Brunner, Martin; Stallasch, Sophie E.; Artelt, Cordula; Lüdtke, Oliver

An Individual Participant Data Meta-Analysis to Support Power Analyses for Randomized Intervention Studies in Preschool : Cognitive and Socio-Emotional Learning Outcomes

Primary publication

Brunner, Martin; Stallasch, Sophie E.; Artelt, Cordula; Lüdtke, Oliver (2025): An Individual Participant Data Meta-Analysis to Support Power Analyses for Randomized Intervention Studies in Preschool : Cognitive and Socio-Emotional Learning Outcomes, in: Educational psychology review, New York, NY: Springer, Vol. 37, Nr. 1, 6, pp. 1–38, doi: 10.1007/s10648-024-09981-z.

Educational Psychology Review (2025) 37:6
https://doi.org/10.1007/s10648-024-09981-z

META-ANALYSIS

An Individual Participant Data Meta-Analysis to Support Power Analyses for Randomized Intervention Studies in Preschool: Cognitive and Socio-Emotional Learning Outcomes

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

Accepted: 19 December 2024 / Published online: 14 January 2025

Abstract

There is a need for robust evidence about which educational interventions work in preschool to foster children's cognitive and socio-emotional learning (SEL) outcomes. Lab-based individually randomized experiments can develop and refine such interventions, and field-based randomized experiments (e.g., cluster randomized trials) evaluate their effectiveness in real-world daycare center settings. Applying reliable estimates of design parameters in the context of a priori power analyses is essential to ensure that the sample size of these studies is adequate to support strong statistical conclusions regarding the strength of the intervention effect. However, there is little knowledge on relevant design parameters with preschool children. We therefore utilized a systematic collection of individual participant data from four German probability samples ($554 \leq N \leq 2928$) with preschool children (aged two to six years) to estimate and meta-analyze design parameters. These parameters are relevant for planning single-level (e.g., in non-clustered lab-based settings), two-level (children nested in daycare centers), and three-level (children nested in groups, with groups nested in daycare centers) randomized intervention studies targeting cognitive and SEL outcomes assessed with three methods (standardized tests, parent ratings, and educator ratings). The design parameters depict between-group and -center differences as well as the proportion of variance in the outcomes explained by different covariate sets (socio-demographic characteristics, baseline measures, and their combination) at the child, group, and center level. In conclusion, this paper provides a rich source of design parameters, recommendations, and illustrations to support a priori power analyses for randomized intervention studies in early childhood education research.

Keywords Power analysis · Preschool · Cognitive skills · Socio-emotional learning · Meta-analysis

Extended author information available on the last page of the article

Introduction

Early educational interventions may be particularly beneficial for fostering preschool children's cognitive, academic, and socio-emotional learning (SEL) outcomes (Barnett, 2011; OECD, 2020). However, not all of these interventions are equally successful or work equally well in all preschools (Barnett, 2011; Sabol et al., 2022). Thus, further research is needed to develop, refine, and evaluate educational interventions aiming at fostering preschool children's development. To this end, randomized experiments are indispensable because they allow for strong causal inferences about the impact of educational interventions (Slavin, 2020). Many randomized experiments have been conducted in the last two decades (Connolly et al., 2018) because major funding agencies (e.g., the UK Education Endowment Foundation and the US Institute of Education Sciences) have emphasized the importance of randomized field experiments to provide strong evidence for what fosters student learning (Hedges & Schauer, 2018; Slavin, 2020). However, a review of large-scale intervention studies (Lortie-Forgues & Inglis, 2019) showed that a large majority of these studies were "underpowered," meaning they were not sensitive enough to detect typical interventions effects.

To tackle this problem, applying reliable estimates of design parameters in a priori power analyses is highly recommended to guarantee that the sample size of an intervention study is large enough to allow strong statistical conclusions regarding the strength of the intervention effect in the target population and to detect a meaningful intervention effect if it exists (Bloom et al., 2007; Hedges & Hedberg, 2007; Hedges & Rhoads, 2010; Raudenbush et al., 2007). To achieve this, researchers in early childhood education require design parameters that align with the specific design features of the planned intervention study. Depending on the research objective, researchers can select from various study designs that employ different strategies for (a) sampling preschool children and (b) randomizing these children to experimental groups (Hedges & Rhoads, 2010). Lab-based randomized experiments conducted under well-controlled conditions by well-trained staff are a valuable methodological tool when the research goal is to develop and refine educational interventions, or test their efficacy. These studies typically assume simple random sampling of preschool children and apply individual random assignment of the sampled children to experimental conditions. In the remainder of this paper, we refer to this type of lab-based randomized experiment as a single-level design. However, evaluating the effectiveness of interventions on a larger scale—such as when implemented during regular preschool days by teachers or educators—requires randomized experiments in field settings, particularly in daycare centers. These randomized experiments utilize hierarchical sampling methods. For example, researchers may first draw a (random) sample of daycare centers, followed by a (random) sample of children within the selected centers. Randomization in these field settings can take various forms, leading to different multilevel designs. Multisite randomized trials (MSRTs) use blocked random assignment to allocate either individual children or entire groups to experimental conditions within daycare centers. Cluster randomized trials (CRTs), also known as group-randomized trials, randomly assign entire daycare centers to different experimental groups. Notably, if randomization is carried out *within* daycare centers, the

validity of causal inferences may be threatened, as children in the control group may be (unintentionally) exposed to the intervention. Moreover, many social interventions operate at the group level, such as reforms affecting whole daycare centers (Hedges & Rhoads, 2010). In such cases, CRTs may be the only reasonable design option.

Crucially, although reliable knowledge of design parameters is essential for a priori power analyses during the planning stage of randomized intervention studies, this knowledge is largely lacking for the target population of preschool children in early childhood education. Except for parameters relevant to the very specific target population of socioeconomically disadvantaged preschool children in the United States (Jacob et al., 2010; Spybrook et al., 2011), there is little empirically-based information on design parameters that researchers can use to plan studies for broader target populations or in other countries. The primary goal of this paper is, therefore, to address this significant research gap. To this end, we provide a comprehensive, systematic collection of reliable design parameters for conducting power analyses for intervention studies in both lab settings (with simple random sampling and individual random assignment) and multilevel field settings (implementing CRTs or MSRTs).[1] These studies may target preschool children's cognitive or SEL outcomes, assessed by standardized tests, parents, and/or educators. To enhance their reliability and generalizability, we derived design parameters from the analysis and meta-analytic integration of individual participant data (IPD) from several large-scale studies involving probability samples of children attending German daycare centers—the preschool setting that most children in Germany and many other countries attend. Notably, our paper is accompanied by extensive online supplementary materials (OSM) in the Open Science Framework (OSF),[2] including details on the applied systematic search for IPD, measures, and methods (OSM A), the R code (R Core Team, 2024) for reproducibility and replicability, and detailed interactive tables of the design parameters that are required for sample size planning (OSM B). By embedding the present design parameters within the robust methodological literature on the statistical power of randomized experiments, we thoroughly discuss and illustrate how these parameters can be applied in power analyses for randomized intervention studies involving preschool children. In summary, this paper makes several significant contributions that can assist researchers in early childhood education to conduct effective and efficient studies on the impact of cognitive and socio-emotional development interventions for preschool children.

A Spotlight on the History of Randomized Field Trials in Education

Building on substantial methodological developments, randomized experiments in education have a long but complex history (Bloom, 2006; Bloom et al., 1999; Boruch, 2005; Cook, 2002, 2005; Hedges & Schauer, 2018; Mosteller & Boruch,

[1] Notably, the present design parameters are also useful for a priori power analyses of multilevel quasi-experimental designs (Bulus, 2022; Dong & Maynard, 2013; Schochet, 2009).

[2] https://osf.io/qz7fy. Tables and Figures presented in OSM A and B are indicated by corresponding letters (e.g., Table A1 in OSM A or Table B.CT.M in OSM B).

2002). Particularly from the 1960s onward, randomized experiments were recognized as a key source of evidence on the impact of educational interventions in the United States. Consequently, from the 1960s to the 1980s, many randomized field trials were conducted to evaluate educational reforms and policies (Hedges & Schauer, 2018). Some of these trials continue to have a lasting impact on U.S. educational policy today. For instance, the Perry Preschool Project involving low-income children demonstrated that high-quality early childhood education can have long-term beneficial effects on key life outcomes, such as higher earnings and reduced criminal activity, for the students (Schweinhart, 2005), as well as positive effects on the lives of their own children (García et al., 2023). However, during the 1980s and 1990s, there was a paradigm shift in educational research towards qualitative approaches. Furthermore, policy funding for randomized trials in the United States was significantly reduced. Consequently, the number of randomized field experiments (but not lab-based experiments) decreased sharply (Hedges & Schauer, 2018). There were, however, two notable exceptions. The policy-relevant MSRT—Project STAR (Student–Teacher Achievement Ratio)—demonstrated a positive impact of small class sizes on students' achievement in elementary school and beyond (Nye et al., 1999). Furthermore, drawing on national probability samples, an MSRT examined whether participation in the Upward Bound program improved high school outcomes for low-income children (Myers & Schirm, 1999). Importantly, with the onset of the new millennium, new laws took effect—the No Child Left Behind Act in 2001 and the Education Sciences Reform Act in 2002—reflecting a strong policy demand for robust scientific evidence on the impact of educational interventions and programs. Randomized experiments were regarded as one of the standard methods for generating this type of evidence. As a result, the newly established U.S. Institute of Education Sciences launched funding strategies that led to numerous randomized experiments (Hedges & Schauer, 2018). An overview of these trials is provided by Spybrook and colleagues (Spybrook & Raudenbush, 2009; Spybrook et al., 2016).

The timing of when randomized experiments have been recognized as an important source of rigorous evidence on the impact of educational interventions—if at all—varies considerably across countries. For example, in the United Kingdom, the Education Endowment Foundation was established in 2011 to initiate major funding schemes for randomized experiments (Dawson et al., 2018). Around the same time, policy-initiated or policy-funded randomized trials were conducted in Denmark, Norway, and Sweden (Pontoppidan et al., 2018). Finally, in Germany, there have only been a few randomized trials (e.g., the CRT by Gaspard et al., 2015), likely because the policy demand for generating this type of robust experimental evidence is much lower than in other countries, particularly the United States (Standing Scientific Commission on Education Policy, 2022). In summary, the importance of randomized experiments for advancing knowledge on the impact of educational interventions appears to be widely recognized by researchers and policymakers, although some cross-national variation exists. This is evident in the significant overall increase in randomized studies conducted in recent years, as well as the varying number of randomized trials across countries (Connolly et al., 2018).

Which Design Parameters Do We Need in Power Analyses?

One key goal in planning a randomized intervention study is to ensure that it can effectively assess the strength of the intervention effect. In this regard, design sensitivity (or simply *sensitivity*) is an umbrella term that covers several interrelated statistical concepts (Hedges & Hedberg, 2013): the precision (i.e., standard error) with which the intervention effect can be estimated, the statistical power to detect the intervention effect if it exists, and the minimum detectable effect size (*MDES*; Bloom, 1995). Power analyses are essential for evaluating and assuring the sensitivity of randomized studies, for example, to determine the sample size needed to achieve a certain *MDES* with confidence (e.g., with level of statistical significance $\alpha = 0.05$ and power of $1 - \beta = 0.80$). To this end, researchers need to set a meaningful value for the *MDES,* which, for a standardized test or questionnaire, is often given in terms of the standardized mean difference *SMD* to depict the effect of an educational intervention (Hedges & Rhoads, 2010). For example, based on his review of over 1,000 randomized intervention studies on student achievement, Kraft (2020) proposes considering an intervention effect of $SMD < 0.05$ as "small", $0.05 \leq SMD < 0.20$ as "medium", and $SMD \geq 0.20$ as "large." Further, when evaluating the meaningfulness of the intervention effect it is also recommended to take into account the cost and scalability of the intervention (Kraft, 2020; Lipsey et al., 2012). Finally, it may be helpful to compare the expected effect to empirical benchmark values, such as normative expectations of academic growth, performance differences between socio-demographic groups, or performance differences between preschools with weak and average performance levels (Brunner et al., 2023b; Dong et al., 2016; Lipsey et al., 2012).

Power analyses of randomized experiments require several design parameters to determine the *MDES* or the required sample size to achieve a certain statistical power (e.g., Dong & Maynard, 2013; Hedges & Rhoads, 2010).[3] First, most field-based randomized studies, including CRTs and MSRTs in preschool settings, implement hierarchical sampling strategies to reflect the multilevel nature of these environments. Children can be nested in daycare centers (two-level design), or they can be nested in groups, with those groups nested in daycare centers (three-level design). Clustering implies that the target outcome measures of children belonging to the same group or daycare center tend to be more similar to each other than to children in other groups or centers. Intraclass correlations (*ICC*s; ρ) quantify the degree of similarity between children in the same group or daycare center, measuring how much the target outcome tend to cluster together. ICC values can range from zero to one. For example, when using vocabulary test scores as the outcome in a single-level lab-based randomized experiment, the scores are not clustered because children's test scores are independently sampled, with no higher-order units

[3] In addition to the design parameters (i.e., R^2s and ICCs) that we present in this paper, power analyses of MSRTs require information on the expected heterogeneity of the treatment effect across daycare centers and the extent to which covariates may explain this heterogeneity (Dong & Maynard, 2013; Hedges & Rhoads, 2010). We further elaborate on this type of design parameters in the Discussion section.

(e.g., daycare centers) involved by definition. Hence, $\rho = 0$. Moreover, in a two-level design where children are nested in daycare centers, a value of $\rho = 0$ indicates that there are no mean differences in test scores between daycare centers, and all variability in the test scores is observed within the centers. Conversely, a value of $\rho = 1$ indicates that all children in the same daycare center achieve the same test score, meaning that the total observed variability in the test scores is due to mean-level differences between daycare centers rather than within. Importantly, when estimating the effect of an intervention, the similarity between children in the same group and/or daycare center makes the clustered data from CRTs less efficient than the unclustered data from single-level lab-based randomized experiment with the same total sample size (Hedges & Hedberg, 2007). In other words, all else being equal, CRTs with clustered data require a larger sample size than a single-level lab-based randomized experiment with unclustered data to achieve the same *MDES* or statistical power. Moreover, with MSRTs, the efficiency of estimating the intervention effect depends on the variance proportions attributable to (a) mean-level differences in the target outcome and (b) variations in the magnitude of the intervention effect across daycare centers. Specifically, due to the use of blocked random assignment in MSRTs, the variance attributable to mean-level differences between daycare centers in the target outcome is not considered when calculating the standard error of the intervention effect. As a result, when the magnitude of the intervention effect is the same (or very similar) across daycare centers, MSRTs can be more efficient than lab-based randomized experiments (Moerbeek & Teerenstra, 2016). However, when the heterogeneity of the intervention effect across daycare centers (and the corresponding variance proportion) becomes sufficiently large, MSRTs are less efficient than lab-based randomized experiments in many practical applications (Hedges & Rhoads, 2010; Moerbeek & Teerenstra, 2016). In summary, clustering has important consequences for power analyses. Specifically, the preschool setting that most children attend in Germany (Autor:innengruppe Bildungsberichterstattung, 2022) and elsewhere (e.g., the United States; U.S. Department of Education, National Center for Education Statistics, 2021) is a group within a daycare center. Therefore, the most relevant design parameters for planning CRTs and MSRTs in preschool settings are *ICC*s, which represent the proportion of total variance in children's outcomes attributable to differences between (a) groups (ρ_{Group}) within daycare centers and (b) daycare centers themselves (ρ_{Center}).

Second, covariates may substantially improve the sensitivity of randomized experiments in general, and single-level, lab-based studies with unclustered data (Porter & Raudenbush, 1987) and CRTs and MSRTs (Dong & Maynard, 2013; Hedges & Hedberg, 2007; Hedges & Rhoads, 2010; Raudenbush et al., 2007) in particular. Covariates remove noise in the variance of the outcome measure, improving the signal of the treatment effect (Raudenbush et al., 2007, p. 18). Covariates are not required for randomized experiments, but when they explain a substantial proportion of variance in outcomes (R^2), they are a very efficient way to improve (i.e., decrease) the *MDES* and to reduce the required sample size to achieve a certain level of statistical power (Hedges & Hedberg, 2007; Porter & Raudenbush, 1987; Raudenbush et al., 2007). Values of R^2 can range from zero to one. In single-level lab-based studies, information is needed on the proportion of total variance in the outcome (R^2_{Total})

that can be explained by covariates to guide researchers' decisions about inclusion. In CRTs and MSRTs, covariates may operate at various levels. To decide about the inclusion of covariates, researchers therefore need information on the proportion of variance in the outcome that can be explained by covariates at the individual child level (R^2_{Child}), the group level (R^2_{Group}), and the daycare center level (R^2_{Center}).

What Can Go Wrong in Power Analyses?

Design parameters, in terms of R^2_{Total} as well as *ICC*s and R^2s at various levels, are essential for power analyses of randomized studies. Two major problems can occur when using the "wrong" estimates for these design parameters. First, the actual degree of clustering may be larger and/or proportion of explained variance may be smaller than expected. This leads to reduced precision in estimating the intervention effect, diminished statistical power to detect that effect if it exists, and a *MDES* that is larger than the actual intervention effect. Consequently, the design sensitivity of the study is compromised, undermining the research team's statistical conclusions regarding the strength of the intervention effect in the target population. Second, the actual degree of clustering may be smaller and/or the proportion of explained variance may be larger than expected. Although this may result in very high design sensitivity with small standard errors, high statistical power, and an estimated *MDES* that is smaller than the actual intervention effect, it may also make the experiment inefficient from a cost perspective due to an unnecessarily large sample size. Hence, the research team and funding agencies invested more resources in the study than necessary. For these reasons, leading methodologists recommend that researchers base power analyses on reliable empirical estimates of design parameters (Bloom et al., 2007; Hedges & Hedberg, 2007; Hedges & Rhoads, 2010; Raudenbush et al., 2007), because research has shown that the value of design parameters depends strongly on the target outcome and target population (Brunner et al., 2018; Stallasch et al., 2021, 2024).

The Empirical Body of Knowledge on Design Parameters for Preschool Children

Cognitive and Socio-Emotional Learning Outcomes

Educational interventions in preschool may have a broad, positive impact on children's cognitive and socio-emotional development (Barnett, 2011; OECD, 2020). We therefore aim to provide reliable design parameters for two broad outcome domains: cognitive and SEL outcomes. We use cognitive outcomes as an umbrella term to cover children's skills and knowledge in various subdomains, including (a) math and science skills (e.g., counting skills), (b) verbal skills (e.g., vocabulary, sentence comprehension), as well as (c) general cognitive skills and knowledge (e.g., working memory, reasoning skills, general knowledge). In addition, drawing on the widely-accepted Cattell-Horn-Carroll (CHC) taxonomy of cognitive abilities (Flanagan & Dixon, 2014), we also consider children's psychomotor skills as a subdomain of cognitive outcomes. This

subdomain covers children's more general psychomotor skills (e.g., throwing a ball, jumping with two feet) as well as psychomotor skills that are relevant in everyday situations, for example, the ability to close a zipper or to walk up stairs (Sparrow et al., 2005).

SEL refers to the learning of knowledge, skills, and attitudes that enable individuals to regulate thoughts, emotions, and behavior; establish and manage interpersonal relationships; and achieve personal, academic, and collective goals (Durlak et al., 2022). According to the taxonomy by Schoon (2021), important SEL outcomes reflect affective, cognitive, or behavioral manifestations of socio-emotional characteristics. These characteristics can be grouped into three broad subdomains: (a) self-orientation (e.g., self-control, emotion regulation, neuroticism, and conscientiousness), (b) other-orientation (e.g., empathy, pro-social behavior, externalizing problems, aggressive behavior, disruptive behavior, extraversion, and agreeableness), and (c) task-orientation (e.g., interest, persistence, and openness). Drawing on these definitions and frameworks, we review research on single- and multilevel design parameters for the target population of preschool children.

Previous Research on Design Parameters: An Overview

Previous research on design parameters has contributed substantial knowledge across various target populations, including students in elementary and secondary school at national (Dong et al., 2016; Hedberg, 2016; Hedges & Hedberg, 2007; Jacob et al., 2010; Stallasch et al., 2021, 2024; Westine et al., 2013) and international levels (Brunner et al., 2018; Kelcey et al., 2016; Zopluoglu, 2012), as well as teachers (Westine et al., 2020) and students in community colleges (Somers et al., 2022). In particular, research on design parameters with student populations in elementary school—the target population most closely related to preschool children, which are the focus of this paper—has shown that design sensitivity can be substantially improved by including two types of covariate sets: (a) socio-demographic (SD) characteristics (e.g., children's age, socioeconomic status [SES], gender, or migration background) and (b) baseline measures of the target outcome (Dong et al., 2016; Hedges & Hedberg, 2007; Jacob et al., 2010; Kelcey et al., 2016; Stallasch et al., 2021, 2024; Westine et al., 2013). This research has also highlighted significant variation in design parameters. Specifically, cross-national variation has been observed, which limits the ability to transfer these parameters from one country to another (Kelcey et al., 2016; Stallasch et al., 2021; Zopluoglu, 2012). Even within the same nation, design parameters for achievement or SEL outcomes vary across achievement (Kelcey et al., 2016; Stallasch et al., 2021) and SEL domains (Dong & Maynard, 2013), particularly when using different types of assessments (e.g., parent and teacher reports; Hedberg, 2016). Consequently, design parameters cannot be easily generalized across domains or types of assessments. In summary, these findings emphasize the importance of developing and applying design parameters that align closely with the target population, outcome, and assessment of the planned study.

Design Parameters for Single-Level Randomized Experiments with Preschool Children

Design parameters for single-level lab-based studies (with individual random assignment) require knowledge about how much covariates enhance their sensitivity. Therefore, researchers need reliable estimates of R^2_{Total} for a target population and outcome, based on a specified set of covariates. In contrast to the body of knowledge for students in elementary or secondary education (Stallasch et al., 2024), there is no compilation of such single-level design parameters for SD characteristics and baseline measures as covariate sets for the target population of preschool children.

Nevertheless, relevant results can be found in several large-scale studies and meta-analyses. First, previous research has shown that SD characteristics are correlated with (and therefore explain variance in) cognitive and SEL outcomes in preschool children. In particular, children's age, especially during preschool, is substantively related to their cognitive and SEL outcomes. When children grow older their cognitive skills generally improve in all domains (Tucker-Drob, 2019), and their socio-emotional characteristics become more differentiated (Caspi et al., 2005), demonstrating a multidimensional age-based developmental pattern (Bleidorn et al., 2022). Further, meta-analytic results by Letourneau et al. (2013) indicate that higher values on family SES measures are associated with better cognitive and somewhat better SEL outcomes in preschool children (e.g., lower levels of externalizing and internalizing behavior problems). This pattern of results was also confirmed by an international large-scale assessment with representative samples from the United States, England, and Estonia (OECD, 2020). Moreover, this latter study also showed that girls in preschool have higher levels of verbal skills and SEL outcomes (e.g., better prosocial and less disruptive behavior). Further, this study also found that preschool children with migration background had lower levels of verbal and mathematical skills. Finally, these children were reported by their educators to demonstrate less prosocial, but also less disruptive behavior (OECD, 2020).

Second, it is well established that preschool children's baseline measures substantially predict their future cognitive performance (e.g., when using prior knowledge as baseline measure; Simonsmeier et al., 2022) and socio-emotional characteristics (e.g., when using other-reports as a baseline measure of children's temperament or personality; see Table S7 in Bleidorn et al., 2022).

Design Parameters for Multilevel Randomized Experiments with Preschool Children

In stark contrast to target student populations in elementary or secondary school, little is known about multilevel design parameters for children attending preschool. The most comprehensive source on design parameters is the (largely unknown) data supplement that comes along with the Optimal Design power analysis software (Spybrook et al., 2011). These data were obtained from three large-scale longitudinal studies of children enrolled in Head Start centers in the US, with several waves

of measurement between 1997 and 2006. In addition, Jacob et al. (2010) provided a few multilevel design parameters for cognitive outcomes using data (collected between 2004 and 2009) with US samples of preschool children from low-income families living in Chicago. Design parameters for populations in other countries are even more difficult to obtain because relevant results are (a) scattered across individual studies, (b) usually available only for between-center differences (i.e., ρ_{Center}) but not for other key design parameters (i.e., ρ_{Group}, R^2_{Child}, R^2_{Group} or R^2_{Center}), and (c) not yet systematically summarized. For example, between-center differences have been reported as auxiliary results in studies involving large-scale samples with preschool children in Germany (Leyendecker et al., 2014; Ulferts, 2017)[4] and the United Kingdom (Sammons et al., 2002, 2003).

Integrating the results on multilevel design parameters from these selected studies reveals the following pattern (see also Figure A1 for a more detailed overview): First, in the United States, there were mostly small between-center differences in cognitive outcomes (e.g., when applying standardized tests: *Mdn* $\rho_{Center} = 0.03$) and SEL outcomes (educator reports: *Mdn* $\rho_{Center} = 0.01$; parent reports: *Mdn* $\rho_{Center} = 0.00$). These between-center differences were (much) more pronounced in Germany and the United Kingdom for both cognitive (*Mdn* $\rho_{Center} = 0.13/0.17$) and SEL outcomes (educator reports: *Mdn* $\rho_{Center} = 0.28/0.05$). Second, the size of between-group differences in the United States depended on the combination of outcome domain and method of assessment. Typical values for cognitive outcomes were *Mdn* $\rho_{Group} = 0.05/0.08$ when using standardized tests/parent reports; typical values for SEL outcomes were *Mdn* $\rho_{Group} = 0.19/0.00$ when using teacher/parent reports. The studies from Germany and the United Kingdom used two-level designs (children nested in daycare centers) and thus cannot provide results for ρ_{Group}. Third, for the United States, information was available for R^2_{Child} and R^2_{Center}, but not for R^2_{Group}. Further, most estimates were available for R^2_{Child} and using a covariate set comprising SD characteristics and a baseline measure. Applying these covariates, typical values for cognitive outcomes were *Mdn* $R^2_{Child} = 0.24/0.21$ when using standardized tests/parent reports, and *Mdn* $R^2_{Child} = 0.26/0.18$ for SEL outcomes when using teacher/parent reports. Design parameters for R^2_{Center} were only available for cognitive outcomes measured with standardized tests: Median values of R^2_{Center} were in the range $0.20 \leq R^2_{Center} \leq 0.89$. The studies from Germany and the United Kingdom did not report results on R^2*s* at different levels.

The Present Study

There is a strong need for early childhood education research to provide robust evidence on which educational interventions and practices are effective in preschools. This requires researchers to conduct lab- and field-based randomized experiments

[4] These studies drew on a subset of the data that we also used for estimating design parameters in the present paper (i.e., data from the BIKS and NUBBEK study). Of note, these previous studies presented only results for ρ_{Center} for a limited set of outcome variables.

to enable causal inferences about intervention effects with preschool children. Using reliable estimates of design parameters in a priori power analyses is essential to ensure that these studies can support strong statistical conclusions regarding the strength of the intervention effect in the target population (Hedges & Hedberg, 2007; Hedges & Rhoads, 2010). However, knowledge of these parameters has been largely lacking in early childhood education research. Moreover, our review of the limited available design parameters for preschool children, along with research on design parameters for elementary school children, highlights that these parameters can vary across target populations, outcomes, and types of assessments. As a result, design parameters specific to the target population, outcome, and type of assessment in early childhood education research are needed.

Therefore, the overarching goals of this paper are to (a) significantly expand the knowledge base on design parameters relevant to preschool children, (b) synthesize these parameters to build a systematic body of knowledge, and (c) integrate this knowledge into the relevant methodological literature, providing a thorough discussion and illustration of how to apply it when conducting a priori power analyses during the planning stage of randomized experiments in both lab- and field-based settings. In doing so, our paper makes the following unique contributions to early childhood education research. First, we address a major gap in knowledge about design parameters for randomized experiments with preschool children. Specifically, there is currently no comprehensive compilation of design parameters (i.e., R^2_{Total}) for single-level (e.g., lab-based) randomized experiments targeting the population of preschool children. Furthermore, aside from the specific population of socioeconomically disadvantaged preschool children in the United States (Jacob et al., 2010; Spybrook et al., 2011), there is limited systematic knowledge about design parameters for two-level or three-level randomized field studies (e.g., CRTs and MSRTs) for broader populations involving preschool children in the United States or in other countries. To address these significant gaps in the literature, we provide a comprehensive collection of single- and multilevel design parameters for planning randomized experiments with preschool children. To this end, we conducted an IPD meta-analysis using several datasets based on probability samples of preschool children attending daycare centers. We focus our analyses on children attending daycare centers, as this is the preschool setting that most children (aged 3 years or older) attend in Germany (see Table C3-3web in Autor:innengruppe Bildungsberichterstattung, 2022) and many other countries (e.g., the United States; U.S. Department of Education, National Center for Education Statistics, 2021). A second contribution of our paper to early childhood education research is to offer design parameters for a large variety of cognitive and SEL outcomes because interventions in preschool may target different dimensions of children's development. A third contribution is our provision of design parameters for cognitive and socio-emotional characteristics, using both parent reports (representing the family context) and educator reports (representing the preschool context), acknowledging that these parameters may differ between sources (Hedberg, 2016) as children may exhibit different behaviors across contexts (Mischel & Shoda, 1995). A fourth contribution is to offer design parameters for vital covariate sets, as covariates may substantially improve the sensitivity of randomized experiments (Hedges & Hedberg, 2007; Porter & Raudenbush,

1987; Raudenbush et al., 2007). In accordance with previous research on design parameters for the school context (Hedges & Hedberg, 2007; Jacob et al., 2010; Stallasch et al., 2021, 2024), we estimate single- and multilevel R^2's for covariate sets including (a) sociodemographic characteristics, (b) baseline measures of the target outcome, and (c) sociodemographic characteristics and baseline measures combined. Given the strong developmental dynamics of cognitive and SEL outcomes in early childhood, it is not always possible to use identical baseline (IB) measures. We therefore also explored the predictive utility of proxy baseline (PB) measures. Specifically, we used children's vocabulary knowledge as a PB measure for cognitive outcomes because it is known to predict future learning in many cognitive domains (Peng & Kievit, 2020). Further, children's problem behavior is theoretically, conceptually, and empirically linked to children's SEL outcomes (De Pauw & Mervielde, 2010; Schoon, 2021; Tackett et al., 2013). We therefore used an important facet of children's (externalizing) problem behavior from the SEL subdomain "other-orientation" as a PB measure (see Schoon, 2021). Specifically, we chose children's disruptive behavior (e.g., a child interrupts or disturbs other children), as obtained from parent or educator reports, as a PB measure for SEL outcomes as assessed with same method. A fifth contribution of our paper is to provide standard errors for all design parameters, as any estimate of a design parameter is subject to sampling error (Jacob et al., 2010; Stallasch et al., 2021). These standard errors are essential for accounting for the statistical uncertainty of design parameters when used in an a priori power analysis. A sixth contribution of our paper to early childhood education research is to (a) build systematic knowledge on design parameters, (b) examine their generalizability, and (c) provide normative values. To achieve this, we meta-analytically integrated single-, two-, and three-level design parameters across domains, subdomains, applied measures, assessment methods, and age groups.

Method

Database Search for Large-Scale Studies

To build systematic knowledge on design parameters for preschool children, we applied a two-stage approach to the meta-analysis of IPD (Brunner et al., 2023a). To this end, we first carried out a systematic search for IPD of German large-scale studies with preschool children (see Fig. 1). Specifically, we sought studies that met the following inclusion criteria: The studies should (a) include probability samples targeting the general population of preschool children in Germany or specific federal states to avoid sample selectivity bias and ensure coverage of the full range of target outcomes, thereby enhancing the generalizability of the results; (b) use daycare centers as the primary sampling unit to facilitate the estimation of multilevel design parameters; (c) be conducted in the year 2000 or later to ensure that the design parameters are current; and (d) apply an observational study design (i.e., not an experimental design) to align with the methodology used in most studies providing design parameters in the school context (e.g., Bloom et al., 2007; Brunner et al., 2018; Hedges & Hedberg, 2007; see Stallasch et al., 2021 and 2024

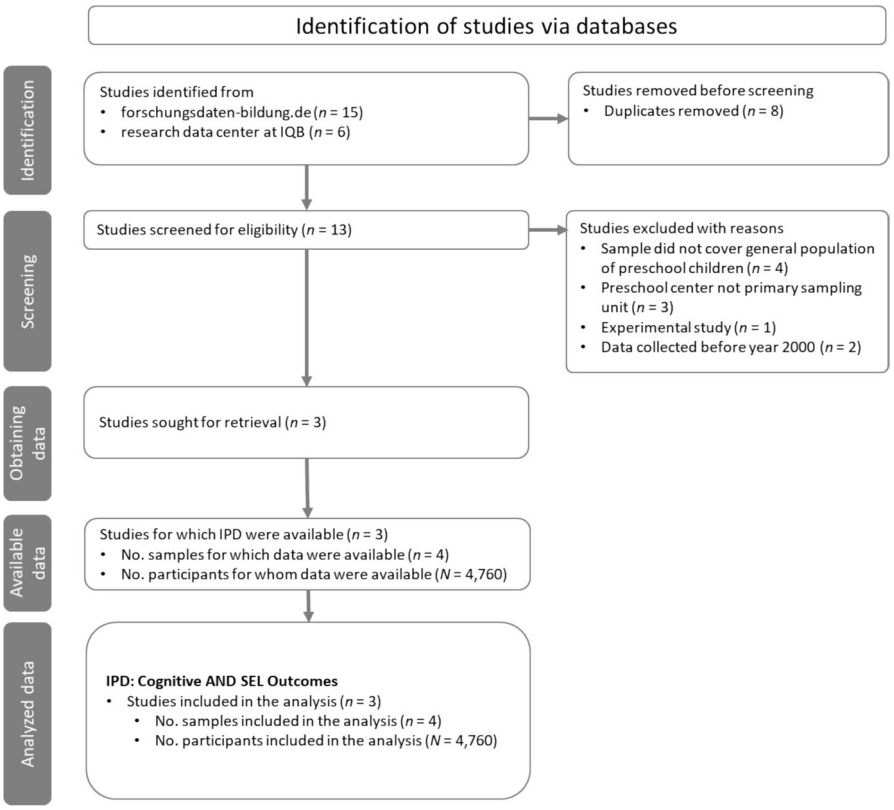

Fig. 1 Flow Diagram to Identify the Large-Scale Studies That Were Used to Estimate Design Parameters for Preschool Children. *Note*. Adapted from the PRISMA 2020 (Page et al., 2021) and PRISMA IPD standards (Stewart et al., 2015)

for comprehensive overviews), which also assures that design parameters are not affected by an educational intervention (see Jacob et al., 2010). To identify studies meeting these inclusion criteria, we searched (on January 29, 2024) two key German electronic data repositories (OSM A2). The search in the German Network of Educational Research Data repository[5] returned a total 15 studies (Tables A1 and A2). The search in the Research Data Centre (FDZ)[6] returned a total of six studies (Table A3). After removing eight duplicate studies (Table A4), we screened the description of 13 studies. We excluded 10 of these 13 studies because four studies did not cover the general target population, three studies were not based on samples with daycare centers as the primary sampling unit, one study implemented an experimental design, and two studies were collected before the year 2000 (Table A5).

[5] https://www.forschungsdaten-bildung.de/en/studies/search
[6] https://www.iqb.hu-berlin.de/fdz/studies/

IPD from the remaining three studies were sought for retrieval and obtained: (a) the National Survey on Education, Care, and Development in Early Childhood (NUBBEK; Tietze et al., 2015), (b) the study on Educational Processes, Competence Development, and Selection Decisions in Preschool and School Age (BIKS; Weinert et al., 2019), and (c) the starting cohort of preschool children (starting cohort 2) of the National Educational Panel Study (NEPS; NEPS Network, 2022; Blossfeld & Roßbach, 2019).

Studies and Samples

NUBBEK (carried out in eight federal states of Germany in 2010 and 2011) is a cross-sectional study with two samples representing two-year and four-year-old children (Leyendecker et al., 2014). BIKS (carried out in two federal states of Germany) is a longitudinal panel study with two cohorts (i.e., preschool children and students in primary education) that started in 2006. We used the data from the cohort of three-year old preschool children that were followed up to age six before they entered primary school. The time intervals between waves of measurement were about six months (Homuth et al., 2024). To estimate design parameters, we defined separate BIKS samples for each wave of measurement because the planned missing data design that was applied in this study resulted in a large number of missing values. Notably, for this reason we excluded data from the second wave of measurement because reliable imputation of missing data was not possible. Further, in each remaining wave of measurement we excluded data from those daycare centers for which no target outcome data were available. Finally, NEPS (carried out in all federal states of Germany) is an ongoing multi-cohort longitudinal panel study (Artelt & Sixt, 2023; Blossfeld & Roßbach, 2019). We used the data for the cohort of 4-year old children from the first two waves of measurement when they were in preschool. Data collection started in 2011; the time interval between waves was about nine months. All three studies followed a multistage sampling procedure where a random selection of daycare centers was drawn in the first stage of sampling. All children who attended the selected centers and who fulfilled the age-based inclusion criteria were invited to participate in a specific study. We used the IPD from the children who actually participated in these studies for our statistical analyses. As information on groups within daycare centers was missing for some children in the second wave of measurement in the NEPS, we defined separate samples for the first and second wave of measurement to estimate design parameters. In summary, the IPD of NUBBEK, BIKS, and NEPS were used to estimate design parameters for children aged 2, 3, 4, 5, and 6 years. Information on key characteristics for each sample can be found in Table 1.[7]

[7] Notably, we did not apply sampling weights when estimating design parameters because (a) information on weights was only available for NEPS and (b) the application of sampling weights is not possible with the lme4 package (Bates et al., 2015) that we used for the multilevel analyses. Therefore, our design parameters based on NEPS are representative only for the population of preschool children included in the present analyses.

Table 1 Description of the Samples that Were Applied to Estimate Design Parameters

Statistic	BIKS[a]					NUBBEK		NEPS	
	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	2-year-olds	4-year-olds	Wave 1	Wave 2
Sample Size									
$n_{Children}$	518	468	460	457	425	564	714	2928	2727
n_{Groups}	–	–	–	–	–	–	–	719	690
$n_{Centers}$	89	81	80	79	72	202	220	277	275
Mdn $n_{Children.Group}$	–	–	–	–	–	–	–	4	3
Mdn $n_{Children.Center}$	6	6	6	6	6	3	3	10	9
Mdn $n_{Group.Center}$	–	–	–	–	–	–	–	2	2
Age (in months)									
M	42.2	54.0	60.0	66.0	72.0	33.1	53.9	57.8	66.7
SD	4.1	4.1	4.1	4.1	4.0	2.0	3.7	3.9	3.8
% girls	48	48	47	47	47	49	51	49	50
% migration	20	18	19	19	16	19	30	31	30
Years of education									
M	15.0	15.0	15.0	15.0	15.1	15.6	14.9	14.5	14.5
SD	2.5	2.5	2.5	2.5	2.5	2.8	3.0	2.4	2.4
% missing data									
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1
25th Percentile	4.1	6.4	1.3	14.4	5.4	0.4	0.1	2.5	1.5
Mdn	5.0	6.4	7.6	14.7	17.5	2.5	2.9	10.3	10.0
75th Percentile	10.2	11.7	29.3	18.3	18.5	4.6	3.2	15.4	14.8
Max	19.7	39.1	29.3	45.3	39.8	18.9	3.4	23.4	28.7

BIKS Educational Processes, Competence Development and Selection Decisions in Preschool and School Age. *NUBBEK* National Survey on Education, Care, and Development in Early Childhood. NEPS National Educational Panel Study (NEPS) – starting cohort 2 with kindergarten children. $n_{Children}$ total number of children. n_{Groups} total number of groups within daycare centers. n_{center} total number of daycare centers. *Mdn* $n_{Children.Group}$ median number of children per group within daycare centers. *Mdn* $n_{Children.Center}$ median number of children per daycare center. *Mdn* $n_{Group.Center}$ median number of groups per daycare center. % migration percentage of children with migration background. Years of education highest educational level of education in the family in terms of the completed years of education. % missing data percentage of missing data per variable

[a]: Data of the original second wave of measurement of BIKS (when children were on average 48 months old) were not used in the present paper because of the large number of planned missing values. Wave 2 in this table refers to the third wave of measurement of BIKS when children were on average 54 months old

Measures

NUBBEK, BIKS, and NEPS utilized standardized tests to assess cognitive outcomes, covering the subdomains of early mathematics/science and verbal skills as well as general cognitive skills. Further, NUBBEK also provided parent and teacher report data on children's verbal skills and psycho-motor skills. Reliabilities for cognitive outcomes were $0.39 \leq r_{tt} \leq 0.90$ ($Mdn = 0.78$) for standardized tests,

$0.68 \leq r_{tt} \leq 0.88$ ($Mdn = 0.83$) for parent reports, and $0.78 \leq r_{tt} \leq 0.95$ ($Mdn = 0.88$) for educator reports. In addition, all studies provided rich data on children's SEL outcomes as assessed by parent and educator reports. SEL outcome measures targeted the subdomains self-orientation, other-orientation, and task-orientation. Reliabilities for SEL outcomes were $0.35 \leq r_{tt} \leq 0.90$ ($Mdn = 0.70$) for parent reports and $0.55 \leq r_{tt} \leq 0.93$ ($Mdn = 0.81$) for educator reports. Tables A6 to A10 in OSM A3 present further details on the applied measures.

Statistical Analyses

The two-stage approach of individual participant data (IPD) meta-analysis applied in this paper comprised two stages: In Stage 1, missing data were imputed, and design parameters were estimated separately for each sample and wave of measurement (when multiple waves were available). In Stage 2, these estimates were summarized meta-analytically.

Stage 1: Treatment of Missing Data

The percentage of missing values per variable varied from 0% to 45.3% (Table 1). To deal with missing data we used an adjusted cluster-means imputation approach for multilevel data (Grund et al., 2023) and generated 50 multiply imputed datasets for each sample (and wave of measurement) using the mice (van Buuren & Groothuis-Oudshoorn, 2011) and miceadds (Robitzsch & Grund, 2023) R packages. Notably, the applied multilevel imputation models were compatible with the models employed for estimating single- and multilevel design parameters. Design parameters were pooled across imputations using Rubin's (1987) rules. We used the mitml package (Grund et al., 2021) to combine the estimates into a single set of results and to obtain standard errors that take into account within and between imputation variance. The methods for estimating the within-imputation variance (i.e., the standard errors for single- and multilevel design parameters) are detailed in OSM A4.

Stage 1: Estimation of Design Parameters

Single-Level Design Parameters To estimate single-level design parameters (i.e., R^2_{Total}) for each sample (and wave of measurement), we used the R function lm (R Core Team, 2024) with ordinary least squares (OLS) to analyze each cognitive or SEL outcome using up to five sets of linear regression models. Model Set 1-SD comprised five SD characteristics, including children's age, gender, migration background, and two measures of parents' socioeconomic status (SES). The available SES measures varied somewhat across studies. We used the highest educational attainment within the family (i.e., the greatest number of years of schooling completed within a family) in all studies, the highest International Socio-Economic Index of Occupational Status within a family (HISEI; Ganzeboom & Treiman, 1996) in BIKS and NEPS, and family income in NUBBEK. Model Set 2-IB comprised an identical baseline (IB) measure of the target cognitive or socio-emotional

outcome. We selected baseline measures with the shortest possible time lag between waves of measurement and the same method of assessment (i.e., standardized test, educator report, or parent report). In Model Set 2-PB we used children's vocabulary knowledge as the PB measure for cognitive outcomes, and children's disruptive behavior (e.g., a child interrupts or disturbs other children) as obtained from parent or educator reports as the PB measure for SEL outcomes assessed with the same method. Notably, most results for Model Set 2-PB were obtained for the second wave of measurement of NEPS. Model Set 3-SD + IB and Model Set 3-SD + PB comprised SD characteristics and the selected IB or PB baseline measure of the target outcome. Using these procedures, we estimated a total of $k = 237$ single-level design parameters.

Two- and Three-Level Design Parameters To ensure reliable estimation of multilevel design parameters, particularly the variance components of random effects, we followed the recommendation of McNeish and Stapleton (2016) by using restricted maximum likelihood estimation (REML) as implemented in the R package lme4 (Bates et al., 2015). Specifically, we analyzed up to six sets of multilevel models for each sample (and wave of measurement). Model Set 0 comprised so-called "empty" multilevel models to estimate ICCs. We estimated five additional sets of multilevel models (Model Sets 1-SD, 2-IB, 2-PB, 3-SD + IB, and 3-SD + PB) for the covariates that we previously applied to estimate single-level design parameters. Given the available data and the study-specific sampling designs, we used the IPD of NEPS, BIKS, and NUBBEK to specify two-level models (children nested in daycare centers) for estimating the two-level design parameters (i.e., ρ_{Center}, R^2_{Child}, and R^2_{Center}) for each outcome and model set. In addition, we drew on the NEPS data to specify three-level models (children nested in groups, with groups nested in daycare centers) for estimating three-level design parameters (i.e., ρ_{Group}, ρ_{Center}, R^2_{Child}, R^2_{Group} and R^2_{Center}) for each outcome and model set. All covariates were assessed at the child level. In addition to covariates at the child level, for each multilevel model set we entered daycare center means into the two-level models, and group and daycare center means into the three-level models. Further, we applied group-mean centering: covariates at the child level were centered around their respective daycare center/group means in the two-/three-level models, and group means were centered around their respective daycare center means in the three-level models (Raudenbush & Bryk, 2002). Using these procedures, we estimated a total of $k = 589/226$ two-level/three-level design parameters.

Stage 2: Meta-Analytic Integration

To synthesize the sample-specific design parameters obtained in Stage 1, we provide numerous meta-analytical summaries (see Tables A11 to A15) using the R package metafor (Viechtbauer, 2010). Because random-effects models cannot be expected to reliably gauge the heterogeneity of a specific (true) design parameter when less than $k = 10$ observed design parameters are available (Langan et al., 2019, p. 95), we applied (multivariate) fixed-effects models (Rice et al., 2018) when $2 \leq k < 10$, and

(multivariate) random-effects models (Hedges, 2019) when $k \geq 10$ (see OSM A3 for further details).

To assess the heterogeneity of design parameters, we computed the 95% prediction interval (95% PI), which provides a plausible range of values in which the true value of a specific design parameter of about 95% of all relevant populations will fall. We also estimated the standard deviation of the random effects (σ), I^2, and the Q statistic as additional heterogeneity measures (Borenstein et al., 2017). Notably, all design parameters have a theoretical range between zero and one. When the lower (or upper) bound value obtained for the 95% confidence interval (95% CI) for the meta-analytic average or the 95% PI was below zero (or above one), we truncated these values. Several design parameters were obtained for the same sample. To take the resulting within-sample dependencies into account we used the R package clubSandwich (Pustejovsky, 2021) to impute a working covariance matrix for the observed effect sizes (Hedges, 2019). We used $r = 0.70$ as a reasonable upper-bound estimate for the within-sample correlations among design parameters. Because we used an estimated working covariance matrix rather than an empirical one, we conducted sensitivity analyses (Hedges, 2019). These analyses corroborated that the meta-analytic statistics were fairly robust against the different values chosen for the correlation among design parameters (see OSM A5 for details).

Results and Discussion

Figure 2 presents the point estimates of design parameters that we estimated in Stage 1 as well as their meta-analytic summaries from Stage 2. Drawing on the empirical results, we discuss the use of the present design parameters for a priori power analyses to plan randomized intervention studies with preschool children, along with specific recommendations for early education researchers.

Match Design Parameters to Key Characteristics of the Target Intervention

When carrying out an a priori power analysis for a randomized intervention study it is important to strive for an ideal match between the selected set of design parameters on the one side and key characteristics of the planned intervention on the other (Bloom et al., 2007; Brunner et al., 2018; Hedges & Hedberg, 2007; Hedges & Rhoads, 2010; Stallasch et al., 2021, 2024). For one, this involves selecting design parameters that offer the best match to the assumptions of how the data of the planned study will be clustered. Specifically, single-level design parameters are required for lab-based randomized experiments where nonclustered data are assumed. Two-level parameters (children in daycare centers) and three-level parameters (children in groups, with groups in daycare centers) are needed for field-based randomized experiments.

In addition, we often observed substantial heterogeneity of design parameters, which was reflected in the 95% PIs in the meta-analytic summaries (Fig. 2). For

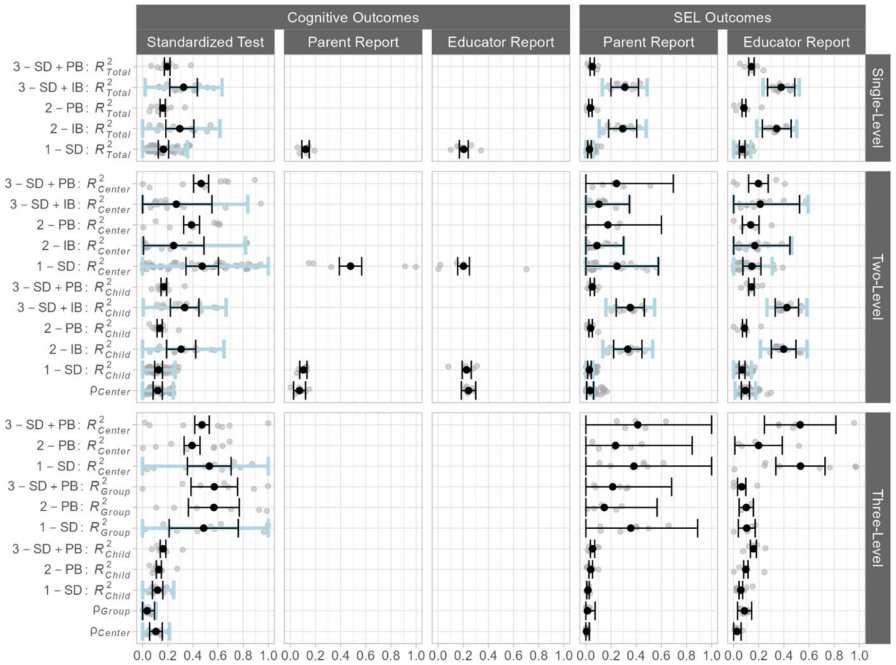

Fig. 2 Single-, Two-, and Three-Level Design Parameters by Outcome Domain and Assessment Method. *Note*. The grey circles show the point estimates of design parameters. The black circles depict the meta-analytic averages; error bars in black/blue color depict their 95% CIs/95% PIs. Lower/upper bound values of these intervals outside the possible value range were truncated at 0/1, respectively. PIs were only computed when $k \geq 10$ (see Method section). 1-SD = Model set 1 with five socio-demographic characteristics as covariates; 2-IB = Model set 2 with one identical baseline measure as covariate; 2-PB = Model set 2 with one proxy baseline measure as covariate; 3-SD + IB = Model Set 3 with five socio-demographic characteristics and one identical baseline measure as covariates; 3-SD + PB = Model Set 3 with five socio-demographic characteristics and one proxy baseline measure as covariates

example, the 95% PI of ρ_{Center} obtained for two-level designs was [0.00, 0.24] for cognitive outcomes assessed with standardized tests, and [0.01, 0.06]/[0.00, 0.18] for SEL outcomes assessed with parent/educator reports. Given their heterogeneity, design parameters should be matched to the target population and outcome of the intervention, ideally using a set of design parameter point estimates (e.g., for a two-level CRT: ρ_{Center}, R^2_{Child}, and R^2_{Center}) that were obtained for the same or a very similar target measure and target age group.

Yet, certain circumstances may limit this endeavor, such as the unavailability of suitable estimates for a specific data structure, measure, age group, or covariate set. Here, two strategies may be helpful. First, virtually all daycare centers in Germany use some sort of grouping for their children (e.g., by age, or by a specific team of educators). Hence, random assignment of daycare centers ideally requires three-level design parameters. When suitable parameters are not available (i.e., information on ρ_{Group} and R^2_{Group} is missing), relevant two-level design parameters can be used because between-group differences in cognitive and SEL outcomes were often

(very) small (Fig. 2), and the variance component attributable to these differences is reflected in the present two-level parameters (Zhu et al., 2012). Thus, planning a randomized experiment based on the available two-level information (though having a three-level data structure) will likely lead to very similar results (in terms of the *MDES* or required sample size), particularly when including covariates at the child level (Zhu et al., 2012).

Second, the meta-analytic results obtained for broader outcome domains, subdomains, measures, or age groups which provide the best match to the application context can be used in power analysis when point estimates of design parameters are not (a) available, (b) useful (e.g., when the intervention targets verbal skills rather than a specific measure), or (c) reliable (i.e., when they have large standard errors). Doing so helps to determine sample sizes, power rates, or *MDES* values assuming typical values (i.e., meta-analytic averages) and plausible lower and upper bounds by drawing on the 95% PIs (if available). The 95% PIs may become especially important when researchers plan randomized experiments with target outcome measures or target populations of preschool children that differ substantially from those applied in the present study.

Beware of Between-Group and Between-Center Differences in the Target Outcome

Everything else being equal, clustering renders multilevel field-based studies less sensitive than single-level lab-based studies, where data are assumed not to be clustered. The most important design parameters for planning field-based randomized experiments in preschool settings therefore comprise information on between-group and between-daycare center differences in children's outcomes. When using two-level designs, the meta-analytic averages for between-center differences lay in the range $\rho_{Center} = 0.03$ (parent report of SEL outcomes) and $\rho_{Center} = 0.24$ (educator report of cognitive outcomes). When using three-level designs, the meta-analytic averages for between-center differences in three-level designs varied between $\rho_{Center} = 0.01$ (parent report of SEL outcomes) and $\rho_{Center} = 0.11$ (standardized tests cognitive outcomes).[8] Meta-analytic averages of between-group differences (within daycare centers) lay in the range $\rho_{Group} = 0.01$ (parent report of SEL outcomes) and $\rho_{Group} = 0.09$ (educator report of SEL outcomes).

The impact of between-group and/or between-center differences on the sensitivity of randomized experiments with preschool children can be illustrated using the *MDES* (two-sided testing; $\alpha = 0.05$; $1 - \beta = 0.80$) as a measure of design sensitivity. For example, when using the meta-analytic average across all SEL outcomes assessed with educator reports, a two-level CRT (with balanced allocation of 50 daycare centers to experimental groups; 20 children per center) was considerably

[8] Of note, the variation in between-center differences between two- and three-level designs can be largely attributed to using only the subset of cognitive or SEL outcomes obtained for NEPS for estimating the three-level parameters.

less sensitive (*MDES* = 0.30) than a corresponding single-level lab-based study with $N = 1{,}000$ children and individual random assignment (*MDES* = 0.18), although the total sample sizes of these studies were the same. Notably, Figure A6 in the OSM further illustrates the impact of between-group and/or between-center differences on the sensitivity of randomized experiments. To sum up, cognitive and SEL outcomes may demonstrate substantial between-group and between-center differences that imply that randomized experiments with preschool children may become noticeably less sensitive when using field-based intervention designs (e.g., two-level CRTs) rather than carrying out single-level lab-based randomized experiments.

Use Pre-Treatment Covariates to Improve the Sensitivity of Randomized Experiments

Adjusting for pre-treatment covariates to estimate the intervention effect in a randomized experiment does not change the size or nature of the estimated treatment effect. However, covariates can substantially improve the sensitivity of randomized experiments (Lin, 2013; Maxwell et al., 2018; Porter & Raudenbush, 1987). For example, everything else being equal, larger values of R^2_{Total}, R^2_{Child}, R^2_{Group}, and R^2_{Center} lead to smaller values of the *MDES*. Importantly, when covariates are to be used it is essential to ensure that these covariates are measured before random assignment (i.e., pre-treatment covariates). Otherwise, covariates could be affected by the treatment and therefore act as "bad controls" when used to estimate an adjusted treatment effect (Cinelli et al., 2022).

When used as pre-treatment covariates, all the covariate sets we investigated have the capacity to improve (at least somewhat) the sensitivity of randomized experiments with preschool children. The point estimates as well as the meta-analytic averages obtained for the various covariate sets showed that covariates typically explained some— and often a considerable proportion— of the variance in the outcome in total as well as at the child, group and center levels (see Fig. 2). Notably, the combination of SD and IB measures or SD and PB measures typically improved the explained amount of variance compared to using either set alone. Thus, when feasible, using Model Set 3-SD + IB or Model Set 3-SD + PB will generally lead to the most significant improvements in design sensitivity in most intervention scenarios compared to designs that do not incorporate covariates. For example, using model Set 3-SD + PB for cognitive outcomes measured with standardized tests improved the *MDES* from 0.18 to 0.16 in single-level designs, and from 0.33 to 0.26 in two-level and three-level CRTs, based on the sample specifications noted above. These improvements were achieved using the meta-analytic averages (shown in Fig. 2) of R^2_{Total} in single-level designs, R^2_{Child} and R^2_{Center} (along with ρ_{Center}) in two-level CRTs, and R^2_{Child}, R^2_{Group} and R^2_{Center} (along with ρ_{Group} and ρ_{Center}) in three-level CRTs. Additional scenarios are illustrated in OSM A6.

Importantly, the ability of covariates to improve the sensitivity of two- or three-level randomized experiments depends on (a) the level at which random assignment is implemented, (b) where the covariate is located, and (c) the degree of clustering in the data (Bulus & Sahin, 2019; Konstantopoulos, 2012). Specifically, CRTs

randomly assign entire daycare centers to experimental groups, whereas MSRTs offer more flexibility since random assignment can apply to individual children or groups within centers. Additionally, power analyses for MSRTs require assumptions about how much the intervention effect varies across daycare centers and how much of this variation can be explained by covariates (see also Discussion section; Dong & Maynard, 2013; Konstantopoulos, 2008). Given these complexities and the limitations of the applied data, which did not allow for the estimation of these additional design parameters, we focus our discussion on CRTs that involve the random assignment of entire daycare centers.[9] When between-center differences in CRTs are $\rho_{\text{Center}} \geq 0.10$, large proportions of explained variance at the daycare center level (in terms of R^2_{Center}) result in significant improvements in design sensitivity. For such clustered data, covariates at the daycare center level outperform (group-mean centered) covariates at the child or group levels in many practical applications (Bulus & Sahin, 2019; Konstantopoulos, 2012). Conversely, when between-center differences in CRTs are negligible to small ($\rho_{\text{Center}} < 0.10$), large proportions of explained variance at the daycare center level do not result in significant improvements in design sensitivity. This becomes evident in Figure A6 when looking at three-level designs with SEL outcomes assessed with parent reports. There is not much of a difference in *MDES* values between the models that do (all *MDES*s = 0.19) and do not use covariates (*MDES* = 0. 20) because the between-center differences were very small ($\rho_{Center} = 0.01$). Moreover, as demonstrated by Konstantopoulos (2012) and Bulus and Sahin (2019), when between-center and between-group differences are very small (e.g., $\rho_{\text{Center}} \leq 0.02$ and $\rho_{\text{Group}} \leq 0.02$), (group-mean centered) covariates at the child level can often outperform those at the daycare center level in improving design sensitivity in many practical settings. In summary, adjusting for pre-treatment covariates in single-level lab-based experiments and multilevel CRTs and MSRTs typically improves the sensitivity of randomized experiments, at least to some extent (Bulus & Sahin, 2019; Konstantopoulos, 2008, 2012; Maxwell et al., 2018). To study the impact of covariates on design sensitivity, we recommend carrying out power analyses both assuming the application of different pre-treatment covariate sets and without covariates. This approach will also allow the detection of the rare situations when covariates do not explain sufficient variance in the outcome to outweigh the loss in degrees of freedom in the statistical tests, for example when using very small samples (Konstantopoulos, 2012; Maxwell et al., 2018).

[9] Bulus and Sahin (2019) offer analytic solutions for evaluating the relative effectiveness of covariates at the child, group, or daycare center levels in enhancing the design sensitivity of two-level and three-level CRTs. For instance, their work identifies the conditions under which daycare center-level covariates are more effective in improving the design sensitivity of a two-level CRT compared to child-level covariates (and vice versa). Notably, Bulus (2022) extends this analysis by providing analytic solutions to determine these conditions for quasi-experimental regression discontinuity designs.

Account for the Statistical Uncertainty of Design Parameters in Power Analyses

Taking advantage of large-scale probability samples implied that most standard errors obtained for the point estimates of the present design parameters were relatively small (see Figure A2).[10] How the statistical uncertainty of design parameters (as reflected by their standard errors and 95% CIs) is taken into account in a power analysis depends on the risk preferences of the intervention researchers and their funding agencies as well as the cost structure of the project (Jacob et al., 2010). For example, for high-profile interventions with strong interest in detecting the intervention effect (e.g., for policy implementation), researchers may want to apply conservative estimates of the *MDES* and required samples sizes. To this end, they can draw on the upper bound estimates of the 95% CI's for ρ_{Group} and/or ρ_{Center} in combination with the lower bound estimates of R^2_{Total}, R^2_{Child}, R^2_{Group}, and $R^2_{\text{Center.}}$ When standard errors are relatively large (e.g., larger than 0.05),[11] we recommend also looking for alternative sets of point estimates of design parameters that (a) match the target outcome and target population (e.g., other outcomes belonging to the same subdomain). In such cases, researchers can also use meta-analytic averages and their 95% CIs that were obtained for higher aggregate levels or (if standard errors for these averages are also large) alternative aggregate levels that could be estimated with higher statistical precision. For example, one could use the age-specific meta-analytic average for a subdomain rather than the age-specific point estimates obtained for a specific measure or apply the domain-specific meta-analytic average for a certain age-group rather than the meta-analytic average for a specific subdomain.

Application

How Should Appropriate Design Parameters be Selected?

To support power analyses for single-level and multilevel randomized experiments with preschool children, we created OSM B as a rich source with (a) point estimates

[10] Nevertheless, some standard errors were relatively large. For example, most standard errors obtained for the point estimates and meta-analytic averages of R^2_{Group} and R^2_{Center} for SEL outcomes were quite large, particularly when using parent report (see Figures A2 and 1). Large standard errors of R^2_{Group}, and R^2_{Center} (but not R^2_{Child}) were often observed for very small values of $_{Center}$ and $_{Group}$ (see Figure A3). Hence, there was not much variance for covariates to explain in the outcomes at the group or daycare center levels. In these cases, variance estimates at the group and daycare center levels became unstable (likely due to chance differences), resulting in large standard errors of the point estimates of R^2_{Group} and R^2_{Center} and their meta-analytic averages.

[11] It may be the case that no suitable alternative point estimates or meta-analytic averages are available for R^2_{Group} and R^2_{Center} that could be estimated with greater statistical precision. In this case, and when corresponding values for both ρ_{Group} and ρ_{Center} were negligible or at most very small (e.g., $\rho_{Group} \leq .02$ and $\rho_{Center} \leq .02$) we recommend drawing on the lower and upper bound estimates for R^2_{Group} and R^2_{Center} because doing so still leads to a plausible range of *MDES* values although these design parameters have large standard errors (see OSM A6 for a discussion).

(Tables B.CT, B.CP, B.CE, B.SP, and B.SE) and (b) meta-analytic summaries (Tables B.CT.M, B.CP.M, B.CE.M, B.SP.M, and B.SE.M) of design parameters. As discussed above, we recommend matching the design parameters to key characteristics of the target intervention. This involves (a) potential clustering of the data, (b) the target age group, (c) the target outcome domain, (d) the method to be used for assessing the outcome, and (e) the target measure. To find appropriate matches in OSM B, intervention researchers can first consult Tables A6 to A10 in OSM A that provide short descriptions of the measures for which point estimates of design parameters are available.

When the target measures or other target characteristics of the planned study (e.g., target age group) do not match well to the characteristics of the studies that were used to estimate the present design parameters or when the available point estimates are associated with large standard errors, we recommend applying meta-analytic summaries of design parameters that match the target intervention. To do so, intervention researchers can consult Tables A11 to A15 that present overviews of available meta-analytic summaries.

Finally, to select the design parameters or their meta-analytic summaries from the spreadsheets in OSM B, we recommend using the interactive filter functions of the spreadsheet software or adapting the R code provided in the OSF.

Application Scenarios

This section presents two scenarios to illustrate the use of the present design parameters for planning the sample size of randomized experiments with preschool children. For each scenario, we assumed a balanced design with children (Scenario 1) and daycare centers (Scenario 2) randomly assigned to the experimental groups in equal shares. Further, we set the desired power at $1-\beta=0.80$ and used a two-tailed test (with significance level $\alpha=0.05$) to allow testing whether the intervention has unexpected negative effects on the outcome (Bland & Altman, 1994). To compute the required sample sizes, we used the R package PowerUpR (Bulus et al., 2021) that is based on the power formulas provided in Dong and Maynard (2013). Notably, other software tools that can be used for this purpose include the Excel worksheets by Dong and Maynard (2013), the PowerUpR shiny app (Ataneka et al., 2023), and the Optimal Design software (Spybrook et al., 2011).[12] Tables A17 (Scenario 1) and A18 (Scenario 2) show the estimates of design parameters that were applied in the power analyses.

[12] The Excel worksheets can be downloaded here: https://www.causalevaluation.org/power-analysis.html. The PowerUpR shiny app can be accessed here: https://powerupr.shinyapps.io/index. The Optimal Design software can be downloaded here: https://wtgrantfoundation.org/optimal-design-with-empirical-information-od.

Scenario 1: How Many Children Are Required for a Lab-Based Randomized Experiment?

A research team developed an intervention to foster five- to six-year-old children's performance on standardized tests of verbal skills. To investigate its impact, the team plans a single-level lab-based randomized experiment to test whether the new method is generally effective under well-controlled conditions in the lab. The team expects that the intervention should have at least an effect that lies in the average range of effects relative to other randomized experiments on cognitive outcomes. Drawing on Kraft (2020), the team therefore considers an intervention effect of $SMD = 0.15$ (i.e., a "medium" effect) as meaningful. The objective of the research team is to ensure that the intervention study can detect a treatment effect of $MDES = 0.15$ (with $1 - \beta = 0.80$ and $\alpha = 0.05$). Because the intervention targets verbal skills (and not a specific measure), the team selects meta-analytic averages of single-level design parameters for verbal skills (see Table B.CT.M in OSM B). Notably, the team uses meta-analytic estimates obtained by averaging across all age groups, as no meta-analytic summary for the target age group is available that includes all covariate sets. The research team begins by considering designs without covariates and learns that 1,398 children would be required to achieve $MDES = 0.15$ (Fig. 3a). Hence, 699 children should be randomly assigned to the intervention group, and 699 children to the control group. The team also tests the influence of different covariate sets (in terms of R^2_{Total}) on the required sample sizes. This results in a total sample size ranging from $N = 916$ when controlling for Set 3-SD + IB to $N = 1{,}150$ when controlling for Set 2-PB. The team also wants to take into account the statistical uncertainty associated with the applied design parameters and therefore determines conservative lower bound/optimistic upper bound estimates for N by drawing on the corresponding lower bounds/upper bounds of the 95% CIs of the meta-analytic averages of R^2_{Total}. When using lower bound values, the estimated total sample size lies between $N = 1{,}034$ (Set 3-SD + IB) and $N = 1{,}200$ (Set 1-SD). When using upper bound values, the research team would need between $N = 798$ (Set 3-SD + IB) and $N = 1{,}118$ (Set 2-PB) children. In sum, when it is not possible to use covariates, a total N of 1,398 preschool children is required to achieve $MDES = 0.15$. When covariates are an option, the required sample size depends on the risk preferences of the team. For example, opting for a conservative approach, the team should recruit a total N of 1,034 preschool children when using Set 3-SD + IB as covariates.

Scenario 2: How Many Daycare Centers Are Required for a Two-Level CRT?

The research team was successful and found that their intervention improved children's verbal skills with $SMD = 0.15$. The team now plans to study whether the intervention also works when it is implemented by educators in the regular preschool context. To avoid unintentionally exposing children in the control group to the intervention, the research team wants to carry out a two-level CRT and to randomize whole daycare centers to the intervention or control group. The team intends to sample 20 children per daycare center. The researchers are now interested in J, the number of daycare centers necessary to detect an intervention effect that is identical to

their lab-based study ($SMD = 0.15$). The team employs again meta-analytic averages of two-level design parameters for verbal skills (as obtained by averaging across all age groups) from Table B.CT.M from OSM B in their power analyses. The research team initially considers designs without covariates, uses the meta-analytic average of $\rho_{Center} = 0.15$, and learns that the minimum number of daycare centers amounts to $J = 278$ (Fig. 3b). The team again tests the influence of different covariate sets on the required sample size using the meta-analytic averages obtained for $\rho_{Center,}$ R^2_{Child} and R^2_{Center}. This leads to a total J in the range of $J = 142$ (Set 1-SD) to $J = 202$ (Set 2-PB). Given the high profile of the planned study, the team also wants to take into account the statistical uncertainty and therefore determines the upper bound estimates for J by using the upper bound of the 95% confidence interval for ρ_{Center}, and the lower bound values for R^2_{Child} and $R^2_{Center.}$ When using these values, the research team needs $J = 238$ when not using covariates, and between $J = 214$ and $J = 294$ when using Set 3-SD + IB or Set 2-IB.[13] To summarize, to achieve $MDES = 0.15$ when covariates are an option, the team should recruit a total of $J = 142$ daycare centers (total $N = 142 \cdot 20 = 2{,}840$; Set 1-SD) when relying on the point estimates of the meta-analytic averages, or $J = 214$ daycare centers ($N = 4{,}280$; Set 3-SD + PB) when opting for a conservative approach.

General Discussion

A New Resource to Support Early Childhood Education Research

"Early learning remains one of the most neglected areas of educational research" (OECD, 2020, p. 18). Particularly, there is a strong need for robust evidence about which educational interventions "work" or "work best" to promote preschool children's cognitive and socio-emotional development. Randomized experiments are key for drawing causal conclusions about the impact of such interventions. Hence, lab-based studies are required to develop and refine interventions, and field-based randomized experiments (e.g., CRTs or MSRTs) are beneficial for evaluating their effectiveness in real-world preschool settings. To ensure that these studies provide strong statistical conclusions regarding the size of the intervention effect in the target population, researchers should apply reliable estimates of design parameters when conducting a priori power analyses. However, there has been little research on relevant design parameters with preschool children. To address this significant research gap, we utilized a systematic collection of IPD from four large-scale German samples of preschool children to (a) estimate and (b) meta-analyze design parameters for single-level (non-clustered data), two-level (children in daycare centers), and three-level (children in groups, with groups nested in daycare centers) experimental designs.

[13] The standard errors for the meta-analytic average of R^2_{Center} obtained for Model Sets 1-SD, 2-IB, and 3-SD + IB were relatively large with $SEs > 0.05$. However, the team did not find alternative meta-analytic summaries with SEs < .05 for Model Sets 2-IB and 3-SD + IB and therefore considered the present values as the best guess.

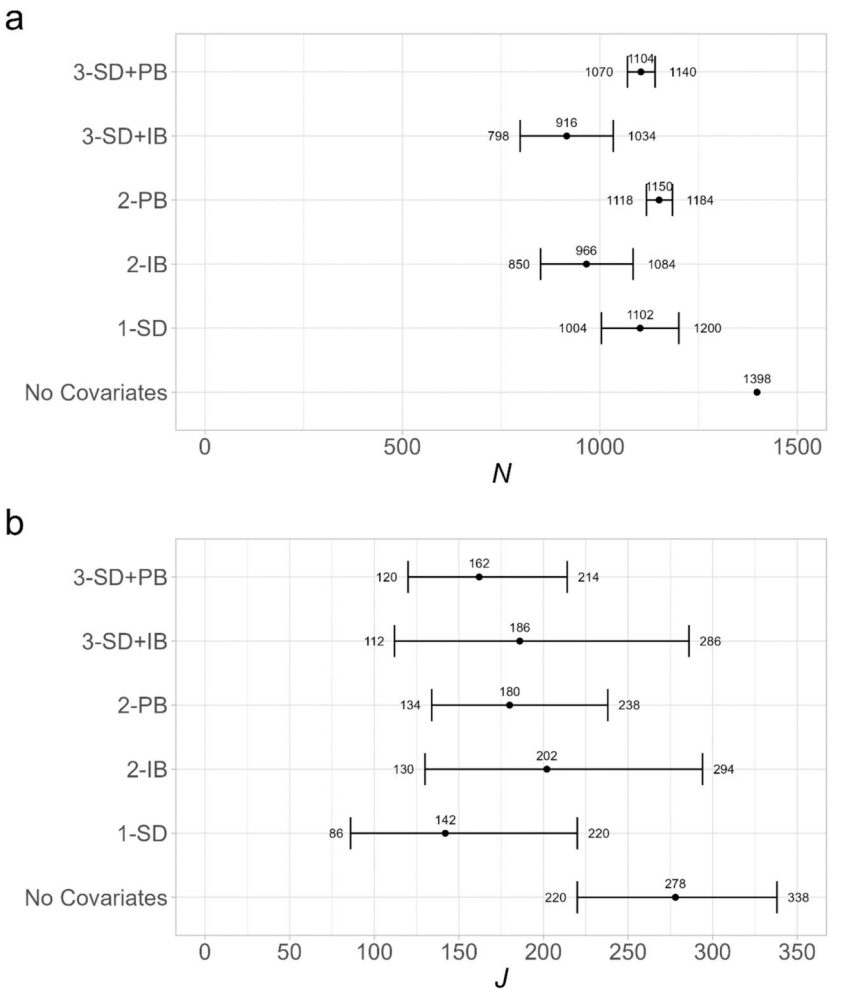

Fig. 3 Results from Power Analyses to Estimate the Required Sample Size to Achieve MDES = 0.15 for (**a**) the Single-Level Randomized Experiment in Scenario 1 and (**b**) the CRT in Scenario 2 When Using Meta-Analytic Summaries of Design Parameters obtained for Different Sets of Covariates. Note. *MDES* = minimum detectable effect size with level of statistical significance α = .05 and power of 1 − β = .80. The value of the *MDES* = 0.15 was chosen based on Kraft (2020), who suggested that an intervention effect size of a standardized mean difference of 0.15 (i.e., a medium" effect) is meaningful. *N* = total sample size. *J* = number of daycare centers. In Scenario 2 we assumed that 20 children are sampled in each daycare center, and, thus, $N = J \cdot 20$. The points show the estimated sample sizes when drawing on the meta-analytic averages of design parameters. The error bars for Scenario 1 represents conservative/optimistic estimates when drawing on lower/upper bound values of the 95% CIs for the meta-analytic averages of R^2_{Total}. The error bars for Scenario 2 represents conservative/optimistic estimates when drawing on lower/upper bound values of the 95% CIs for the meta-analytic averages of R^2_{Child} and R^2_{Center} in combination with the upper/lower bound values of the 95% CI for the meta-analytic average of $\rho_{\text{Center.}}$ The values of the applied design parameters are shown in Tables A17 and A18 in OSM A. 1-SD = Model set 1 with five socio-demographic characteristics as covariates; 2-IB = Model set 2 with one identical baseline measure as covariate; 2-PB = Model set 2 with one proxy baseline measure as covariate; 3-SD + IB = Model Set 3 with five socio-demographic characteristics and one identical baseline measure as covariates; 3-SD + PB = Model Set 3 with five socio-demographic characteristics and one proxy baseline measure as covariates

These parameters target cognitive and SEL outcomes, assessed using three methods: standardized tests, parent ratings, and educator ratings. The design parameters depict between-group and -center differences as well as the proportion of variance in the outcomes explained by five covariate sets including SD, IB, and PB measures, and the combination of SD and IB as well as SD and PB measures. Furthermore, we embedded the results on design parameters in the relevant methodological literature to discuss and illustrate their application in a priori power analyses. In summary, this paper offers a unique and rich resource to support researchers in early childhood education in carrying out a priori power analyses when planning lab-based and field-based randomized experiments.

Generalizability of the Present Design Parameters Across Countries?

The present design parameters address the target population of children attending daycare centers – the preschool setting that most children aged 3 years or older attend in Germany and many other countries. Comparing the multilevel design parameters from previous research and the present study shows that some country-specific estimates were similar, whereas others were (strikingly) different. For example, average between-center differences were $\rho_{Center} = 0.17/0.12/0.03$ for cognitive outcomes (assessed by standardized tests) and $\rho_{Center} = 0.05/0.09/0.01$ for SEL outcomes (assessed by educator reports) in the UK/Germany/US. One explanation for the large differences between the US and the other countries are the differences in the underlying samples. Design parameters for the United States targeted the population of socioeconomically disadvantaged preschool children (Jacob et al., 2010; see Appendix A.2 in Spybrook et al., 2011). By contrast, the studies from the United Kingdom and Germany were based on more heterogeneous samples. Importantly, even small differences in the value of design parameters may make a substantive difference in the required sample sizes. For example, consider a two-level CRT (with a balanced design, 20 children sampled per center; two-sided testing; $\alpha = 0.05$; $1 - \beta = 0.80$) to test the effectiveness of an intervention on cognitive outcomes (assessed with standardized tests) where the researchers expect a $SMD = 0.15$ and cannot use covariates. Whereas the UK average estimate of $\rho_{Center} = 0.17$ would result in a required number of daycare centers of $J = 298$ (total N = 5,960), the German average estimate of $\rho_{Center} = 0.12$ would result in a requirement of $J = 232$ daycare centers (total N 4,640), a difference of over 1,000 participants.

This example illustrates that the present design parameters based on German samples should be applied very cautiously in countries other than Germany. To apply design parameters that align with their local context, researchers in the US or UK could draw on estimates from previous research (Jacob et al., 2010; Sammons et al., 2002, 2003; Spybrook et al., 2011) that we used to generate Figure A1 and which are also available on our OSF. Furthermore, researchers can conduct a pilot study or apply the current analytic strategy using IPD from preschool children in their country (e.g., by utilizing our R code in the OSF as a template). If such

alternatives are not feasible, it seems reasonable to use the present design parameters in a priori power analyses (e.g., employing conservative values). Considering their theoretical range from zero to one, the current design parameters define a plausible range relevant to the preschool context and certainly provide better estimates than nonspecific conventional benchmarks, such as categorizing $R^2_{\text{Total}} = 0.02/0.13/0.26$ as "small", "medium", or "large" (see Cohen, 1988, who provided these values as well as a critical discussion).

Limitations and Outlook

Our study has several limitations that should be addressed in future research and nuances that should be considered when applying the present design parameters. First, the present paper provides design parameters that are highly relevant for power analyses of randomized intervention studies with preschool children. However, it was beyond the scope of our paper to elaborate on the statistical background of a priori power analyses of randomized experiments (e.g., Dong & Maynard, 2013; Hedges & Rhoads, 2010) or the process for estimating causal treatment effects once the empirical data have been collected (e.g., Ding, 2023; Lin, 2013).

Second, we present design parameters necessary for planning randomized intervention studies with (a) simple random assignment of preschool children in single-level lab-based settings (assuming that the data are not clustered) and (b) random assignment in multilevel field settings, particularly in cluster randomized trials (CRTs) where entire daycare centers (including all sampled children or groups) are allocated either to the experimental or control group. Importantly, the present design parameters are also needed to plan MSRTs with blocked random assignment. In multi-site experiments the sample of children or groups is randomly assigned to experimental conditions within daycare centers. In addition to the design parameters (i.e., R^2s and ICCs) that we presented in this paper, power analyses of such experiments require information on the expected heterogeneity of the treatment effect across groups or daycare centers, as well as the extent to which covariates may explain this heterogeneity. General benchmarks for the potential magnitude of heterogeneity in treatment effects can be found in Weiss et al. (2017). A vital task for future research is to provide systematic knowledge of these parameters for the preschool context, for example by using the analytic approach by Sabol et al. (2022).

Third, our applied methodology as well as the characteristics of the applied studies were associated with some limitations (see Stallasch et al., 2021, 2024). Specifically, we analyzed and meta-analyzed design parameters utilizing IPD from several large-scale studies with probability samples, which mitigates potential bias due to sample selectivity and ensures coverage of the full range of target outcomes without variance restrictions. Overall, this approach supports reliable parameter estimation and the generalizability of results. However, we did not apply sampling weights when estimating design parameters because (a) information on weights was only available for NEPS and (b) the application of sampling weights is not possible with the lme4 package (Bates et al., 2015) that we used for the multilevel analyses. Therefore, our design parameters based on NEPS are representative only for

the preschool children included in the present analyses and are likely somewhat less accurate than estimates derived from analyses using sampling weights (e.g., Wenger et al., 2018). Moreover, the NUBBEK and NEPS studies contained a relatively small number of children/groups per daycare center (see Table 1). However, robust evidence from simulation studies shows that accurate estimates of ρ_{Group} and ρ_{Center} can still be obtained in such data constellations, as the number of daycare centers exceeded 100 in the two-level models (McNeish, 2014) and 200 in the three-level models (Zhang et al., 2024). Nevertheless, larger sample sizes at all levels would have further enhanced the quality of the model parameters (McNeish & Stapleton, 2016). Additionally, we drew on rather heterogeneous samples. Higher homogeneity may lead to smaller values for all design parameters under investigation due to range restrictions (Miciak et al., 2016). Thus, between-center and between-group differences— but also the amount of variance explained by covariates— may become smaller in more homogenous samples (Hedges & Hedberg, 2007). Moreover, the time lag between baseline and outcome measures was between six and 12 months. Longer time lags are typically associated with (somewhat) smaller values of R^2 at all levels of analyses (Bleidorn et al., 2022; Stallasch et al., 2024). Thus, relative to the present design parameters, somewhat larger values of R^2 should be expected for shorter time intervals in the planned randomized experiment, and smaller values of R^2 should be expected for longer time intervals. Finally, most measures in the social and behavioral sciences are affected by measurement error. The reliabilities (r_{tt}) for the measures of cognitive and SEL outcomes that we applied to estimate design parameters were fairly typical for applied (experimental) research, where $r_{tt} \geq 0.70$ is desirable, but even smaller values may suffice for many research purposes (see Schmitt, 1996 for a discussion). Notably, fallible measures usually lower R^2 in total as well as at the child, group, or center levels (Cochran, 1970; Raudenbush & Bryk, 2002). Thus, our estimates can be considered conservative estimates that may generalize well to empirical data of randomized studies. Of note, measurement error in pre-treatment covariates does not introduce bias in the estimated treatment effect but rather improves the precision with which it can be estimated compared to analyses that do not adjust for covariates (Maxwell et al., 2018).

Fourth, we explored the explanatory power of two PB measures, namely vocabulary knowledge as a proxy baseline measure for cognitive outcomes and children's problem behavior (as assessed by parent or educator reports) as a proxy baseline measure for SEL outcomes. Our results showed that these measures may explain small to substantial proportions of variance in the outcome measures at all levels of analyses. Hence, PB measures may be a useful alternative to IB measures in the preschool context, where it may not be possible to apply identical pretest measures because of the strong developmental dynamics in cognitive and SEL outcomes among very young children. However, our design parameters for PB are confined to the applied or very similar measures. Thus, future research may profit from examining the explanatory power of a broader range of PB measures of cognitive or SEL outcomes.

Fifth, we provided guidance and illustrative examples on how to account for the statistical uncertainty of design parameters in power analyses of randomized studies. Accordingly, we offered 95% CIs for point estimates and meta-analytic

averages of design parameters, along with 95% PIs to depict the spread of the distribution of design parameters (in random-effects meta-analyses). The meta-analytic summaries—but not the point estimates—of design parameters were based on estimated working covariances which we computed using a reasonable upper-bound estimate of the correlation *r* among design parameters. Our sensitivity analyses showed that the standard errors of the meta-analytic averages and the standard deviations of the random effects (σ) increased with increasing values of *r*. However, the observed increase of these parameters was in most cases (very) small. The standard errors of the meta-analytic averages enter the estimation of 95% CIs and 95% PIs, and the standard deviations enter 95% PIs. Thus, most—but not all—95% CIs and 95% PIs can be considered to be fairly robust against the different values chosen for *r*. Notably, all meta-analytic results presented in this paper, along with the meta-analytic design parameters in OSM B, are based on using $r = 0.70$ as a reasonable upper bound for the within-sample correlations among design parameters. Thus, the 95% CIs and 95% PIs included in these tables represent conservative estimates of these intervals. Consequently, using the lower bound values for R^2s and/or upper bounds values of *ICC*'s from these intervals represents a conservative approach to account for the statistical uncertainty of these design parameters in a priori power analyses. When researchers wish to apply less conservative approaches, they can use the R code available in our OSF to extract meta-analytic design parameters based on smaller values for the within-sample correlation.

Finally, we provided numerous meta-analytic summaries of design parameters to support the planning of randomized intervention studies with plausible meta-analytic averages and prediction intervals. However, it was beyond the scope of the present paper to investigate moderator variables that may explain the variability among design parameters. Further, such moderator analyses require a considerably larger number of studies than the three (i.e., BIKS, NUBBEK, and NEPS) we identified with our systematic search for the present paper. Nevertheless, an important next step for future research is to conduct meta-regression analyses to examine how moderator variables at the level of (a) effect sizes (e.g., specific characteristics of the applied measures, time lag between pre- and posttest, reliability of measures) or (b) studies (e.g., country, coverage of the target population, quality of the sampling process, observational or experimental study design, year of data collection) may explain the observed heterogeneity among design parameters.

Conclusion

This paper offers a unique and comprehensive resource for conducting a priori power analyses to plan sample sizes for both lab-based and field-based randomized experiments in early childhood education research. We hope that the design parameters, along with our recommendations for their application and the illustrative examples, will assist researchers in conducting randomized intervention studies that provide rigorous evidence to support preschool children's cognitive and socio-emotional development.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-024-09981-z.

Authors' Contribution All authors contributed to the study conception and design. Data preparation and analyses were performed by Martin Brunner and Sophie Stallasch. The first draft of the manuscript was written by Martin Brunner and all authors commented on subsequent versions of the manuscript. All authors read and approved the final manuscript.

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

Data Availability This paper uses data from the National Educational Panel Study (NEPS; see Blossfeld & Roßbach, 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network. The dataset for the study titled "Educational Processes, Competence Development and Selection Decisions in Preschool and School Age (BIKS-3–10, Weinert et al., 2024)" was made available by the Research Data Centre at the Institute for Educational Quality Improvement (FDZ at IQB). Permission from the dataset owners was granted to use these datasets for the research objectives of the present paper. The dataset for the National Survey on Education, Care, and Development in Early Childhood (NUBBEK) was made available by the Leibniz Institute for the Social Sciences in Mannheim (GESIS). The R code for reproducing all results as well as the data with the design parameters used in the present paper can be accessed via the Open Science Framework at https://osf.io/qz7fy.

Declarations

Ethics Approval We used scientific use files of BIKS, NEPS, and NUBBEK. All ethical issues related to these data were handled by the scientific consortia who were responsible for collecting the data.

Consent All authors agreed with the content and all authors gave explicit consent to submit.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Artelt, C., & Sixt, M. (2023). The National Educational Panel Study (NEPS)—Framework, design, and research potential. *Zeitschrift Für Erziehungswissenschaft, 26*(2), 277–298. https://doi.org/10.1007/s11618-023-01156-w

Ataneka, A., Kelcy, B., Dong, N., Bulus, M., & Bai, F. (2023). PowerUp R Shiny App (v. 0.9) Manual. https://www.causalevaluation.org/uploads/7/3/3/6/73366257/r_shinnyapp_manual_0.9.pdf

Autor:innengruppe Bildungsberichterstattung. (2022). Bildung in Deutschland 2022 [Education in Germany 2022]. wbv Media. https://doi.org/10.3278/6001820hw

Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science, 333*(6045), 975–978. https://doi.org/10.1126/science.1204534

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. gcrnkw.

Bland, J. M., & Altman, D. G. (1994). Statistics Notes: One and two sided tests of significance. *BMJ, 309*(6949), 248. https://doi.org/10.1136/bmj.309.6949.248

Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological Bulletin, 148*(7–8), 588–619. https://doi.org/10.1037/bul0000365

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. (2006). The core analytics of randomized experiments for social research. MDRC. 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

Blossfeld, H.-P., & Roßbach, H.-G. (Eds.). (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd ed.). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-23162-0

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Boruch, R. (2005). Better evaluation for evidence-based policy: Place randomized trials in education, criminology, welfare, and health. *The ANNALS of the American Academy of Political and Social Science, 599*(1), 6–18. https://doi.org/10.1177/0002716205275610

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. gd4q25.

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

Bulus, M. (2022). Minimum detectable effect size computations for cluster-level regression discontinuity studies: Specifications beyond the linear functional form. *Journal of Research on Educational Effectiveness, 15*(1), 151–177. https://doi.org/10.1080/19345747.2021.1947425

Bulus, M., & Sahin, S. G. (2019). Estimation and standardization of variance parameters for planning cluster-randomized trials: A short guide for researchers. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 2. https://doi.org/10.21031/epod.530642

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2021). *PowerUpR: Power analysis tools for multilevel randomized experiments* (Version 1.1.0) [Computer software]. https://cran.r-project.org/web/packages/PowerUpR/index.html

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology, 56*, 453–484. cx7kjq.

Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research, 53*(3), 1071–1104. https://doi.org/10.1177/00491241221099552

Cochran, W. G. (1970). Some effects of errors of measurement on multiple correlation. *Journal of the American Statistical Association, 65*(329), 22–34. mmb8.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum.

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. gjrpjc.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175–199. https://doi.org/10.3102/01623737024003175

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*. https://doi.org/10.1177/0002716205275738

Dawson, A., Yeomans, E., & Brown, E. R. (2018). Methodological challenges in education RCTs: Reflections from England's Education Endowment Foundation. *Educational Research, 60*(3), 292–310. https://doi.org/10.1080/00131881.2018.1500079

De Pauw, S. S. W., & Mervielde, I. (2010). Temperament, personality and developmental psychopathology: A review based on the conceptual dimensions underlying childhood traits. *Child Psychiatry & Human Development, 41*(3), 313–329. fthv76.

Ding, P. (2023). A first course in causal inference (No. arXiv:2305.18793). arXiv. https://doi.org/10.48550/arXiv.2305.18793

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. gd4q27.

Dong, N., Reinke, W. M., Herman, K. C., Bradshaw, C. P., & Murray, D. W. (2016). Meaningful effect sizes, intraclass correlations, and proportions of variance explained by covariates for planning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review, 40*(4), 334–377. https://doi.org/10.1177/0193841X16671283

Durlak, J. A., Mahoney, J. L., & Boyle, A. E. (2022). What we know, and what we need to find out about universal, school-based social and emotional learning programs for children and adolescents: A review of meta-analyses and directions for future research. *Psychological Bulletin, 148*, 765–782. kx88.

Flanagan, D. P., & Dixon, S. G. (2014). *The Cattell-Horn-Carroll theory of cognitive abilities*. John Wiley & Sons, Ltd. mmcb.

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 International Standards Classification of Occupations. *Social Science Research, 25*, 201–239.

García, J. L., Heckman, J. J., & Ronda, V. (2023). The lasting effects of early-childhood education on promoting the skills and social mobility of disadvantaged African Americans and their children. *Journal of Political Economy, 131*(6), 1477–1506. https://doi.org/10.1086/722936

Gaspard, H., Dicke, A.-L., Flunger, B., Brisson, B. M., Häfner, I., Nagengast, B., & Trautwein, U. (2015). Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology, 51*(9), 1226–1240. https://doi.org/10.1037/dev0000028

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

Grund, S., Lüdtke, O., & Robitzsch, A. (2023). Handling missing data in cross-classified multilevel analyses: An evaluation of different multiple imputation approaches. *Journal of Educational and Behavioral Statistics, 48*, 454–489. kx9d.

Hedberg, E. C. (2016). Academic and behavioral design parameters for cluster randomized trials in kindergarten: An analysis of the Early Childhood Longitudinal Study 2011 kindergarten cohort (ECLS-K 2011). *Evaluation Review, 40*(4), 279–313. https://doi.org/10.1177/0193841X16655657

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. https://doi.org/10.7758/9781610448864.16

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., & Rhoads, C. (2010). Statistical power analysis in education research. NCSER 2010–3006. In *National Center for Special Education Research*. National Center for Special Education Research. http://files.eric.ed.gov/fulltext/ED509387.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., & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research, 60*(3), 265–275. https://doi.org/10.1080/00131881.2018.1493350

Homuth, C., Lehrl, S., Volodina, A., Weinert, S., & Rossbach, H.-G. (2024). From Preschool to Vocational Training and Tertiary Education—Study Design of the BiKS-3–18 Study. In S. Weinert, H.-G. Rossbach, J. Von Maurice, H.-P. Blossfeld, & C. Artelt (Eds.), *Educational Processes, Decisions, and the Development of Competencies from Early Preschool Age to Adolescence* (Vol. 16, pp. 21–53). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-43414-4_2

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

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

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

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

Letourneau, N. L., Duffett-Leger, L., Levac, L., Watson, B., & Young-Morris, C. (2013). Socioeconomic status and child development: A meta-analysis. *Journal of Emotional and Behavioral Disorders, 21*(3), 211–224. c9sdc3.

Leyendecker, B., Agache, A., & Madsen, S. (2014). Nationale Untersuchung zur Bildung, Betreuung und Erziehung in der frühen Kindheit (NUBBEK) – Design, Methodenüberblick, Datenzugang und das Potenzial zu Mehrebenenanalysen [NUBBEK – a national German study on early childhood education and care: Design, methods overview, data access, and the potential for multilevel analyses]. *ZfF – Zeitschrift für Familienforschung / Journal of Family Research*, *26*(2), 2. https://www.budrich-journals.de/index.php/zff/article/view/16528

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

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

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher, 48*(3), 158–166. ggm84b.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (Third edition). Routledge. https://doi.org/10.4324/9781315642956

McNeish, D. M. (2014). Modeling sparsely clustered data: Design-based, model-based, and single-level methods. *Psychological Methods, 19*(4), 552–563. https://doi.org/10.1037/met0000024

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

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. mmcc.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*(2), 246–268.

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

Mosteller, F., & Boruch, R. (2002). *Evidence matters: Randomized trials in education research*. Brookings Institution Press. https://www.jstor.org/stable/10.7864/j.ctvc16n69

Myers, D., & Schirm, A. (1999). *The Impacts of Upward Bound: Final Report for Phase I of the National Evaluation*. https://eric.ed.gov/?id=ED432621

NEPS Network. (2022). *National Educational Panel Study, Scientific Use File of Starting Cohort Kindergarten* (Version 10.0.0) . LIfBi Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/NEPS:SC2:10.0.0

Nye, B., Hedges, L. V., & Konstantopoulos, S. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Educational Evaluation and Policy Analysis, 21*(2), 127–142. https://doi.org/10.3102/01623737021002127

OECD. (2020). *Early learning and child well-being: A study of five-year olds in England, Estonia, and the United States*. OECD. https://doi.org/10.1787/3990407f-en

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., …, Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. gjkq9b.

Peng, P., & Kievit, R. A. (2020). The development of academic achievement and cognitive abilities: A bidirectional perspective. *Child Development Perspectives, 14*(1), 15–20. ggxvw3.

Pontoppidan, M., Keilow, M., Dietrichson, J., Solheim, O. J., Opheim, V., Gustafson, S., & Andersen, S. C. (2018). Randomised controlled trials in Scandinavian educational research. *Educational Research, 60*(3), 311–335. https://doi.org/10.1080/00131881.2018.1493351

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. fn8rhp.

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. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raudenbush, S. W., Martinez, 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., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Sage.

Rice, K., Higgins, J. P. T., & Lumley, T. (2018). A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 181*(1), 205–227. https://doi.org/10.1111/rssa.12275

Robitzsch, A., & Grund, S. (2023). *miceadds: Some additional multiple imputation functions, especially for "mice" (Version 3.16-18)* [Computer software].

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. J. Wiley & Sons. https://doi.org/10.1002/9780470316696

Sabol, T. J., McCoy, D., Gonzalez, K., Miratrix, L., Hedges, L., Spybrook, J. K., & Weiland, C. (2022). Exploring treatment impact heterogeneity across sites: Challenges and opportunities for early childhood researchers. *Early Childhood Research Quarterly, 58*, 14–26. mmcd.

Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., & Elliot, K. (2002). *Measuring the impact of pre-school on children's cognitive progress over the preschool period: Technical Paper 8a.* https://discovery.ucl.ac.uk/id/eprint/10005295/1/Sammons2003Effective(Tech.Paper8A).pdf

Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., & Elliot, K. (2003). The Effective Provision of Pre-School Education (EPPE) Project: Measuring the Impact of Pre-School on Children's Social/Behavioural Development over the Pre-School Period. In *Institute of Education, University of London/ Department for Education and Skills: London.* [Report]. Institute of Education, University of London/ Department for Education and Skills. https://discovery.ucl.ac.uk/id/eprint/10005288/

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353. cjqw3x.

Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics, 34*(2), 238–266. https://doi.org/10.3102/1076998609332748

Schoon, I. (2021). Towards an integrative taxonomy of social-emotional competences. *Frontiers in Psychology, 12*, 515313. https://doi.org/10.3389/fpsyg.2021.515313

Schweinhart, L. J. (Ed.). (2005). *Lifetime effects: The High/Scope Perry preschool study through age 40*. High/Scope Press.

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2022). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist, 57*(1), 31–54. jhm4.

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist, 55*(1), 21–31. gf4hzh.

Somers, M.-A., Weiss, M. J., & Hill, C. (2022). Design parameters for planning the sample size of individual-level randomized controlled trials in community colleges. *Evaluation Review*, 0193841X221121236. https://doi.org/10.1177/0193841X221121236

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales, Second Edition (Vineland-II): Survey forms manual*. Pearson Assessments.

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., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal Design plus empirical evidence: Documentation for the "Optimal Design" software*. http://hlmsoft.net/od/od301.zip. http://hlmsoft.net/od/od-manual-20111016-v300.pdf

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. mmcf.

Stallasch, S. E., Lüdtke, O., Artelt, C., Hedges, L. V., & Brunner, M. (2024). Single- and multilevel perspectives on covariate selection in randomized intervention studies on student achievement. *Educational Psychology Review, 36*(4), 112. https://doi.org/10.1007/s10648-024-09898-7

Standing Scientific Commission on Education Policy. (2022). *Impulspapier: Entwicklung von Leitlinien für das Monitoring und die Evaluation von Förderprogrammen im Bildungsbereich [Position paper on the development of guidelines for the monitoring and evaluation of support programs in the education sector]*. https://www.swk-bildung.org/content/uploads/2024/02/SWK-2022-Impulspapier_Monitoring.pdf

Stewart, L. A., Clarke, M., Rovers, M., Riley, R. D., Simmonds, M., Stewart, G., & Tierney, J. F. (2015). Preferred reporting items for a systematic review and meta-analysis of individual participant data: The PRISMA-IPD statement. *JAMA, 313*(16), 1657–1665. f7bhcv.

Tackett, J. L., Kushner, S. C., De Fruyt, F., & Mervielde, I. (2013). Delineating personality traits in childhood and adolescence: Associations across measures, temperament, and behavioral problems. *Assessment, 20*(6), 738–751. https://doi.org/10.1177/1073191113509686

Tietze, W., Becker-Stoll, F., Bensel, J., Haug-Schnabel, G., Kalicki, B., Keller, H., & Leyendecker, B. (2015). *NUBBEK - Nationale Untersuchung zur Bildung, Betreuung und Erziehung in der frühen Kindheit [NUBBEK - National survey on education, care, and development in early childhood]* (Version 3.0.0) . GESIS Data Archive. https://doi.org/10.4232/1.12297

Tucker-Drob, E. M. (2019). Cognitive aging and dementia: A life-span perspective. *Annual Review of Developmental Psychology, 1*(1), 177–196. gh3gxg.

U.S. Department of Education, National Center for Education Statistics. (2021). *Early childhood program participation: 2019 (NCES 2020–075REV), Table 1*. National Center for Education Statistics. https://nces.ed.gov/fastfacts/display.asp?id=4

Ulferts, H. (2017). *Komponenten und Auswirkungen der Qualität mathematischer Bildung in frühkindlichen Bildungs- und Betreuungseinrichtungen [Components and impact of the quality of math education in early childhood education and care centers]* [Freie Universität Berlin]. https://refubium.fu-berlin.de/handle/fub188/5432

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

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

Weinert, S., Roßbach, H.-G., Faust, G., Blossfeld, H.-P., Artelt, C., & Otto-Friedrich-Universität Bamberg. (2019). Educational Processes, Competence Development and Selection Decisions in

Preschool and School Age (BiKS-3-10) (Version 6) [Dataset]. IQB - Institute for Educational Quality Improvement. https://doi.org/10.5159/IQB_BIKS_3_10_V6

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. mmcg.

Wenger, M., Lüdtke, O., & Brunner, M. (2018). Übereinstimmung, Variabilität und Reliabilität von Schülerurteilen zur Unterrichtsqualität auf Schulebene [Interrater agreement, variability, and reliability of student ratings of instructional quality at the school-level]. *Zeitschrift für Erziehungswissenschaft, 21*, 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

Westine, C. D., Unlu, F., Taylor, J., Spybrook, J., Zhang, Q., & Anderson, B. (2020). Design parameter values for impact evaluations of science and mathematics interventions involving teacher outcomes. *Journal of Research on Educational Effectiveness, 13*(4), 816–839. https://doi.org/10.1080/19345747.2020.1821849

Zhang, H., Shen, Z., & Leite, W. L. (2024). The impacts of small teacher-level sample sizes in cluster-randomized trials. *The Journal of Experimental Education, 0*(0), 1–20. https://doi.org/10.1080/00220973.2024.2376627

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). Across-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Journal of Measurement and Evaluation in Education and Psychology, 3*(1), 242–278.

Authors and Affiliations

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

✉ Martin Brunner
martin.brunner@uni-potsdam.de

[1] Department of Educational Sciences, University of Potsdam, Potsdam, Germany

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

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

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

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