderately affected by the covariate types; the median $MDES_{\text{IRT|IP/CP/Gf/SC}}$ equaled .10/.11/.12/.11 in elementary, .09/.10/.11/.11 in lower secondary, and .10/.11/.12/.12 in upper secondary school. Notably, percentage $MDES$ reduction for a certain covariate type remained constant across IRT sizes. Since precision is a positive function of sample size, absolute $MDES$ shrinkage was 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 raised precision in an IRT with $N = 644$ ($MDES_{\text{IRT}} = .22$ vs. $Mdn(MDES_{\text{IRT|SC}}) = .20$) but not with $N = 4600$ ($MDES_{\text{IRT}} = Mdn(MDES_{\text{IRT|SC}}) = .08$).

Multilevel Perspective. In a medium CRT, median $MDES_{\text{CRT}} = .35/.53/.32$ (i.e., unconditional) in elementary/lower secondary/upper secondary school. In lower secondary school, all covariate types strongly boosted median precision, first and foremost IP ($MDES_{\text{3L-CRT|IP}} = .15$), followed by CP ($MDES_{\text{3L-CRT|CP}} = .20$), but also Gf ($MDES_{\text{3L-CRT|Gf}} = .25$) and SC ($MDES_{\text{3L-CRT|SC}} = .28$). In upper secondary school, IP markedly reduced the $MDES_{\text{2L-CRT}}$ to around .19, twice as much as CP/Gf/SC, which averaged .26/.27/.26. In elementary school, particularly SC ($MDES_{\text{3L-CRT|SC}} = .26$), but also IP ($MDES_{\text{3L-CRT|IP}} = .27$) and CP ($MDES_{\text{3L-CRT|CP}} = .29$), evoked reasonable average precision improvements, while Gf performed more poorly ($MDES_{\text{3L-CRT|Gf}} = .33$). 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 ($Mdn(MDES_{\text{3L-CRT}}) = .68$ vs. $Mdn(MDES_{\text{3L-CRT|SC}}) = .52$) and to 27% in large CRTs ($Mdn(MDES_{\text{3L-CRT}}) = .24$ vs. $Mdn(MDES_{\text{3L-CRT|SC}}) = .17$). 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 diminished the $MDES$ over and above IP, but only in elementary/lower secondary school ($Mdn(MDES_{\text{IRT|IP+CP}})$ = .09/.08). In these grade levels, no other combination resulted in further improvements. In upper secondary school, only the complete covariate battery resulted in genuine precision benefits beyond IP alone ($Mdn(MDES_{\text{IRT|IP+CP+Gf+SC}})$ = .09).

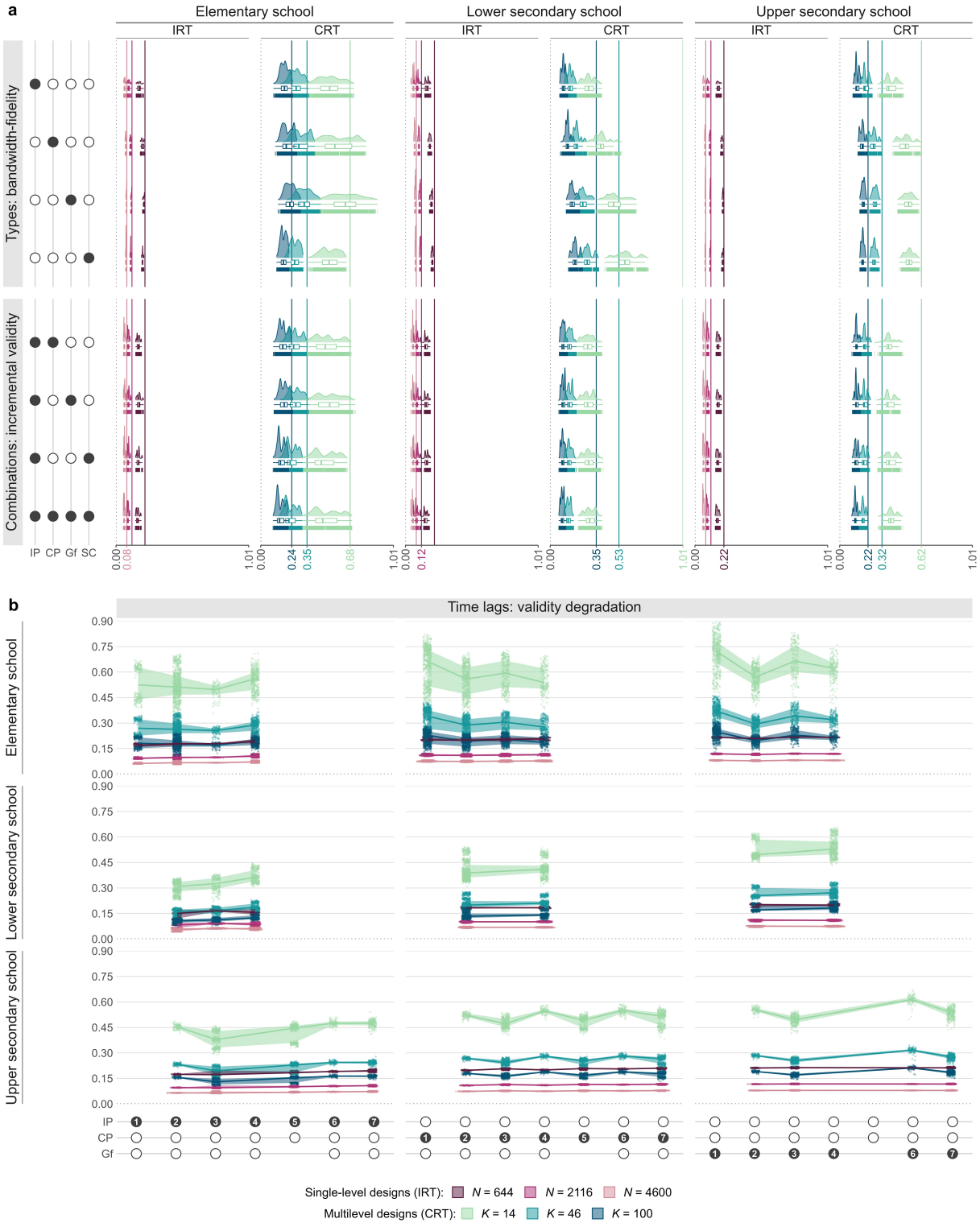
Multilevel Perspective. In a medium CRT targeted at 1st–4th graders, adding CP to IP led to notable $MDES$ drops ($Mdn(MDES_{\text{3L−CRT|IP+CP}})$ = .25), but IP plus SC raised precision the most ($Mdn(MDES_{\text{3L−CRT|IP+SC}})$ = .22), with no further gains through the full covariate array. From Grade 5 on, we did not detect any improvements in the $MDES$ by pairing IP with CP or Gf; the addition of SC, alone or with CP and Gf, returned only miniscule $MDES$ declines averaging .14/.18 in lower/upper secondary school.

Covariate Time Lags: Validity Degradation

Single-Level Perspective. In a medium IRT, precision was slightly affected by temporal validity losses in IP ($\Delta Mdn(MDES_{\text{IRT|IP}})$ = +.02/+.01/+.01 from the shortest to the longest time lag in elementary/lower secondary/upper secondary school), and in CP only after 7 years in upper secondary school ($\Delta Mdn(MDES_{\text{IRT|CP}-7})$ = +.01). Of note, precision was more prone to validity deterioration in IP and CP in small rather than large IRTs (e.g., $Mdn(MDES_{\text{IRT|IP}-2})$= .17/.07 and $Mdn(MDES_{\text{IRT|IP}-7})$ = .19/.07 with N = 644/4600 in upper secondary school). By contrast, $MDES_{\text{IRT|Gf}}$ consistently remained highly stable.

Multilevel Perspective. In a medium CRT, median $MDES_{\text{CRT}}$ = .35/.54/.32 (i.e., unconditional) 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, $\Delta MDES_{\text{CRT|IP}}$ = +.02/+.03/+.01, $\Delta MDES_{\text{CRT|CP}}$ = −.06/+.01/±.00, and $\Delta MDES_{\text{CRT|Gf}}$ = −.05/+.02/±.00. As for IRTs, cross-time precision decay appeared more pronounced in small rather than large CRTs (e.g., $Mdn(MDES_{\text{2L−CRT|IP}-2})$ = .45/.16 and $Mdn(MDES_{\text{2L−CRT|IP}-7})$ = .48/.16 for K = 14/100 in upper secondary school).

Figure 5. *MDES Distributions for Small, Medium, and Large IRTs and CRTs*



Note. Figure 5a: Vertical lines show the unconditional (i.e., unadjusted) $MDES_{\text{IRT}}$ (by definition, only varying by sample size) and $Mdn(MDES_{\text{CRT}})$. Figure 5b: Lines connect $Mdn(MDES)$ values of consecutive time lags; ribbons depict interquartile ranges; for small/medium/large RTs, unconditional $Mdn(MDES_{3\text{L}-\text{CRT}})$ = .68/.35/.24 (elementary school) and 1.05/.54/.36 (lower secondary school), $Mdn(MDES_{2\text{L}-\text{CRT}})$ = .62/.32/.21 (upper secondary school). In multilevel designs, $n_{\text{L2}} = 23$ and $J_{\text{L3}} = 2$ for 3L-CRTs (elementary and lower secondary school), n_{L3} = 46 for 2L-CRTs (upper secondary school). On the axes, 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. SC = Sociodemographic characteristics (gender, migration background, socioeconomic status).

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 representative 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, Shi, 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, irrespective 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, pp. 25–26).

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 $R^2_{|\mathrm{CP/Gf/SC}}$ 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 1st–4th 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, pp. 123–124).

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. U.S. 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 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. Thus, 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 implemented in lower secondary school (i.e., Grades 5–10).
4. Sociodemographic measures may be considered in RTs implemented in elementary school (i.e., Grades 1–4).
5. If a pretest in the outcome domain is available, additional covariates should be avoided, except for point 4.
6. In IRTs, a pretest in the outcome domain should be granted priority, in spite of its potential temporal validity degradation. In CRTs—especially implemented in secondary school (i.e., 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. In small RTs, covariate adjustment comes with two threats: (a) Every covariate at the top hierarchical level costs 1 df (and $df \propto MDES^{-1}$); so, too many covariates are detrimental (Liu, 2011; Moerbeek & Teerenstra, 2016, p. 85). (b) The likelihood of violating covariate-treatment orthogonality is amplified (Konstantopoulos, 2012; Moerbeek & Teerenstra, 2016, p. 83). Risk (b) may be compensated through further balancing methods (e.g., minimization, matching, stratification; Moerbeek & Teerenstra, 2016, pp. 87–90), albeit, in turn, thwarting the attempt to prevent risk (a).

9. Reliable covariates (i.e., with low measurement error), and in CRTs, aggregated L1 covariates that demonstrate large ρ values at the implementation level of the intervention, are advantageous. Thus, in case of newly-developed covariate measures for the RT, it may be worthwhile to add items to improve score reliability or to use pilot data of ρ estimates for each item to construct multi-item scales that optimize between-group differentiation (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 output is most relevant to RTs whose target populations and outcomes are similar to those analyzed here. More precisely, our results ideally apply to the German school context, but may still be valuable for RTs conducted in other school systems characterized by early performance-based tracking such as Austria, Czech Republic, Hungary, Slovak Republic, and Turkey (Salchegger, 2016). Further, our findings optimally match measures resembling those used in NEPS, PISA, or DESI. Caution is warranted when designing RTs relying on (substantively) divergent measures. Second, our selection of covariates was oriented towards a theoretical and empirical rationale. Nevertheless, as noteworthy amounts of variance remained unexplained for many outcomes, further individual- or group-level attributes (e.g., motivation or instruction quality; Haertel et al., 1983; Levy et al., 2023) might also function as profitable covariates. 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, $R^2_{|\mathrm{Gf}}$ values should be interpreted as lower-bound estimates and would possibly have been stronger with a broader spectrum of subtests. Fourth, the applicability of our results may suffer from range restriction (Miciak et al., 2016). Although we offered design parameters specific to the academic and non-academic track (differing in students' mean achievement), they may be inappropriate to plan RTs exclusively targeted at low-performers.

Conclusion

Inspired by psychometric heuristics and capitalizing on representative 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.

References

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences* (2nd Edition). Routledge, Taylor & Francis Group.

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

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.

Bailey, D. H., Oh, Y., Farkas, G., Morgan, P., & Hillemeier, M. (2020). Reciprocal effects of reading and mathematics? Beyond the cross-lagged panel model. *Developmental Psychology*, *56*(5), 912–921. https://doi.org/10.1037/dev0000902

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), 48. https://doi.org/10.18637/jss.v067.i01

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.

Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., & Zhao, L. (2013). Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review*, *37*(3–4), 170–196. https://doi.org/10.1177/0193841X13513025

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. (2008). The core analytics of randomized experiments for social research. In P. Alasuutari, L. Bickman, & J. Brannen, *The SAGE Handbook of Social Research Methods* (pp. 115–133). SAGE Publications Ltd. https://doi.org/10.4135/9781446212165.n9

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

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

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., 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., 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., 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*, 1–39. 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. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

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

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. North-Holland ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co.

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

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

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

Dahlke, J. A., Kostal, J. W., Sackett, P. R., & Kuncel, N. R. (2018). Changing abilities vs. changing tasks: Examining validity degradation with test scores and college performance criteria both assessed longitudinally. *Journal of Applied Psychology*, *103*(9), 980–1000. https://doi.org/10/gd2wpd

DESI-Konsortium (Ed.). (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). 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.

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

Georgiou, G. K., Inoue, T., & Parrila, R. (2021). Do reading and arithmetic fluency share the same cognitive base? *Frontiers in Psychology*, *12*, 709448. https://doi.org/10.3389/fpsyg.2021.709448

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*. https://cran.r-project.org/web/packages/mitml/mitml.pdf

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

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

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

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<627::AID-JOB2828>3.0.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) (Version 1) [Dataset]*. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://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

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

Li, L., Valiente, C., Eisenberg, N., Spinrad, T. L., Johns, S. K., Berger, R. H., Thompson, M. S., Southworth, J., Pina, A. A., Hernández, M. M., & Gal-Szabo, D. E. (2022). Longitudinal relations between behavioral engagement and academic achievement: The moderating roles of socio-economic status and early achievement. *Journal of School Psychology*, *94*, 15–27. https://doi.org/10.1016/j.jsp.2022.08.001

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

Liu, X. S. (2011). The effect of a covariate on standard error and confidence interval width. *Communications in Statistics - Theory and Methods*, *40*(3), 449–456. https://doi.org/10.1080/03610920903391337

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. (2021). Bootstrapping clustered data in R using lmeresampler. *arXiv*. https://doi.org/10.48550/ARXIV.2106.06568

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective* (Third edition). Routledge.

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. (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

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

Morrison, K. (2020). *Taming randomized controlled trials in education: Exploring key claims, issues and debates* (1st ed.). Routledge. https://doi.org/10.4324/9781003042112

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

National Research Council (Ed.). (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. (2019a). National Educational Panel Study, Scientific Use File of Starting Cohort Grade 5. *Leibniz Institute for Educational Trajectories (LIfBi), Bamberg*. https://doi.org/doi:10.5157/NEPS:SC3:9.0.0

NEPS Network. (2019b). National Educational Panel Study, Scientific Use File of Starting Cohort Grade 9. *Leibniz Institute for Educational Trajectories (LIfBi), Bamberg*. https://doi.org/doi:10.5157/NEPS:SC4:10.0.0

NEPS Network. (2020). National Educational Panel Study, Scientific Use File of Starting Cohort Kindergarten. *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

Passolunghi, M. C., & Lanfranchi, S. (2012). Domain-specific and domain-general precursors of mathematical achievement: A longitudinal study from kindergarten to first grade. *British Journal of Educational Psychology*, *82*(1), 42–63. https://doi.org/10.1111/j.2044-8279.2011.02039.x

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

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., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rost, J., & Schiefele, U. (Eds.). (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) [Dataset]*. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_I_Plus_v1

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.

Pustejovsky, J. E. (2021). *ClubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. R package version 0.5.3.* [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., & Butcher, I. (2005). Randomization inference for balanced cluster-randomized trials. *Clinical Trials, 2*(2), 130–140. https://doi.org/10.1191/1740774505cn075oa

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. (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. https://doi.org/10/cn6grx

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) [Dataset]*. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_Plus_2012-13_v2

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

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*(feb10 2), d549–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*. Wiley.

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

Saß, S., Schütte, K., Kampa, N., & Köller, O. (2021). Continuous time models support the reciprocal relations between academic achievement and fluid intelligence over the course of a school year. *Intelligence*, *87*, 101560. https://doi.org/10.1016/j.intell.2021.101560

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.

Spencer, M., Fuchs, L. S., Geary, D. C., & Fuchs, D. (2022). Connections between mathematics and reading development: Numerical cognition mediates relations between foundational competencies and later academic outcomes. *Journal of Educational Psychology*, *114*(2), 273–288. https://doi.org/10.1037/edu0000670

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

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

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *2*(1), 15. https://doi.org/10.1177/2332858415625975

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*. Organisation for Economic Co-operation and Development.

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

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., Toby Prevost, A., & 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

U.S. 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

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

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

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

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

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

Zhang, Q., Spybrook, J., Kelcey, B., & Dong, N. (2023). Foundational methods: Power analysis. In *International Encyclopedia of Education(Fourth Edition)* (pp. 784–791). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10088-0

4

GENERAL DISCUSSION

4.1 Compendia and Guidance for Power Analysis: Contributions, Key Results, and Design Implications

Stakeholders in education policy, practice, and research worldwide committed in prioritizing evidence-based reform to improve schooling (e.g., OECD, 2007; Pellegrini & Vivanet, 2021; Slavin et al., 2021). This also holds for Germany (BMBF, 2018; KMK, 2016). Building useable evidence on what works to shape learning trajectories presupposes strong RTs that are sensitive (i.e., sufficiently powered and precise) to draw valid inferences on the effectiveness of educational innovations, products, and services. Crucially, if—and only if—RTs are well-designed, they can become indispensable methodological tools for generating empirical evidence to develop and refine theory, practice, and policies.

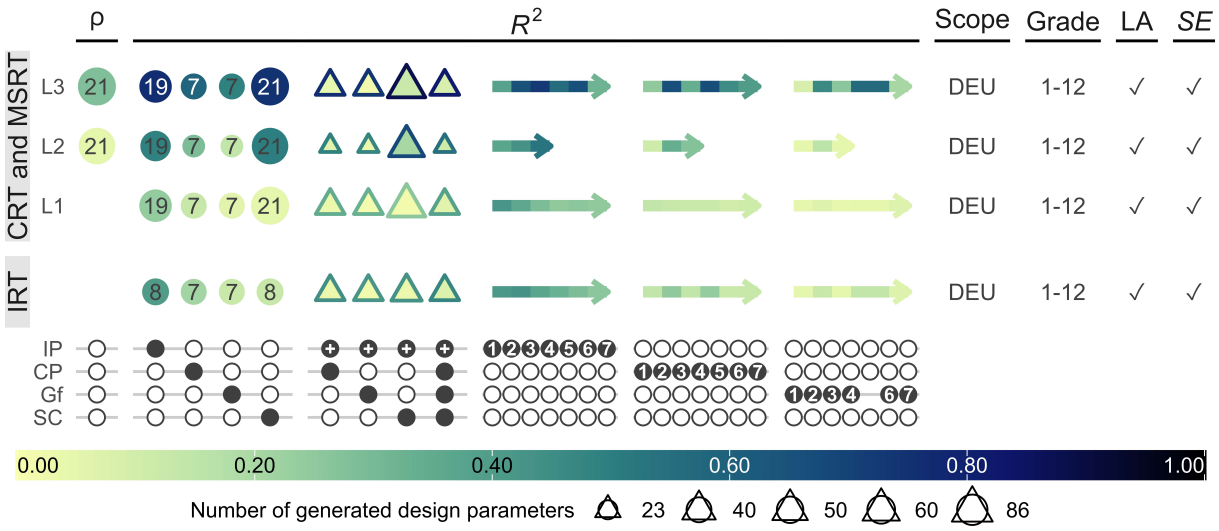
Thus, the overarching goal of the present doctoral thesis was to support the evolution of rigorous RT designs in the German (and similar) school context through the analysis of design parameters for student achievement. In doing so, this thesis developed reliable and versatile compendia and guidance to tweak power analysis.

For decades leading experts have been stressing that the estimates informing power analysis should optimally reflect the realities around the target RT: its population and outcome, applied covariates, and the concrete design as well as planned analysis (e.g., Bloom et al., 2007; Brunner et al., 2018; M. Campbell et al., 2000; Cohen, 1988; Hedges & Hedberg, 2007; Lipsey et al., 2012; Murray, 1998; Schochet, 2008; Spybrook, Westine, et al., 2016; Zhang et al., 2023). At the same time, ρ and R^2 can only lead to locally optimal designs (Moerbeek & Teerenstra, 2016, p. 203). Quantifying their (random and true) variation is consequently also of utmost relevance (e.g., Donner & Klar, 2000; Hedges et al., 2012; Jacob et al., 2010; Turner et al., 2004). However, as the research reviews in the preceding chapters showed (see Figure 7 in Chapter 1, Figure 1 in Chapter 2, and Figure 2 in Chapter 3), most existing design parameters for student achievement are limited to (1) the United States, (2) core outcome domains (mathematics, reading, science), (3) few selected sets of covariate types (domain-specific pretests and sociodemographics), combinations, and short time lags (1 to 3 years) as well lacking theoretical and empirical justification, (4) 2L-CRTs, (5) manifest analysis models, and (6) rarely report uncertainty and heterogeneity estimates (e.g., standard errors, meta-analytic CIs/PIs).

To address these gaps, I realized two complementary studies. In juxtaposition to Figure 7 in Chapter 1, Figure 1 visualizes how the present thesis contributes to the current knowledge base on design parameters for student achievement by summarizing the joint output of Study I

and II: Taken together, the studies bundle (meta-analytically integrated) design parameters for (1) four German student (sub)populations across Grades 1 to 12; (2) 21 achievement (sub)domains; (3) 11 covariate sets of varying types, combinations, and times lags, provided with concrete guidelines; (4) six RT designs, (5) manifest and latent analysis models, and (6) corresponding standard errors and (meta-analytic) CIs/PIs. These resources are complemented by a precision simulation study, a plethora of illustrative application examples as well as flow charts guiding the choice of appropriate design parameters.

Figure 1. *Overview on the Contributions of the Present Doctoral Thesis*



Note. The color code corresponds to the median ρ or (Δ)R^2 value. The number in a bubble counts the achievement (sub)domains analyzed. Outer triangles map the R^2 value (absolute) for a combination, inner triangles map the ΔR^2 value (increment) for a covariate over and above a domain-identical pretest. 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. LA = ρ and/or R^2 values are also suitable for a latent variable modeling target analysis of the treatment effect. *SE* = Standard errors and meta-analytic 95% confidence and prediction intervals of ρ and/or R^2 were reported. DEU = Germany (ISO 3166-1 ALPHA-3 classification). IP = Domain-identical pretest. CP = Cross-domain pretest (reading for STEM [i.e., mathematical-scientific] outcomes, mathematics for verbal outcomes). Gf = Fluid intelligence pretest. SC = Sociodemographic characteristics (gender, migration background, socioeconomic status).

The remainder of this chapter is organized as follows. Alongside the six major dimensions in which crucial research gaps have been identified and that are addressed in this dissertation, Section 4.1 presents the concrete contributing features broken down by Study I and II, puts the spotlight on respective key results and situates them in the existing body of research, as well as elaborates on important implications for the design of RTs.[48] Section 4.2 reflects on further vital challenges that educational researchers and psychologists face when planning RTs that seek to inform evidence-based education. Section 4.3 outlines strengths and

[48] Throughout Section 4.1, I refer to ρ and R^2 estimates as obtained from three-level models in Grades 1 to 10 and two-level models in Grades 11 to 12 for the total student population, unless otherwise stated.

limitations of the present thesis, and highlights directions for future research. Section 4.4 closes the present doctoral thesis by formulating some final concluding remarks.

4.1.1 Design Parameters Tailored to the German School Context

German Student Populations as a Whole and Within School Tracks Across the School Career

Although cross-country research suggests that design parameters are not unambiguously interchangeable across nations (e.g., Brunner et al., 2018; Kelcey et al., 2016), previous ρ and R^2 collections devoted to inform power analysis for RTs on student achievement show a clear dysbalance in the *target populations*, in favor of the U.S. student population. In particular, reliable ρ and R^2 values for 1st to 12th graders in German schools have not yet been systematically accumulated.

Studies I and II in the present dissertation used nationally representative, longitudinal data from three central large-scale assessments (NEPS, PISA, DESI) to analyze design parameters for four (sub)populations across elementary (Grades 1 to 4), lower secondary (Grades 5 to 10), and upper secondary (Grades 11 to 12) school in the German school context. Study I drew on five samples (starting cohorts 2, 3, 4 from NEPS, the follow-up of the 2003 PISA cycle, DESI). In Study II, a systematic database search was carried out to (meta-analytically) integrate empirically estimated ρ and R^2 values from the six available national probability samples (samples used in Study I plus the follow-up of the 2012 PISA cycle) which were suitable to study design parameters in Grades 1 to 12. Both studies purposefully took core features of the German school system (e.g., early onset of ability-based school type tracking) into account, by covering not only the total German student population as a whole, but also the subpopulations in the academic (i.e., "Gymnasium") and non-academic track. In Study I, I additionally adjusted the ICCs and explained variances for mean-level achievement differences between the various school types offered in German secondary education to support RT design with the populations of students attending specific school types within the non-academic track. In the following, I briefly discuss key results and implications for RT design with regard to the ICCs; the coverage of the explained variances is embedded in Section 4.1.3.

Key Results

Total Population: Between-School Achievement Differences Outweighed Between-Classroom Achievement Differences, Following a Non-Linear Trend Across the School Career. For the total German student population (i.e., analyzed as a whole), ρ_{L3} clearly surpassed ρ_{L2} in magnitude—irrespective of the grade level and the achievement domain. In concrete terms: The combined results from Studies I and II, across all 86 outcomes, indicated that typically between 14% and 37% of the total variance in student achievement was located at the school level, but only around 3% to 6% at the classroom level (see Figure 1).

Whereas ρ_{L2} remained fairly robust throughout the entire school career, ρ_{L3} hinged on the grade level. Specifically, Study I revealed that between-school achievement differences emerged considerably smaller in elementary ($Mdn(\rho_{L3}) = .11$) than secondary school ($Mdn(\rho_{L3}) = .35$; see Table 3 in Chapter 2). This pattern of results widely mirrors those previously documented for Germany (e.g., Haag & Roppelt, 2012; Knigge & Köller, 2010; see Figure 8 in Chapter 1 and Table 1 in Chapter 2). At the same time, it stands in stark contrast to the conclusions drawn from the three-level variance decompositions based on U.S. secondary school samples, where ρ_{L2} usually outweighed ρ_{L3} (Xu & Nichols, 2010; Zhu et al., 2012).

The meta-analytic integration from Study II further clarified the picture in secondary school: the large ICCs at L3 were mainly driven by variation between lower rather than upper secondary schools. Indeed, ρ_{L3} for 5th to 10th graders emerged three times larger than ρ_{L3} for 11th to 12th graders. This marked difference may reflect the generally more homogenous composition of the student body in upper secondary school where permeability is usually granted only for higher performing students. Moreover, especially the German lower secondary school system is characterized by an extensive tracking occurring at L3 via the establishment of various school types that cater to students with different ability levels. This kind of tracking induces further variation between schools.

On International Trial: German L3 ICCs as Outliers? The observed discrepancies between the German and the U.S. ICCs align well with the findings from international research suggesting considerable cross-country heterogeneity in ρ_{L3} (e.g., Brunner et al., 2018; Zopluoglu, 2012). Although achievement differences attributed to schools emerged strikingly large in Germany when compared to their equivalents registered for the United States, they still lay well below the international average (e.g., 40% in Brunner et al., 2018). Hence, in international comparison, the German ICCs at L3 are no exception. The specificity of design parameters by country may first and foremost point to differences in the structure and the

characteristics of the school systems such as the rationale for (e.g., ability-based, interest-based) or the onset and degree of tracking (Reichelt et al., 2019; Salchegger, 2016).

Subpopulations by School Track and Adjustments by School Type: Between-School Achievement Differences Appeared More Pronounced in the Non-Academic Track Than in the Academic Track and Could Largely Be Captured by Differences Between School Types. German secondary education pursues extensive school type tracking—like in many other countries (Reichelt et al., 2019; Salchegger, 2016). The analyses conducted separately by school track in Grades 5 to 12 revealed considerable deviations for ρ_{L2} and ρ_{L3}. Study I suggests that within the academic track, ρ_{L3} practically converged with ρ_{L2} at a low level. In contrast, within the non-academic track, the share of total variation at L3 was still far more pronounced than at L2. The meta-analytic results from Study II largely substantiate these findings. Recall that all school types except the most demanding "Gymnasium" were subsumed under the non-academic track. Thus, its higher L3 ICCs may to a certain degree depict remaining variation between the various less demanding school types.

This idea is supported by the fact that ρ_{L3} dramatically dropped when students' mean achievement within the (up to) five major school types in German secondary education had been taken into account: As Study I showed, school types removed around two thirds of the unconditional variance at L3. The tendency was even slightly increasing in the course of lower secondary school. This is a well-studied phenomenon pointing to so-called differential learning and developmental environments by school type structurally created through the early onset of performance-based tracking in Germany (Baumert et al., 2003, 2006; Maaz et al., 2008): students within the same school type are (intendedly) more similar in their proficiency level than students between distinct school types; and these between-school-type differences have been proven to manifest throughout educational trajectories (Baumert et al., 2003), in the vein of a Matthew effect (see e.g., Baumert et al., 2012; Ceci & Papierno, 2005). On a side note, median fractions of between-classroom differences equaled 6%, and thus, remained virtually unaffected by the adjustments (see Table 3 in Chapter 2).

Implications for RT Design

First, the school system in Germany differs markedly from those in other countries, and so do its ICCs. Thus, borrowing those estimates from, for instance, the United States will most likely lead to severely flawed power analysis.

Second, as stressed in Section 1.4.4, the variance component at the top hierarchical (i.e., school) level (together with its corresponding sample size) predominates power and precision

in any of the multilevel RT designs considered: All else being equal, the larger ρ_{L3}, the greater the efficiency loss in a multilevel RT (as compared to a single-level IRT). Hence, given the pronounced achievement differences between (especially lower secondary) German schools, it may be wise in RT design to give precedence to considerations on strategies that may cushion the respective detrimental effects (e.g., blocking, matching, stratification, or—ideally—covariate adjustment; Moerbeek & Teerenstra, 2016; Raudenbush et al., 2007).

Third, German secondary school types are associated with different aspiration levels. Since academic track students in the German "Gymnasium" represent a comparatively homogenous subpopulation (in particular with respect to their academic ability), researchers will find that it is easier to achieve adequate design sensitivity in RTs with this student body as opposed to that in the non-academic track; at least provided the absence of covariates: Of importance, greater homogeneity may also be associated with a loss of covariate performance, as the respective outcomes should demonstrate increased balance with respect to potential predictive factors (see Hedges & Hedberg, 2007). Therefore, strong covariates may even offset the expected differences in power and precision between RTs implemented within these two school tracks.

Fourth, the adjusted ρ values may be especially helpful when planning RTs to be run in a specific school type within the non-academic track as they correct for the heterogeneity engendered by achievement differences between the various school types. However, researchers should understand these values as a kind of school type average, which may not perfectly reflect the state of affairs in a specific school type of the non-academic track: Since they involve the adjustment for the academic track, they tend to underestimate ρ_{L3}. Here, sensitivity analyses using the corresponding normative distributions across design parameters from Study I should be performed.

In sum, this work once again demonstrates the limited generalizability of ρ estimates across national contexts (e.g., Brunner et al., 2018). It even testifies that ρ_{L3} values are not universally applicable in each of the various subpopulations as defined by grade and performance level. Hence, the present findings can be interpreted as another confirmation that variance estimates used for power analysis should accurately match the RT's target (sub)population. For this purpose, the present dissertation provides novel design parameters for the German student population as a whole, in the academic, and (a specific school type of the) non-academic track across Grades 1 to 12.

4.1.2 Design Parameters Matched to a Wide Array of Outcome Domains

Mathematical-Scientific, Verbal, and Domain-General Skills

Given the demands presented to student learning in the 21st century (OECD, 2018), many relevant *target outcome domains* beyond the core subjects (mathematics, reading, and at best also science) are severely underrepresented in existing compilations of ICCs and explained variances to be employed in the design of RTs on student achievement.

Studies I and II significantly enlarged this spectrum. Study I is special in amassing design parameters for 21 different subdomains (apart from the core domains, many verbal skills in German as first language and English as foreign language, multifarious domain-general outcomes such as declarative metacognition or problem solving), easily accessible in lucid interactive Excel worksheets. An important contribution of Study II is the meta-analytic integration of ρ and R^2 across eight STEM[49] and German verbal domains to inform power analysis for RTs that target multiple and/or not covered target outcomes (and grades). In the following, I shortly recapitulate key results and RT design implications with regard to the ICCs, the explained variances will be separately discussed in Section 4.1.3.

Key Results

Between-School and—Albeit to a Smaller Degree—Between-Classroom Achievement Differences Emerged Strictly Domain-Specific. Studies I and II returned that values of ρ_{L2} and ρ_{L3} clearly differed by achievement domain. Three aspects should be recorded: (a) ρ_{L3} showed larger variation across domains than ρ_{L2}. (b) The shares of the total variance in student achievement that could be attributed to differences between schools were typically most pronounced for verbal skills in English and the least pronounced for domain-general skills. (c) The median shares that could be attributed to differences between classrooms also maximized for English, but did, apart from that, not vary in a noteworthy manner across the remaining domains. These findings corroborate former research in Germany (e.g., Knigge & Köller, 2010; see Figure 8 in Chapter 1 and Table 1 in Chapter 2), in the United States (e.g., Westine et al., 2013), as well as at an international level (e.g., Brunner et al., 2018).

Domain-Specific Skill Acquisition Requires Domain-Specific Learning Opportunities. The acquisition of domain-specific skills basically hinges on the availability of suitable learning opportunities with tasks tailored to the specific domain in question (Baumert

[49] Recall that the term STEM subsumes domains of science/technology/engineering/mathematics.

et al., 2009). Hence, the variability in the magnitudes of the ICCs by achievement domain may depict differences in the schools' effectiveness in teaching certain subjects (e.g., due to differences in didactic approaches, instructional quality, or the processes determining the composition of the student and teacher body; see Snijders & Bosker, 2012, p. 35) or their prioritization of certain subjects over others (e.g., due to differences in school policies or curricula, as is the case for schools with a special focus on a certain domain such as schools with an emphasis on mathematics and science or schools with an emphasis on languages).

Implications for RT Design

First, researchers who clearly define a specific target outcome domain for their planned RT should use the best possible matching input variance estimates in power analysis that are available. Specifically, when designing RTs aimed at fostering students' foreign language skills in English with samples from Germany, evaluators should anticipate that the respective between-classroom and between-school achievement differences can be pronounced as compared to those in other domains.

Second, when the prospective RT targets multiple domains, it may be promising to conduct sensitivity power analysis. To establish plausible ranges of required sample sizes, power or precision rates, the normative design parameter distributions offered in Study I or the meta-analytic integrations accumulated in Study II may be of high practical value.

Overall, the present results coincide with former works in emphasizing that ICCs do also not generalize well across achievement domains (e.g., Brunner et al., 2018; Westine et al., 2013). Responding to this circumstance, this thesis considerably widens the range of potential target outcome domains by compiling multifarious (normatively and meta-analytically summarized) design parameters for mathematical-scientific, verbal, and domain-general contents.

4.1.3 Design Parameters and Guidelines for Covariate Adjustment

Various Covariate Types (Including Fluid Intelligence), Combinations, and Time Lags, with Selection Guidelines Based on Psychometric Heuristics

Experts and agencies have repeatedly highlighted that covariate decisions should be based on both theoretical and empirical considerations, and should ideally be preregistered (e.g., Cook, 2005; EMA, 1998, 2015; Raab et al., 2000). Yet, past research on design parameters for student achievement shows crucial gaps concerning the coverage and the selection guidance of *target covariates*. If at all, (joint) explained variances and precision-enhancing impacts were

quantified for domain-identical pretests, sociodemographic characteristics, and occasionally also cross-domain pretests. Importantly, existing works made generally no attempt to ground the choice of covariates and their hypothesized unique, incremental, and relative effects on theoretical and empirical considerations.

Study I accumulated R^2 values for the previously most often applied covariate sets (domain-identical pretests, sociodemographics, the combination of both) to replicate the former results for the German school context. Study II is the first that suggests deriving predictions on the unique, incremental, and relative covariate effects from influential psychometric heuristics. In doing so, I studied a large battery of 11 distinct covariate sets: (a) covariate types (domain-identical, cross-domain, fluid intelligence pretests, sociodemographics) as embedded in the bandwidth-fidelity dilemma theory (Cronbach & Gleser, 1957), (b) their combinations as embedded in the validity degradation concept (Sechrest, 1963), and (c) varying time lags (1 to 7 years) for the cognitive covariates as embedded in the validity degradation principle (Ghiselli, 1956; Humphreys, 1960). To evaluate how precision is actually affected by the various covariates, I carried out simulations which followed a hybrid Bayesian-classical approach to power analysis. The meta-analytically integrated results were finally used to formulate so far unique empirically supported guidelines on covariate selection when planning RTs on student achievement.

Key Results

Fidelity Covariates, Especially When Assessed in the Target Outcome Domain, Outperformed Bandwidth Covariates in Explaining Achievement Differences. Study II provides strong evidence that covariates should psychometrically and conceptually be well-aligned to the RT's target outcome in order to optimize design sensitivity. Specifically, in line with the theoretical predictions implied by the bandwidth-fidelity dilemma (Cronbach & Gleser, 1957), a narrowly measured domain-identical pretest that faithfully maps the substantive peculiarities of the outcome (Ackerman & Lohman, 2006; Hogan & Roberts, 1996; Salgado, 2017) raised precision most in an RT which addresses achievement in a specific domain. For example, when predicting reading achievement, a fidelity baseline measure of previous reading skills appeared generally superior to any divergent fidelity variable capturing antecedent cross-domain performance, or to bandwidth measures on fluid intelligence or sociodemographics.

Overall, the predominance of a domain-identical pretest over another domain-specific fidelity covariate or domain-general bandwidth covariates was proven robust across German

student (sub)populations, grade levels, achievement domains, as well as RT designs, and hierarchical levels. Consequently, Study II connects to a well-documented finding (e.g., Bloom et al., 2007; Hedges & Hedberg, 2013; Spybrook, Westine, et al., 2016; see Figure 7 in Chapter 1), expanding it to the German school context. Of note, these observations not only substantiate the bandwidth-fidelity heuristic but may also empirically support related theories such as the specificity-matching principle (Swann et al., 2007) or the well-known Brunswik summetry (Wittmann, 1988, 2011; Wittmann & Süß, 1999) which has also been explicitly discussed under an experimental lens. All these mentioned paradigms amalgamate in the common gist that maximal predictive validity necessitates the psychometric and conceptual harmonization of outcome and covariate.

The Incremental Validity of Additional Covariates Over and Above a Pretest in the Target Outcome Domain Tended to Be Small. Generally spoken, the present doctoral thesis provides little support for noteworthy incremental benefits in power and precision when accounting for additional covariates over and above a domain-identical pretest: Study I mostly documented rather small incremental returns in the empirical estimates of R^2, and Study II undergirds this pattern of results via both meta-analytic R^2 as well as simulated $MDES$ values. Thus, our findings largely mirror the state of knowledge on this topic (e.g., Hedges & Hedberg, 2007; Jacob et al., 2010; Xu & Nichols, 2010; see Figure 7 in Chapter 1). It seems reasonable to interpret the observed low incremental validities of cross-domain pretests and the usually high-dimensional fluid intelligence pretests as a reflection of their substantial intercorrelation with domain-identical pretests (Baumert et al., 2009; Cattell, 1987; Jensen, 1993; Neisser et al., 1996).[50]

The models involving the array of sociodemographics or the full battery of covariates were associated with the largest yields in R^2, first and foremost in multilevel designs at L2, but also at L3, and especially in elementary school. These observations are congruent with the expectations derived from the incremental validity concept (Sechrest, 1963; see also Haynes & Lench, 2003; Hunsley & Meyer, 2003), which may rationalize the ubiquitous, though often not questioned assumption that in RTs, additional outcome-related covariates beyond known prognostic factors (such as a baseline measure of the outcome itself) should incrementally add to the prediction of the outcome, and thus, incrementally increase design sensitivity (e.g., Bloom et al., 2007; Kahan et al., 2014; Tafti & Shmueli, 2020).

[50] Note that if the assumption of covariate-treatment orthogonality in RTs holds, any potential detrimental impacts on the estimation of the treatment effect through (multi)collinearity could theoretically be ruled out.

Importantly, even when non-negligible, the observed increments did (apart from minor exceptions) barely translate into substantial precision improvements when simulating power analysis in Study II. Perhaps, this can partly be explained by the fact that the RT sample sizes were rather large in all simulation conditions (as covariate effects tend to be stronger in small-sized RTs); however, the assumed RT sizes mimic the empirical state of affairs for extant educational large-scale RTs (see Lortie-Forgues & Inglis, 2019). Nevertheless, there was considerable heterogeneity among simulated $MDES$ values. Occasionally, incremental validities of additional covariates were indeed strikingly large, depending on the (sub)population, grade level, achievement domain, and in multilevel designs also on the hierarchical level. Due to the lack of substantive benchmarks for these increments and given their strict connection to a specific design and research context, any judgements on their actual practical significance for study design must remain conditional, and thus, are difficult to generalize to a broader application scope (Hunsley & Meyer, 2003).

Why "The More, the Better" Still Often Holds. Against the background of the results just discussed, it might seem far-fetched that it is common practice in effectiveness research to collect as much auxiliary information on the studied sample as financially, logistically, and ethically feasible (Balzer et al., 2023; Lin, 2013; Wright et al., 2015). The charm of having multifarious covariates is actually intelligible in that covariates fulfill (at least) a triple function in RTs. Apart from their potential value in raising the sensitivity to detect a (true) treatment effect by eliminating variation in the target outcome (e.g., Kahan et al., 2014; Raudenbush, 1997; Shadish et al., 2002), covariates have two further important advantages: They help to inform on the population to which the RT's findings may be generalized by creating the setup to report on the key characteristics of the sample based on which inferences on intervention effectiveness shall be drawn (Bausell & Li, 2002). Moreover, covariates often serve as mediators when exploring operating mechanisms of treatment effects (Kelcey et al., 2021; Lynch et al., 2008), as moderators when investigating variation in these treatment effects (Dong et al., 2021, 2023), or as blocking factors (Bausell & Li, 2002).

Validity Degradation Was Most Pronounced for a Pretest in the Outcome Domain and the Least Pronounced for a Pretest in Fluid Intelligence. Study II produced mixed findings on a potential validity degradation of cognitive covariates. The prognostic properties of domain-identical pretests emerged the most negatively affected from developmental dynamics across time, often showing a clear—albeit still relatively mild—simplex time series pattern (Humphreys, 1960). In its essence, this pattern of results buttress the validity degradation principle (Ghiselli, 1956; Humphreys, 1960), according to which a pretest's predictive power

should gradually decline with growing time gap to the achievement outcome. However, given the rather moderate deterioration, this conclusion best holds under a less strict interpretation of this hypothesis (see Reeve & Bonaccio, 2011). Similar tendencies were also registered previously based on U.S. samples (Bloom et al., 2007; Westine et al., 2013; Xu & Nichols, 2010; see Figure 7 in Chapter 1). Their comparability with the present findings is limited though, as these studies investigated validity decay in domain-identical pretests over only two or three years. In contrast, the explanatory power of fluid intelligence pretests (and also partly cross-domain pretests) often leaned towards striking stability over time. This corresponds well to existing single-level research in this vein (see Figure 2 in Chapter 3; see Reeve & Bonaccio, 2011, for a broader review in terms of outcomes).

The precision simulations imply that threats by validity degradation may be more pertinent to the smaller-sized RTs. Here, the detrimental effects of the small sample sizes plus the weakened amounts of explained variance cumulatively added up and led to decreased design sensitivity as compared to larger-sized RTs.

Overall, the Explanatory Power of Covariates Was Strongest at the Top Hierarchical School Level. Irrespective of the concrete covariate set, one of the most consistent findings across Studies I and II was the predictive superiority of the aggregated[51] cluster-level covariates over their (disaggregated) individual-level equivalents. Indeed, in virtually every constellation, the (latent) school and—albeit to a smaller degree—classroom means significantly outweighed both the multi- and single-level student-level scores in explaining variance in an achievement outcome (i.e., $R^2_{\mathrm{L3}} > R^2_{\mathrm{L2}} > R^2_{\mathrm{L1}}$ and $R^2_{\mathrm{L3}} > R^2_{\mathrm{L2}} > R^2_{\mathrm{T}}$). With some minor exceptions, this was true across designs, (sub)populations, grade levels, outcome domains as well as covariate sets. Yet, as the meta-analytic integrations from Study II accentuated, this pattern was most unequivocal in lower secondary school: here, pooled amounts of explained L3 variance consistently exceeded thresholds of 76%, while corresponding shares at L2 widely ranged between 12% and 67%, and at L1 always fell below 36% (see Figure 3a in Chapter 3). In both elementary and upper secondary school, similar gradations emerged, although in general somewhat less clear-cut. This general pattern has also previously been proven highly robust, both at an international level in general (e.g., Brunner et al., 2018), and in the United States in particular (e.g., Hedges & Hedberg, 2007; Jacob et al., 2010; see Figure 7 in Chapter 1), as well as in Germany (e.g., Baumert et al., 2003). Notably, the values of single-level $R^2_{\mathrm{T}} \leq .53$

[51] Recall that only cluster-level covariates aggregated from L1 covariates were analyzed. Such aggregated cluster-level variables has been referred to as contextual or analytical variables and has to be distinguished from so-called global or integral variables which represent cluster-level variables by definition (i.e., that are measured directly and cannot be disaggregated; e.g., school size; Lüdtke et al., 2008).

appeared in general slightly larger than multilevel R^2_{L1}, yet, still consistently smaller than R^2_{L3} (see Figure 3a in Chapter 3).

Inflated Cluster-Level R^2 as Statistical Artefacts Through Aggregation? The tendency of aggregated entities to intercorrelate higher than their individual-level counterparts has been discussed by several scholars for a long time (e.g., Bloom et al., 2007; Härnqvist et al., 1994; Ostroff, 1993; Raudenbush et al., 2007; W. S. Robinson, 1950; Snijders & Bosker, 2012). As early as in 1950, Robinson introduced the term "ecological correlations" to describe the phenomenon that correlations at the aggregate level will virtually always differ from those at the individual level (when based on the very same variables). He also offered mathematical proofs. Although his calculations have been criticized, mainly for being technically insufficient and partly erroneous (see Oakes, 2009; Subramanian et al., 2009; Te Grotenhuis et al., 2011 for debates and re-analyses), the conclusion still applies: (very) high R^2 values for covariate measures that were aggregated to higher hierarchical levels occur frequently (see e.g., Bloom et al., 2007; Klar & Darlington, 2004; Teerenstra et al., 2012 for some notes on this).

Of importance, from a mathematical point of view, the relations $R^2_{L3} > R^2_{L2} > R^2_{L1}$ or $R^2_{L3} > R^2_{L2} > R^2_{T}$ are neither generally true nor are necessary consequences following from aggregation (Snijders & Bosker, 2012); rather, these relations are determined by a more complex interplay of several factors, such as the correlations between the coefficients/residuals in (multilevel) models specified for both the outcome and the cluster-level covariate(s) as well as those reliabilities (for proofs, see Snijders & Bosker, 2012, pp. 32–33). Everything else being equal, higher reliabilities of both the outcome and the cluster-level covariate(s) at L2 and/or L3 should trigger higher R^2_{L2} and/or R^2_{L3} values (Bloom et al., 2007; Teerenstra et al., 2012).

The Influence of Measurement Error. It is widely acknowledged that measurement error (in both the outcome and the covariate) in general[52] negatively affects the amounts of explained variance in regression modeling (Cohen et al., 2003; Raudenbush & Bryk, 2002). This means that higher reliabilities are typically associated with higher R^2 estimates. In the multilevel context, the reliability of a cluster-level covariate is a function of its ICC and the cluster size (Raudenbush & Bryk, 2002, p. 46; Snijders & Bosker, 2012, p. 26): The reliability of an L3 covariate quickly improves with (a) increasing L2 and L3 cluster sizes, and (b) growing values of the L3 covariates' ICC. From this follows that the reliability of an L3 covariate

[52] More precisely, measurement error in the outcome always diminishes R^2, in both bivariate as well as multiple regression. In bivariate regression, this also applies when the covariate is unreliable (Cohen et al., 2003; Raudenbush & Bryk, 2002). However, in multiple regression, R^2 is *most likely* reduced with unreliable covariates; but under certain circumstances, R^2 may be inflated or even "unbiased" (see Cohen et al., 2003, pp. 121–124 for proofs and illustrations).

approaches 1 with either (a) large schools (provided that its ICC is nonzero and positive), or (b) high values of its ICC (see also Figure 3.3 in Snijders & Bosker, 2012, p. 27). The relations for L2 covariates are equivalent: The reliability of an L2 covariate is a positive function of its (positive) ICC and the L2 cluster size. Since the classroom size is smaller than the school size (as soon as there is more than one classroom per school), the reliability of an L2 covariate should also be smaller than the reliability of an L3 covariate. This might be one reason why R^2_{L2} tends to be lower than R^2_{L3}.

Implications for RT Design

In general, covariate adjustment has great potentials to boost power and precision in RTs on student achievement, and thus, addresses a major threat of statistical conclusion validity (Shadish et al., 2002). The practical implications for RT planning that can be derived from the analyses on covariate impacts are numerous and sweeping. A useful summary of general recommendations on covariate selection is given in the Section "Essentials of Covariate Adjustment in RTs on Student Achievement at a Glance" in Chapter 3. So, which are further, more integrative issues to primarily consider with covariate inclusion?

First, given that that the concrete R^2 estimates not only widely varied as a function of the covariate type, combination, and time lag, but also by (sub)population, achievement outcome, hierarchical level and so forth, researchers should take care they optimally match the specific peculiarities of the target RT when used for power analysis.

Second, the above results may perhaps create the impression that, in principle, the target outcome's baseline measure is the ultimate linchpin of any meaningful RT design to study interventions on student achievement. It is not. There are various realistic scenarios that allow for or even require the shift to other covariates. Such a strategy may be either intended (e.g., when the RTs' target outcome heavily depends on dynamic and/or cumulative developmental processes in students' learning trajectories so that there might not exist any test battery that accurately captures or yields enough variance in the abilities under investigation; Shadish et al., 2002, p. 118), or may be a logical consequence of emergent circumstances under which a particular RT is implemented (e.g., when the RT's target outcome changes during realization due to logistical factors or political decisions; Bloom et al., 2007, p. 32).

Third, the unique, incremental, and relative effects of the studied covariates on power and precision in RTs on student achievement usually intensify with small sample sizes. Hence, a problematic issue, particularly in long-term RTs, is sample attrition: Apart from the fact that sample attrition poses threats to internal and external validity, the loss of students (or their

responses) basically implies shrinkage of the effective total sample size. From the designs addressed by this thesis—unless entire schools drop out of the study—IRTs will typically suffer most from sample attrition (Rickles et al., 2018).[53]

Fourth, in multilevel RTs, especially in lower secondary school, the orientation towards the three psychometric heuristics may become somewhat less relevant given the pervasive predictive strength of cluster-level covariates that usually translated well into power and precision returns. Of note, L3 covariates can also be exclusively used in RTs, that is, without simultaneously adjusting for covariates at L1 or L2 (e.g., Bloom et al., 2007; Westine et al., 2013). There is no absolute need to introduce covariates at all hierarchical levels in any RT design (although it still remains best-practice to do so) and some planning scenarios even necessitate this strategy (e.g., when resources are firmly restricted). Administrative data from local authorities or schools themselves (provided as aggregated information, e.g., on shares of female students within a school, or representing global information, e.g., on the school size), or archive data from former cohorts (e.g., the nation-wide "Vergleichsarbeiten" [VERA]) may be easily and cheaply obtained (Bloom et al., 2007). Hence, utilizing L3 covariates derived from such data usually pledges high cost-effectiveness (see e.g., W. Li et al., 2020; Moerbeek, 2006 for discussions on cost-effectiveness under covariate inclusion), because researchers may curtail or even skip a good deal of steps during RT implementation (e.g., construction and validation of test batteries and questionnaires, multiple testing sessions).

Fifth, high reliabilities are generally advantageous for the design sensitivity in RTs. Yet, it should be noticed that even (highly) unreliable covariates still boost power and precision, compared to an RT design without covariates (Maxwell et al., 2017, p. 481); unless in—very rare—scenarios of very small RTs where the reduction in error variance is offset by the loss in degrees of freedom through too many covariates (Konstantopoulos, 2012; Liu, 2011; Moerbeek & Teerenstra, 2016).[54]

Taken together, the present doctoral thesis considerably expands the knowledge and guidance on covariate adjustment and selection in RTs on student achievement. Not only does it provide a vast collection of reliable R^2 estimates to inform power analysis and to adjust required sample sizes, or prospectively achieved power and precision rates. It also presents a so far unique theoretical framework based on the psychometric heuristics of the bandwidth-fidelity dilemma,

[53] Note that in a long-term IRT, even when small-sized, a strongly predictive (domain-identical) pretest may still level off mild harmful impacts induced by temporal validity decay. A prominent example is the Perry preschool program of noble prize winner James Heckman and colleagues (Heckman et al., 2013).

[54] Covariates—perfectly reliable or not—affect the standard error of the treatment effect, i.e., its precision, but not its estimate itself (in magnitude or direction); therefore, treatment effects are not biased through unreliable covariates (Borenstein & Hedges, 2019; Maxwell et al., 2017, pp. 471, 481; Porter & Raudenbush, 1987).

the incremental validity concept, and the validity degradation principle to derive hypotheses on the potential unique, incremental, and relative precision-enhancing impacts of a broad spectrum of varying covariate types, combinations, and time lags. Not least, a special feature of this dissertation is the hierarchical decomposition of fluid intelligence effects on student achievement which is, to the best of my knowledge, the very first of its kind.

4.1.4 Design Parameters Adapted to a Broad Range of Experimental Designs

Single-Level IRT, Two- and Three-Level CRT and MSRT Designs

So far, design parameters for several *target experimental designs* commonly implemented in applied educational and psychological research are widely lacking (e.g., Connolly et al., 2018; Spybrook, Shi, et al., 2016). Compendia largely remain restricted to estimates informative for planning two-level RT designs (students within schools).

Study I applied two- and three-level modeling to estimate ρ_{L2} and ρ_{L3} as well as R^2_{L1}, R^2_{L2}, and R^2_{L3} values. Study II used single-, two-, and three-level modeling to estimate ρ_{L2} and ρ_{L3} as well as R^2_{T}, R^2_{L1}, R^2_{L2}, and R^2_{L3}. In doing so, the present doctoral thesis significantly augments the range of designs by covering six different kinds of RTs (see Figure 4 in Chapter 1): IRTs (individually sampled students who are assumed to represent stochastically independent experimental units, that is, disregarding their classroom or school affiliation, with randomization at an individual basis), 2L-CRTs (students at L1 within schools at L3, and randomization at L3), 3L-CRTs (students at L1 within classrooms at L2 within schools at L3, and randomization at L3), 2L-MSIRTs (students at L1 within schools at L3, and randomization at L1), 3L-MSIRTs (students at L1 within classrooms at L2 within schools at L3, and randomization at L1), as well as 3L-MSCRTs (students at L1 within classrooms at L2 within schools at L3, and randomization at L2).[55]

[55] Note that, however, for MSRT designs, estimates of the treatment effect heterogeneity (i.e., the degree of variation among site-specific treatment effects) are additionally required, which are not covered by the present thesis (see also Section 1.4.4). To reliably estimate treatment effect heterogeneities, and also corresponding R^2 values (i.e., quantifying the degree to which heterogeneity can be explained by covariates), a re-analysis with data of (a pool of) MSRTs would be necessary. Weiss et al. (2017), for instance, were among the first systematically compiling such heterogeneity design parameters based on 16 MSRTs on educational and professional interventions, suitable for the U.S. context. Schochet et al. (2014) and Weiss et al. (2014) thoroughly discuss the conceptual underpinnings of treatment effect heterogeneity. The high relevance of treatment effect heterogeneity in MSRT design is further highlighted by the fact that in 2017, the Journal of Research on Educational Effectiveness devoted a whole issue (Volume 10, Issue 4) to this topic.

Key Results

Design Parameters Varied Considerably by (Hierarchical) Level. The findings from both Study I and Study II underline that the design parameter estimates hinged on the assumptions on the data's underlying variance structure, and if assumed to be clustered, emerged strictly tied to the hierarchical level. These phenomena have been proven robust across all analyses and were thoroughly discussed in several sections above: When the variances were decomposed at the various hierarchical levels, (a) differences in student achievement located at L2 usually appeared (clearly) smaller than those located at L3 (depending on the subpopulation), and (b) explained variances by L1, L2, and L3 covariates were typically smallest at L1, somewhat higher (but widely varying) at L2, and the largest at L3. When the variances were not hierarchically decomposed, R^2_{T} (meant to inform IRT designs) typically surpassed the corresponding R^2_{L1}. This should be due to the fact that the estimates of R^2_{T} conflate variation between students with variation between classrooms and schools.

The Omission of the Classroom-Level Variance Component Barely Impacted the ICCs, but Occasionally Affected the Explained Variances at the School Level. To inform the design of both 3L-CRTs/3L-MSRTs (students at L1 within classrooms at L2 within schools at L3) as well as 2L-CRTs/2L-MSIRTs (students at L1 within schools at L3), this dissertation offers values of ρ_{L3}, R^2_{L1}, and R^2_{L3} as estimated via both three-level models (i.e., treating classrooms at L2 as random effects) as well as two-level models (i.e., neglecting the random effects of classrooms at L2) for Grades 1 to 10.[56] In Study I, I explicitly juxtaposed these three- and two-level design parameters to learn about the degree of deviation in the estimates that may result from a misspecification due to the omission of L2 variance. I observed that ρ_{L3} values as estimated in two-level models exceed their three-level equivalents, which held across all (sub)populations. This is as it should be: Since the total variance is identical in both models, the variance located between classrooms not modeled shifts to both L1 and L3 instead (Moerbeek, 2004; Zhu et al., 2012).[57] However, generally, differences in ρ_{L3} were trivial. A similar pattern of results—concerning both direction and degree of the deviations—also applied to R^2_{L1}. In contrast, discrepancies in the R^2_{L3} values were more apparent. Compared to three-level models, few R^2_{L3} values were slightly overstated in two-level models; yet, most R^2_{L3} values

[56] In upper secondary school, exclusively two-level models were specified since 11th and 12th grade students are typically not grouped into intact classrooms, but courses where student compositions vary by the subject taught.

[57] Recall that $\rho_{\mathrm{L3}} = \sigma^2_{\mathrm{L3}}/\sigma^2_{\mathrm{T}}$. σ^2_{L3} consists of a true component depicting the "real" variation across schools plus an error component associated with σ^2_{L1}. The latter is underestimated in the absence of σ^2_{L2}, and consequently, the true variation in σ^2_{L3} is overestimated, resulting in larger ρ_{L3} estimates (Zhu et al., 2012, pp. 48–49).

were understated in two-level models, occasionally even substantially. These results largely replicated those previously reported based on U.S. samples (Zhu et al., 2012).

When to Rely on Two- Instead of Three-Level Design Parameters? In practice, there may be various application scenarios in which researchers may assume a two-level variance structure ignoring the variance component at L2 when planning (in reality actually three-level) RTs on student achievement. First, as the research reviews above highlighted, reliable estimates of ρ_{L2} and R^2_{L2} are still scarce. If such values were reported previously, they were limited to the application in the U.S. school context, and were largely restricted to selected grades, and few core achievement domains and covariate sets. Therefore, RT designs relying on previously generated design parameters often do not explicitly take into account the nesting of students into classrooms within schools (Jacob et al., 2010; Konstantopoulos, 2009; Zhu et al., 2012).

Second, the exact identification of clusters might not (or hardly) be feasible. For instance, sometimes the grouping of students into learning groups is not static, but rather varies, for instance, depending on the subject taught. Usually this does not eliminate the random effects at the middle level, but rather leads to a cross-classified cluster variance structure. Note that such "imperfect hierarchies" (Snijders & Bosker, 2012, p. 205) modulating cluster effects in fact routinely occur in educational settings, for a variety of reasons including the blending of students from distinct classrooms in neighborhoods, sports clubs, tutoring classes and so forth (Raudenbush & Bryk, 2002, pp. 373–375). Thus, it is—strictly speaking—implausible to categorize grouping contributions on student achievement exclusively into classroom or school effects. In German upper secondary school, students are often enrolled into courses differentiated by the aspiration level chosen for a certain subject (e.g., basic vs. advanced mathematics courses). Since there are usually limited combination options for these courses, students frequently learn within one and the same group, but not at all times. Although design parameters that accurately reflect such cross-classified structures technically represent a relatively straightforward extension (or, rather simply an internal differentiation; Raudenbush & Bryk, 2002, pp. 377–378) of the design parameters that neglect multiple group memberships, their implementation and meaning in power analysis have rarely been explicitly studied so far (but see e.g., Moerbeek & Safarkhani, 2018). Furthermore, such cross-classified ρ and R^2 estimates would even more strongly depend on the actual nature, target outcome, and specific inferential goal of the RT, and as a consequence, might miss adaptability in most application scenarios.

Third, researchers often use administrative data or official statistics to measure outcomes or covariates for their RT. Most of these records, however, lack an identifier for the

classroom a student attends (Zhu et al., 2012). Consequently, the effectiveness of the RT's intervention is planned to be evaluated by specifying a two-level model only, with students at L1 nested within schools at L3. And finally, even if such an identifier is available, the number of classrooms per school should be adequately large to specify a robust three-level model (Lee & Hong, 2021) or, at least, to be able to differentiate classroom from school effects (Opdenakker & Van Damme, 2000, p. 283). All in all, the motivations to use two- instead of three-level design parameters are manifold.

Implications for RT Design

First, researchers should be aware of the fact that the level-specific ρ and R^2 values are not interchangeable: For instance, L2 estimates cannot be used as substitute for lacking L3 estimates; likewise, R^2_{T} as obtained from single-level models should not be entered into power analysis for multilevel RTs, and so on.

Second, under some specific circumstances, the general rule of an accurate matching between design parameters and target experimental design may be relaxed: Employing values of ρ_{L3}, R^2_{L1}, and R^2_{L3} values as estimated in two-level models omitting L2 information in combination with data from samples that actually have a three-level variance structure may barely affect power analysis. At best, power might be somewhat understated when employing R^2_{L3} obtained from two-level instead of three-level models. The practical relevance of this problem might be insignificant though, given that power calculations for large-scale RTs tend to be rather optimistic than pessimistic (Konstantopoulos, 2008b; see e.g., Lortie-Forgues & Inglis, 2019; Spybrook & Raudenbush, 2009). Moreover, it has been empirically proven that L1 covariates may compensate for potentially underestimated R^2_{L3} estimates in two-level RTs (Zhu et al., 2012). Importantly, as described above, the samples used in the present studies produced rather small estimates of ρ_{L2}. The deviation between two- and three-level estimates is expected to be larger with growing ρ_{L2} (Moerbeek, 2004; Zhu et al., 2012). In either case, unless $\rho_{\mathrm{L2}} = 0$, dropping the L2 variance components from power analysis will always overestimate design sensitivity to a certain degree (Konstantopoulos, 2008b).

In summary, the present analyses reinforce calls for design parameters that accurately reflect the actual (hierarchical) variance structure of the prospective RT (e.g., Konstantopoulos, 2009; Westine, 2016). By generating single, two-, and three-level ρ and R^2, the present thesis helps to fill an important research gap to enable the planning of RTs of six different experimental designs.

4.1.5 Design Parameters Suitable for Manifest and Latent Analysis Models

Manifest and Latent Variable Modeling Frameworks

Despite the fact that the applied *target analysis models* for the test of treatment effect in psychological and educational RTs are diverse (e.g., Blanca, Alarcón, & Bono, 2018; Schochet, 2008), most former collections of ρ and R^2 values drew exclusively on manifest variable modeling; precluding potential applications of latent variable modeling techniques to analyze treatment effects in RTs (e.g., Lüdtke et al., 2008; Mayer et al., 2016).

In Study I, the multilevel variance components were estimated within the general latent variable modeling framework implemented in M*plus* (Muthén & Muthén, 2017) to compute ρ and R^2. In particular, the multilevel latent covariate models (Lüdtke et al., 2008) implied the latent aggregation of L1 covariates to their L2 and L3 cluster means, partialing out measurement error. In Study II, a manifest approach to multilevel modeling in R (R Core Team, 2023) was followed using the lme4 package (Bates et al., 2015). Therefore, in Study II, the cluster-level variables underlying the R^2_{L2} and R^2_{L3} estimates represent manifest cluster means, including measurement error. Thus, the present doctoral thesis facilitates a flexible adaption of design parameters to the concrete statistical analysis procedure to be used for mean comparisons of the outcomes between the experimental groups. Moreover, since Study II re-analyzed ρ and R^2 values for selected outcomes and covariate sets based on the same data (see also Footnote 40 in Chapter 3), it becomes possible for researchers to directly contrast ρ and R^2 values for student achievement obtained via M*plus* vs. R when planning RTs.

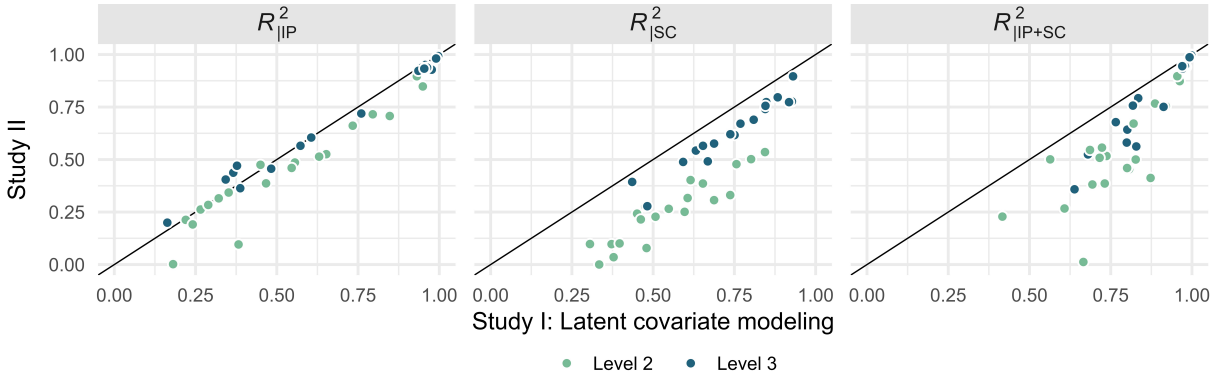
Key Results

Cluster-Level R^2 Values Based on Latent Covariate Means Clearly Surpassed Their Counterparts Based on Manifest Covariate Means. The present thesis demonstrates that values of R^2_{L2} and R^2_{L3} calculated from variance components estimated in latent models correcting for unreliability consistently exceeded their equivalents derived from manifest models in magnitude, often considerably. Figure 2 directly juxtaposes the R^2_{L2} and R^2_{L3} values for the duplicate covariate sets including a domain-identical pretest, sociodemographics, and their combination from Study I and Study II.[58] It is obvious that the decreased reliabilities of the manifest cluster means in Study II largely attenuate the amounts of explained variance estimated for all covariate sets—this is as expected given that that measurement error (in both

[58] Note that classroom means at L2 were group-mean centered at their respective L3 school means in both studies.

the outcome and the covariate) has been proven to negatively influence R^2 estimates in regression modeling (Cohen et al., 2003; Raudenbush & Bryk, 2002; see also the discussion in Section 4.1.3). Furthermore, the systematic understatement of R^2_{L2} and R^2_{L3} in Study II is exacerbated with multiple fallible covariates (i.e., for the sets involving the battery of sociodemographics). Of importance, this pattern of results emerged highly robust irrespective of the (sub)population, grade level, outcome, and concrete covariate set. Notably, R^2_{L2} values near zero in Study II can probably be explained by estimation error induced by low variance components at L2 (see Jacob et al., 2010, p. 177). On a side note, the equivalent ICCs from Study I and II practically converged, and the explained variances at L1 only occasionally showed noteworthy—albeit very rarely severe—deviations; thus, the different functioning of the M*plus* and R software did barely affect these design parameters.

Figure 2. *Comparison of Explained Variances at the Classroom (R^2_{L2}) and School Level (R^2_{L3}) As Estimated in Study I and Study II for Equivalent Covariate Models*



Note. R^2 values for the total student population as estimated in three-level models (students at L1 within classrooms at L2 within schools at L3). R^2 values below the diagonal line were higher in Study I employing latent cluster means than in Study II employing manifest cluster means. IP = Domain-identical pretest. SC = Sociodemographic characteristics (gender, migration background, socioeconomic status).

Implications for RT Design

Unless researchers intend to forego covariate adjustment, they should consider the following when applying the present design parameters to perform power analysis. First, when some sort of structural equation model is intended for the test of the treatment effect and/or M*plus* will be used, it is advisable to draw on the ρ and R^2 values provided in Study I. In such cases, these design parameters modeled within the latent variable framework serve as unbiased and valid input estimates, also probably avoiding overpowered RTs. Otherwise—which I anticipate will still be most often the case (Blanca, Alarcón, & Bono, 2018; Luo et al., 2021)—, especially

when measurement error in the outcome and covariate is not partialed out, the design parameters from Study II represent reliable values for power calculations.

Second, special caution is warranted in scenarios, where a certain covariate set to be applied in the RT was not covered in Study I, but the analysis stage will involve latent modeling. Here, the design parameters from Study II should be considered rather conservative, lower bound estimates. Nevertheless, evaluators will ideally perform sensitivity analyses around plausible ranges of design parameters; either by using CIs/PIs or by running simulations.

As a whole, this dissertation again puts the spotlight on the relevance of the target analysis model when choosing design parameters for power analysis (e.g., Ahn et al., 2020; Kleinman & Huang, 2017; Schochet, 2008). In doing so, the present work offers novel insights to which extent ρ and R^2 estimates to design RTs on student achievement may differ between manifest and latent variable modeling techniques. In particular, by applying both approaches to multilevel estimation, the present design parameters should also be helpful to adequately power tests for the treatment effect conducted outside the conventional manifest procedures, which substantively expands the potential scope of applications.

4.1.6 Quantifications of Uncertainties and Meta-Analytic Heterogeneities

Standard Errors and Meta-Analytic Heterogeneity Estimates, With Application Illustrations and Simulation Study

Although required for sensitivity analyses across power analysis outputs (e.g., Liu, 2014; Moerbeek & Teerenstra, 2016), the reporting of *uncertainties and heterogeneities* in form of standard errors or CIs for ICCs and explained variances has been rather a rare practice so far, leave alone meta-analytic PIs.[59]

In both Study I and II, standard errors and/or 95% CIs were consistently documented for all ρ and R^2 estimates. Guidance on how to use them when explicitly incorporating uncertainties or heterogeneities into power analysis was offered via manifold realistic application scenarios. Study II additionally provided 95% PIs obtained from random-effects meta-analysis. A simulation study to assess covariate impacts on statistical precision showcased how to implicitly allow for uncertainties following a Bayesian rationale when setting priors based on the joint empirical distributions of ρ and R^2.

[59] For instance, to date only Hedberg and Hedges (2014) propounded random-effects meta-analytic summaries of (within-district) ρ_{L3} values, which inform on their random (i.e., due to sampling error) and true variation shares, but which are suitable for (rather specific) RT applications with students from the United States and do likely not generalize well to the German school context.

Key Results

Statistical Uncertainties in the Design Parameters Were of Practical Significance in RT Design. Drawing on large-scale assessment data from large probability samples, most ρ and R^2 values could be estimated with an adequate degree of precision in terms of low standard errors. Notwithstanding, already small adjustments in ρ and R^2 occasionally translated to substantial shifts in the required sample size, or the expected power, or $MDES$ when performing sensitivity analysis across power analyses; a tendency that has been well-acknowledged previously (e.g., Donner & Klar, 2000; Williamson et al., 2023). Of course, this behavior aggravated when simultaneously considering uncertainties in ρ and R^2. This held true under both explicit and implicit handling of uncertainty.

A relatively consistent observation was that the statistical uncertainties in R^2_{L2} and R^2_{L3} appeared often (much) more pronounced than in R^2_{L1}. This may not come as a surprise given that the closed-form solution for the large sample variances of R^2 (see Equation (C5) in the Appendix C) exclusively involves the respective total sample sizes at L2 and L3, which are inevitably (by far) smaller than the total sample size at L1. Of note, Equation (C5) also implies that the clustered nature of the variances was neglected; other methods might have provided more exact results. Nevertheless, our approach facilitated connectivity to and comparability with the standard errors of multilevel R^2 values as previously—and so far uniquely—recorded in Hedges and Hedberg (2013). In addition, large amounts of between-imputation variance induced by high missing rates in certain measures probably further inflated the standard errors. Finally, large standard errors of R^2_{L2} may also be attributable to the very small size of most L2 variance components (see Jacob et al., 2010, p. 177).

Meta-Analytic Heterogeneities Across Design Parameters Were Considerable. One leading edge of meta-analyzing the design parameters in Study II was the opportunity to explore sources of their inherent heterogeneity. Specifically, the (multivariate multilevel) random-effects models[60] (Borenstein et al., 2021; Pustejovsky & Tipton, 2022) allowed to distinguish variation due to sampling error from true heterogeneity among ρ and R^2 values.

A highly robust finding from these pieces of analysis states that the great majority (often over 90%; see OSM F for Study II) of the total variation among design parameters did not represent sampling error but was rather attributed to true heterogeneity. This implies "considerable heterogeneity" according to the benchmarks established by Higgins (2022). This

[60] Recall that random-effects modeling was applied with at least 10 ρ or R^2 estimates to be integrated (Langan et al., 2019, p. 95); otherwise, fixed-effect modeling was applied.

may partly be a direct consequence of the high precision (in terms of low standard errors) of the ρ and R^2 estimates, but may also indirectly point to proper study designs of the large-scale assessments. Thereby, true heterogeneity was, as a rule, mainly at the level of effect sizes (within samples), which seems credible given that ρ and R^2 are expected to show larger variation by exact grade and outcome than between samples assumed to be randomly drawn from, and thus, representative for a common population of students.

The absolute magnitudes of true variation among ρ and R^2 estimates were in general substantial, translating into 95% PIs of large width (see Figure 3a in Chapter 3). This led in consequence to broad ranges of expected power when relying on the lower and upper bounds of these PIs (see Figure C46 in OSM C for Study II). Several attributes may explain the true heterogeneity among values of ρ and R^2. These should include the exact grade or age group as well as the achievement subdomain, but presumably also specific characteristics of the samples (e.g., composition, cohorts), schools (e.g., curriculum, instruction quality), or test instruments (e.g., conceptualization, psychometric properties).

Fixed-Effect vs. Random-Effects Meta-Analysis. One vital goal of the meta-analytic integration in Study II was to provide average estimates of design parameters, along with their standard errors. Both the conducted fixed-effect and the random-effects meta-analyses produced these quantities. However, these two methods critically differ in their assumptions on the underlying true design parameters—whether there is a common (ergo, "fixed") population ρ or R^2 value or a distribution of (ergo, "random") population ρ or R^2 values. This assumption determines the type of inference that can be drawn: conditional vs. unconditional (Hedges & Vevea, 1998). In a nutshell, while fixed-effect models allow for *conditional* inferences on ρ or R^2 values across those (large-scale assessment) samples actually meta-analyzed, random-effects models support *unconditional* inferences on ρ or R^2 values in a population of samples from which the samples actually meta-analyzed were drawn (Konstantopoulos & Hedges, 2019). Therefore, claims on the true heterogeneity among ρ or R^2 values are only possible on the basis of random-effects models.

Implications for RT Design

First, ignoring uncertainty in the empirically derived input parameters may severely distort power analysis (e.g., Liu, 2014). In RT planning, researchers should therefore regard sensitivity analyses as a standard procedure, realized by (explicitly or implicitly) incorporating uncertainty measures to assess plausible ranges of the desired power analysis output.

Second, the meta-analytically generated $Pooled\ \rho$ and $Pooled\ R^2$ values and corresponding standard errors are meant to inform power analysis for RTs whose target achievement outcomes differ to some extent to those covered by the present analyses (see the illustrative scenario in Chapter 3 and the Application Scenario 6 in OSM C for Study II for a didactic illustration of power analysis in such a case). Thus, this ambition anticipates some kind of generalization beyond the observed design parameters (see also Hedberg & Hedges, 2014). However, as Hedges and Vevea (1998, p. 488) pointed out, any generalization is a two-part process: the first is statistical, rationalized by sampling theory, in both fixed-effect and random-effects models; but the second is "extrastatistical" in fixed-effect models. Precisely, when using our meta-analyzed ρ or R^2 values in power analysis for a prospective RT, the first part is to generalize from the samples included in our meta-analysis to a universe of samples considered to be identical (where uncertainty is addressed by sampling error), and the second part is to generalize from the universe of samples considered to be identical to a universe of target RT samples considered to be nonidentical (albeit similar/representative; Hedges & Vevea, 1998). Random-effects meta-analysis is superior with respect to the second part of generalization because it attempts to statistically quantify this kind of uncertainty via estimates of true heterogeneity. Fixed-effect approaches cannot afford this. Nevertheless, the fixed-effect meta-analytic results that were generated in the few cases where unbiased heterogeneity parameter estimates could not be guaranteed (due to a too small number of effect sizes to be aggregated; see Langan et al., 2019, p. 95) are still reliable, and therefore also valuable for RT design, but with the restriction that they remain conditional on the NEPS, PISA, and DESI samples analyzed in the present thesis. Thus, generalizations based on these fixed-effect meta-analytic design parameters should be made with great caution and have to be accordingly justified (Hedges & Vevea, 1998).

In conclusion, quantifications of uncertainties in the ICCs and explained variances are integral fundaments of robust a priori power analysis to design RTs. This dissertation satisfies this requirement by supplying all design parameters along with their standard errors and estimating meta-analytic heterogeneity parameters. The manifold illustrative examples as well as the simulation study propounded may serve as helpful guidance on how to incorporate uncertainty into power analysis. In doing so, I addressed a sixth major gap in the current research state on design parameters for student achievement.

4.2 Further Challenges in the Design of Randomized Experiments to Reliably Inform Evidence-Based Education

Beyond the estimation of reliable ICCs and explained variances, there are plenty other issues—well familiar ones and more recently emergent ones—in the design stage of RTs on student achievement. Challenges include attrition, complexities of educational interventions, cost-effectiveness, endogeneity of design, generalizability, $MDES$ benchmarking and justification, noncompliance, reliability of measures, replication, spillover/contamination, and many more (Hedges, 2018; Maxwell et al., 2017; Moerbeek & Teerenstra, 2016; Raudenbush & Schwartz, 2020). Several of them have already been more or less addressed up to here (and for certain all of them deserve thorough examination). Next, I briefly elaborate on two topics that I think are among the most directly relevant to the present doctoral thesis, namely $MDES$ benchmarking and justification, and generalizability.

4.2.1 $MDES$ Benchmarking and Justification

"How big is big?" asked Robert Slavin (2018) in his blog post on the practical meaningfulness of effect sizes, showing an image of an oversized mouse next to a tiny elephant. The message: How big is big? *It depends.* To be precise, it depends on the reference point whether a mouse appears large or an elephant small, and so it also depends on the research context under investigation how small, medium, or large a (standardized) treatment effect observed for certain intervention is deemed. This crucial discernment has been reiterated by numerous scholars and experts in the field (Baird & Pane, 2019; Bloom, Hill, et al., 2008; Brunner, Stallasch, et al., 2023; C. J. Hill et al., 2008; Konstantopoulos & Hedges, 2008; Kraft, 2020; Lipsey et al., 2012; Valentine, 2019), even Cohen (1988) himself whose rules of thumb for the interpretation of effect sizes—which are not the only ones, but certainly the most prominent—are still granted priority in many educational and psychological studies (C.-Y. J. Peng et al., 2013).

Identifying a reasonable magnitude of the $MDES$, or degree of statistical precision, therefore is as an integral as demanding task in power analysis (see Section 1.4.3). When assessing the meaningfulness of the $MDES$, various rationales may (simultaneously) play a role (see Bloom, 2006; Brunner et al., 2018; Schochet, 2008, for thorough examinations). First, a cost-effectiveness or economic rationale would prioritize an $MDES$ that translates into (monetary) earnings large enough to offset the costs of the RT (see Schochet, 2008, p. 66, for a computational illustration).

Second, a programmatic rationale would prefer an $MDES$ that is attainable given the nature of the intervention and the specific context of the RT, including the target population and outcome. Attainability may be oriented towards previous empirical evidence quantifying treatment effects for similar RTs. If available, meta-analyses may offer promising reference points. However, as Brunner et al. (2018) cautioned, it is essential to carefully compare the treatment protocols, to study the metric of the reported pooled effect size, and to take into account potential discrepancies in the variances of the outcome measure.

Third, a political rationale would favor an $MDES$ that satisfies the demands and expectations of policymakers and other important stakeholders in education. For instance, ministries may be interested in how much the intervention may improve students' learning outcomes as measured by their typical or expected annual achievement growth. Or, they may expect from a program to be implemented in daily school routine that it facilitates closing the achievement gaps between relevant demographically defined student groups (e.g., in terms of gender, migration background, and socioeconomic status) or between weak- and average-performing schools. Respective benchmarks have been made available for the United States (e.g., Bloom, Hill, et al., 2008; C. J. Hill et al., 2008; Konstantopoulos & Hedges, 2008; Lipsey et al., 2012; Scammacca et al., 2015). Recently, we added a vast compilation of meta-analytically summarized effect size benchmarks for various achievement outcomes across the entire school career from Grade 1 to 12 in the German school system to this body of knowledge (Brunner, Stallasch, et al., 2023). A common key result from the U.S. works and our study was the large heterogeneity of effect size benchmarks across subpopulations (e.g., as defined by school types), grade levels, and achievement domains, emphasizing that the substantive relevance of a chosen $MDES$ is strongly tied to the specific research context in question. For example, according to our analysis for the student population in Germany, a desired $MDES = .20$ for an RT to enhance students' reading achievement would, on average, translate into a learning gain of almost one and a half years from Grade 5 to 6 in lower secondary school, but only into a learning gain of half a year from Grade 11 to 12 in upper secondary school (Brunner, Stallasch, et al., 2023, Table 1). Likewise, when related to mean differences in mathematics achievement between weak and average German schools, the gap could be closed for upper secondary schools, but not so for elementary and lower secondary schools if the RT's mathematics intervention produced a standardized treatment effect of $d = .20$ (Brunner, Stallasch, et al., 2023, Table 5).

To conclude, when designing an RT on student achievement, setting a reasonable precision level to compute the required sample size or the prospectively achieved power rate,

or evaluating the attainable precision level given a specified sample size and power in terms of the $MDES$ requires researchers' substantive knowledge on the practical importance of a (standardized) treatment effect size and involves adopting economic, programmatic, and/or political rationales.

4.2.2 Generalizability

As stressed throughout this dissertation, educational RTs seek to provide a basis for policy and practice to decide on whether to launch or continue, or to cancel programs in real-life schooling (Stuart et al., 2017). In doing so, an RT usually tests a particular intervention in a sample drawn from some relevant population.[61] In other words: it is intended to generalize the results of an RT beyond the sample studied to a certain target population (which may include the sample or not; Raudenbush & Schwartz, 2020; Shadish et al., 2002). However, this endeavor is anything but trivial.

Under this notion[62] of "causal generalization" (Shadish et al., 2002), and more precisely, external validity, two crucial premises collude: (a) the target population has to be well-defined (O'Muircheartaigh & Hedges, 2014; Tipton & Olsen, 2018), and (b) treatment effects have to be assumed to vary (Hedges, 2018). It is not reasonable to discuss the generalizability of RTs without specifying to whom the results should be generalized (e.g., an RT on a university preparation training may generalize to students in the academic track but not to students in less demanding school types of the non-academic track). Nor is the question of generalizability relevant at all if one does not expect that the intervention may function differentially (e.g., a new teaching method works fine in one school but not in another school).

In regard to the first point: The target population may be defined rather narrowly (e.g., as for a [small-scale] efficacy [single-level] RT, see also Section 1.3.1) or rather broadly (e.g., as for a [large-scale] effectiveness [multilevel] RT; Stuart et al., 2011; Tipton & Olsen, 2018)—the challenges and implications associated with complex multilevel designs have been intensively discussed above. Ideally, the definition of the target population then guides the recruitment of those sample units that support the desired generalizations (Tipton et al., 2014; Tipton & Olsen, 2018). Importantly, although the definition of the target population should be oriented towards the rationale which students, classrooms, or schools will finally be affected

[61] There are also RTs relying on dual- or multiple-frame sampling with the aim to generalize to two or multiple target populations, for instance, when addressing two or several different research questions (Hedges, 2018).

[62] Some have discussed generalizability rather through the lens of the underlying mechanisms of the intervention, focusing on how a treatment may work (e.g., Deaton & Cartwright, 2018). I follow Raudenbush and Schwartz (2020) and conceive generalizability as a matter of the sampling process and the transferability and applicability of results from an RT sample to a specified target population.

by decisions made based on the findings of the RT (Tipton & Olsen, 2018), there are also examples of RTs whose conclusions have been used to inform program decisions for populations that were actually not targeted by RT (see Hedges, 2018).

In regard to the second point—the anticipation of treatment effect heterogeneity: The implausibility of a contrary assumption (i.e., that treatment effects are static or homogeneous) has also been empirically proven. For instance, Weiss et al. (2017) found that treatment effects show considerable heterogeneity across schools. Here, MSRTs are highly promising tools to tackle generalizability: As noted above, an MSRT can be seen as replicating an intervention study multiple times, that is, once in each site (Liu, 2014). This not only lays the foundation for meta-analysis but also gives a formal test of generalizability across the varying settings of the included sites (Raudenbush & Liu, 2000), whereby the more heterogeneous the sites are, the more generalizability may be improved (Bloom & Spybrook, 2017).

In the past, there has been quite a bit of concern about the generalizability and the extrapolations of results from educational RTs, partly questioning their widely praised value to shape evidence-based policies and practices (Deaton & Cartwright, 2018; Morrison, 2020; D. H. Robinson et al., 2013; Sullivan, 2011; Thomas, 2016). One major plea that has been raised is that most large-scale RTs in education (but this holds also true for other areas, such as medicine or health service; see M. J. Campbell & Walters, 2014; Eldridge & Kerry, 2012) are based nonrandom convenience samples (Stuart et al., 2017) selected under relevant criteria of restriction and inclusion (e.g., the number of students per school as determined through a priori power analysis; Tipton et al., 2014). Bell and Stuart (2016), for example, argue that bias in external validity (and also in the results from the RTs) through non-representative sample selection may reach an intolerable degree, at least when applying criteria for internal validity.

Raudenbush and Schwartz (2020) recapitulated three currently followed design-based approaches to tackle generalizability under sample selection, whose common logic is to link the RT data to some kind of extern auxiliary data to model (a) the selection process via (stratified) weighting schemes using propensity score methods (Kern et al., 2016; O'Muircheartaigh & Hedges, 2014; Stuart et al., 2001, 2011; Tipton, 2013; Tipton et al., 2014), or—though so far less prevalent—(b) the outcome of the RT given the covariates using (Bayesian) response surface methods (J. Hill et al., 2020; Kern et al., 2016), or (c) both sampling and outcome using doubly robust methods (Kern et al., 2016). The auxiliary data often stem from large-scale assessments, other surveys, or administrative or census records (Tipton & Olsen, 2018). Pivotal for either approach, these should reflect the inference population *and*

contain the same covariates as the RT (Raudenbush & Schwartz, 2020).[63] Put simply, the overarching goal is to minimize observed discrepancies in the compositions of the RT sample and the target population (Kern et al., 2016; Tipton et al., 2014). In a related strand of research, methodologists then also established techniques and measures for quantifying generalizability in terms of the accuracy of the predictions based on RTs and their extrapolations to the populations which will probably be affected by political or practical decisions on the educational innovations, products, and services in question (Orr et al., 2019; Stuart et al., 2011; Tipton, 2014).

Finally, causal generalizations from RTs basically rest on assumptions not directly testable with the data; therefore calls for stronger emphasis on sensitivity analyses have been brought forward (Raudenbush & Schwartz, 2020; see e.g., Nguyen et al., 2017, 2018). Furthermore, the above mentioned techniques to estimate treatment effects with convenience samples hinge on (strong) assumptions on treatment effect heterogeneity and its correlates (Hedges, 2018). Therefore, a focal mission of experimental education research remains the exploration of the context characteristics that either favor or hamper impacts of interventions, preferably via strong, and ideally via representative MSRTs (e.g., Yeager et al., 2019). Bryan et al. (2021) go even that far to call for a "heterogeneity revolution."

4.3 Strengths, Limitations, and Future Directions

Central Strengths

Strong Databases. With the high-quality and ample data of several national probability samples from three German large-scale assessments (NEPS, PISA, DESI), the present thesis capitalizes on the strongest database to generate design parameters for achievement outcomes of 1st to 12th graders, thus, across the entire school career, which accurately map the specific features of the (tracked) German school system. In particular, in Study II, it was possible to take full advantage of the available longitudinal datasets that are suitable for tackling the objectives of this dissertation, as identified by a systematic search.

State-of-the-Art Methods. I consistently applied state-of-the-art methods to handle the special challenges associated with the analysis goals. Specifically, to properly reflect the data's

[63] This also points to the assumption of sampling ignorability (Tipton & Olsen, 2018): The generalizability hinges on the extent to which the covariates explain treatment effect variation in the population.

underlying cluster sampling strategy, I used (group-wise) multilevel multiple imputation to handle missing data; advanced (latent) multilevel modeling to estimate the (un)conditional variance components; and non-parametric cluster bootstrapping to obtain robust standard errors of the unconditional variance components. Further, (multivariate multilevel) meta-analysis and meta-regression allowed reliable and generalizable syntheses (see Findley et al., 2021) of the results based on individual participant data, while accounting for stochastic dependencies in the design parameters. Moreover, precision simulations following a hybrid Bayesian-classical approach to power analysis facilitated a proper implicit allowance of design parameter uncertainties via priors representing the joint empirical distribution of the ICCs and explained variances.

Large Multiplicity. Given the extensive range of student samples and grades, as well as the so far broadest diversity of achievement domains and largest variety of covariate sets, which were involved to estimate ρ and R^2 values for several RT designs by means of both latent and manifest (multilevel) modeling techniques, the present design parameters are highly robust, reliable, and versatile in use. Taking the perspective of a "critical multiplism" sensu Shadish (1993), the strategy followed in the present dissertation helps to prevent constant, unidirectional bias, which increases the credibility of the results. At the same time, it substantively broadens the scope of applications in RT design. Thus, the present dissertation helps to fill many important gaps of previous research on design parameters for student achievement, from both an international perspective (see Figure 7 in Chapter 1 by contrast with Figure 1 in Chapter 4), as well as specifically a German perspective (Ständige Wissenschaftliche Kommission der Kultusministerkonferenz, 2022).

Rich Output. This work aims at contributing to an improved quality and rigor of RTs on student achievement in the German school context by directly supporting researchers in designing their studies. Therefore, this thesis has developed (a) lucid, interactive Excel workbooks amassing broadly applicable design parameters (broken down by population, achievement domain, grade, and enriched with flow charts guiding the choice of the appropriate set of estimates), and (b) thorough guidance for power analysis in general (via multifarious illustrative planning scenarios and simulations) and covariate selection in particular (via concrete guidelines inspired by three influential psychometric heuristics). In doing so, this thesis strives to build a bridge between the research in the methodological underpinnings of RT design and applied experimental research.

Overall Limitations

The results of the present doctoral thesis should be interpreted in the light of several shortcomings. Most of the more (study-)specific ones among these were addressed in the respective limitation sections in Studies I and II. I therefore now focus on overall limitations relevant for both studies.

Generalizability I. The narrow definition and high specificity of the generated design parameters (e.g., in terms of the target population or target outcome) come at the cost of generalizability to other contexts. The present estimates are most suitable for RTs carried out in the German school system. Yet, they may also support the planning of RTs in school systems that share vital characteristics with the German one (e.g., early onset of ability-based school type tracking, as is the case in, e.g., Austria, Czech Republic, Hungary, Slovakia, or Turkey; Reichelt et al., 2019; Salchegger, 2016), when more appropriate design parameters are lacking.

The same logic applies to RTs that draw on measures that deviate from those analyzed: the poorer the design parameters' match to the target measure of the intervention, the worse their adequacy for respective power analysis. Mismatch may occur with divergent test instruments (e.g., with different psychometric properties), but should be more severe with substantively divergent measures of achievement, such as school grades instead of standardized tests (Borghans et al., 2016; Brookhart, 2015). With this in mind, it is best practice to perform sensitivity analyses to pointedly take into account uncertainty and heterogeneity in the design parameters (see e.g., Liu, 2014; Moerbeek & Teerenstra, 2016). Of course, a realistic appraisal of applied experimental research anticipates the complexity and diversity in prospective RTs; thus, slight deviations between empirical estimates of ρ and R^2 and the context features of the target RTs may be natural. However, there might be situations where these deviations are more severe (e.g., when an intervention aims at enhancing student achievement in a domain not analyzed in the present thesis). My hope is that, in such cases, the normative distributions from Study I or the meta-analytic estimates from Study II can serve as valuable approximations. In either case, such aggregates should still be more reliable values than conventional (e.g., Cohen, 1988), quite arbitrary benchmarks (LeBreton & Senter, 2008). To better judge which kind of estimates may be most optimal for a certain planning purpose, I invite researchers to consult the flow charts enclosed in each study.

Generalizability II. A related issue, but from an opposite perspective: Sometimes, the target context defined for the compiled design parameters might be too broad. Consider, for example, a prospective RT implemented within a specific school type within the non-academic track in

German secondary education. Although it would have been, in principle, possible to estimate school-type specific ρ and R^2 values, the potential value of such more specific design parameters is subject to an accuracy-uncertainty trade-off (Hedges & Hedberg, 2007): Although school-type specific design parameters are expected to demonstrate less bias in reflecting the distinctive peculiarities of a respective student population, lower sample sizes induce greater uncertainty in the estimates (i.e., higher standard errors). Thus, the accuracy of the more specific estimates may be offset by their higher uncertainty. Similarly, the range restrictions associated with more selective populations (other examples include populations in a certain federal state in Germany, impoverished regions, or low-performing schools) may also result in diminished covariate-outcome correlations, translating into attenuated statistical power rates (Miciak et al., 2016). When designing RTs with more homogenous samples, the compiled ρ and R^2 values should therefore be interpreted rather as upper bound estimates.

Assumptions of Power Analysis. The applied power analysis formulae to determine the required sample size, power, or $MDES$ for two-sample independent t-tests hinge on strong assumptions. One of them is balance in sample allocation, meaning that the TG and the CG are equal in sample size, and in multilevel RTs, additionally fixed cluster sizes. With varying cluster sizes, statistical power has been demonstrated to decrease (everything else held constant, and as long as $0 \leq \rho \leq 1$), where the more pronounced the variation in cluster size, the greater the efficiency loss (e.g., Lauer et al., 2015; van Breukelen et al., 2007). Another assumption is homoscedasticity, that is, the TG and the CG share a common variance of the target outcome. A (mild) violation of homoscedasticity, however, may be more the rule than the exception (especially with comparably heterogeneous populations, as is typically the case in educational and psychological research; Blanca, Alarcón, Arnau, et al., 2018; Delacre et al., 2017; P. Thompson et al., 2023), and may occur due to non-normality (which is then often solvable through score transformations; Aberson, 2019), but also due to treatment effect heterogeneity (Bloom, 2005; Bryk & Raudenbush, 1988; P. Thompson et al., 2023): For instance, when a reading program shows greater effects for poor readers, the error variance at the individual level in the TG can decrease, relative to the CG. Likewise, in multilevel RTs, the reading program may exert greater influences in high-performing schools, probably decreasing the respective school-level variance component in the TG. Power is affected by heteroscedasticity (i.e., unequal variances) because it can flaw the treatment effects' standard error (and thus, also the $MDES$; Bloom, 2005), inflating Type I error rates (Aberson, 2019; Bryk & Raudenbush, 1988; P. Thompson et al., 2023). In balanced designs, statistical tests for the treatment effect estimate that assume common variances have been shown robust even under heteroscedasticity at either

level (Blanca, Alarcón, Arnau, et al., 2018; Gail et al., 1996; Korendijk et al., 2008). Of note, this does not hold for unbalanced designs, though.[64] For all power analyses in this dissertation, I consistently assumed completely balanced designs (i.e., equal sample sizes in the TG and the CG, and in multilevel designs, fixed cluster sizes); hence, the results of the illustrative application scenarios as well as of the precision simulations may be most valid for prospective RT that (at least roughly) mimic these design characteristics, even when homoscedasticity is violated.

Outcome Reliability. Most power analysis tools and software solutions implicitly assume that the RT's target outcome demonstrates perfect reliability (Cox & Kelcey, 2019), and so do the present power analyses carried out with the R package PowerUpR (Bulus et al., 2021).[65] As already alluded in Section 4.1.3, as is the case for covariates, low reliabilities of the outcomes can also negatively influence the design sensitivity in RTs (Cohen et al., 2003; Cox & Kelcey, 2019; Raudenbush & Bryk, 2002; Raudenbush & Sadoff, 2008). For 2L-CRTs and 2L-MSIRTs, Cox and Kelcey (2019) also showed that measurement error in the target outcome may undermine the efficiency of conventional optimal sampling allocations. As a consequence, the power analysis illustrations may depict RT planning under rather ideal circumstances of highly reliable outcome measures. It should be noted, however, that fallibility in the outcomes is, at least partly, accounted for through the design parameters from Study II as estimated within a manifest rather than latent variable modeling framework, indirectly incorporating measurement error (at all hierarchical levels) into power analysis (Cox & Kelcey, 2019).

Deviating Nesting Structures. The present design parameter compendia cover RT designs with three sampling schemes of up to three hierarchical levels (i.e., students within classrooms within schools). The offered ρ and R^2 values, however, may not be appropriate for designing multilevel RTs with other nesting structures (e.g., students within teachers within schools or

[64] Gail et al. (1996) presented simulations showing that heteroscedasticity distorts the t-tests in unbalanced CRT designs, both when the variance in the TG is larger than in the CG (if the sample size in the CG surpass the sample size in the TG) as well as when the variance in the TG is smaller than in the CG (if the sample size in the TG surpass the sample size in the CG). The extent of inferential error hinges on the degrees of imbalance in the sample sizes between the TG and the CG and their variances. Nevertheless, the authors claim that this problem may be of less practical relevance since the differences in the variances between the TG and the CG are usually not expected to be as large (at least in clinical trials; Gail et al., 1996). In cases where it is not feasible or less cost-efficient to attain balance in sample allocation to the experimental conditions (e.g., when the treatment is expensive and it would be cheaper to assign more units to the CG), researchers may adjust for heteroscedasticity; for this, several methods and tests have been proposed (see e.g., Aberson, 2019; Blanca, Alarcón, Arnau, et al., 2018; Bloom, 2005 for some listings). Delacre et al. (2017) even call for the use of Welch's t-test (instead of Student's t-test) as the default procedure.

[65] Note that PowerUpR currently allows the specification of an individual-level outcome reliability coefficient in power computations for 2L-CRTs and 2L-MSIRTs as based on the formulae given in Cox and Kelcey (2019); however, so far, a congruent strategy is implemented neither for IRTs nor the various three-level RT designs.

students within schools within districts; Shen et al., 2023; Spybrook, Westine, et al., 2016). Nevertheless, the six different RT designs considered in this dissertation apply to those RT designs which are most frequently implemented to evaluate interventions on student achievement (Connolly et al., 2018; Spybrook, Shi, et al., 2016; Spybrook & Raudenbush, 2009).

Future Directions

Increasing the Diversity of Design Parameters. Educational interventions are diverse in their goals (Morrison, 2020), and so are the RTs that test their actual impact (Connolly et al., 2018; Lortie-Forgues & Inglis, 2019; Spybrook, Shi, et al., 2016; Spybrook & Raudenbush, 2009). Therefore, design parameters for many other targets are required. These include further populations and age groups (e.g., preschool children, university students, teachers), non-cognitive outcomes (e.g., socio-emotional, behavioral, professional characteristics), covariate sets (e.g., motivational predictors, global variables such as school size or instructional quality), experimental designs and hierarchical levels (e.g., stepped-wedge or cross-classified designs, four- or even five-level designs with students nested within classrooms nested within teachers nested within schools nested within cities/regions), as well as innovative, auspicious analysis models and statistical procedures for estimating treatment effects (e.g., latent repeated measures ANOVA; Langenberg et al., 2022; or the "EffectLiteR" approach to latently model conditional, interindividual treatment effect differences; Mayer et al., 2020).

Providing Design Parameters for Binary Outcomes. Many educational outcomes are binary in nature (e.g., obtaining a certain certificate or not) or are dichotomized from on a continuous scale (e.g., achieving a certain proficiency standard or not, having a low to medium frustration tolerance or not). To plan RTs targeted at binary outcomes, researchers need appropriate design parameters (e.g., in log odds), which are widely lacking to date (but see Schochet, 2013 for an exception), again, especially for the German school context.

Studying Treatment Effect Heterogeneity. Most educational RTs use methods such as stratification or blocking (Spybrook, Shi, et al., 2016). To plan such MSRTs, researchers heavily rely on reliable estimates of treatment effect heterogeneity, which are widely lacking so far (but see Weiss et al., 2017 for an exception), especially for the German school context. Hence, to generate and accumulate these values will be one of the pressing issues for future research. Apart from this, studying variation in the impacts of interventions is of high scientific and practical relevance it its own right.

Upgrading Simulation-Based Sensitivity Analysis Methods. It cannot be overemphasized that sensitivity analyses across a range of plausible design parameter estimates are key to rigorous RT design; and simulation-based methods within a hybrid Bayesian-classical approach to power analysis making use of (empirically informed) priors are highly promising developments to tackle uncertainty inherent in ρ and R^2 (Pek & Park, 2019; Spiegelhalter et al., 2004; Williamson et al., 2023; as well as in effect sizes, see Du & Wang, 2016). However, the adoption of such approaches in applied experimental research is comparatively rare. This is probably due to the fact that little practical guidance exists on how to properly implement these techniques in RT planning. Therefore, it would be helpful if methodologists could provide such resources in the future. Relatedly, recent innovations in this area should be expanded further, for instance, by building on works that integrate multiple ρ and R^2 estimates by specifying commensurate priors (Turner et al., 2005; Zheng et al., 2023), by diligently elaborating on the plausibility of distributional assumptions for the priors, or by also embracing fully Bayesian approaches to power analysis to simulate posterior densities of design parameters as well as of the various power analysis outcomes themselves, both with continuous (see e.g., Spiegelhalter, 2001) and binary (see e.g., Turner et al., 2001) outcomes.[66]

Building an RT Infrastructure in Germany. Germany lacks a firmly established infrastructure that couples support for the design of RTs with the dissemination of the resulting evidence, perhaps based on the model of the IES together with the WWC in the United States. Probably, such an endeavor would also contribute to a better connection between methodological advancements and applied evaluation research. This also includes enabling even more easy access to versatile design parameter compilations, for instance via an online tool, similar to the "Variance Almanac" built by Hedges and Hedberg (2023). The latter aspect seems important since even though funding agencies stress the need to carefully document relevant estimates from RTs (e.g., Education Endowment Foundation, 2022; German Research Foundation, 2022; Institute of Education Sciences, 2023), comprehensive summaries of the data from these reports are not regularly made available to researchers who are possibly planning similar research, at least in Germany.

[66] Generally, Bayesian approaches to power analysis increasingly gain ground (Kruschke, 2010); promising developments encompass, for instance, Bayes factor design analysis to determine sample size (Schönbrodt & Wagenmakers, 2018).

4.4 Conclusion

"A power analysis is only as good as the formulae and parameter estimates that are used (…) Without good estimates, power analysis is only guesswork" (Murray, 1998, pp. 349–350), as claimed at the very beginning of this doctoral thesis. Indeed, power analysis is by no means a surefire success for strong experimental study designs. Exactly in this spirit, this dissertation pursued the systematic analysis of reliable and versatile design parameters to plan meaningful randomized experiments on student achievement. In doing so, I generated extensive compendia of ρ and R^2 estimates, enriched with nuanced guidance to deliberately optimize power analysis.

My overarching endeavor was to support evaluators in realizing sensitive—adequately powered and precise—RT designs that allow for conclusive, unbiased, and valid causal inferences on the impact of innovative programs, novel developments, and promising services devoted to foster student learning.

The quintessence of the analyses at hand is that the design parameter estimates to rely on when determining required sample sizes, attainable power, or precision rates should mirror as accurately as possible the context features and idiosyncrasies of the prospective RT. To be precise: the RT's target population and achievement outcome domain, the (preregistered) covariates, as well as its planned experimental design and statistical analysis model. The design parameters considerably varied across all these dimensions. Hence, a close match is crucial.

The emerging materials are appropriate to design robust single- and multilevel RTs with several German and similar student (sub)populations across the entire school career, multifaceted achievement (sub)domains, varied covariate sets, diverse single- and multilevel experimental designs, as well as manifest and latent analysis models. When the fit between design parameters and target RT is less than perfect (which I anticipate will virtually always be the case to some degree), the additionally supplied quantifications of the estimates' uncertainties and heterogeneities facilitate expedient (simulation-based) sensitivity analyses.

I hope that the resources accumulated in my doctoral thesis function as valuable toolkits for power analysis in RT design, and thus, contribute to the quality and rigor of our randomized experiments in psychology and education. Because these studies represent the indispensable fundaments of wise and profound decisions in evidence-based policies and practices whose ultimate aim is to improve schooling, and therefore, every students' personal life as well as the whole societies' prosperity.

References

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences* (2nd Edition). Routledge, Taylor & Francis Group.

Ackerman, P. L., & Lohman, D. F. (2006). Individual differences in cognitive function. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (2nd ed., pp. 139–161). Lawrence Erlbaum Associates.

Ahn, C., Heo, M., & Zhang, S. (2020). *Sample size calculations for clustered and longitudinal outcomes in clinical research* (First issued in paperback). CRC Press.

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher*, *48*(4), 217–228. https://doi.org/10.3102/0013189X19848729

Balzer, L. B., Van Der Laan, M., Ayieko, J., Kamya, M., Chamie, G., Schwab, J., Havlir, D. V., & Petersen, M. L. (2023). Two-stage TMLE to reduce bias and improve efficiency in cluster randomized trials. *Biostatistics*, *24*(2), 502–517. https://doi.org/10.1093/biostatistics/kxab043

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

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

Baumert, J., Nagy, G., & Lehmann, R. (2012). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary schools?: Cumulative advantages in reading and math? *Child Development*, *83*(4), 1347–1367. https://doi.org/10.1111/j.1467-8624.2012.01779.x

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and the formation of differential learning and developmental environments]. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* (pp. 95–188). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90082-7_4

Baumert, J., Trautwein, U., Artelt, C., Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten—Institutionelle Bedingungen des Lehrens und Lernens [School contexts—Institutional conditions for teaching and learning]. In Deutsches PISA-Konsortium, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 261–331). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-97590-4_11

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

Bell, S. H., & Stuart, E. A. (2016). On the "where" of social experiments: The nature and extent of the generalizability problem. *New Directions for Evaluation*, *2016*(152), 47–59. https://doi.org/10.1002/ev.20212

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

Blanca, M. J., Alarcón, R., & Bono, R. (2018). Current practices in data analysis procedures in psychology: What has changed? *Frontiers in Psychology, 9*, 2558. https://doi.org/10.3389/fpsyg.2018.02558

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., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289–328. https://doi.org/10.1080/19345740802400072

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., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness, 10*(4), 877–902. https://doi.org/10.1080/19345747.2016.1271069

Borenstein, M., & Hedges, L. V. (2019). Effect sizes for meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine, *Handbook of research synthesis and meta-analysis* (3rd ed., pp. 207–243). Russell Sage Foundation.

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

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences, 113*(47), 13354–13359. https://doi.org/10.1073/pnas.1601135113

Brookhart, S. M. (2015). Graded achievement, tested achievement, and validity. *Educational Assessment, 20*(4), 268–296. https://doi.org/10.1080/10627197.2015.1093928

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*, 1–39. https://doi.org/10.1080/19345747.2023.2175753

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

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104*(3), 396–404. https://doi.org/10.1037/0033-2909.104.3.396

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

Bundesministerium für Bildung und Forschung (Ed.). (2018). *Rahmenprogramm empirische Bildungsforschung [Framework program educational research]*.

https://www.empirische-bildungsforschung-bmbf.de/img/Rahmenprogramm%20empirische%20Bildungsforschung_barrierefrei_NEU(1).pdf

Campbell, M., Grimshaw, J., Steen, N., & Changing Professional Practice in Europe Group (EU BIOMED II Concerted Action). (2000). Sample size calculations for cluster randomised trials. *Journal of Health Services Research & Policy*, *5*(1), 12–16. https://doi.org/10.1177/135581960000500105

Campbell, M. J., & Walters, S. J. (2014). *How to design, analyse and report cluster randomised trials in medicine and health related research*. John Wiley & Sons.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. North-Holland ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co.

Ceci, S. J., & Papierno, P. B. (2005). The rhetoric and reality of gap closing: When the "have-nots" gain but the "haves" gain even more. *American Psychologist*, *60*(2), 149–160. https://doi.org/10.1037/0003-066X.60.2.149

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. (Eds.). (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). L. Erlbaum Associates.

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, K., & Kelcey, B. (2019). Optimal design of cluster- and multisite-randomized studies using fallible outcome measures. *Evaluation Review*, *43*(3–4), 189–225. https://doi.org/10.1177/0193841X19870878

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

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

Dong, N., Kelcey, B., & Spybrook, J. (2021). Design considerations in multisite randomized trials probing moderated treatment effects. *Journal of Educational and Behavioral Statistics*, *46*(5), 527–559. https://doi.org/10.3102/1076998620961492

Dong, N., Kelcey, B., & Spybrook, J. (2023). Experimental design and power for moderation in multisite cluster randomized trials. *The Journal of Experimental Education*, 1–17. https://doi.org/10.1080/00220973.2023.2226934

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

Du, H., & Wang, L. (2016). A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate Behavioral Research*, *51*(5), 589–605. https://doi.org/10.1080/00273171.2016.1191324

Education Endowment Foundation. (2022). *Statistical analysis guidance for EEF evaluations*. https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1698086955

Eldridge, S., & Kerry, S. (2012). *A practical guide to cluster randomised trials in health services research: Eldridge/A practical guide to cluster randomised trials in health services research*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119966241

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

Findley, M. G., Kikuta, K., & Denly, M. (2021). External validity. *Annual Review of Political Science*, *24*(1), 365–393. https://doi.org/10.1146/annurev-polisci-041719-102556

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<1069::AID-SIM220>3.0.CO;2-Q

German Research Foundation (Ed.). (2022). *Proposal preparation instructions. Project proposals*. https://www.dfg.de/formulare/54_01/54_01_en.pdf

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

Haag, N., & Roppelt, A. (2012). Der Ländervergleich im Fach Mathematik [National Assessment Study in Mathematics]. In P. Stanat, H. A. Pant, K. Böhme, & D. Richter (Eds.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011 [Competencies of students at the end of 4th grade in German and mathematics. Results of the National Assessment Study 2011]* (pp. 117–127). Waxmann. https://www.iqb.hu-berlin.de/bt/LV2011/Bericht

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

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, *103*(6), 2052–2086. https://doi.org/10.1257/aer.103.6.2052

Hedberg, E. C., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics results from a meta-analysis of district-specific values. *Evaluation Review*, *38*(6), 546–582. https://doi.org/10.1177/0193841X14554212

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. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

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. (2023). *Variance Almanac (VA) of Academic Achievement* [Computer software]. Center for Advancing Research & Communication. https://arcdata.uchicago.edu/search.php

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2022). *Cochrane handbook for systematic reviews of interventions. Version 6.3 (updated February 2022)*. Cochrane. www.training.cochrane.org/handbook

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hill, J., Linero, A., & Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application, 7*(1), 251–278. https://doi.org/10.1146/annurev-statistics-031219-041110

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<627::AID-JOB2828>3.0.CO;2-F

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

Institute of Education Sciences. (2023). *Education research grants program. Request for applications.* (ALN: 84.305A). https://ies.ed.gov/funding/pdf/2021_84305A.pdf

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

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

Kelcey, B., Xie, Y., Spybrook, J., & Dong, N. (2021). Power and sample size determination for multilevel mediation in three-level cluster-randomized trials. *Multivariate Behavioral Research, 56*(3), 496–513. https://doi.org/10.1080/00273171.2020.1738910

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness, 9*(1), 103–127. https://doi.org/10.1080/19345747.2015.1060282

Klar, N., & Darlington, G. (2004). Methods for modelling change in cluster randomization trials. *Statistics in Medicine, 23*(15), 2341–2357. https://doi.org/10.1002/sim.1858

Kleinman, K., & Huang, S. S. (2017). Calculating power by bootstrap, with an application to cluster-randomized trials. *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *4*(1), 32. https://doi.org/10.13063/2327-9214.1202

Knigge, M., & Köller, O. (2010). Effekte der sozialen Zusammensetzung der Schülerschaft [Impact of the social classroom composition of schools]. In O. Köller, M. Knigge, & B. Tesch, *Sprachliche Kompetenzen im Ländervergleich [Verbal competencies in the National Assessment Study]* (pp. 227–244). Waxmann.

Konstantopoulos, S. (2008). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, *1*(1), 66–88. https://doi.org/10.1080/19345740701692522

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation Review*, *33*(4), 335–357. https://doi.org/10.1177/0193841X09337991

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

Konstantopoulos, S., & Hedges, L. V. (2008). How large an effect can we expect from school reforms? *Teachers College Record: The Voice of Scholarship in Education*, *110*(8), 1611–1638. https://doi.org/10.1177/016146810811000803

Konstantopoulos, S., & Hedges, L. V. (2019). Statistically analyzing effect sizes: Fixed- and random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine, *Handbook of research synthesis and meta-analysis* (3rd ed., pp. 245–280). Russell Sage Foundation.

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

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kruschke, J. K. (2010). Bayesian data analysis. *WIREs Cognitive Science*, *1*(5), 658–676. https://doi.org/10.1002/wcs.72

Kultusministerkonferenz. (2016). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Wolters Kluwer. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

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

Langenberg, B., Helm, J. L., & Mayer, A. (2022). Repeated measures ANOVA with latent variables to analyze interindividual differences in contrasts. *Multivariate Behavioral Research*, *57*(1), 2–19. https://doi.org/10.1080/00273171.2020.1803038

Lauer, S. A., Kleinman, K. P., & Reich, N. G. (2015). The effect of cluster size variability on statistical power in cluster-randomized trials. *PLOS ONE*, *10*(4), e0119074. https://doi.org/10.1371/journal.pone.0119074

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lee, E., & Hong, S. (2021). Adequate sample sizes for a three-level growth model. *Frontiers in Psychology*, *12*, 685496. https://doi.org/10.3389/fpsyg.2021.685496

Li, W., Dong, N., & Maynard, R. A. (2020). Power analysis for two-level multisite randomized cost-effectiveness trials. *Journal of Educational and Behavioral Statistics*, *45*(6), 690–718. https://doi.org/10.3102/1076998620911916

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

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research. http://eric.ed.gov/?id=ED537446

Liu, X. S. (2011). The effect of a covariate on standard error and confidence interval width. *Communications in Statistics - Theory and Methods*, *40*(3), 449–456. https://doi.org/10.1080/03610920903391337

Liu, X. S. (2014). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. Taylor&Francis. http://site.ebrary.com/id/10801501

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

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

Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting practice in multilevel modeling: A revisit after 10 years. *Review of Educational Research*, *91*(3), 311–355. https://doi.org/10.3102/0034654321991229

Lynch, K. G., Cary, M., Gallop, R., & Ten Have, T. R. (2008). Causal mediation analyses for randomized trials. *Health Services and Outcomes Research Methodology*, *8*(2), 57–76. https://doi.org/10.1007/s10742-008-0028-9

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, *2*(2), 99–106. https://doi.org/10.1111/j.1750-8606.2008.00048.x

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective* (Third edition). Routledge.

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

Mayer, A., Zimmermann, J., Hoyer, J., Salzer, S., Wiltink, J., Leibing, E., & Leichsenring, F. (2020). Interindividual differences in treatment effects based on structural equation models with latent variables: An EffectLiteR tutorial. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(5), 798–816. https://doi.org/10.1080/10705511.2019.1671196

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. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, *39*(1), 129–149. https://doi.org/10.1207/s15327906mbr3901_5

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

Moerbeek, M., & Safarkhani, M. (2018). The design of cluster randomized trials with random cross-classifications. *Journal of Educational and Behavioral Statistics*, *43*(2), 159–181. https://doi.org/10/gd4q3b

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

Morrison, K. (2020). *Taming randomized controlled trials in education: Exploring key claims, issues and debates* (1st ed.). Routledge. https://doi.org/10.4324/9781003042112

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

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide*. Muthén & Muthén.

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

Nguyen, T. Q., Ackerman, B., Schmid, I., Cole, S. R., & Stuart, E. A. (2018). Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PLOS ONE*, *13*(12), e0208795. https://doi.org/10.1371/journal.pone.0208795

Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., & Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, *11*(1). https://doi.org/10.1214/16-AOAS1001

Oakes, J. M. (2009). Commentary: Individual, ecological and multilevel fallacies. *International Journal of Epidemiology*, *38*(2), 361–368. https://doi.org/10.1093/ije/dyn356

O'Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *63*(2), 195–210. https://doi.org/10.1111/rssc.12037

Opdenakker, M.-C., & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement*, *11*, 103–130.

Organisation for Economic Co-operation and Development. (2007). *Evidence in education: Linking research and policy*. OECD Publishing. https://doi.org/10.1787/9789264033672-en

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

Orr, L. L., Olsen, R. B., Bell, S. H., Schmid, I., Shivji, A., & Stuart, E. A. (2019). Using the results from rigorous multisite evaluations to inform local policy decisions. *Journal of Policy Analysis and Management*, *38*(4), 978–1003. https://doi.org/10.1002/pam.22154

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology*, *78*(4), 569–582. https://doi.org/10.1037/0021-9010.78.4.569

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

Pellegrini, M., & Vivanet, G. (2021). Evidence-based policies in education: Initiatives and challenges in Europe. *ECNU Review of Education*, *4*(1), 25–45. https://doi.org/10.1177/2096531120924670

Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review*, *25*(2), 157–209. https://doi.org/10.1007/s10648-013-9218-2

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

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: expanding the range of working models. *Prevention Science*, *23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

R Core Team. (2023). *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. (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., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). SAGE Publications, Inc.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. https://doi.org/10.1037/1082-989X.5.2.199

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., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, *1*(2), 138–154. https://doi.org/10.1080/19345740801982104

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. https://doi.org/10/cn6grx

Reichelt, M., Collischon, M., & Eberl, A. (2019). School tracking and its role in social reproduction: Reinforcing educational inheritance and the direct effects of social origin. *The British Journal of Sociology*, *70*(4), 1323–1348. https://doi.org/10.1111/1468-4446.12655

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

Robinson, D. H., Levin, J. R., Schraw, G., Patall, E. A., & Hunt, E. B. (2013). On going (way) beyond one's data: A proposal to restrict recommendations for practice in primary educational research journals. *Educational Psychology Review*, *25*(2), 291–302. https://doi.org/10.1007/s10648-013-9223-5

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

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

Scammacca, N. K., Fall, A.-M., & Roberts, G. (2015). Benchmarks for expected annual academic growth for students in the bottom quartile of the normative distribution. *Journal of Research on Educational Effectiveness*, *8*(3), 366–379. https://doi.org/10.1080/19345747.2014.952464

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

Schochet, P. Z. (2013). Statistical power for school-based RCTs with binary outcomes. *Journal of Research on Educational Effectiveness*, *6*(3), 263–294. https://doi.org/10.1080/19345747.2012.725803

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods*. Institute of Education Sciences (IES). https://ies.ed.gov/ncee/pubs/20144017/pdf/20144017.pdf

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

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

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New Directions for Program Evaluation*, *1993*(60), 13–57. https://doi.org/10.1002/ev.1660

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

Shen, Z., Curran, F. C., You, Y., Splett, J. W., & Zhang, H. (2023). Intraclass correlations for evaluating the effects of teacher empowerment programs on student educational outcomes. *Educational Evaluation and Policy Analysis*, *45*(1), 134–156. https://doi.org/10.3102/01623737221111400

Slavin, R. E. (2018, April 12). Effect sizes: How big is big? *Robert Slavin's Blog*. https://robertslavinsblog.wordpress.com/2018/04/12/effect-sizes-how-big-is-big/

Slavin, R. E., Cheung, A. C. K., & Zhuang, T. (2021). How could evidence-based reform advance education? *ECNU Review of Education*, *4*(1), 7–24. https://doi.org/10.1177/2096531120976060

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

Spiegelhalter, D. J. (2001). Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine*, *20*(3), 435–452. https://doi.org/10.1002/1097-0258(20010215)20:3<435::AID-SIM804>3.0.CO;2-E

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

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the institute of education sciences. *Educational Evaluation and Policy Analysis*, *31*(3), 298–318. https://doi.org/10.3102/0162373709339524

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

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *2*(1), 15. https://doi.org/10.1177/2332858415625975

Ständige Wissenschaftliche Kommission der Kultusministerkonferenz (Ed.). (2022). *Entwicklung von Leitlinien für das Monitoring und die Evaluation von Förderprogrammen im Bildungsbereich. Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz (SWK) [Development of guidelines for the monitoring and evaluation of funding programs in education. Impulse paper of the Standing Conference of the Ministers of Education and Cultural Affairs]*. https://doi.org/10.25656/01:26147

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of school districts that participate in rigorous national educational evaluations. *Journal of Research on Educational Effectiveness*, *10*(1), 168–206. https://doi.org/10.1080/19345747.2016.1205160

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2001). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Subramanian, S. V., Jones, K., Kaddour, A., & Krieger, N. (2009). Revisiting Robinson: The perils of individualistic and ecologic fallacy. *International Journal of Epidemiology*, *38*(2), 342–360. https://doi.org/10.1093/ije/dyn359

Sullivan, G. M. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Education*, *3*(3), 285–289. https://doi.org/10.4300/JGME-D-11-00147.1

Swann, W. B., Chang-Schneider, C., & Larsen McClarty, K. (2007). Do people's self-views matter? Self-concept and self-esteem in everyday life. *American Psychologist*, *62*(2), 84–94. https://doi.org/10/cm3gpc

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

Te Grotenhuis, M., Eisinga, R., & Subramanian, S. (2011). Robinson's ecological correlations and the behavior of individuals: Methodological corrections. *International Journal of Epidemiology*, *40*(4), 1123–1125. https://doi.org/10.1093/ije/dyr081

Teerenstra, S., Eldridge, S., Graff, M., Hoop, E., & Borm, G. F. (2012). A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in Medicine*, *31*(20), 2169–2178. https://doi.org/10.1002/sim.5352

Thomas, G. (2016). After the gold rush: Questioning the "gold standard" and reappraising the status of experiment and randomized controlled trials in education. *Harvard Educational Review*, *86*(3), 390–411. https://doi.org/10.17763/1943-5045-86.3.390

Thompson, P., Owen, K., & Hastings, R. P. (2023). Examining heterogeneity of education intervention effects using quantile mixed models: A re-analysis of a cluster-randomized controlled trial of a fluency-based mathematics intervention. *International Journal of Research & Method in Education*, 1–16. https://doi.org/10.1080/1743727X.2023.2215699

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*(3), 239–266. https://doi.org/10.3102/1076998612441947

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501. https://doi.org/10.3102/1076998614558486

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*(1), 114–135. https://doi.org/10.1080/19345747.2013.831154

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/gd32gj

Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine*, *20*(3), 453–472. https://doi.org/10.1002/1097-0258(20010215)20:3<453::AID-SIM803>3.0.CO;2-L

Turner, R. M., Thompson, S. G., & Spiegelhalter, D. J. (2005). Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials*, *2*(2), 108–118. https://doi.org/10.1191/1740774505cn072oa

Turner, R. M., Toby Prevost, A., & 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

Valentine, J. C. (2019). Interpreting effect sizes. In A. M. Aloe & S. J. Wilson, *Handbook of research synthesis and meta-analysis* (3rd ed., pp. 433–452). Russell Sage Foundation.

van Breukelen, G. J. P., Candel, M. J. J. M., & Berger, M. P. F. (2007). Relative efficiency of unequal *versus* equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*, *26*(13), 2589–2603. https://doi.org/10.1002/sim.2740

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808. https://doi.org/10.1002/pam.21760

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876. https://doi.org/10.1080/19345747.2017.1300719

Westine, C. D. (2016). Finding efficiency in the design of large multisite evaluations: Estimating variances for science achievement studies. *American Journal of Evaluation*, *37*(3), 311–325. https://doi.org/10.1177/1098214015624014

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*(6), 490–519. https://doi.org/10.1177/0193841X14531584

Williamson, S. F., Tishkovskaya, S. V., & Wilson, K. J. (2023). *Hybrid sample size calculations for cluster randomised trials using assurance* (arXiv:2308.11278). arXiv. http://arxiv.org/abs/2308.11278

Wittmann, W. W. (1988). Multivariate reliability theory. Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 505–560). Plenum Press.

Wittmann, W. W. (2011). Principles of symmetry in evaluation research with implications for offender treatment. In T. Bliesener, A. Beelmann, & M. Stemmler (Eds.), *Antisocial behavior and crime. Contributions of developmental and evaluation research to prevention and intervention* (pp. 357–368). Hogrefe.

Wittmann, W. W., & Süß, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik symmetry. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and*

individual differences: Process, trait, and content determinants. (pp. 77–108). American Psychological Association. https://doi.org/10.1037/10315-004

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

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

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., … Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369. https://doi.org/10.1038/s41586-019-1466-y

Zhang, Q., Spybrook, J., Kelcey, B., & Dong, N. (2023). Foundational methods: Power analysis. In *International Encyclopedia of Education(Fourth Edition)* (pp. 784–791). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10088-0

Zheng, H., Jaki, T., & Wason, J. M. S. (2023). Bayesian sample size determination using commensurate priors to leverage preexperimental data. *Biometrics*, *79*(2), 669–683. https://doi.org/10.1111/biom.13649

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2012). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Educational Evaluation and Policy Analysis, 34*(1), 45–68. https://doi.org/10.3102/0162373711423786

Zopluoglu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(1), 233–270.

Appendix A: Variance Inflation Factor in a Two-Stage Clustered Sample

This appendix introduces the variance inflation factor for two-stage clustered samples and the respective effective sample size derived from it (see also e.g., Hedges & Rhoads, 2010).

Let Y be an achievement outcome, which is normally distributed with variance σ_T^2. For a simple random sample of N students, the variance of the sample mean is: $Var(\bar{Y}) = \sigma_T^2/N$. Now consider a two-stage (i.e., clustered) random sample of the same total sample size N, but where first, $k \in \{1, 2, \dots, K\}$ schools, and second, within each school the same number of $i \in \{1, 2, \dots, n_k\}$ students are selected, so that $N = Kn$.

There are two important differences between this clustered sample and the simple sample described before. First, although the total sample size is still N, there may be various possible configurations of Kn (e.g., $K = 50$ and $n = 10$ results in $N = 500$, but also $K = 5$ and $n = 100$; see Hedges & Rhoads, 2010). Second as students mutually influence each other in the context of school-specific attributes, norms and standards (Kreft & de Leeuw, 1998; Murray, 1998), their scores resemble each other when they belong to the same school (Donner & Klar, 2000). Statistically, this implies a correlated error structure (Kreft, 1993; Schochet, 2008). This correlation between individuals within clusters, or put differently, the degree of between-group differences, is expressed by the intraclass correlation (ICC) ρ. These two properties of a clustered sample are reflected in the variance inflation factor (VIF; Donner et al., 1981; in survey sampling theory best known as the "design effect"; Kish, 1965):

$$\text{VIF} = 1 + (n_k - 1)\rho \tag{A1}$$

A derivation of the VIF is, for example, given in M. J. Campbell and Walters (2014, p. 23).

The sampling variance of the mean in a clustered sample has to be multiplied by the VIF: $Var(\bar{Y}) = [\sigma_T^2/N][1 + (n_k - 1)\rho]$. This way, the VIF can be conceived as the ratio of the variance in a clustered random sample of size $N = Kn$ to the variance in a simple random sample of the same size N. Recall that this adjustment of $Var(\bar{Y})$ through the VIF in its basic form of Equation (A1) assumes constant cluster sizes, all well as neither covariate adjustment, nor other balancing techniques such as stratification, blocking, or matching, and Y being a continuous (or also binary) response measure (Eldridge & Kerry, 2012). It follows that as soon as $\rho > 0$, the correctly computed variance of the sample mean (i.e., by taking the hierarchical structure into account) is always larger with a clustered than a simple random sample. For instance, for $N = 500$, $K = 50$, $n_k = 10$, and $\rho = .01/.25/.50$, the variance of the sample mean increases by 9%/225%/450%.

Note that this very same VIF also augments the required sample size (M. J. Campbell & Walters, 2014); or equivalenty, decreases the effective sample size in a clustered sample. The effective sample size is computed as (Snijders & Bosker, 2012, Equation 3.18):

$$N_{\text{effective}} = \frac{Kn}{\text{VIF}} \tag{A2}$$

For the example of $K = 50$ and $n_k = 10$, with $\rho = .25$, the effective sample size is $N_{\text{effective}} = (50 * 10) \,/\, 1 + (10 - 1) * .25 \approx 154$. Thus, the clustered sample with $Kn = 500$ is equivalent to a simple random sample of $N = 154$ students.

With covariates, the VIF may be adjusted by a factor $1 - R^2$ sensu Teerenstra et al. (2012) in an ANCOVA framework (see also Hayes & Moulton, 2017); however, notably with the restriction that ρ is assumed to be unadjusted, making this correction most useful when a researcher wishes to adjust the required sample size of a CRT for baseline covariates subsequent to having corrected for the VIF (M. J. Campbell & Walters, 2014; Teerenstra et al., 2012).

Appendix B: Statistical Models of the Experimental Designs

This appendix deals with the statistical formulations of the six RT designs for which the present doctoral thesis has generated appropriate design parameters to be used in power analysis (see Figure 4 in Chapter 1). The discussion is limited to two-arm RTs with one single TG and one single CG, although, in principle, extensions to multi-arm RTs are possible (see e.g., Liu, 2014a). The models yield estimates of the average treatment effect on an achievement outcome Y (i.e., $\bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}}$; Bloom, 2006). For simplicity, I assume that TG and CG share a common variance in Y (i.e., $\sigma^2_{\mathrm{TG}} = \sigma^2_{\mathrm{CG}} = \sigma^2_{\mathrm{T}}$; see also Bloom, 2006). See Section 4.3 in this dissertation, or also Bloom (2005), for a discussion of the implications when $\sigma^2_{\mathrm{TG}} \neq \sigma^2_{\mathrm{CG}}$.

For each design, first the expression for an unadjusted model that does not contain any covariates is given, which I refer to as unconditional model. Second, the expression for an adjusted model that does contain one or more covariates is given, which I refer to as conditional model. For multilevel designs, both all models are formulated in a combined form as well as decomposed at the various hierarchical levels.

In the MSRTs, the sites can be treated as random or fixed (see Dong & Maynard, 2013; Schochet, 2008). This choice basically affects statistical power through the noncentrality parameter, and determines whether the RT can be generalized to a superpopulation of sites (in the case of random site effects) or only to the sites actually included in the RT (in the case of fixed site effects; Spybrook & Raudenbush, 2009).

Individually Randomized Trial

Suppose that $i \in \{1, 2, \ldots, N\}$ students are sampled independently of each other (i.e., disregarding any grouping into classrooms and schools) and are randomly assigned to the TG or CG (see Figure 4a in Chapter 1). For such an IRT, the unconditional single-level model can be written as (Bloom, 2006, Equation 12):

$$Y_i = \beta_0 + \psi_1 T_i + \mathrm{e}_i \tag{B1}$$

Y_i is the achievement of the ith student and T_i is the treatment indicator of the ith student, with $T_i = .50/{-}.50$ for students in the TG/CG. β_0 is the intercept and ψ_1 is the treatment effect (i.e., $\widehat{\psi}_1 = \bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}}$). e_i is the residual of the ith student, with $e_i \sim N(0, \sigma^2_{\mathrm{T}})$, where σ^2_{T} is the total variance of Y within experimental groups.

Adding $q_{\mathrm{T}} \in \{1, 2, \ldots, Q_{\mathrm{T}}\}$ covariates C_{T} yields the conditional single-level model (Bloom, 2006, Equation 13):

$$Y_i = \beta_0 + \psi_1 T_i + \sum_{q_{\mathrm{T}}=1}^{Q_{\mathrm{T}}} \beta_{q_{\mathrm{T}}} C_{\mathrm{T}_{q_{\mathrm{T}}i}} + e_i \tag{B2}$$

$\beta_{q_{\mathrm{T}}}$ is the coefficient of the q_{T}th covariate C_{T} of the ith student. $e_i \sim N(0, \sigma^2_{\mathrm{T}|\mathrm{C_T}})$, where $\sigma^2_{\mathrm{T}|\mathrm{C_T}}$ is the covariate-adjusted total variance of Y.

Two-Level Cluster-Randomized Trial

Suppose that $i \in \{1, 2, \ldots, n_k\}$ students at L1 are nested within $k \in \{1, 2, \ldots, K\}$ schools at L3 and schools are randomly assigned to the TG or CG. Therefore, schools are treated as random effects. Since treatment allocation occurs at the top hierarchical school level, the treatment is *not* crossed with these random school effects; rather, schools are nested within experimental conditions (Liu, 2014b; see Figure 4b in Chapter 1). For such a 2L-CRT, the unconditional two-level model can be written as (Bloom, 2006, Equation 17):

$$Y_{ik} = \pi_{00} + \psi_{01} T_k + u_{0k} + e_{ik} \,, \tag{B3}$$

where

L1: $$Y_{ik} = \beta_{0k} + e_{ik} \tag{B4}$$

L3: $$\beta_{0k} = \pi_{00} + \psi_{01} T_k + u_{0k} \tag{B5}$$

Y_{ik} is the achievement outcome of the ith student in the kth school and T_k is the treatment indicator of the kth school, with $T_k = .50/-.50$ for schools in the TG/CG. β_{0k} is the intercept of the kth school, π_{00} is the grand mean, and ψ_{01} is the treatment effect (i.e., $\widehat{\psi}_{01} = \bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}}$). e_{ik} is the residual of the ith student in the kth school, with $e_{ik} \sim N(0, \sigma^2_{\mathrm{L1}})$, where σ^2_{L1} is the between-student-within-school variance of Y at L1. u_{0k} is the residual of the kth school, with $u_{0k} \sim N(0, \sigma^2_{\mathrm{L3}})$, where σ^2_{L3} is the between-school variance of Y at L3.

Adding $q_{\mathrm{L1}} \in \{1, 2, \ldots, Q_{\mathrm{L1}}\}$ covariates C_{L1} at L1 and $q_{\mathrm{L3}} \in \{1, 2, \ldots, Q_{\mathrm{L3}}\}$ covariates C_{L3} at L3 yields the conditional two-level model (Dong & Maynard, 2013, pp. 50–51):

$$Y_{ik} = \pi_{00} + \psi_{01} T_k + \sum_{q_{\mathrm{L3}}=1}^{Q_{\mathrm{L3}}} \pi_{0 q_{\mathrm{L3}}+1} C_{\mathrm{L3}_{q_{\mathrm{L3}} k}} + \sum_{q_{\mathrm{L1}}=1}^{Q_{\mathrm{L1}}} \pi_{q_{\mathrm{L1}} 0} C_{\mathrm{L1}_{q_{\mathrm{L1}} ik}} + u_{0k} + e_{ik} \tag{B6}$$

where

L1: $$Y_{ik} = \beta_{0k} + \sum_{q_{\mathrm{L1}}=1}^{Q_{\mathrm{L1}}} \beta_{q_{\mathrm{L1}} k} C_{\mathrm{L1}_{q_{\mathrm{L1}} ik}} + e_{ik} \tag{B7}$$

L3: $$\begin{gathered} \beta_{0k} = \pi_{00} + \psi_{01} T_k + \sum_{q_{\mathrm{L3}}=1}^{Q_{\mathrm{L3}}} \pi_{0 q_{\mathrm{L3}}} C_{\mathrm{L3}_{q_{\mathrm{L3}} k}} + u_{0k} \\ \beta_{q_{\mathrm{L1}} k} = \pi_{q_{\mathrm{L1}} 0} \end{gathered} \tag{B8}$$

$\beta_{q_{\mathrm{L1}} k}$ is the coefficient of the q_{L1}th covariate C_{L1} of the ith student in the kth school and $\pi_{0 q_{\mathrm{L3}}}$ is the coefficient of the q_{L3}th covariate C_{L3} of the kth school. $e_{ik} \sim N(0, \sigma^2_{\mathrm{L1}|\mathrm{C}_{\mathrm{L1}}})$, where $\sigma^2_{\mathrm{L1}|\mathrm{C}_{\mathrm{L1}}}$ is the covariate-adjusted between-student-within-school variance of Y. $u_{0k} \sim N(0, \sigma^2_{\mathrm{L3}|\mathrm{C}_{\mathrm{L3}}})$, where $\sigma^2_{\mathrm{L3}|\mathrm{C}_{\mathrm{L3}}}$ is the covariate-adjusted between-school variance of Y.

Three-Level Cluster-Randomized Trial

Suppose that $i \in \{1, 2, \ldots, n_{jk}\}$ students at L1 are nested within $j \in \{1, 2, \ldots, J_k\}$ classrooms at L2 which are, in turn, nested within $k \in \{1, 2, \ldots, K\}$ schools at L3 and schools are randomly assigned to the TG or CG (see Figure 4c in Chapter 1). For such a 3L-CRT, the unconditional three-level model can be written as:

$$Y_{ijk} = \pi_{000} + \psi_{001} T_k + u_{00k} + r_{0jk} + e_{ijk} \tag{B9}$$

where

L1: $$Y_{ijk} = \beta_{0jk} + e_{ijk} \tag{B10}$$

L2: $$\beta_{0jk} = \gamma_{00k} + r_{0jk} \tag{B11}$$

L3: $$\gamma_{00k} = \pi_{000} + \psi_{001} T_k + u_{00k} \tag{B12}$$

Y_{ijk} is the achievement outcome of the ith student in the jth classroom in the kth school and T_k is the treatment indicator of the kth school, with $T_k = .50/-.50$ for schools in TG/CG. β_{0jk} is the intercept of the jth classroom in the kth school, γ_{00k} is the intercept of school k, π_{000} is the grand mean, and ψ_{001} is the treatment effect (i.e., $\widehat{\psi}_{001} = \bar{Y}_{\mathrm{TG}} - \bar{Y}_{\mathrm{CG}}$). e_{ijk} is the residual of the ith student in the jth classroom in the kth school, with $e_{ijk} \sim N(0, \sigma^2_{\mathrm{L1}})$, where σ^2_{L1} is the between-student-within-classroom variance of Y. r_{0jk} is the residual of the jth classroom in the kth school, with $r_{0jk} \sim N(0, \sigma^2_{\mathrm{L2}})$, where σ^2_{L2} is the between-classroom-within-school variance of Y. u_{00k} is the residual of the kth school with $u_{00k} \sim N(0, \sigma^2_{\mathrm{L3}})$, where σ^2_{L3} is the between-school variance of Y.

Adding $q_{\mathrm{L1}} \in \{1, 2, \ldots, Q_{\mathrm{L1}}\}$ covariates C_{L1} at L1, $q_{\mathrm{L2}} \in \{1, 2, \ldots, Q_{\mathrm{L2}}\}$ covariates C_{L2} at L2, and $q_{\mathrm{L3}} \in \{1, 2, \ldots, Q_{\mathrm{L3}}\}$ covariates C_{L3} at L3 yields the conditional three-level model (Dong & Maynard, 2013, p. 51):

$$Y_{ijk} = \pi_{000} + \psi_{001}T_k + \sum_{q_{\mathrm{L3}}=1}^{Q_{\mathrm{L3}}} \pi_{00q_{\mathrm{L3}}} C_{\mathrm{L3}q_{\mathrm{L3}}k} + \sum_{q_{\mathrm{L2}}=1}^{Q_{\mathrm{L2}}} \pi_{0q_{\mathrm{L2}}0} C_{\mathrm{L2}q_{\mathrm{L2}}jk} + \sum_{q_{\mathrm{L1}}=1}^{Q_{\mathrm{L1}}} \pi_{q_{\mathrm{L1}}00} C_{\mathrm{L1}q_{\mathrm{L1}}ijk} + u_{00k} + r_{0jk} + e_{ijk}\,, \tag{B13}$$

where

L1:
$$Y_{ijk} = \beta_{0jk} + \sum_{q_{\mathrm{L1}}=1}^{Q_{\mathrm{L1}}} \beta_{q_{\mathrm{L1}}jk} C_{\mathrm{L1}q_{\mathrm{L1}}ijk} + e_{ijk} \tag{B14}$$

L2:
$$\begin{aligned} \beta_{0jk} &= \gamma_{00k} + \sum_{q_{\mathrm{L2}}=1}^{Q_{\mathrm{L2}}} \gamma_{0q_{\mathrm{L2}}k} C_{\mathrm{L2}q_{\mathrm{L2}}jk} + r_{0jk} \\ \beta_{q_{\mathrm{L1}}jk} &= \gamma_{q_{\mathrm{L1}}0k} \end{aligned} \tag{B15}$$

L3:
$$\begin{aligned} \gamma_{00k} &= \pi_{000} + \psi_{001}T_k + \sum_{q_{\mathrm{L3}}=1}^{Q_{\mathrm{L3}}} \pi_{00q_{\mathrm{L3}}} C_{\mathrm{L3}q_{\mathrm{L3}}k} + u_{00k} \\ \gamma_{0q_{\mathrm{L2}}k} &= \pi_{0q_{\mathrm{L2}}0} \\ \gamma_{q_{\mathrm{L1}}0k} &= \pi_{q_{\mathrm{L1}}00} \end{aligned} \tag{B16}$$

$\beta_{q_{\mathrm{L1}}jk}$ is the coefficient of the q_{L1}th covariate C_{L1} of the ith student in the jth classroom in the kth school, $\gamma_{0q_{\mathrm{L2}}k}$ is the coefficient of the q_{L2}th covariate C_{L2} of the jth classroom in the kth school, and $\pi_{00q_{\mathrm{L3}}}$ is the coefficient of the q_{L3}th covariate C_{L3} of the kth school. $e_{ijk} \sim N(0, \sigma^2_{\mathrm{L1|C_{L1}}})$, where $\sigma^2_{\mathrm{L1|C_{L1}}}$ is the covariate-adjusted between-student-within-classroom variance of Y. $r_{0jk} \sim N\left(0, \sigma^2_{\mathrm{L2|C_{L2}}}\right)$, where $\sigma^2_{\mathrm{L2|C_{L2}}}$ is the covariate-adjusted between-classroom-within-school variance of Y. $u_{00k} \sim N(0, \sigma^2_{\mathrm{L3|C_{L3}}})$, where $\sigma^2_{\mathrm{L3|C_{L3}}}$ is the covariate-adjusted between-school variance of Y.

Two-Level Multisite Individually Randomized Trials

Suppose that $i \in \{1, 2, \ldots, n_k\}$ students at level (L) 1 are nested within $k \in \{1, 2, \ldots, K\}$ schools at L3 and individual students are randomly assigned to the TG or CG (see Figure 4d in Chapter 1). For such a 2L-MSIRT, the unconditional two-level model with random between-student-within-school site effects can be written as (Raudenbush & Liu, 2000, Equation 4):

$$Y_{ik} = \pi_{00} + \pi_{10}T_{ik} + u_{0k} + u_{1k}T_{ik} + e_{ik}\,, \tag{B17}$$

where

L1:
$$Y_{ik} = \beta_{0k} + \psi_{1k}T_{ik} + e_{ik} \tag{B18}$$

$$\text{L3:} \quad \begin{aligned} \beta_{0k} &= \pi_{00} + u_{0k} \\ \psi_{1k} &= \pi_{10} + u_{1k} \end{aligned} \tag{B19}$$

Y_{ik} is the achievement outcome of the ith student in the kth school and T_{ik} is the treatment indicator of the ith student in the kth school, with $T_{ik} = .50/-.50$ for students in TG/CG. β_{0k} is the intercept of the kth school, π_{00} is the grand mean, ψ_{1k} is the treatment effect in the kth school (i.e., $\widehat{\psi}_{1k} = \bar{Y}_{\text{TG}} - \bar{Y}_{\text{CG}}$), and π_{10} is the average treatment effect. e_{ik} is the residual of the ith student in the kth school, with $e_{ik} \sim N(0, \sigma^2_{\text{L1}})$, where σ^2_{L1} is the between-student-within-school variance of Y at L1. u_{0k} and u_{1k} are the residuals of the kth school, with

$$\begin{pmatrix} u_{0k} \\ u_{1k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\text{L3}} & \sigma_{\text{L3}\delta_{\text{L3}}} \\ \sigma_{\text{L3}\delta_{\text{L3}}} & \sigma^2_{\delta_{\text{L3}}} \end{bmatrix} \right), \tag{B20}$$

where σ^2_{L3} is the between-school variance of Y at L3, $\sigma^2_{\delta_{\text{L3}}}$ is the between-school variance in the treatment effect at L3, and $\sigma_{\text{L3}\delta_{\text{L3}}}$ is the covariance between u_{0k} and u_{1k}.

Adding $q_{\text{L1}} \in \{1, 2, \ldots, Q_{\text{L1}}\}$ covariates C_{L1} at L1 and $q_{\text{L3}} \in \{1, 2, \ldots, Q_{\text{L3}}\}$ covariates C_{L3} at L3 yields the conditional within two-level model with random site effects (Dong & Maynard, 2013, p. 47):

$$\begin{aligned} Y_{ik} = \pi_{00} + \pi_{10}T_{ik} + \textstyle\sum_{q_{\text{L3}}=1}^{Q_{\text{L3}}} \pi_{0q_{\text{L3}}+1} C_{\text{L3}_{q_{\text{L3}}k}} + \sum_{q_{\text{L3}}=1}^{Q_{\text{L3}}} \pi_{1q_{\text{L3}}} C_{\text{L3}_{q_{\text{L3}}k}} T_{ik} + \\ \textstyle\sum_{q_{\text{L1}}=1}^{Q_{\text{L1}}} \pi_{q_{\text{L1}}0} C_{\text{L1}_{q_{\text{L1}}ik}} + u_{0k} + u_{1k}T_{ik} + e_{ik} \end{aligned} \tag{B21}$$

where

$$\text{L1:} \quad Y_{ik} = \beta_{0k} + \psi_{1k}T_{ik} + \textstyle\sum_{q_{\text{L1}}=1}^{Q_{\text{L1}}} \beta_{q_{\text{L1}}k} C_{\text{L1}_{q_{\text{L1}}ik}} + e_{ik} \tag{B22}$$

$$\text{L3:} \quad \begin{aligned} \beta_{0k} &= \pi_{00} + \textstyle\sum_{q_{\text{L3}}=1}^{Q_{\text{L3}}} \pi_{0q_{\text{L3}}} C_{\text{L3}_{q_{\text{L3}}k}} + u_{0k} \\ \psi_{1k} &= \pi_{10} + \textstyle\sum_{q_{\text{L3}}=1}^{Q_{\text{L3}}} \pi_{1q_{\text{L3}}} C_{\text{L3}_{q_{\text{L3}}k}} + u_{1k} \\ \beta_{q_{\text{L1}}k} &= \pi_{q_{\text{L1}}0} \end{aligned} \tag{B23}$$

$\beta_{q_{\text{L1}}k}$ is the coefficient of the q_{L1}th covariate C_{L1} of the ith student in the kth school and $\pi_{0q_{\text{L3}}}$ is the coefficient of the q_{L3}th covariate C_{L3} of the kth school. $e_{ik} \sim N(0, \sigma^2_{\text{L1}|\text{C}_{\text{L1}}})$, where $\sigma^2_{\text{L1}|\text{C}_{\text{L1}}}$ is the covariate-adjusted between-student-within-school variance of Y.

$$\begin{pmatrix} u_{0k} \\ u_{1k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\text{L3}|\text{C}_{\text{L3}}} & \sigma_{\text{L3}\delta_{\text{L3}}|\text{C}_{\text{L3}}} \\ \sigma_{\text{L3}\delta_{\text{L3}}|\text{C}_{\text{L3}}} & \sigma^2_{\delta_{\text{L3}}|\text{C}_{\text{L3}}} \end{bmatrix} \right), \tag{B24}$$

where $\sigma^2_{\text{L3}|\text{C}_{\text{L3}}}$ is the covariate-adjusted between-school variance of Y, $\sigma^2_{\delta_{\text{L3}}|\text{C}_{\text{L3}}}$ is the covariate-adjusted between-school variance in the treatment effect, and $\sigma_{\text{L3}\delta_{\text{L3}}|\text{C}_{\text{L3}}}$ is the covariate-adjusted covariance between u_{0k} and u_{1k}. Note that C_{L3} may be either aggregated C_{L1} variables or cluster characteristics by definition (e.g., school size). In the first case, group-mean centering is recommended which ensures that the covariates explain variance exclusively at the level of their specification (Konstantopoulos, 2008). Note further that the expected effect of covariate adjustment on the treatment effect equals zero, assuming covariate-treatment orthogonality (Konstantopoulos, 2008).

If the site effects are treated as fixed, u_{0k} and u_{1k} are fixed effects with a mean constrained to zero. Equation (B21) reduces to (Dong & Maynard, 2013, pp. 46–47):

$$Y_{ik} = \pi_{00} + \pi_{10}T_{ik} + \textstyle\sum_{q_{\text{L1}}=1}^{Q_{\text{L1}}} \pi_{q_{\text{L1}}0} C_{\text{L1}_{q_{\text{L1}}ik}} + u_{0k} + u_{1k}T_{ik} + e_{ik} \tag{B25}$$

Equation (B23) reduces to (Dong & Maynard, 2013, pp. 46–47):

$$\text{L3:} \quad \begin{aligned} \beta_{0k} &= \pi_{00} + u_{0k} \\ \psi_{1k} &= \pi_{10} + u_{1k} \\ \beta_{q_{\text{L1}}k} &= \pi_{q_{\text{L1}}0} \end{aligned} \tag{B26}$$

Three-Level Multisite Individually Randomized Trials

Suppose that $i \in \{1, 2, \ldots, n_{jk}\}$ students at L1 are nested within $j \in \{1, 2, \ldots, J_k\}$ classrooms at L2 which are, in turn, nested within $k \in \{1, 2, \ldots, K\}$ schools at L3 and individual students are randomly assigned to the TG or CG (see Figure 4e in Chapter 1). For such a 3L-MSIRT, the unconditional within three-level model with random site effects can be written as:

$$Y_{ijk} = \pi_{000} + \pi_{100}T_{ijk} + u_{00k} + u_{10k}T_{ijk} + r_{0jk} + r_{1jk}T_{ijk} + e_{ijk}\,, \tag{B27}$$

where

$$\text{L1:} \quad Y_{ijk} = \beta_{0jk} + \psi_{1jk}T_{ijk} + e_{ijk} \tag{B28}$$

$$\text{L2:} \quad \begin{aligned} \beta_{0jk} &= \gamma_{00k} + r_{0jk} \\ \psi_{1jk} &= \gamma_{10k} + r_{1jk} \end{aligned} \tag{B29}$$

$$\text{L3:} \quad \begin{aligned} \gamma_{00k} &= \pi_{000} + u_{00k} \\ \gamma_{10k} &= \pi_{100} + u_{10k} \end{aligned} \tag{B30}$$

Y_{ijk} is the achievement outcome of the ith student in the jth classroom in the kth school and T_{ijk} is the treatment indicator of the ith student in the jth classroom in the kth school, with $T_{ijk} = .50/-.50$ for students in TG/CG. β_{0jk} is the intercept of the jth classroom in the kth school, γ_{00k} is the intercept of school k, π_{000} is the grand mean, ψ_{1jk} is the treatment effect in the jth classroom in the kth school (i.e., $\widehat{\psi}_{1jk} = \bar{Y}_{\text{TG}} - \bar{Y}_{\text{CG}}$), γ_{10k} is the average treatment effect in the kth school, and π_{010} is the average treatment effect. e_{ijk} is the residual of the ith student in the jth classroom in the kth school, with $e_{ijk} \sim N(0, \sigma^2_{\text{L1}})$, where σ^2_{L1} is the between-student-within-classroom variance of Y at L1. r_{0jk} and r_{1jk} are the residuals of the jth classroom in the kth school, with

$$\begin{pmatrix} r_{0jk} \\ r_{1jk} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\text{L2}} & \sigma_{\text{L2}\delta_{\text{L2}}} \\ \sigma_{\text{L2}\delta_{\text{L2}}} & \sigma^2_{\delta_{\text{L2}}} \end{bmatrix}\right), \tag{B31}$$

where σ^2_{L2} is the between-classroom-within-school variance of Y at L2, $\sigma^2_{\delta_{\text{L2}}}$ is the between-classroom-within-school variance in the treatment effect at L2, and $\sigma_{\text{L2}\delta_{\text{L2}}}$ is the covariance between r_{0jk} and r_{1jk}. u_{00k} and u_{10k} are the residuals of the kth school, with

$$\begin{pmatrix} u_{00k} \\ u_{10k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\text{L3}} & \sigma_{\text{L3}\delta_{\text{L3}}} \\ \sigma_{\text{L3}\delta_{\text{L3}}} & \sigma^2_{\delta_{\text{L3}}} \end{bmatrix}\right), \tag{B32}$$

where σ^2_{L3} is the between-school variance of Y at L3, $\sigma^2_{\delta_{\text{L3}}}$ is the between-school variance in the treatment effect at L3, and $\sigma_{\text{L3}\delta_{\text{L3}}}$ is the covariance between u_{00k} and u_{10k}.

Adding $q_{\text{L1}} \in \{1, 2, \ldots, Q_{\text{L1}}\}$ covariates C_{L1} at L1, $q_{\text{L2}} \in \{1, 2, \ldots, Q_{\text{L2}}\}$ covariates C_{L2} at L2, and $q_{\text{L3}} \in \{1, 2, \ldots, Q_{\text{L3}}\}$ covariates C_{L3} at L3 yields the conditional within three-level model with random site effects (see Dong & Maynard, 2013, p. 48; Konstantopoulos, 2008, pp. 279–280):

$$Y_{ijk} = \pi_{000} + \pi_{100}T_{ijk} + \sum_{q_{\mathrm{L3}}=1}^{Q_{\mathrm{L3}}} \pi_{00q_{\mathrm{L3}}} C_{\mathrm{L3}_{q_{\mathrm{L3}}k}} + \sum_{q_{\mathrm{L3}}=1}^{Q_{\mathrm{L3}}} \pi_{10q_{\mathrm{L3}}} C_{\mathrm{L3}_{q_{\mathrm{L3}}k}} T_{ijk} + \sum_{q_{\mathrm{L2}}=1}^{Q_{\mathrm{L2}}} \pi_{0q_{\mathrm{L2}}0} C_{\mathrm{L2}_{q_{\mathrm{L2}}jk}} + \sum_{q_{\mathrm{L2}}=1}^{Q_{\mathrm{L2}}} \gamma_{1q_{\mathrm{L2}}k} C_{\mathrm{L2}_{q_{\mathrm{L2}}jk}} T_{ijk} + \sum_{q_{\mathrm{L1}}=1}^{Q_{\mathrm{L1}}} \pi_{q_{\mathrm{L1}}00} C_{\mathrm{L1}_{q_{\mathrm{L1}}ijk}} + u_{00k} + u_{10k}T_{ijk} + r_{0jk} + r_{1jk}T_{ijk} + e_{ijk} \quad \text{(B33)}$$

where

L1:
$$Y_{ijk} = \beta_{0jk} + \psi_{1jk}T_{ijk} + \sum_{q_{\mathrm{L1}}=1}^{Q_{\mathrm{L1}}} \beta_{q_{\mathrm{L1}}jk} C_{\mathrm{L1}_{q_{\mathrm{L1}}ijk}} + e_{ijk} \quad \text{(B34)}$$

L2:
$$\begin{gathered} \beta_{0jk} = \gamma_{00k} + \sum_{q_{\mathrm{L2}}=1}^{Q_{\mathrm{L2}}} \gamma_{0q_{\mathrm{L2}}k} C_{\mathrm{L2}_{q_{\mathrm{L2}}jk}} + r_{0jk} \\ \psi_{1jk} = \gamma_{10k} + \sum_{q_{\mathrm{L2}}=1}^{Q_{\mathrm{L2}}} \gamma_{1q_{\mathrm{L2}}k} C_{\mathrm{L2}_{q_{\mathrm{L2}}jk}} + r_{1jk} \\ \beta_{q_{\mathrm{L1}}jk} = \gamma_{q_{\mathrm{L1}}0k} \end{gathered} \quad \text{(B35)}$$

L3:
$$\begin{gathered} \gamma_{00k} = \pi_{000} + \sum_{q_{\mathrm{L3}}=1}^{Q_{\mathrm{L3}}} \pi_{00q_{\mathrm{L3}}} C_{\mathrm{L3}_{q_{\mathrm{L3}}k}} + u_{00k} \\ \gamma_{10k} = \pi_{100} + \sum_{q_{\mathrm{L3}}=1}^{Q_{\mathrm{L3}}} \pi_{10q_{\mathrm{L3}}} C_{\mathrm{L3}_{q_{\mathrm{L3}}k}} + u_{10k} \\ \gamma_{0q_{\mathrm{L2}}k} = \pi_{0q_{\mathrm{L2}}0} \\ \gamma_{1q_{\mathrm{L2}}k} = \pi_{1q_{\mathrm{L2}}0} \\ \gamma_{q_{\mathrm{L1}}0k} = \pi_{q_{\mathrm{L1}}00} \end{gathered} \quad \text{(B36)}$$

$\beta_{q_{\mathrm{L1}}jk}$ is the coefficient of the q_{L1}th covariate C_{L1} of the ith student in the jth classroom in the kth school, $\gamma_{0q_{\mathrm{L2}}k}$ is the coefficient of the q_{L2}th covariate C_{L2} of the jth classroom in the kth school, and $\pi_{00q_{\mathrm{L3}}}$ is the coefficient of the q_{L3}th covariate C_{L3} of the kth school. $e_{ijk} \sim N(0, \sigma^2_{\mathrm{L1}|\mathrm{C_{L1}}})$, where $\sigma^2_{\mathrm{L1}|\mathrm{C_{L1}}}$ is the covariate-adjusted between-student-within-school variance of Y.

$$\begin{pmatrix} r_{0jk} \\ r_{1jk} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\mathrm{L2}|\mathrm{C_{L2}}} & \sigma_{\mathrm{L2}\delta_{\mathrm{L2}}|\mathrm{C_{L2}}} \\ \sigma_{\mathrm{L2}\delta_{\mathrm{L2}}|\mathrm{C_{L2}}} & \sigma^2_{\delta_{\mathrm{L2}}|\mathrm{C_{L2}}} \end{bmatrix} \right), \quad \text{(B37)}$$

where $\sigma^2_{\mathrm{L2}|\mathrm{C_{L2}}}$ is the covariate-adjusted between-classroom-within-school variance of Y at L2, $\sigma^2_{\delta_{\mathrm{L2}}|\mathrm{C_{L2}}}$ is the covariate-adjusted between-classroom-within-school variance in the treatment effect at L2, and $\sigma_{\mathrm{L2}\delta_{\mathrm{L2}}|\mathrm{C_{L2}}}$ is the covariate-adjusted covariance between r_{0jk} and r_{1jk}.

$$\begin{pmatrix} u_{00k} \\ u_{10k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\mathrm{L3}|\mathrm{C_{L3}}} & \sigma_{\mathrm{L3}\delta_{\mathrm{L3}}|\mathrm{C_{L3}}} \\ \sigma_{\mathrm{L3}\delta_{\mathrm{L3}}|\mathrm{C_{L3}}} & \sigma^2_{\delta_{\mathrm{L3}}|\mathrm{C_{L3}}} \end{bmatrix} \right), \quad \text{(B38)}$$

$\sigma^2_{\mathrm{L3}|\mathrm{C_{L3}}}$ is the covariate-adjusted between-school variance of Y at L3, $\sigma^2_{\delta_{\mathrm{L3}}|\mathrm{C_{L3}}}$ is the covariate-adjusted between-school variance in the treatment effect at L3, and $\sigma_{\mathrm{L3}\delta_{\mathrm{L3}}|\mathrm{C_{L3}}}$ is the covariate-adjusted covariance between u_{00k} and u_{10k}. Note that C_{L2} and C_{L3} may be either aggregated C_{L1} variables or cluster characteristics by definition (e.g., classroom or school size). In the first case, group-mean centering is recommended which ensures that the covariates explain variance exclusively at the level of their specification (Konstantopoulos, 2008). Note further that the expected effect of covariate adjustment on the treatment effect equals zero, assuming covariate-treatment orthogonality (Konstantopoulos, 2008).

If the site effects are treated as fixed, r_{0jk} and r_{1jk} as well as u_{00k} and u_{10k} are fixed effects with means constrained to zero. Equation (B33) reduces to:

$$Y_{ijk} = \pi_{000} + \pi_{100}T_{ijk} + \sum_{q_{\mathrm{L1}}=1}^{Q_{\mathrm{L1}}} \pi_{q_{\mathrm{L1}}00} C_{\mathrm{L1}_{q_{\mathrm{L1}}ijk}} + u_{00k} + u_{10k}T_{ijk} + r_{0jk} + r_{1jk}T_{ijk} + e_{ijk} \quad \text{(B39)}$$

Equations (B35) and (B36) reduce to:

$$\text{L2:} \quad \begin{matrix} \beta_{0jk} = \gamma_{00k} + r_{0jk} \\ \psi_{1jk} = \gamma_{10k} + r_{1jk} \\ \beta_{q_{\text{L1}}jk} = \gamma_{q_{\text{L1}}0k} \end{matrix} \tag{B40}$$

$$\text{L3:} \quad \begin{matrix} \gamma_{00k} = \pi_{000} + u_{00k} \\ \gamma_{10k} = \pi_{100} + u_{10k} \\ \gamma_{q_{\text{L1}}0k} = \pi_{q_{\text{L1}}00} \end{matrix} \tag{B41}$$

Three-Level Multisite Cluster-Randomized Trials

Suppose that $i \in \{1, 2, \ldots, n_{jk}\}$ students at L1 are nested within $j \in \{1, 2, \ldots, J_k\}$ classrooms at L2 which are, in turn, nested within $k \in \{1, 2, \ldots, K\}$ schools at L3, and classrooms are randomly assigned to the TG or CG, while schools form the sites (see Figure 4f in Chapter 1). For such a 3L-MSCRT, the unconditional three-level model with random between-classroom-within-school site effects can be deduced from the formulas given in Dong and Maynard (2013, pp. 54–55) and Konstantopoulos (2008, pp. 270–272) as follows:

$$Y_{ijk} = \pi_{000} + \pi_{010}T_{jk} + u_{00k} + u_{01k}T_{jk} + r_{0jk} + e_{ijk} \tag{B42}$$

where

$$\text{L1:} \quad Y_{ijk} = \beta_{0jk} + e_{ijk} \tag{B43}$$

$$\text{L2:} \quad \beta_{0jk} = \gamma_{00k} + \psi_{01k}T_{jk} + r_{0jk} \tag{B44}$$

$$\text{L3:} \quad \begin{matrix} \gamma_{00k} = \pi_{000} + u_{00k} \\ \psi_{01k} = \pi_{010} + u_{01k} \end{matrix} \tag{B45}$$

Y_{ijk} is the achievement outcome of the ith student in the jth classroom in the kth school and T_{jk} is the treatment indicator of the jth classroom in the kth school, with $T_{jk} = .50/-.50$ for classrooms in TG/CG. β_{0jk} is the intercept of the jth classroom in the kth school, γ_{00k} is the intercept of school k, π_{000} is the grand mean, ψ_{01k} is the treatment effect in the kth school (i.e., $\widehat{\psi}_{01k} = \bar{Y}_{\text{TG}} - \bar{Y}_{\text{CG}}$), and π_{010} is the average treatment effect. e_{ijk} is the residual of the ith student in the jth classroom in the kth school, with $e_{ijk} \sim N(0, \sigma^2_{\text{L1}})$, where σ^2_{L1} is the between-student-within-classroom variance of Y at L1. r_{0jk} is the residual of the jth classroom in the kth school, with $r_{0jk} \sim N(0, \sigma^2_{\text{L2}})$, where σ^2_{L2} is the between-classroom-within-school variance of Y at L2. u_{00k} and u_{01k} are the residuals of the kth school, with

$$\begin{pmatrix} u_{00k} \\ u_{01k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\text{L3}} & \sigma_{\text{L3}\delta_{\text{L3}}} \\ \sigma_{\text{L3}\delta_{\text{L3}}} & \sigma^2_{\delta_{\text{L3}}} \end{bmatrix} \right), \tag{B46}$$

where σ^2_{L3} is the between-school variance of Y at L3, $\sigma^2_{\delta_{\text{L3}}}$ is the between-school variance in the treatment effect at L3, and $\sigma_{\text{L3}\delta_{\text{L3}}}$ is the covariance between u_{00k} and u_{01k}.

Adding $q_{\text{L1}} \in \{1, 2, \ldots, Q_{\text{L1}}\}$ covariates C_{L1} at L1, $q_{\text{L2}} \in \{1, 2, \ldots, Q_{\text{L2}}\}$ covariates C_{L2} at L2, and $q_{\text{L3}} \in \{1, 2, \ldots, Q_{\text{L3}}\}$ covariates C_{L3} at L3 yields the conditional within three-level model with random site effects (see Dong & Maynard, 2013, pp. 54–55; Konstantopoulos, 2008, pp. 270–272):

$$\begin{matrix} Y_{ijk} = \pi_{000} + \pi_{010}T_{jk} + \sum_{q_{\text{L3}}=1}^{Q_{\text{L3}}} \pi_{00q_{\text{L3}}} C_{\text{L3}_{q_{\text{L3}}k}} + \sum_{q_{\text{L3}}=1}^{Q_{\text{L3}}} \pi_{01q_{\text{L3}}} C_{\text{L3}_{q_{\text{L3}}k}} T_{jk} + \\ \sum_{q_{\text{L2}}=1}^{Q_{\text{L2}}} \pi_{0q_{\text{L2}}0} C_{\text{L2}_{q_{\text{L2}}jk}} + \sum_{q_{\text{L1}}=1}^{Q_{\text{L1}}} \pi_{q_{\text{L1}}00} C_{\text{L1}_{q_{\text{L1}}ijk}} + u_{00k} + u_{01k}T_{jk} + r_{0jk} + e_{ijk} \end{matrix} \tag{B47}$$

where

$$\text{L1:} \quad Y_{ijk} = \beta_{0jk} + \sum_{q_{\text{L1}}=1}^{Q_{\text{L1}}} \beta_{q_{\text{L1}}jk} C_{\text{L1}q_{\text{L1}}ijk} + e_{ijk} \tag{B48}$$

$$\text{L2:} \quad \begin{aligned} \beta_{0jk} &= \gamma_{00k} + \psi_{01k} T_{jk} + \sum_{q_{\text{L2}}=1}^{Q_{\text{L2}}} \gamma_{0q_{\text{L2}}k} C_{\text{L2}q_{\text{L2}}jk} + r_{0jk} \\ \beta_{q_{\text{L1}}jk} &= \gamma_{q_{\text{L1}}0k} \end{aligned} \tag{B49}$$

$$\text{L3:} \quad \begin{aligned} \gamma_{00k} &= \pi_{000} + \sum_{q_{\text{L3}}=1}^{Q_{\text{L3}}} \pi_{00q_{\text{L3}}} C_{\text{L3}q_{\text{L3}}k} + u_{00k} \\ \psi_{01k} &= \pi_{010} + \sum_{q_{\text{L3}}=1}^{Q_{\text{L3}}} \pi_{01q_{\text{L3}}} C_{\text{L3}q_{\text{L3}}k} + u_{01k} \\ \gamma_{0q_{\text{L2}}k} &= \pi_{0q_{\text{L2}}0} \\ \gamma_{q_{\text{L1}}0k} &= \pi_{q_{\text{L1}}00} \end{aligned} \tag{B50}$$

$\beta_{q_{\text{L1}}jk}$ is the coefficient of the q_{L1}th covariate C_{L1} of the ith student in the jth classroom in the kth school, $\gamma_{0q_{\text{L2}}k}$ is the coefficient of the q_{L2}th covariate C_{L2} of the jth classroom in the kth school, and $\pi_{00q_{\text{L3}}}$ and $\pi_{01q_{\text{L3}}}$ are the coefficients of the q_{L3}th covariate C_{L3} of the kth school. $e_{ijk} \sim N(0, \sigma^2_{\text{L1}|\text{C}_{\text{L1}}})$, where $\sigma^2_{\text{L1}|\text{C}_{\text{L1}}}$ is the covariate-adjusted between-student-within-classroom variance of Y at L1. $r_{0jk} \sim N(0, \sigma^2_{\text{L2}|\text{C}_{\text{L2}}})$, where $\sigma^2_{\text{L2}|\text{C}_{\text{L2}}}$ is the covariate-adjusted between-classroom-within-school variance of Y at L2.

$$\begin{pmatrix} u_{00k} \\ u_{01k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\text{L3}|\text{C}_{\text{L3}}} & \sigma_{\text{L3}\delta_{\text{L3}}|\text{C}_{\text{L3}}} \\ \sigma_{\text{L3}\delta_{\text{L3}}|\text{C}_{\text{L3}}} & \sigma^2_{\delta_{\text{L3}}|\text{C}_{\text{L3}}} \end{bmatrix} \right), \tag{B51}$$

where $\sigma^2_{\text{L3}|\text{C}_{\text{L3}}}$ is the covariate-adjusted between-school variance of Y at L3, $\sigma^2_{\delta_{\text{L3}}|\text{C}_{\text{L3}}}$ is the covariate-adjusted between-school variance in the treatment effect at L3, and $\sigma_{\text{L3}\delta_{\text{L3}}|\text{C}_{\text{L3}}}$ is the covariate-adjusted covariance between u_{00k} and u_{01k}. Note that C_{L2} and C_{L3} may be either aggregated C_{L1} variables or cluster characteristics by definition (e.g., classroom or school size). In the first case, group-mean centering is recommended which ensures that the covariates explain variance exclusively at the level of their specification (Konstantopoulos, 2008). Note further that the expected effect of covariate adjustment on the treatment effect equals zero, assuming covariate-treatment orthogonality (Konstantopoulos, 2008).

If the site effects are treated as fixed, u_{00k} and u_{01k} are fixed effects with a mean constrained to zero. Equation (B47) reduces to (Dong & Maynard, 2013, p. 53):

$$\begin{aligned} Y_{ijk} = \pi_{000} + \pi_{010} T_{jk} + \sum_{q_{\text{L2}}=1}^{Q_{\text{L2}}} \pi_{0q_{\text{L2}}0} C_{\text{L2}q_{\text{L2}}jk} + \sum_{q_{\text{L1}}=1}^{Q_{\text{L1}}} \pi_{q_{\text{L1}}00} C_{\text{L1}q_{\text{L1}}ijk} + \\ u_{00k} + u_{01k} T_{jk} + r_{0jk} + e_{ijk} \end{aligned} \tag{B52}$$

Equation (B50) reduces to (Dong & Maynard, 2013, p. 53):

$$\text{L3:} \quad \begin{aligned} \gamma_{00k} &= \pi_{000} + u_{00k} \\ \psi_{01k} &= \pi_{010} + u_{01k} \\ \gamma_{0q_{\text{L2}}k} &= \pi_{0q_{\text{L2}}0} \\ \gamma_{q_{\text{L1}}0k} &= \pi_{q_{\text{L1}}00} \end{aligned} \tag{B53}$$

Appendix C: Sampling Variances of ρ and R^2

This appendix provides the expressions for the sampling variances of the ρ and R^2 estimates at the various hierarchical levels, which were used to derive the standard errors offered in the present doctoral thesis.

Intraclass Correlation Coefficients

The sampling variances of the ICCs (and therefore their standard errors) can be analytically approximated, using the derivations provided in Hedges et al. (2012) and Donner and Koval (1980). In an unbalanced three-level design with varying cluster sizes at both L2 (i.e., $n_{jk} \neq n_{j'k'}$) and L3 (i.e., $J_k \neq J_{k'}$), the large-sample variance of ρ_{L2} is given by (Hedges et al., 2012, Equations 7 and 8):

$$\frac{(1-\rho_{L2})^2 Var(\sigma^2_{L2})}{\sigma^4_T} + \frac{\rho^2_{L2} Var(\sigma^2_{L3})}{\sigma^4_T} - \frac{2\rho_{L2}(1-\rho_{L2}) Cov(\sigma^2_{L2},\sigma^2_{L3})}{\sigma^4_T} \quad \text{(C1)}$$

The large-sample variance of ρ_{L3} is given by (Hedges et al., 2012, Equations 7 and 9):

$$\frac{\rho^2_{L3} Var(\sigma^2_{L2})}{\sigma^4_T} + \frac{(1-\rho_{L3})^2 Var(\sigma^2_{L3})}{\sigma^4_T} - \frac{2\rho_{L3}(1-\rho_{L3}) Cov(\sigma^2_{L2},\sigma^2_{L3})}{\sigma^4_T} \quad \text{(C2)}$$

where

$$Cov(\sigma^2_{L2},\sigma^2_{L3}) = -\frac{Var(\sigma^2_{L2}) \sum_{k=1}^{K} a_k/(1+b_k\sigma^2_{L3})}{\sum_{k=1}^{K} b_k^2/(1+b_k\sigma^2_{L3})} \quad \text{(C3)}$$

with $a_k = \sum_{j=1}^{J_k} n_{jk}^2 / \left(n_{jk}\sigma^2_{L2} + \sigma^2_{L1}\right)^2$ and $b_k = \sum_{j=1}^{J_k} n_{jk} / (n_{jk}\sigma^2_{L2} + \sigma^2_{L1})$.

In an unbalanced two-level design with varying cluster sizes at L3 (i.e., $n_k \neq n_{k'}$), the large-sample variance of ρ_{L3} is given by (Donner & Koval, 1980, Equation 3):

$$\frac{2N(1-\rho_{L3})^2}{N\sum_{k=1}^{K} n_k(n_k-1)[1+(n_k-1)\rho^2_{L3}]/[1+(n_k-1)\rho_{L3}]^2 - \rho^2_{L3}\left[\sum_{k=1}^{K} n_k(n_k-1)/[1+(n_k-1)\rho_{L3}]\right]^2} \quad \text{(C4)}$$

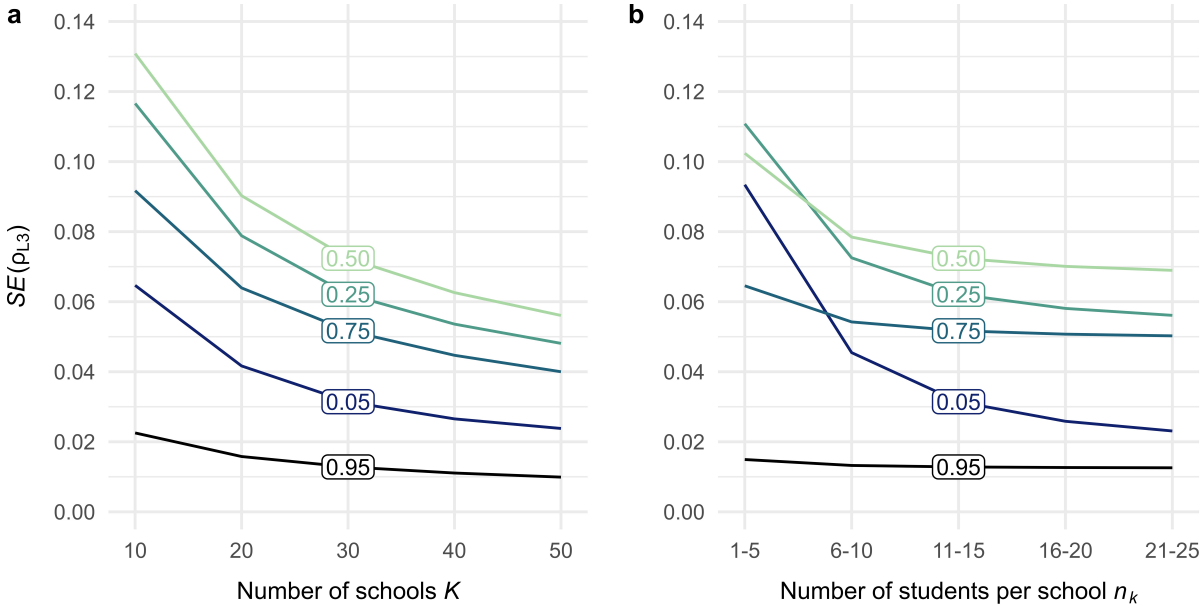
where $N = \sum_{k=1}^{K} n_k$ is the total sample size. It becomes immediately clear from Equations (C1) to (C4) that increasing the sample sizes decreases the variance of ρ values.

For such an unbalanced two-level design, Figure B1 visualizes the relations between the standard error of ρ_{L3} and (a) the number of schools, and (b) the number of students per school for $\rho_{L3} \in \{.05, .25, .50, .75, .95\}$. Three aspects are noteworthy: When holding the total sample size constant, (1) adding schools reduces the standard error more than adding students per school, (2) unless the cluster size is very small, the standard error maximizes with $\rho_{L3} = .50$, and declines both with low as well as high values of ρ_{L3}, and (3) increasing the cluster size is most efficient in reducing the standard error when ρ_{L3} is small. Albeit more complex, this general pattern of results can be transferred to the standard errors of ρ_{L2} and ρ_{L3} in three-level designs.

Note that respective expressions for the large-sample variances of the ICCs under the assumption of balanced designs have also been made available (e.g., Hedges et al., 2012; Jacob et al., 2010). Here, constant cluster sizes (i.e., $n_{jk} = n_{j'k'}$, $J_k = J_{k'}$, or $n_k = n_{k'}$) may be depicted through their averages such as their (harmonic) means. These formulae are generally less complex, and facilitate quick and straightforward computation of standard errors. Yet, simulations in two-level designs indicate that the respective confidence intervals may be distorted, especially with few large-sized clusters and small σ^2_{L1} values (Ukoumunne, 2002).

Other methods to obtain the sampling variances of the ICCs include bootstrapping (e.g., Xiao et al., 2012) or Bayesian estimation (e.g., Turner et al., 2006). However, Eldridge and Kerry (2012, p. 193) stated that the benefits of these approaches over the analytic approximations are in general negligible, given their additional computational burden.

Figure B1. *Standard Error of ρ_{L3} as a Function of the (a) the Number of Schools, (b) the Number of Students per School*



Note. The figure shows the standard errors of $\rho_{L3} \in \{.05, .25, .50, .75, .95\}$ given a total sample size of $N =$ 100/231/390/544/679 when (a) $K = 10/20/30/40/50$ with $4 \leq n_k \leq 13/8 \leq n_k \leq 18/11 \leq n_k \leq 15/7 \leq n_k \leq 24/6 \leq n_k \leq 26$ and (b) $1 \leq n_k \leq 5/6 \leq n_k \leq 10/11 \leq n_k \leq 15/16 \leq n_k \leq 20/21 \leq n_k \leq 25$ for $K = 30$.

Squared Multiple Correlation Coefficients

The simplified analytic approximation of the sampling variances of R^2 uses Fisher's (1925) formulation and is given in Hedges and Hedberg (2013, p. 451):

$$\frac{4R^2(1-R^2)^2}{N^*}, \tag{C5}$$

where N^* is the total sample size at the respective level. That is, for R^2_{L3}, N^* is the total number of schools K, for R^2_{L2}, N^* is the total number of classrooms J, and for R^2_{L1} and R^2_T, N^* is the total number of students N. Note that the expression in Equation (C5) does not take into account the clustering of students within classroom in schools, and the clustering of classrooms within schools. Therefore, in multilevel designs, the standard errors for R^2_{L2}, and especially for R^2_{L1} computed from these sampling variances are only approximations.

Note that the reporting of sampling variances for the R^2 design parameters is very scarce; in fact, only Hedges and Hedberg (2013) provided standard errors for estimates of R^2_{L1} and R^2_{L3} so far. As Jacob et al. (2010) speculated, the reason for this might be that the distributional parameters of R^2 are unkown (Ohtani, 2000; Press & Zellner, 1978), and therefore the attempt to find more accurate analytic solutions largely failed (but see Helland, 1987, for another approximation based on the *F* distribution).

Glossary: Terms and Abbreviations As Frequently Used Throughout the Present Doctoral Thesis

Cluster-randomized trial (CRT). Generally, a class of multilevel experimental studies in which entire clusters are randomly assigned to the experimental groups. From a statistical perspective, this means that the observations within clusters can no longer be assumed stochastically independent; rather, their errors are (typically) correlated (Schochet, 2008). In a straight CRT, treatment allocation always occurs at the top hierarchical level (Hedges & Rhoads, 2010).

Two-level cluster-randomized trial (2L-CRT). Students at Level 1 are nested within schools at Level 3, and schools are randomly assigned to the experimental groups.

Three-level cluster-randomized trial (3L-CRT). Students at Level 1 are nested within classrooms at Level 2, which are, in turn, nested within schools at Level 3, and schools are randomly assigned to the treatment and control condition.

Control group (CG). The randomly composed experimental group (here, in a two-arm experiment) which is not delivered with the treatment to be studied, and instead often doing "business as usual."

Cross-domain pretest (CP). A pretest assessed in a different domain than the outcome. In Study II, reading served as a predictor of STEM (i.e., mathematical-scientific) achievement outcomes, and mathematics served as a predictor of verbal achievement outcomes.

Design sensitivity. Umbrella term used to embrace the concepts of statistical power and statistical precision (Hedges & Hedberg, 2013). Thus, the sensitivity of a design is its probability (i.e., statistical power) to detect a real contrast between the experimental groups on the studied outcome at a given level of statistical significance with a low standard error (i.e., statistical precision; Lipsey, 1990).

Domain-identical pretest (IP). A pretest assessed in the same domain as the outcome, that is, the baseline achievement score of the outcome itself (e.g., prior mathematics skills predicting future mathematics skills).

Estimand. Population target quantity to be estimated or to formulate hypotheses about (e.g., the average treatment effect).

Estimate. The sample value of a population quantity; the result of the estimator.

Estimator. Method (or "recipe") applied to compute an estimate (e.g., the difference between the means observed for the treatment and control group).

Fluid intelligence (Gf). One multifaceted, integral component of general intelligence that encompasses, for instance, reasoning, perception speed, accuracy, and problem solving. Gf is distinguished from crystallized intelligence (Gc; i.e., cumulative knowledge and learned skills; Cattell, 1963).

Individually randomized trial (IRT). Students are sampled and randomly assigned to the experimental groups completely independently of each other, regardless of their membership to a classroom and/or school. Statistically, this means that all observations

are assumed stochastically independent. IRTs most often use (non-representative) convenience samples (Stuart et al., 2011).

Intraclass correlation coefficient (ICC; ρ). Generally, the degree of redundancy in an achievement outcome, due to cluster membership, or equivalently, the extent of variation between clusters.

Intraclass correlation coefficient at Level 2 (ρ_{L2}). Between-classroom (within-school) achievement differences. The ratio of the variance located at the classroom level to the total variance.

Intraclass correlation coefficient at Level 3 (ρ_{L3}). Between-school achievement differences. The ratio of the variance located at the school level to the total variance.

Level 1 (L1). Student level.

Level 2 (L2). Classroom level.

Level 3 (L3). School level.

Minimum detectable effect size (MDES). As a multiple of the standardized standard error of the treatment effect a measure of statistical precision (Bloom, 2005): The $MDES$ quantifies the smallest possible standardized effect that reaches statistical significance in a given design (typically, at a stated alpha level of .05 in a two-tailed test, a statistical power of 80%, and a given sample size).

Multisite cluster-randomized trial (**MSCRT).** Generally, a class of multilevel experimental studies that combines cluster randomization and blocking. Randomization does not occur at the top hierarchical level, but at an intermediate level. Thus, entire clusters are randomly assigned to experimental groups within superordinate clusters, forming the sites. This implies that experimental conditions are crossed with the random effects of the superordinate clusters within which cluster randomization occurs (Konstantopoulos, 2008). Put differently, in an MSCRT, one and the same cluster-randomized trial is replicated in several superordinate clusters (Liu, 2014b).

Three-level multisite cluster-randomized trial (3L-MSCRT). Students at Level 1 are nested within classrooms at Level 2, which are, in turn, nested within schools at Level 3, and classrooms are randomly assigned to the experimental groups within schools. Thus, schools form the sites.

Multisite individually randomized trial (MSIRT). Generally, a class of multilevel experimental studies in which individuals are randomly assigned to the experimental groups within clusters, forming the sites.

Two-level multisite individually randomized trial (2L-MSIRT). Students at Level 1 are nested within schools at Level 3, and students are randomly assigned to the experimental groups within schools. Thus, schools form the sites.

Three-level multisite individually randomized trial (3L-MSIRT). Students at Level 1 are nested within classrooms at Level 2, which are, in turn, nested within schools at Level 3, and students are randomly assigned to the experimental groups within classrooms within schools. Thus, classrooms and schools form (nested) sites.

Multisite randomized trial (MSRT). Generally, a class of multilevel experimental studies in which randomization does not occur at the top hierarchical level, but at any subordinate level. This subordinate level may be composed of either individuals or clusters which

are randomly assigned to experimental groups within superordinate clusters, forming the sites. In either case, this implies that experimental conditions are crossed with the random effects of the superordinate clusters within which randomization occurs (Konstantopoulos, 2008). Put differently, in an MSRT, one and the same experiment is replicated in several superordinate clusters (Liu, 2014b).

Power Analysis. Statistical procedure to determine the required sample size, or the statistical power, or the statistical precision of the treatment effect estimate (here, minimum detectable effect size; Bloom, 2005). These three quantities are interrelated concepts: computing one of them requires assumptions on the two remaining; and on the α level (i.e., the Type I error rate) as well as the type of the test (often a t-test, but sometimes also F-test, Mann-Whitney U test, etc.; e.g., Lipsey, 1990). In multilevel designs, power analysis also involves assumptions on the sample allocation among hierarchical levels and the variance design parameters at the various hierarchical levels (i.e., necessarily the intraclass correlation coefficient[s], and possibly also the amounts of explained variance; Hedges & Rhoads, 2010).

Prediction Interval (PI). Quantifies the total dispersion (sampling variance plus true heterogeneity) around the meta-analytic average of ρ and/or R^2. Thus, it provides a plausible range of ρ and/or R^2, that is the range in which an ρ and/or 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).

Randomization. The selection of units to be assigned to experimental conditions, fully by chance (e.g., via coin toss or lottery).

Randomized trial (RT). A study under controlled conditions, where units (e.g., individual students or entire schools) are allocated by chance (like by tossing a coin) to receive some deliberate intervention (i.e., a treatment) or not, in order to test its effect.

Sociodemographic characteristics (SC). Gender, migration background and socioeconomic status in terms of the highest International Socio-Economic Index of Occupational Status within a family (HISEI; Ganzeboom & Treiman, 1996) and the highest educational attainment within the family. In Study I, also the NEPS starting cohort.

Squared multiple correlation coefficient (R^2). Generally, the amount of explained variance by covariates. In multilevel designs, covariates can act at either hierarchical level, although not necessary. When within-cluster covariates are group-mean centered, the covariates explain variance only at the level at which they are introduced (Konstantopoulos, 2008).

Squared multiple correlation coefficient at Level 1 (R^2_{L1}). The amount of explained variance at L1 by L1 covariates. The ratio of the difference between the unconditional (i.e., not covariate-adjusted) and the conditional (i.e., covariate-adjusted) between-student-within-classroom variance components to the unconditional between-student-within-classroom variance component.

Squared multiple correlation coefficient at Level 2 (R^2_{L2}). The amount of explained variance at L2 by L2 covariates. The ratio of the difference between the unconditional (i.e., not covariate-adjusted) and the conditional (i.e., covariate-adjusted) between-classroom-within-school variance components to the unconditional between-classroom-within-school variance component.

Squared multiple correlation coefficient at Level 3 (R^2_{L3}). The amount of explained variance at L3 by L3 covariates. The ratio of the difference between the unconditional (i.e., not covariate-adjusted) and the conditional (i.e., covariate-adjusted) between-school variance components to the unconditional between-school variance component.

Squared multiple correlation coefficient, in total (R^2_T). The total amount of explained variance among all individual students (i.e., not decomposed). The ratio of the difference between the unconditional (i.e., not covariate-adjusted) and the conditional (i.e., covariate-adjusted) total variance to the unconditional total variance.

Statistical Power ($1-\beta$). The (long-term) probability of rejecting the null hypothesis when it is actually false; or, put differently, the likelihood of a statistical test to detect an effect, if it exists in the population.

Statistical Precision. Basically, the standard error of the treatment effect. Here, precision is quantified via the minimum detectable effect size, which is conceived as a multiple of the standardized standard error of the treatment effect (Bloom, 2005).

Total (T). Not hierarchically decomposed, among all individual students. T is used to index design parameters or other quantities that define a single-level RT design. For instance, R^2_T denotes the total amount of explained variance by covariates in an individually randomized trial, across all individual students.

Treatment. A deliberate measure or intervention.

Treatment effect, average. Difference between the means observed on some achievement outcome Y for the treatment group (TG) and the control group (CG; $\bar{Y}_{TG} - \bar{Y}_{CG}$; Bloom, 2006).

Treatment Group (TG). The randomly composed experimental group (here, in a two-arm experiment) which receives the treatment to be studied.

***Y*.** Achievement outcome variable.

References

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

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

Campbell, M. J., & Walters, S. J. (2014). *How to design, analyse and report cluster randomised trials in medicine and health related research.* John Wiley & Sons.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22. https://doi.org/10.1037/h0046743

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., Birkett, N., & Buck, C. (1981). Randomization by cluster. Sample size requirements and analysis. *American Journal of Epidemiology, 114*(6), 906–914. https://doi.org/10.1093/oxfordjournals.aje.a113261

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

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

Eldridge, S., & Kerry, S. (2012). *A Practical Guide to Cluster Randomised Trials in Health Services Research.* John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119966241

Fisher, R. A. (1925). *Statistical methods for research workers.* Oliver & Boyd.

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

Hayes, R. J., & Moulton, L. H. (2017). *Cluster randomised trials* (2nd ed.). CRC Press.

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

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

Helland, I. S. (1987). On the interpretation and use of R2 in regression analysis. *Biometrics, 43*(1), 61. https://doi.org/10.2307/2531949

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

Kish, L. (1965). *Survey sampling.* Wiley.

Konstantopoulos, S. (2008). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness, 1*(4), 265–288. https://doi.org/10.1080/19345740802328216

Kreft, I. G. G. (1993). Using Multilevel Analysis to Assess School Effectiveness: A Study of Dutch Secondary Schools. *Sociology of Education, 66*(2), 104. https://doi.org/10.2307/2112796

Kreft, I. G. G., & de Leeuw, J. (1998). Introducing multilevel modeling. SAGE Publications.

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

Liu, X. S. (2014a). A note on statistical power in multi-site randomized trials with multiple treatments at each site. *British Journal of Mathematical and Statistical Psychology, 67*(2), 231–247. https://doi.org/10.1111/bmsp.12016

Liu, X. S. (2014b). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. Taylor&Francis. http://site.ebrary.com/id/10801501

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

Ohtani, K. (2000). Bootstrapping R2 and adjusted R2 in regression analysis. *Economic Modelling, 17*(4), 473–483. https://doi.org/10.1016/S0264-9993(99)00034-6

Press, S. J., & Zellner, A. (1978). Posterior distribution for the multiple correlation coefficient with fixed regressors. *Journal of Econometrics, 8*(3), 307–321. https://doi.org/10.1016/0304-4076(78)90050-7

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213. https://doi.org/10.1037/1082-989X.5.2.199

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

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

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the institute of education sciences. *Educational Evaluation and Policy Analysis, 31*(3), 298–318. https://doi.org/10.3102/0162373709339524

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society, 174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Teerenstra, S., Eldridge, S., Graff, M., Hoop, E., & Borm, G. F. (2012). A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in Medicine, 31*(20), 2169–2178. https://doi.org/10.1002/sim.5352

Turner, R. M., Omar, R. Z., & Thompson, S. G. (2006). Constructing intervals for the intracluster correlation coefficient using Bayesian modelling, and application in cluster randomized trials. *Statistics in Medicine, 25*(9), 1443–1456. https://doi.org/10.1002/sim.2304

Ukoumunne, O. C. (2002). A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine, 21*(24), 3757–3774. https://doi.org/10.1002/sim.1330

Xiao, Y., Liu, J., & Bhandary, M. (2012). Resampling approaches for common intraclass correlation coefficients. *Journal of Statistical Computation and Simulation, 82*(9), 1357–1366. https://doi.org/10.1080/00949655.2011.581668

Eigenständigkeitserklärung

Hiermit bestätige ich, dass ich die vorliegende Dissertation

Optimizing Power Analysis for Randomized Experiments: Design Parameters for Student Achievement

selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht. Die Dissertation ist in keinem früheren Promotionsverfahren angenommen oder angelehnt worden.

Potsdam, Dezember 2023

Sophie E. Stallasch