







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

Martin Brunner¹  | Lena Keller¹  | Sophie E. Stallasch¹  |
Julia Kretschmann¹  | Andrea Hasl¹  | Franzis Preckel² | Oliver Lüdtke^{3,4} |
Larry V. Hedges⁵ 

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

²Department of Psychology, University of
Trier, Trier, Germany

³Leibniz Institute for Science and
Mathematics Education, Kiel, Germany

⁴Centre for International Student
Assessment, Munich, Germany

⁵Department of Statistics, Northwestern
University, Evanston, Illinois, USA

Correspondence

Martin Brunner, Department of
Educational Sciences, Faculty of Human
Sciences, University of Potsdam, Karl-
Liebknecht-Str. 24-25, 14476 Potsdam,
Germany.

Email: martin.brunner@uni-potsdam.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/
Award Number: 442358899

Abstract

Descriptive analyses of socially important or theoretically interesting phenomena and trends are a vital component of research in the behavioral, social, economic, and health sciences. Such analyses yield reliable results when using representative individual participant data (IPD) from studies with complex survey designs, including educational large-scale assessments (ELSAs) or social, health, and economic survey and panel studies. The meta-analytic integration of these results offers unique and novel research opportunities to provide strong empirical evidence of the consistency and generalizability of important phenomena and trends. Using ELSAs as an example, this tutorial offers methodological guidance on how to use the two-stage approach to IPD meta-analysis to account for the statistical challenges of complex survey designs (e.g., sampling weights, clustered and missing IPD), first, to conduct descriptive analyses (Stage 1), and second, to integrate results with three-level meta-analytic and meta-regression models to take into account dependencies among effect sizes (Stage 2). The two-stage approach is illustrated with IPD on reading achievement from the Programme for International Student Assessment (PISA). We demonstrate how to analyze and integrate standardized mean differences (e.g., gender differences), correlations (e.g., with students' socioeconomic status [SES]), and interactions between individual characteristics at the participant level (e.g., the interaction between gender and SES) across several PISA cycles. All the datafiles and R scripts we used are available online. Because complex social, health, or economic survey and panel studies share many methodological features with ELSAs, the guidance offered in this tutorial is also helpful for synthesizing research evidence from these studies.

KEYWORDS

complex survey designs, educational large-scale assessments, individual participant data, meta-analysis, Programme for International Student Assessment

Highlights**What is already known**

- Descriptive analyses based on studies with complex survey designs, such as educational large-scale assessments (ELSAs), provide reliable findings for socially important or theoretically interesting phenomena that can be generalized to well-defined populations
- Many important phenomena can be examined by drawing on readily available individual participant data (IPD) from one study cycle or sample as well as by integrating results from multiple sets of IPD
- IPD meta-analysis is a powerful tool for integrating research evidence and for examining heterogeneity of results at the participant and study level

What is new

- IPD meta-analyses of descriptive results from studies with complex survey designs offer unique and novel research opportunities to provide strong empirical evidence of the consistency and generalizability of socially important and theoretically interesting phenomena and trends.
- Using ELSAs as an example, this tutorial introduces a two-stage approach to IPD meta-analysis that is tailored to the methodological challenges of studies with complex survey designs (e.g., sampling weights, clustered and missing IPD, dependent effect sizes), first, for conducting descriptive analyses (Stage 1), and second, for integrating the results with meta-analytic and meta-regression models (Stage 2).
- We provide thoroughly annotated R syntaxes; all datasets are available online

Potential impact for *Research Synthesis Methods* readers outside the authors' field

- The guidance offered in this tutorial is useful for synthesizing research findings from studies with complex survey designs from a variety of fields, because health, social and economic survey and panel studies share many methodological features (e.g., sampling weights, clustered and missing IPD, dependent effect sizes) with ELSAs

1 | INTRODUCTION

Quantitative descriptive analyses are a vital component of research in the behavioral and social sciences (e.g., education, psychology, sociology, and economics) and many other disciplines (e.g., health sciences). Such analyses help to “characterize the world” or a “phenomenon” because they answer questions “about who, what, where, when, and to what extent.”¹ Quantitative descriptive analyses provide robust results when being based on representative individual participant data (IPD) from studies with complex survey designs, such as educational

large-scale assessments (ELSAs)² or social, health, and economic survey and panel studies. Many studies with complex survey designs have been conducted at both international and national levels, involving the World Health Surveys, the European Health Interview Survey, the European Values Study, the Household Finance and Consumption Survey, or the US National Longitudinal Surveys of the Youth. Well-known examples for ELSAs are the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), the Programme for the

International Assessment of Adult Competencies (PIAAC), and the National Assessment of Educational Progress (NAEP). Since studies with complex survey designs have been carried out for several decades, it is now possible to answer research questions not only with IPD from one study cycle or sample (e.g., PISA 2018) but by integrating results from multiple sets of IPD (e.g., PISA 2000–2018). Given the methodological complexities (e.g., sampling weights, clustered and missing IPD, dependent effect sizes) of studies with complex survey designs in general, and ELSAs in specific, a key question is how to do this. IPD meta-analysis may be a powerful tool for this purpose as it offers unique and novel opportunities to synthesize evidence from descriptive analyses.³ In this tutorial, we use the example of ELSAs to offer methodological guidance on how to conduct IPD meta-analyses with studies using complex survey designs to contribute to the cumulative body of knowledge on socially important or theoretically interesting phenomena and trends. Notably, because complex health, social, or economic surveys and panel studies share many methodological features with ELSAs (e.g., sampling weights, clustered and missing IPD, dependent effect sizes), the guidance offered in this tutorial may also be helpful to use IPD meta-analyses to synthesize research evidence from these studies.⁴

2 | ADVANTAGES OF COMBINING META-ANALYTIC TECHNIQUES AND IPD FROM ELSAS

IPD meta-analyses of ELSAs may significantly enrich the extant body of knowledge with reliable and widely generalizable evidence for many research topics, for example, (temporal trends in) gender differences in achievement,^{5–7} or the relationship between students' socioeconomic status (SES) and achievement.⁸ Perhaps even more importantly, IPD meta-analyses of ELSAs open up new and unique research opportunities to synthesize evidence. Such research opportunities involve in-depth meta-analyses of policy-relevant subgroups, for example, to provide reliable empirical evidence of the consistency and generalizability of the magnitude of gender differences among top-performing students (i.e., the top 5%) in mathematics.⁹ Likewise, IPD meta-analyses of ELSAs allow researchers to examine the generalizability of (novel) theoretical propositions across countries, for example, to test predictions about nonlinear relationships between achievement and academic self-concepts,¹⁰ relationships between innovative school environments and teaching practices,¹¹ relationships between epistemic beliefs and educational outcomes in science,¹² year-in-

school effects on academic self-concept,¹³ or, as we illustrate in the present paper, intersectional research questions,¹⁴ such as how gender differences in students' achievement are moderated by their SES.

2.1 | Advantages of IPD from ELSAs

The wealth of data collected in ELSAs yields manifold insights into (a) the distributions of important individual characteristics involving achievement and skills in various domains (e.g., reading, mathematics), socioemotional characteristics (e.g., achievement motivation, academic self-concept, personality), well-being and health (e.g., satisfaction with life, body mass index), or sociodemographic background, (b) learning and home environments, and (c) the relational patterns between these variables. The measures applied in ELSAs are very high in quality because they are based on expert review panels, pilot-tested in large field trials, and thoroughly examined to ensure that they provide unbiased assessments of key constructs.¹⁵ Using such measures, ELSAs allow researchers to describe distributions and patterns of relationships in a certain target population or subpopulation. ELSA data are therefore obtained from large, representative probability samples,¹⁶ thus implying that descriptive analyses of these data meet the gold standard for obtaining reliable knowledge about vital research topics that can be generalized to well-defined populations.^{17,18} More specifically, the random sampling process effectively responds to threats of sample selection bias because it ensures the representativeness of the sample on observable and unobservable dimensions of a certain well-defined population.^{17,18} Further, given their large sample sizes, ELSAs provide very precise estimates of effect sizes for these populations¹⁶ with “effect size” broadly defined as any quantitative estimate used to address a certain research question.¹⁹

2.2 | Advantages of IPD meta-analyses of ELSAs

Because ELSAs have been conducted regularly at both international and national levels for many years, they provide a very large volume of data—Big Data—for carrying out descriptive analyses. For example, PISA provides IPD from over 400 independent samples with almost three million students taking part in one of seven PISA cycles across a time span of 18 years (see Figure 1). When descriptive analyses from several samples from the same or related ELSAs are available, important questions arise concerning the consistency, replicability, and

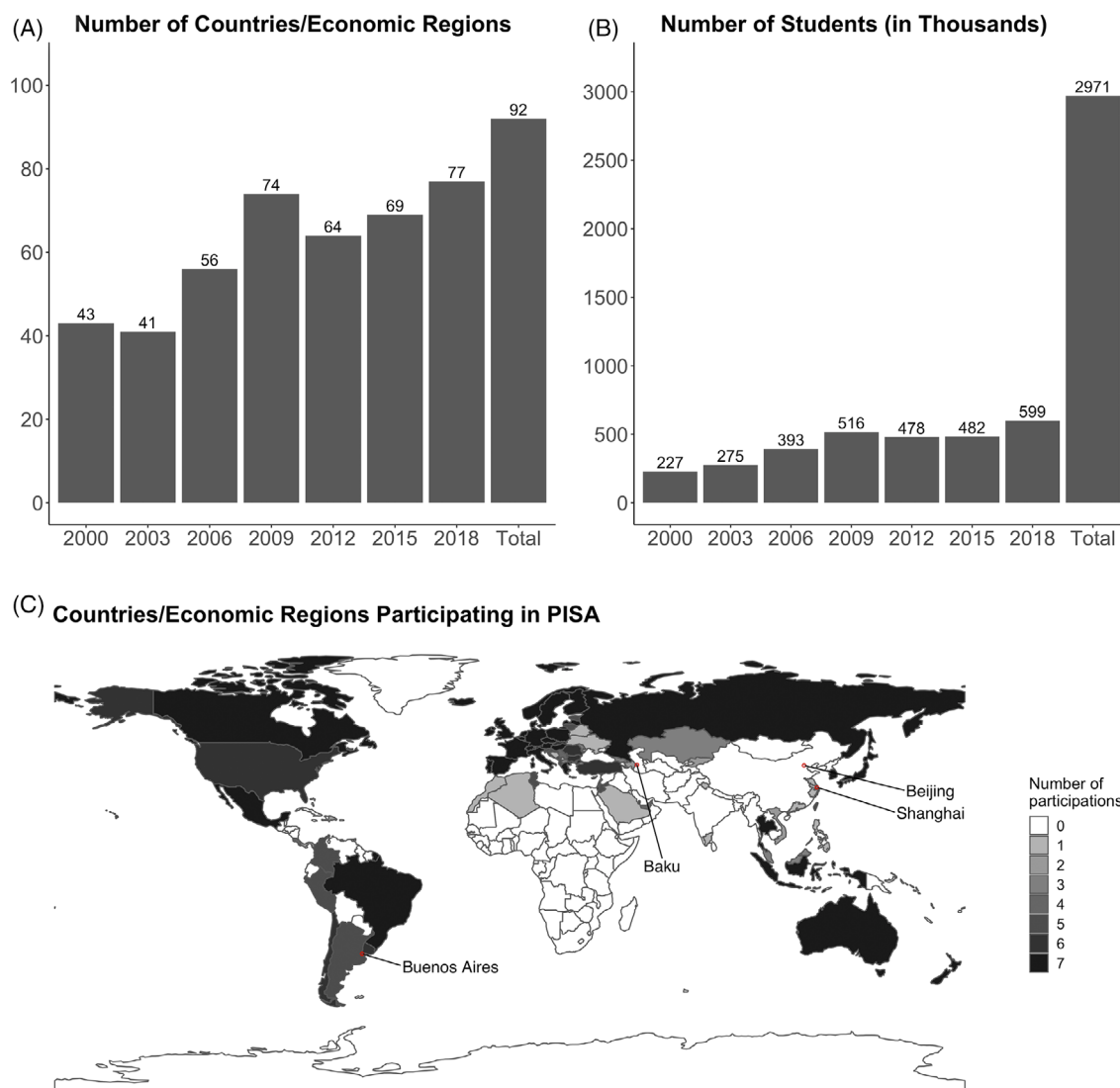


FIGURE 1 Sample sizes (a, b) and participating countries (c) in the full Programme for International Student Assessment (PISA) sample per cycle and in total. Across PISA cycles, individual participant data (IPD) from $k = 433$ independent student samples were collected. IPD from $k = 424$ student samples could be used in the present analyses (depicted in this figure). Color categories in (c) indicate how often a certain country/economic region participated in PISA. Countries colored in white indicate no participation. [Colour figure can be viewed at wileyonlinelibrary.com]

generalizability of the results.²⁰ Such questions can be tackled with IPD meta-analyses.

IPD meta-analysis is a specific type of meta-analysis that was developed primarily in biomedical research.²¹ Instead of extracting summary, aggregate data (AD) from published or unpublished studies,^{22,23} researchers search for original data that can be re-analyzed and then integrated in meta-analyses.²⁴ IPD meta-analyses can improve the generalizability of results because IPD from previously unpublished studies may become available.³ Access to IPD also allows the meta-analyst to apply a standardized analysis protocol to control the quality of the data and statistical analyses to estimate effect sizes.^{3,25} Such advantages of IPD meta-analyses help

mitigate bias and unwanted heterogeneity in effect sizes, and as a consequence, they help improve the precision of the meta-analytic results.^{3,25} Furthermore, IPD allow for the application of types of analyses that would not be possible with aggregate data.²⁴ For these reasons, IPD meta-analyses are considered the “gold standard” of evidence synthesis.^{21,24}

To sum up, IPD meta-analyses that integrate the results of descriptive analyses of ELSA data draw on the strengths of two gold-standard methods: meta-analyses of (a) representative probability samples and (b) the raw data recorded for each participant. However, meta-analytic models have only rarely been applied to results from ELSAs in general^{5,7} and to IPD from ELSAs in

particular.^{6,9–11,26*} Why? One reason may be that there is ample guidance available for either conducting descriptive analyses of studies with complex survey designs^{16,27–29} or carrying out IPD meta-analyses.^{25,30} But only a single guidance paper is available that describes important steps to meta-analytically integrate descriptive analyses from health surveys.⁴ This paper, however, does not focus on IPD meta-analyses of studies with complex survey designs.

3 | THE PRESENT TUTORIAL

To address this gap in the literature, we aim to provide researchers with methodological guidance on how to take advantage of IPD meta-analyses to synthesize the empirical evidence obtained from descriptive analyses of studies with complex survey designs. These studies share several central methodological features (e.g., sampling weights, clustered and missing IPD, dependent effect sizes) that need to be taken into account when analyzing the IPD and integrating the results.⁴ To this end, we use ELSAs as an example to offer a robust and versatile work flow that can be applied to (a) carry out quantitative descriptive analyses with IPD from studies with complex survey designs, which are then (b) integrated by means of meta-analytic models. Specifically, we first discuss general characteristics of IPD meta-analyses of ELSAs. We then elaborate on how the two-stage approach of IPD meta-analysis^{31,32} can be tailored to account for the methodological complexities of ELSAs when estimating effect sizes in Stage 1 and integrating the results by means of meta-analytic models in Stage 2. We illustrate the potential of the two-stage approach for integrating results of descriptive analyses of ELSA data by drawing on PISA public use files encompassing data from over 400 independent samples with about three million students (see Figure 1). Particularly, we show how to estimate effect sizes that are often synthesized in meta-analyses in Stage 1²²: standardized mean differences between groups (i.e., gender differences in reading achievement) and bivariate correlations (i.e., the correlation of students' SES with reading achievement). Further, we discuss and illustrate a key strength of IPD meta-analysis—the analysis of heterogeneity in effect sizes at the participant level (i.e., how the size of gender differences in reading achievement is moderated by students' SES). We also elaborate on and showcase Stage 2 in which meta-analytic and meta-regression models are applied to integrate the (dependent) effect sizes as estimated in Stage 1. Finally, we discuss the opportunities, challenges, and limitations that come from using IPD meta-analyses for descriptive analyses of ELSA data. We also offer extensive

online supplementary material (OSM; see Supporting Information) where we provide details on the applied methods. Further, we share the syntax for reproducing all results on the Open Science Framework (<https://osf.io/wfd6p/>). In doing so, we want to facilitate other researchers' use of IPD meta-analyses of studies with complex survey designs.

4 | IPD META-ANALYSES OF ELSAS: GENERAL CHARACTERISTICS

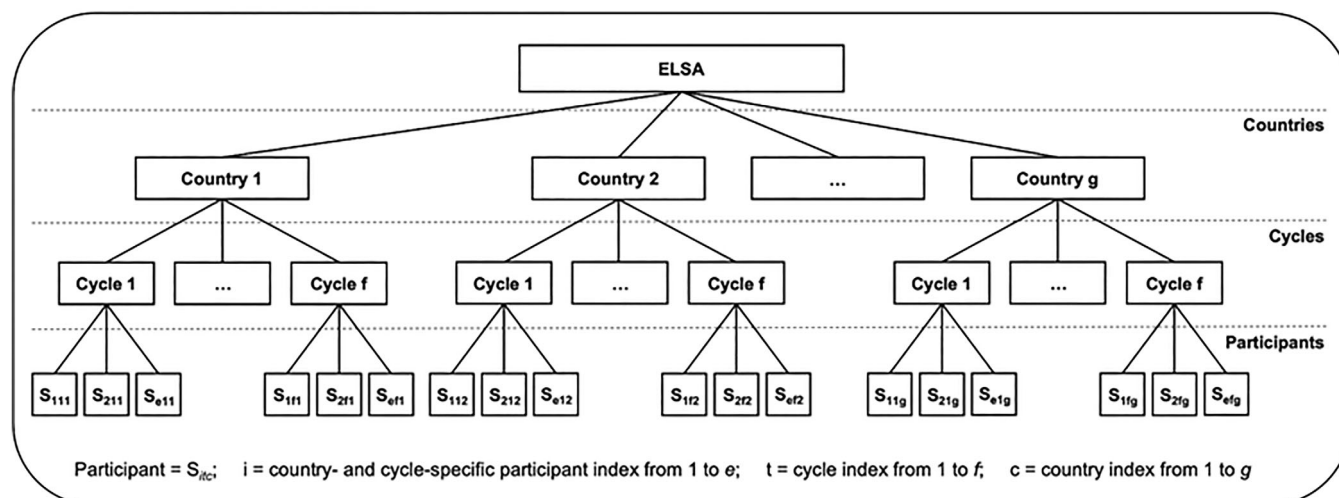
4.1 | IPD meta-analyses of ELSAs and systematic reviews

We begin our discussion of IPD meta-analyses of ELSAs by distinguishing meta-analyses from systematic reviews (also known as research syntheses) because these terms have often been used synonymously even though they refer to different methodological concepts.³³ Specifically, the term meta-analysis concerns the statistical analysis (e.g., of IPD from ELSAs or AD from published and unpublished studies) with the aim of integrating the findings.^{33(p532)} The term systematic review, on the other hand, refers to the entire process of using systematic methods to identify, select, collect, analyze, integrate, and critically appraise relevant research. In systematic reviews, meta-analytic models might or might not be applied to integrate the results.^{33(p535)} Importantly, systematic reviews aim to cover the complete body of empirical data and results that are relevant to a certain research question.³⁴ On the other hand, IPD meta-analyses of descriptive analyses of ELSA data are aimed at covering relevant—but not the complete body of—empirical evidence. They may therefore significantly contribute to the cumulative evidence in the behavioral and social sciences and become an important component, but they are not a substitute for systematic reviews (see Section 6).

4.2 | One-stage and two-stage approaches to IPD meta-analyses of ELSAs

IPD from ELSAs are typically hierarchically structured. For example, in PISA and other international ELSAs (e.g., TIMSS and PIRLS), IPD from several independent student samples are available for most countries (see Figure 2a). This data structure needs to be taken into account when conducting an IPD meta-analysis of ELSA data. There are two major approaches that can accomplish this goal: the two-stage and the one-stage approach. In Stage 1 of a two-stage IPD meta-analysis, the IPD for

(A) Data Structure of ELSAs



(B) One-Stage and Two-Stage IPD Meta-Analyses with ELSAs (Using PISA 2000–2018 as an Example)

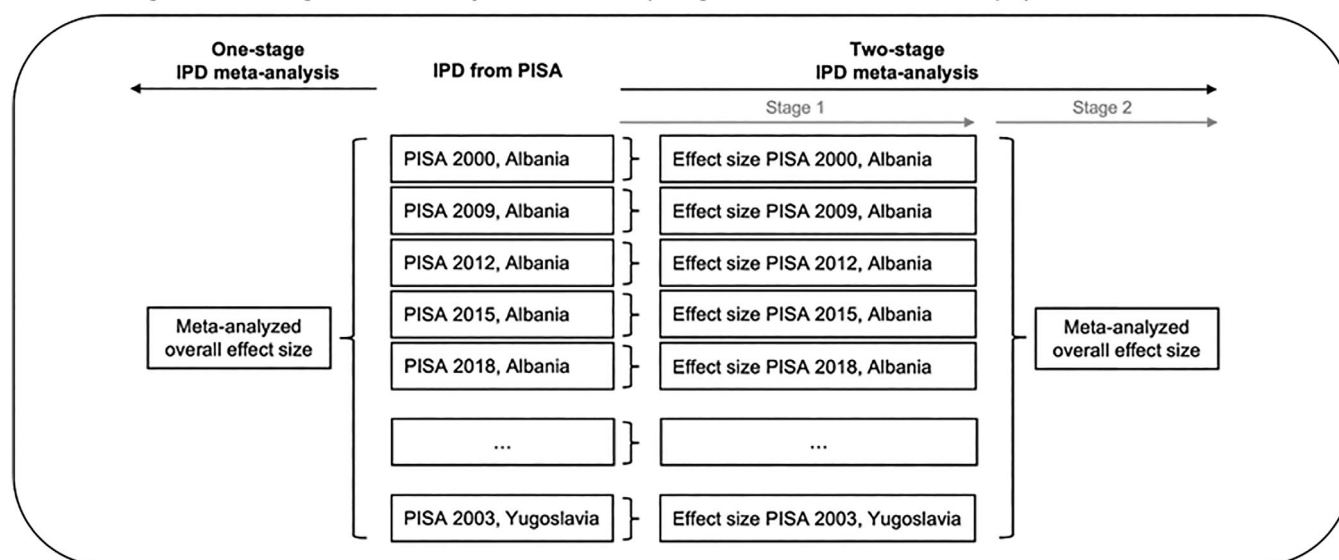


FIGURE 2 Schematic representation of individual participant data (IPD) meta-analyses of ELSAs using Programme for International Student Assessment (PISA) as an example: (a) Hierarchical data structure of international educational large-scale assessments (ELSAs) and (b) one-stage and two-stage meta-analyses applied to these IPD

each sample from an ELSA are used to estimate effect sizes and corresponding sampling variances or their square roots, the SEs (see Figure 2b). In many applications of the two-stage approach, the estimation of effect sizes (and their sampling variances) draws on statistical models from the large family of generalized linear models that are specified at the participant level, for example, regression models for continuous, binary, ordinal, or count data.^{27,28} In Stage 2 of a two-stage IPD meta-analysis, the effect sizes are integrated with meta-analytic models that are also applied in AD meta-analyses, involving common-effect models (fixed-effect models) or random-effects models.^{35,36} In these models, more precise effect size estimates (in terms of their sampling variances) obtain larger weights (inverse variance weighting).

The one-stage approach, on the other hand, combines the two stages of the two-stage approach into a single statistical model to meta-analyze the data. To this end, the IPD from all samples of one specific or several related ELSAs are combined in a single large data set and analyzed simultaneously in this model. For example, using all IPD from PISA, a one-stage approach would involve creating a data set with about three million students (see Figure 2b). Given the hierarchical nature of these IPD, the statistical model is typically a multilevel model taken from the family of generalized linear mixed models.²⁵

Both one-stage and two-stage approaches can be used to fit statistical models that draw on the same set of assumptions for the same set of IPD. So which approach should be used for IPD meta-analyses of ELSAs? The major

difference between the two approaches lies in the procedures they use to estimate the meta-analytic parameters, and the largest differences can be expected for sparse data.^{32,37} For example, when studying differences between independent groups (e.g., male and female students), the most notable differences in meta-analytic estimates between these approaches can be expected for IPD meta-analyses of ELSAs when most samples have IPD from (a) only a few participants (e.g., <30 per group) or (b) only a few events (e.g., <10 per group) demonstrating a rare experience (e.g., winning an international math Olympiad).^{37(p216)} Further, when examining moderating effects at the participant level (e.g., how gender differences are moderated by students' SES), differences between these approaches can be expected to be largest, particularly for binary outcomes (e.g., obtaining a college degree or not), with only a few samples (e.g., 5) and only a few participants per sample (e.g., <50).^{38(p11)} In such applications, the one-stage approach is preferred because it will likely give more precise and less biased meta-analytic estimates.^{25,31,37}

Sparse data occur more often when synthesizing evidence, for example, in the biomedical sciences,³⁹ than when drawing on ELSAs (and other complex survey data) because key characteristics of these studies are large sample sizes and a large number of available samples (see Figure 1). Drawing on such an ensemble of IPD, empirical, theoretical, and simulation studies have shown that the one- and two-stage approaches provide very similar results when statistical models that draw on the same set of assumptions are fit to these data.^{31,32} In the present tutorial, we therefore focus on and recommend the two-stage approach that also offers several additional advantages.

First, the one-stage approach requires a commensurable metric for the applied measures, whereas the two-stage approach requires a commensurable metric for the effect sizes. In this respect, the two-stage approach is easier to apply. For example, different ELSAs (e.g., PISA and PIRLS) typically measure the same target construct (e.g., reading achievement) with different instruments. To allow for comparison of results across ELSAs, the two-stage approach draws on standardized effect sizes that are estimated in Stage 1 and meta-analytically integrated in Stage 2. The one-stage approach, however, requires to first establish a commensurable metric for the applied measures to allow for the meta-analytic integration of results within a single model. To this end, researchers need to harmonize the target measures, for example by using advanced psychometric models or missing data procedures.^{40,41} The success of these harmonization procedures depends strongly on the availability and quality of the applied measures (e.g., a large number of common items that are psychometrically equivalent

across ELSAs), which may even preclude the application of the one-stage approach.^{40,41}

Second, the two-stage approach requires considerably less expertise to account for the statistical complexities that are inherent to IPD from ELSAs (e.g., sample weights, clustered and missing IPD, dependent effect sizes) than the one-stage approach, which requires all these complexities to be incorporated in the specification of a single model.^{31,32,37} This advantage becomes even more valuable when the applied survey methodologies vary across ELSAs (e.g., number of plausible values provided for students' achievement [see Section 5.1.4 on "Plausible Values"], methods used to estimate sampling variances for effect sizes).

Third, IPD meta-analyses of ELSAs often require analyses of complex, hierarchical data structures with IPD from several million individuals (see Figures 1 and 2). This may result in an inability to analyze the data due to the insufficient random-access memory of regular computers—a typical problem with Big Data analyses.⁴² In such cases, the two-stage approach may be the only feasible approach because it takes advantage of an analytic strategy that has been developed for these Big Data problems. This strategy entails first splitting a single large data set (e.g., a data set comprising about 3,000,000 IPD points from students from PISA) into considerably smaller subsets (e.g., 424 data sets comprising student samples from a certain country and PISA cycle). A standardized protocol for managing and analyzing the data is subsequently applied to each subset. Then, the results obtained for each subset are combined, for example, by using meta-analytic methods.⁴²

Finally, the two-stage approach can be used to integrate a broader spectrum of effect sizes, involving effect sizes that are typically applied in meta-analyses (e.g., standardized mean differences, e.g., Cohen's *d* or correlations)^{22,43} as well as less commonly used ones. For example, Keller et al.⁹ used the two-stage approach to meta-analyze the overlap between the distribution of achievement profile scores as obtained from female and male students belonging to the group of top-performing math students in their respective countries. This kind of effect size cannot be specified in generalized linear mixed models that are typically applied in the one-stage approach.

4.3 | Explaining heterogeneity in effect sizes in IPD meta-analyses of ELSAs

4.3.1 | Heterogeneity in effect sizes at multiple levels

A major goal of a meta-analysis is to estimate the extent to which the magnitude of an effect size depends on

further moderator variables.^{36,44,45} In IPD meta-analyses of ELSAs, heterogeneity in effect sizes can be investigated at different levels by applying regression and meta-regression models. These models help to examine the extent to which one or more categorical or continuous moderators explain differences in effect sizes (a) across countries (Level 3 in Figure 2a), (b) across assessment cycles within countries (Level 2 in Figure 2a), or (c) at the participant level (Level 1 in Figure 2a). For example, a moderator analysis at the country level may explain heterogeneity in gender differences in reading achievement between countries by country-level characteristics, involving socioeconomic, educational, cultural, or political factors.^{5,9} A moderator analysis at the cycle level may explain heterogeneity in gender differences in reading achievement across time.^{6,7} Finally, a moderator analysis at the participant level may examine how gender differences in reading achievement are moderated by students' SES within countries.

4.3.2 | Ecological fallacies

Compared with AD meta-analyses, IPD meta-analyses offer one major advantage: moderator analyses at the participant level. IPD meta-analyses of ELSAs may therefore provide unique insights into important phenomena that cannot (or can hardly) be provided with AD meta-analyses because the effect sizes that depict interactions between individual characteristics at the participant level are often not reported at all or are available in only a small subset of the relevant primary studies. Yet, unless data on these effect sizes are available, it is impossible to learn about moderating effects at the participant level. Intuitively, one might expect that moderating effects as observed at higher hierarchical levels (e.g., the country level)—which is possible with AD meta-analyses—would allow for valid inferences about participant-level moderating effects. However, this intuition is wrong and has been referred to by different terms, such as, *ecological fallacy*,⁴⁶ *ecological bias*, or *aggregation bias*.⁴⁷ We illustrate this problem in Figure 3 with artificial ELSA data from 30 countries. Figure 3 displays how the magnitude of gender differences in students' reading achievement may be moderated by (a) students' SES at the participant level and (b) students' SES aggregated on the country level. Figure 3a,b show that the magnitude of gender differences is related to average SES at the country level ($r_{\text{between}} = 0.70$) with larger gender differences observed in countries with higher average SES. Further, with increasing levels of students' SES at the participant level within each country, the magnitude of gender differences increases ($r_{\text{within}} = 0.50$) in Figure 3a, whereas it decreases in Figure 3b ($r_{\text{within}} = -0.50$). Taken together, these examples emphasize that moderating effects may have different sizes and even directions at the

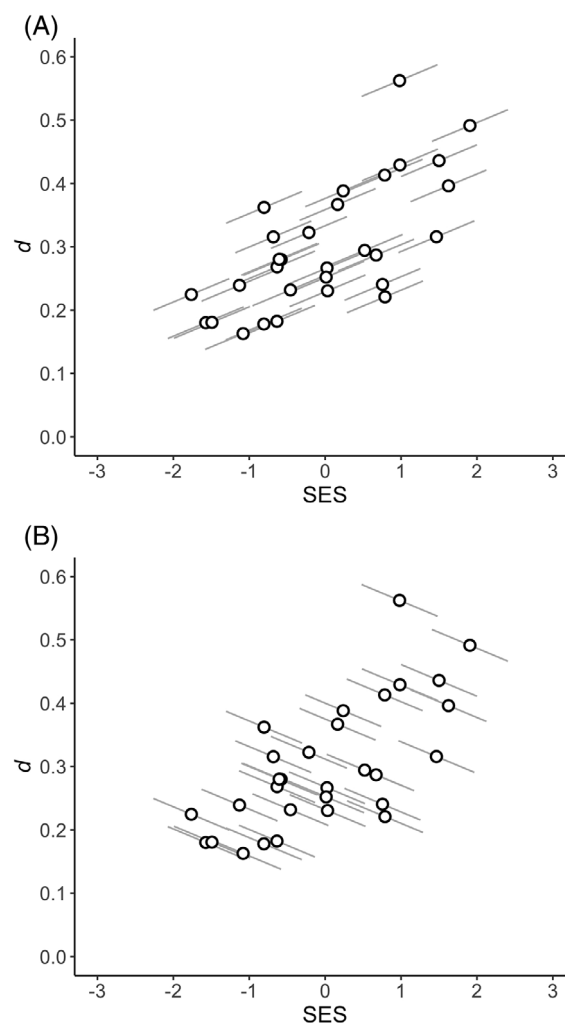


FIGURE 3 Illustration of the ecological fallacy. Hypothetical relationships between gender differences (in terms of d) and students' socioeconomic status (SES) within countries at the participant level (represented by lines) and aggregated at the country level (represented by dots). Average gender differences are substantially related to average SES ($r_{\text{between}} = 0.70$) in both panels. (a) The magnitude of gender differences is substantially positively related to students' SES within countries ($r_{\text{within}} = 0.50$). (b) The magnitude of gender differences is substantially negatively related to students' SES within countries ($r_{\text{within}} = -0.50$). Adapted from Thompson and Higgins⁴⁷

participant and country levels. Thus, it is not possible to draw valid conclusions about moderating effects at the participant level from results obtained at higher hierarchical levels (and vice versa).

4.3.3 | Some guidance for examining the heterogeneity of effect sizes at various levels

The examples of ecological fallacies highlight that the two-stage approach to IPD meta-analyses of ELSAs requires researchers to pay special attention to several

key issues when specifying, estimating, and interpreting moderating effects to explain heterogeneity in effect sizes.^{25,38,48,49} Arguably the most important issue here is that it is important to match the substantive research question on the one hand with the appropriate level of analysis and form of centering of the moderating variables on the other.^{48–52}

First, when the research goal is to learn about heterogeneity in effect sizes at the participant level, it is essential to directly model the interaction between individual characteristics at this level in Stage 1. In particular, when the research goal is to study moderating effects for the populations represented by each sample (see Model Set 3 in Section 5.1.5), the regression models that need to be specified for each sample must include the target interaction effect(s) between the predictor variables depicted as multiplicative terms (e.g., Gender \times SES) in addition to the main effects of these predictors (e.g., Gender and SES).^{25,48,49,53,54} In Stage 2, the multiplicative terms are integrated with meta-analytic and meta-regression models.^{25,48,49,53,54} Of note, the regression models used in Stage 1 can be extended (e.g., by using cubic spline functions) to explore the extent to which the magnitude of the moderating effect varies along the distribution of a certain predictor variable at the participant level.⁴⁸ Integrating the results of these extended models may require multivariate meta-analytic and meta-regression models in Stage 2.^{48,55}

Second, for regression analyses involving interactions as described above, it is recommended that the predictor variables in these models are centered or *z*-standardized to reduce collinearity and ill-conditioning in the data.^{51,52} Further, to facilitate the interpretation of regression coefficients for interaction effects, each predictor variable should have a meaningful zero point. Such zero points can be achieved, for example, by dummy-coding nominal variables (e.g., gender) or *z*-standardizing or centering continuous variables (e.g., occupational status as a measure of SES).⁵² Of note, dichotomizing continuous predictor or outcome variables when examining interaction effects (e.g., with the goal of improving the interpretation of results) is not recommended because doing so decreases the precision with which the interaction term can be estimated.⁴⁸

Third, when applying the two-stage approach to IPD meta-analyses of ELSAs to examine the heterogeneity of effect sizes, we strongly recommend that researchers use the analytic strategies (e.g., meta-analyzing interaction terms) as described above because these strategies can help researchers avoid the statistical pitfalls that are likely to occur when using alternative approaches.^{48,49,53} In particular, when examining how effect sizes vary across participant subgroups (e.g., male and female students), it might be tempting to estimate effect sizes for

each subgroup separately for each sample in Stage 1, meta-analytically combine these effect sizes within these subgroups in Stage 2, and then compare the meta-analytic results across subgroups. However, this strategy should not be used because it is prone to ecological biases as it combines moderating effects that may be observed within samples with moderating effects that may be observed across samples (e.g., across countries).^{48,49}

Fourth, many ELSAs apply stratified multistage random sampling schemes (see Section 5.1.1), for example, by first drawing random samples of schools within a country and then sampling students within schools. This sampling process leads to clustered data at the sample level (e.g., students nested in schools). It is important to always consider the clustered data when estimating the SEs of the effect sizes in Stage 1 (see Section 5.1.2). However, depending on the substantive research question, the clustered nature of the data might or might not be considered when estimating the effect size. For example, when the research goal is to study the extent to which the magnitude of gender differences is moderated by students' SES in the total population that is represented by a certain sample, it is sufficient to specify a linear regression model for each sample as described above in Stage 1. Notably, the resulting effect size represents an amalgam of the within-cluster interaction between gender and SES (e.g., within schools) and the between-cluster interaction between these variables aggregated to the cluster level (e.g., the proportion of female students in schools and the average SES at school level). Thus, when the research goal is to disentangle within-cluster relationships from between-cluster relationships in Stage 1, statistical models (e.g., multilevel models) that provide separate effect sizes for each hierarchical level in each sample need to be specified. To this end, it is necessary to center all variables (including dummy-coded variables, e.g., gender) within clusters within each sample, for example, by subtracting the school-specific proportion of female students from each student's value on the gender variable or subtracting the average SES at the school level from each student's value on the SES variable.⁵⁰ The interaction term is then computed by multiplying the predictor variables after centering them within clusters. When using this approach, it is important to (a) use appropriate approaches (e.g., weights at the participant and cluster levels) that take into account the sampling process of the ELISA when estimating the effect sizes at the participant and cluster levels in Stage 1 (see also Section 5.1.1)^{56,57} and (b) consider the possibility that the magnitude of effect sizes may vary between clusters (e.g., by specifying a multilevel model in Stage 1 that includes a random effect for the interaction term).⁴⁸

Fifth, in IPD meta-analyses of international ELSAs, some moderator variables may vary within and between countries when meta-regression models are used to

examine the heterogeneity of effect sizes in Stage 2. For example, some studies have found that gender differences in mathematics achievement are moderated by gender equality indicators at the country level (e.g., the percentage of women in tertiary education in a country).^{5,9} These factors may vary within countries across time (e.g., because of political initiatives to reduce gender disparities in a certain country) or between countries (e.g., because of stable between-country differences in the educational and economic opportunities offered to men and women). Because moderating effects may vary in magnitude and direction within and between countries, including a certain moderator as a single variable in the meta-regression model yields a meta-regression coefficient that comprises a mixture of the within- and between-country relationships.^{58,59} To disentangle these within- and between-country relationships, the meta-regression model should contain (a) the aggregated country-specific mean of the moderator (e.g., the country-specific value of a certain gender equality indicator averaged across time) to estimate the moderating effect at the country level and (b) the centered value of the moderator (e.g., the deviation of a gender equality indicator as observed at a certain point in time from the country-specific average) to estimate the moderating effect within countries.⁵⁹

4.4 | Missing data in ELSAs

4.4.1 | Sporadically and systematically missing data

All empirical analyses face the problem that missing values reduce the information available on effect sizes, moderator variables, or both. Missing data can take different forms in IPD meta-analyses of ELSAs: (a) sporadically, (b) planned systematically, and (c) unplanned systematically missing data.⁶⁰ Sporadically missing data occur at the participant level (Figure 2a) and refer to missing data for those participants who were included in a certain sample and who were administered a certain instrument (e.g., a test booklet or questionnaire). For example, such missing data may result when participants intentionally ignore or forget to answer certain questions or test items. By contrast, planned systematically missing data occur in ELSAs at the participant level because planned missing designs (also called booklet designs) are applied (see Section 5.1.4 on “Plausible Values”), thus implying that some data are missing for random subsamples of participants. Finally, unplanned systematically missing data may occur in ELSAs at the cycle level (see Figure 2a) when data are not available for all individuals

in a certain sample.⁶⁰ For example, in PISA 2006, all students in the United States had (unplanned) systematically missing data on reading achievement because the applied paper-pencil test contained severe printing errors.^{61(p281)} Thus, it was not possible to estimate, for example, gender differences in reading achievement for that cycle in the United States. Moreover, unplanned systematically missing data may also occur at the cycle or country level when information on a certain moderator variable is not available for a specific cycle or country (e.g., information on country-specific socioeconomic or political factors).

4.4.2 | Missingness mechanisms

Fortunately, (most) missing data in ELSAs can be handled effectively. The possibility of imputation and the choice of imputation method depends on the type of missing values. Three different missingness mechanisms can be distinguished: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).^{62,63} For example, if students happen by chance to be ill on the day of testing, their achievement data can be considered to be MCAR. Missing data in ELSAs are MAR when missingness on a target variable X can be explained by other observed data in the data set but does not depend on unobserved data, including unobserved values of X itself. For example, MAR occurs for achievement data if school personnel discourage students with lower SES from taking the tests. However, when students' SES is observed and taken into account, missingness does not depend on the unobserved achievement scores. Finally, missing data in ELSAs are MNAR when the missingness on X depends on unobserved data (e.g., unobserved values of X itself). For example, MNAR occurs for achievement data if school personnel discourage students they know will perform poorly on the test from coming to school on testing days. If the school personnel's knowledge is not observed, and can thereby not be taken into account, the missingness of the achievement data depends on an unobserved variable.

The missingness mechanism describes the conditions under which a certain missing data method works best to provide unbiased and precise estimates of statistical parameters.^{62,63(p526)} Most modern missing data methods, involving multiple imputation or full information maximum likelihood, were developed for situations in which MAR or MCAR is assumed.^{60,63} These methods substantially outperform traditional methods (e.g., listwise deletion) or work at least equally well.^{62–64} In Section 5.1.4, we explain how modern missing data methods can be applied in the context of the two-stage approach to IPD

meta-analysis of ELSAs to deal with sporadically and planned systematically missing data in Stage 1. In Section 5.2.4, we discuss how to treat unplanned systematically missing data in Stage 2.

5 | INTRODUCING THE TWO-STAGE APPROACH TO IPD META-ANALYSIS TAILORED TO ELSAS

In the following, we discuss and illustrate how to tailor the two-stage approach to IPD meta-analysis to the methodological complexities of ELSAs. In Stage 1, the tailoring involves addressing the statistical challenges of using IPD from complex survey designs (e.g., sampling weights, clustered and missing IPD) when estimating effect sizes for descriptive analyses. Because ELSAs have been conducted regularly at both international and national levels for many years, several effect sizes may be available for the same country. These country-specific effect sizes are likely to be correlated (i.e., dependent); that is, they are more similar to each other than to effect sizes from other countries. In Stage 2, the tailoring therefore involves taking the dependent structure among effect sizes into account when integrating the results with meta-analytic and meta-regression models.

5.1 | Stage 1: Effect size estimation for IPD from ELSAs

ELSAs implement various methodological state-of-the-art features, for example, to ensure that data are representative at the population level and to cover as much content as possible with surveys that are as brief as possible. In addition, as in any other study, missing values occur in ELSA data. These features and characteristics need to be taken into account when estimating effect sizes and their sampling variances in IPD meta-analyses.

5.1.1 | Stratified multistage random sampling and (final) sample weights

Stratified multistage random sampling

ELSAs are designed to obtain precise estimates of effect sizes for a well-defined population. The sampling processes in ELSAs achieve this by two means: representative, random selection of study participants and large sample sizes.¹⁶ To this end, ELSAs often apply stratified multistage sampling procedures. For example, using two-stage sampling, the primary sampling units (PSUs; e.g., schools) are randomly selected within strata (e.g., geographic regions, types of schools) in the first stage

of sampling. In the second stage, a random sample of individuals (e.g., students) are selected from a certain PSU. Stratification offers at least two advantages. First, it guarantees that a prespecified number of individuals who belong to a certain (policy-relevant) subgroup (i.e., the stratum) is included in the sample. Second, stratified sampling helps to substantially improve the precision needed to estimate statistical parameters when the strata variables are related to the key target outcomes (e.g., student achievement).^{28,65}

Stratified multistage random sampling as applied in PISA

To illustrate the stratified multistage random sampling process that is typically applied in ELSAs, we use PISA as an example. PISA is aimed at evaluating education systems worldwide at the end of compulsory education by assessing skills and knowledge in the target population of 15-year-old students. PISA sets high-quality standards for collecting representative probability samples and obtaining precise estimates of effect sizes.¹⁵ Specifically, at least 4500 students in each country participate in each PISA cycle, or the full student population is included if it is smaller than 4500. To this end, most countries apply a stratified two-stage sampling design. In the first sampling stage, individual schools with 15-year-old students are systematically sampled from a stratified list of all schools with sampling probabilities proportional to the number of 15-year-old students enrolled. Strata are specific for each country and include characteristics that explain performance differences between schools, for example, geographic regions or school types. In each country, a minimum of 150 schools have to be selected; if a country has fewer than 150 schools, all schools are selected. Also, in most countries, 35 students who are 15 years old are randomly selected within schools; if a school has fewer than 35 students at age 15, all students in this age group are selected.

(Final) sample weights

After drawing a stratified sample of participants, the next step is to weight the data to be able to calculate unbiased estimates of effect sizes for the target population (i.e., a well-defined finite population). A well-defined finite population is a population for which the number of members is known. For inferences to finite populations, each member of the population has a non-zero probability of being selected into the sample. To this end, a sampling frame is required to estimate the total number of population members. Based on the sampling frame, the data from the individuals selected from the sample are weighted when estimating the effect size. The weight can then be interpreted as the number of people in the population who are represented by a certain individual who was

selected from the sample. OSM.2 provides a didactic example for the computation of weights in stratified two-stage random sampling. Specifically, the weights are computed such that they reflect the joint (i.e., multiplied) probabilities of the inclusion of the PSUs (e.g., schools) as well as of the individuals in the PSUs (e.g., students). Notably, weights in multistage probability samples often require some adjustments to obtain participants' final weights. Adjustments involve accounting/correcting for (a) the over- or undersampling of some strata of the population (e.g., when a larger number of minority students were sampled to obtain more precise effect sizes for this subpopulation), (b) issues in the sampling frame because of inaccuracies in estimating the total number of population members (e.g., because the sampling frame was determined 2 years before the data were collected), (c) nonresponse at the level of PSUs (e.g., a school selected for the sample did not participate), and (d) nonresponse at the individual participant level (e.g., a selected student did unexpectedly not take part in the assessment).^{66,67} The final weights to which such adjustments are applied are then used to estimate the effect size.[†]

5.1.2 | Clustered data and the estimation of sampling variances

Clustered data

The multistage random sampling procedure makes the computation of sampling variances more complex than in studies with simple random sampling. In simple random sampling, individuals are selected independently of each other. Thus, in standard methods for computing sampling variances of effect sizes, selected individuals are treated as independent observations.⁵⁷ By contrast, in multistage sampling, individuals selected from the same PSU do not represent independent observations. Instead, the multistage sampling process induces dependencies among the selected individuals, thus leading to so-called clustered data.^{57,65,66} In particular, students in the same school are often found to be (much) more similar to each other than to students from other schools.⁶⁹ These similarities may result from, for example, performance-based tracking into a certain school (or type of school) or sharing the same teachers. Importantly, the dependencies resulting from multistage sampling need to be taken into account because they substantially increase the sizes of the sampling variances of the effect sizes.⁵⁷

Estimation of sampling variance

Several methods have been developed to account for the estimation of the sampling variance of clustered data, involving sandwich estimators, multilevel models, linearization methods, and replication techniques.^{65,68,70} For IPD

from ELSAs, we recommend replication techniques because they (a) are asymptotically consistent for the true sampling variance with increasing sample size, (b) provide sampling variance estimates that are very similar to those obtained from using more complex procedures, (c) are robust and can be flexibly applied to estimate sampling variances for a very large variety of effect sizes (e.g., mean differences, correlations, and regression coefficients), and (d) are well supported by relevant information (i.e., replicate weights) in public use files because they are the standard method applied in ELSAs.^{65,67,70} Notably, sometimes only replication techniques can take full advantage of the stratified sampling design because the explicit information on strata membership needed to obtain better (i.e., smaller) sampling variances with sandwich estimators, multilevel models, or linearization methods is not provided, or details are left out to guarantee the confidentiality and anonymity of individuals (e.g., students or teachers).^{15(p198)}

Two replication techniques are commonly applied in ELSAs for which replicate weights are provided in public use files⁷¹: the Jackknife (JK) method, which is used (with different modifications, e.g., JK2) to estimate sampling variances (e.g., in TIMSS, PIRLS, and NAEP), and the balanced repeated replication (BRR) method, which is applied in a modified form (i.e., Fay BRR), for example, in PISA.^{66,67} The common idea behind all replication techniques is to use computational intensity to tackle the problem that an analytical technique for estimating the sampling variance of a certain effect size T is not available (for a certain statistical model) or has not yet been developed.^{29,70} Each replication approach therefore estimates the sampling variance $v(T)$ by using a large number of somewhat different subsamples of PSUs—the replicates—taken from the original sample. The subsamples are derived from applying the replicate weights to the original data. The variability of the resulting effect size estimates around T (as obtained from the original sample) is then used as an estimate of $v(T)$.^{29,66,70} OSM.3 provides details on how the replicate weights are derived for JK2, BRR, and Fay BRR and how these weights are used to obtain the sampling variance.

5.1.3 | Estimating effect sizes

Unstandardized and standardized effect sizes

A key feature of ELSAs is that many of the applied measures provide a metric that is commensurable across time or countries (e.g., achievement measures). Thus, meta-analysts can compute unstandardized effect sizes that are based on this original or raw metric, involving mean differences between groups or unstandardized regression coefficients. Alternatively, the meta-analyst can compute standardized effect sizes, involving Cohen's d or Hedges'

g, correlations (r), and standardized regression coefficients (β). Both the standardized and unstandardized options have advantages and disadvantages.^{72,73}

One advantage of unstandardized effect sizes is that their sampling variance can be expected to be somewhat smaller than that of standardized effect sizes. This implies that unstandardized effect sizes can be estimated more precisely because standardized effect sizes are computed as ratios where both the numerator and the denominator are subject to sampling variability.^{74(p175)} For example, d is the mean difference between two groups divided by an estimate of the population SD. Further, measures with a commensurable metric are also necessary to compute unstandardized effect sizes that address research questions that focus on a comparison of effect sizes in absolute terms (e.g., To what extent has the mean level of the reading achievement of a certain country changed over time?). Finally, if the raw metric is intuitive (e.g., number of days absent in school, income in Euro or US dollars) or well-established (e.g., body mass index), unstandardized effect sizes are advantageous because their meaning can be conveyed much more easily to a broader audience comprising researchers as well as practitioners and policy makers.^{72,73}

A major advantage of standardized effect sizes is that they allow results to be compared across studies even when different measures were used because the results are converted to a standardized scale.⁷² For example, PIRLS and PISA use different reading achievement tests. Nevertheless, gender differences in terms of standardized mean differences (e.g., d or g) can be compared and combined across studies because they are expressed in terms of SD units. Further, standardized effect sizes are well-established in the behavioral and social sciences.^{22,72} Hence, when standardized effect sizes are computed for ELSAs, their magnitudes can be compared with relevant prior research as well as with established benchmarks that are expressed in terms of SDs. Such benchmarks involve typical intervention effects, learning gains, differences between policy-relevant subgroups (e.g., gender, ethnicity), or differences between low-achieving and average-achieving schools.⁷⁵ Because of these advantages, we focus on standardized effect size measures in the present tutorial. Notably, when the research goal is to meta-analyze effect sizes that are based on measures with a commensurable metric, meta-analysts can capitalize on the advantages of both unstandardized and standardized effect sizes.

Linear regression models to estimate standardized effect sizes

When estimating effect sizes for IPD meta-analyses of ELSAs, it is important that researchers specify a statistical model that matches the substantive research question. Many substantive research questions (e.g., for ELSAs and

other complex surveys) require researchers to estimate effect sizes for the total population that is represented by a certain sample. Such effect sizes (and their sampling variances) can often be estimated by drawing on the large family of generalized linear models. In particular, linear regression models provide a versatile statistical framework for estimating the key standardized effect size measures that are applied in many meta-analyses,²² involving standardized mean differences between two independent groups and correlations. Moreover, linear regression models can also be applied to estimate standardized effect sizes that depict moderating effects at the participant level. Figure 4 illustrates the model specifications that we used in our empirical examples.

When using linear regression models, obtaining standardized effect sizes requires the standardization of some or all of the variables that are involved in estimating the effect sizes. First, computing standardized mean differences between independent groups requires the outcome (e.g., reading achievement) to be z -standardized ($M = 0$ and $SD = 1$) by using estimates of the population mean and SD. The indicator variable that provides information about group membership should be dummy-coded (e.g., 0 = male students, 1 = female students). The standardized regression coefficient (e.g., β_{Gender}) then depicts a standardized mean difference (equivalent to d). Second, a bivariate regression model where both the outcome (e.g., reading achievement) and the predictor variable (e.g., the SES measure) are z -standardized is needed to compute correlations. The bivariate standardized regression coefficient (e.g., β_{SES}) is identical in size and interpretation to the correlation between the outcome and predictor.⁴³ Third, examining moderating effects at the participant level requires a multiplicative term between the predictor variables to represent their interaction (e.g., Gender \times SES; see Section 4.3.3). To standardize regression coefficients for interaction terms, nominal variables (e.g., gender) can be dummy-coded, and continuous predictor variables (e.g., SES) can be z -standardized.

The models illustrated in Figure 4 do not provide separate effect sizes to disentangle relationships within and between clusters within a certain sample. However, for some substantive research questions researchers may need to obtain effect sizes that depict relationships at different hierarchical levels within a certain sample (e.g., separating gender differences in reading achievement as observed within and between schools within a certain country). To this end, researchers can, for example, apply a two-stage approach to IPD meta-analysis in which multilevel models are used to estimate effect sizes for the within-cluster and between-cluster relationships in each sample in Stage 1. These effect sizes can then be integrated with separate meta-analytic models in Stage 2 (see also Section 4.3.3).

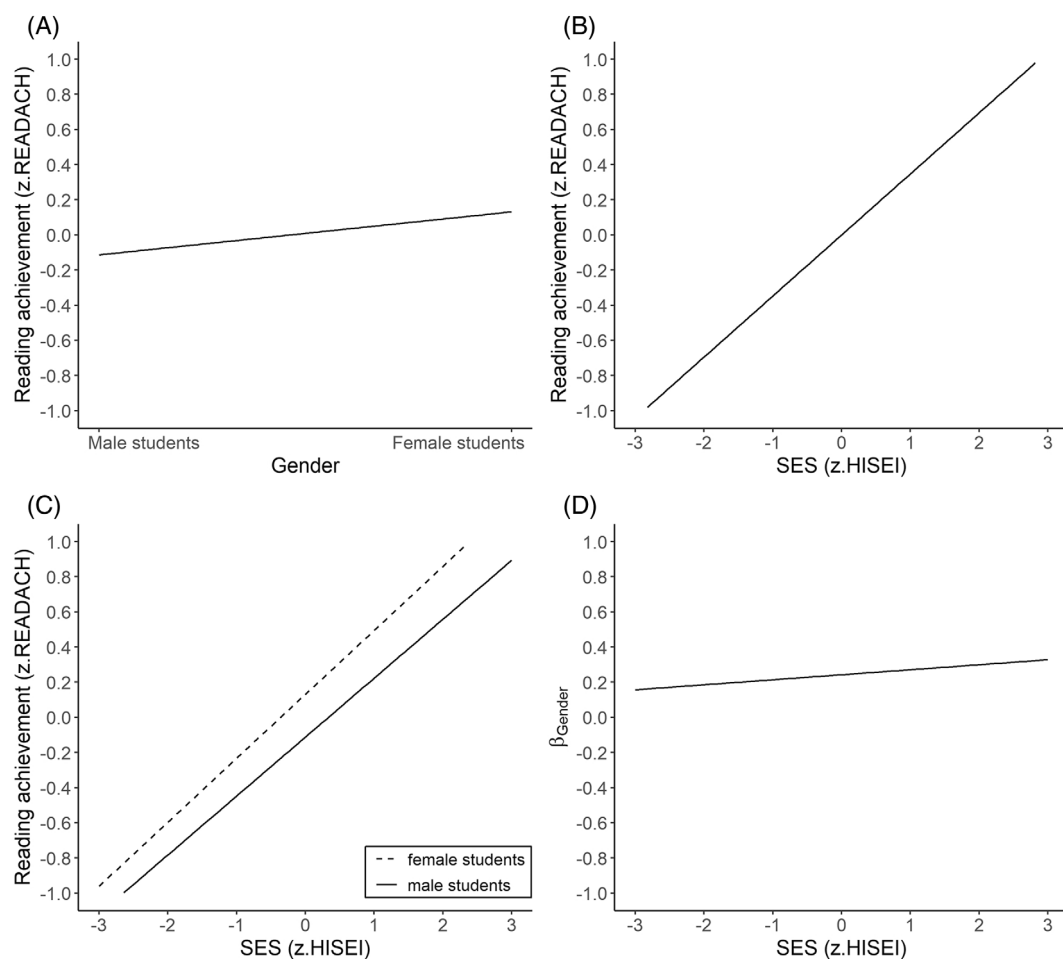


FIGURE 4 Illustration of the model sets applied to estimate standardized effect sizes (and their sampling variances) using data from Germany (Programme for International Student Assessment [PISA] 2018) as an example: (a) Gender differences in reading achievement (β_{Gender} ; Model Set 1), (b) the relationship between students' socioeconomic status (SES) and reading achievement (β_{SES} ; Model Set 2), (c) the relationships between reading achievement and gender and SES, and the interaction between gender and SES ($\beta_{\text{Gender} \times \text{SES}}$; Model Set 3), (d) gender differences in reading achievement as a function of students' SES (β_{Gender}^* as implied by Model Set 3). Specification of model sets and pooled results as obtained for the German PISA 2018 sample: (a) $\widehat{z.\text{READACH}}_{ict} = \beta_{0ct} + \beta_{\text{Gender},ct} \cdot \text{Gender}_{ict} = -0.11 + 0.25 \cdot \text{Gender}_{ict}$; (b) $\widehat{z.\text{READACH}}_{ict} = \beta_{0ct} + \beta_{\text{SES},ct} \cdot z.\text{HISEI}_{ict} = 0.35 \cdot z.\text{HISEI}_{ict}$; (c) $\widehat{z.\text{READACH}}_{ict} = \beta_{0ct} + \beta_{\text{Gender},ct} \cdot \text{Gender}_{ict} + \beta_{\text{SES},ct} \cdot z.\text{HISEI}_{ict} + \beta_{\text{Gender} \times \text{SES},ct} \cdot \text{Gender}_{ict} \cdot z.\text{HISEI}_{ict} = -0.11 + 0.24 \cdot \text{Gender}_{ict} + 0.34 \cdot z.\text{HISEI}_{ict} + 0.03 \cdot \text{Gender}_{ict} \cdot z.\text{HISEI}_{ict}$; (d) $\beta_{\text{Gender}}^* = 0.24 + 0.03 \cdot z.\text{HISEI}_{ict}$. Indices: i , student; c , country; t , cycle. $z.\text{READACH}$, z -standardized reading achievement score ($M = 0$, $SD = 1$ for the German sample in the year 2018 cycle). $z.\text{HISEI}$, z -standardized SES measure ($M = 0$, $SD = 1$ for the German sample in the year 2018 cycle)

5.1.4 | Missing data in ELSAs

Missing data can take different forms in IPD meta-analyses of ELSAs. In the following, we discuss how to handle planned systematically and sporadically missing data.

Plausible values

To cover a broad range of content, ELSAs often assess participants' characteristics by applying a planned missing design (so-called booklet designs) where each participant is randomly assigned to a booklet that comprises a selected subset of test items or questions taken out of a

pool of several hundred test items or questions. To handle the resulting systematically missing data to estimate unbiased group-level effect sizes, ELSAs draw on a methodology that is referred to as plausible values (PVs).⁶⁷ PVs reflect the idea that individual characteristics (e.g., achievement) are latent variables that cannot be observed directly with the applied measures but need to be estimated.^{66,76} To reflect the uncertainty underlying the estimation process, a range of possible values, the PVs, are estimated for each individual rather than a single point estimate (e.g., sum score). Specifically, PVs are derived as random draws from a distribution that is

estimated for each individual. These distributions are typically estimated by using item response theory (IRT) based on individuals' responses (e.g., to the test items) and further background information (e.g., individuals' socio-demographic background).⁷⁶ Using these procedures, for each participant between $5 \leq N \leq 20$ PVs are typically provided in public use files from ELSAs. Applying these PVs in descriptive analyses yields unbiased group-level estimates of effect sizes in a certain population (e.g., a country's total student population) or subpopulation (e.g., male or female students). However, PVs do not allow researchers to reliably assess each individual's characteristics.⁶⁷

The substantive-model-compatible sequential modeling approach

ELSAs may also contain planned, systematically missing values for which PVs are not provided in public use files or are only available for a subsample of participants (e.g., mathematics achievement in PISA 2000). Further, many items or scales in ELSAs have unplanned, sporadically missing values (e.g., the SES measure that we applied in the empirical examples). There is on-going research on how to best handle such missing data in IPD meta-analyses.^{60,77,78} On the basis of the logic of the two-stage approach, we suggest that missing data be imputed separately for each student sample at the participant level (see Figure 2). In using this approach, it is possible to impute (a) sporadically and (b) planned systematically missing data that occur within a certain sample in Stage 1. This imputation strategy is well-suited for examining moderating effects at the participant level,⁷⁹ and it does not require any assumptions about how the effect sizes vary across samples.^{77(pp509,510)} Further, this approach draws on the strength of Big Data analytic strategies⁴² to avoid plausible convergence problems when using a large volume of IPD demonstrating a complex hierarchical data structure (see Figure 2).

However, because missing values are imputed separately for each sample, in this approach, strength in the imputation process cannot be borrowed from the results obtained from other samples to impute unplanned systematically missing values that are missing for whole samples.^{60,77,78,80–82} For example, it would not be possible to impute the unplanned systematically missing data on students' reading achievement for the United States in PISA 2006 (see Section 4.4.1). Further, it is not possible to incorporate higher level covariates at the cycle or country level into the imputation model (see Figure 2). Hence, the imputation strategy that we present here may somewhat underestimate moderating effects at these levels in meta-regression models.^{60,77,78,80–82}

To implement the imputation strategy separately for each sample, we suggest that researchers apply the

substantive-model-compatible sequential modeling (SMC-SM) approach for two reasons.^{79,83} First, the SMC-SM approach is a member of the larger family of multiple imputation procedures, which all draw on an imputation model to replace missing values with imputed values to provide complete data. Because the true values of the missing data are unknown, the imputation model uses the observed data (from participants for whom data are not missing) to create a predictive distribution from which the complete data are sampled (for participants for whom data are missing). More specifically, by using a Markov chain Monte Carlo algorithm, the predictive distribution is applied to impute the missing values in the original data M times to obtain M complete data sets.⁷⁹ Each of these M imputed data sets is then used to estimate an effect size with a certain target model. The resulting variability in effect sizes across multiply imputed data sets reflects the uncertainty of the imputation process.^{62,84} Importantly, with these multiply imputed, complete data, it is possible to take full advantage of all standard statistical tools that take the methodological features of ELSAs into account when estimating effect sizes (e.g., applying sample weights) and their sampling variances (e.g., replication techniques) in Stage 1 of an IPD meta-analysis.

Second, it is important that the imputation model that is used to generate the imputed values is at least as general as the target model that is used to estimate the effect sizes. For example, if the target model includes interaction effects (e.g., a linear regression model with an interaction term), the imputation model needs to take into account these interaction effects. The imputation model of the SMC-SM approach is therefore specified such, that a joint predictive distribution is created that is based on a sequence of conditional models. Importantly, the target model (or a more general model) is included in this sequence. This guarantees that the imputation model is compatible with the target model (SMC).⁸⁵ One particular advantage of the SMC-SM approach is that it can accommodate normally distributed and skewed continuous and binary predictor variables as well as the multilevel structure of the data in these conditional models. Hence, it is well suited for imputing missing data to estimate effect sizes in ELSAs with clustered data (see Section 5.1.2) that draw on standard regression models (e.g., to estimate mean differences and bivariate regression relations) but also regression models involving moderating effects at the participant level (with highly skewed predictor variables).

As noted above, the imputation model is needed to impute the missing values in the original data M times to obtain M complete data sets. One general rule of thumb is that the number of imputed data sets M should be at least as large as the percentage of individuals in a sample

with one or more (sporadically) missing values.^{86(p388)} However, because multiple imputations can be computationally intensive, this rule of thumb might not be feasible with a large proportion of missing values. Often, a smaller number of imputations with a reasonable number of iterations may also be sufficient for the imputation algorithm to converge. Furthermore, it is recommended that auxiliary variables be included in the imputation model.^{79,87} Auxiliary variables are not part of the target model itself but are related to the propensity of missing data and to the variables with missing values themselves. Using auxiliary variables may (a) enhance the plausibility of the MAR assumption⁷⁹ and (b) substantially reduce bias in the parameter estimates.⁸⁷ Finally, it is strongly recommended that researchers check the quality of the imputed data by examining imputation diagnostics.⁶⁴

Applying PVs, multiple imputations, and nested multiple imputations to estimate effect sizes and their sampling variances

All multiple imputation approaches involving PVs and the SMC-SM approach comprise three steps to handle missing values. In Step 1, missing values are imputed by using some imputation model to obtain N PVs or M multiply imputed data sets. In Step 2, once the PVs or the multiply imputed data are available, the meta-analyst can apply the target model to estimate the target effect size T (by using the final weights) and its sampling variance $v(T)$ (e.g., by using JK2 or Fay BRR) from each PV or each imputed data set (e.g., when an ELSA provides 10 PVs, the target effect size is estimated 10 times). In Step 3, the average effect size and its sampling variance are computed by pooling the effect size estimates obtained in the second step. The uncertainty that results from imputing unknown missing values induces some additional variability to the results. In addition to the sampling variance for each PV or each multiply imputed data set, there is between-imputation variance of effect sizes around the average effect size. The total sampling variability for the average effect size is then computed as the sum of the average sampling variances and the between-imputation variance (see OSM.4 and OSM.5 for details). In accordance with Burgess et al.'s⁸⁸ recommendations, the average effect size (and its sampling variance) is then used in Stage 2 of the IPD meta-analysis.

As noted above, key outcome variables in ELSAs (e.g., achievement) are often represented as a set of N PVs rather than a single variable. When using PVs, the SMC-SM approach is extended as a nested multiple imputation approach.⁸⁹ Effect size calculation with nested multiply imputed data also implies three steps. First, for each PV, a set of M imputed data sets is obtained by using the SMC-SM approach as described

above. Second, the target model is run on all M imputed data sets for each of the N PVs, resulting in $N \cdot M$ effect size estimates and sampling variances (e.g., imputing a data set $M = 10$ times that provides $N = 10$ PVs would yield 100 effect size estimates and sampling variances). Third, the average effect size and its sampling variance is pooled across all $N \cdot M$ effect size estimates (see OSM.5 for details).

Finally, it is important to note that using a “shortcut” where participants’ average PV or average multiply imputed value is used rather than conducting the analyses separately for each PV or multiply imputed data set will lead to biased results and statistical inferences. The shortcut will underestimate the heterogeneity of the distribution (e.g., the distribution of students’ achievement) and the sampling variance of the resulting effect size estimate.⁶⁷

5.1.5 | Empirical examples: Estimation of effect sizes and their sampling variances

Research questions

We illustrate the potential of IPD meta-analyses of ELSAs by taking advantage of international PISA data (see Figure 1). In doing so, we demonstrate how to estimate the kind of effect sizes that are typically applied in meta-analyses.^{22,43} First, we estimate standardized mean differences to depict average gender differences in reading achievement. This analysis expands on previous research by Nowell and Hedges⁶ and Reilly et al.⁷ that also used meta-analytic models to integrate gender differences in reading achievement as observed in ELSAs from the United States. Second, we estimate correlations to depict how reading achievement and students’ SES are related. This analysis extends Sirin’s⁸ meta-analysis, which examined the association between achievement and SES in the United States. Third, we illustrate a key strength of IPD meta-analysis by analyzing the heterogeneity in effect sizes at the participant level. To this end, we examine how the magnitude of gender differences in reading achievement is moderated by students’ SES. This analysis connects to current quantitative intersectionality research that posits that each individual is simultaneously influenced by multiple social identities (e.g., gender and SES) that interact in explaining educational outcomes.¹⁴

Samples and measures

In our applications, we drew on international IPD from PISA (2000–2018) with over 400 independent student samples from 92 countries or economic regions (total $N = 2,970,892$ students; see Figure 1). To estimate the various effect sizes, we used the PISA reading

achievement scores that were provided as PVs. Further, we used information on students' gender as obtained from student questionnaires (0/1 = male/female students). Finally, as a measure of students' SES, we applied parents' highest occupational status (HISEI), which corresponds to the higher occupational status of either parent or to the only available parent.¹⁵ Higher HISEI values indicate higher SES.

Handling missing values

Data OSM.6 provides a detailed account of how we handled the missing data. We briefly summarize the key points here. Due to (unplanned) systematically missing values on reading achievement or the SES index, we had usable IPD from (a) 424 samples to estimate gender differences and from (b) 422 samples to estimate the relationship between reading achievement and students' SES and the interaction between gender and SES. Given that we used PVs for reading achievement as the outcome, we used a nested multiple imputation strategy combined with the SMC-SM approach and imputed each of the 422 student sample data sets $M = 10$ times separately for each PV using the raw data provided in the public use files. To this end, we used the R package "mdmb" (version 1.5.8).⁹⁰ The imputation model for the SMC-SM approach comprised a sequence of conditional models that took into account the distribution of the variables to be imputed (e.g., continuous distribution for reading achievement) and the functional form for how these variables are interrelated (e.g., multilevel regression). For example, the conditional (multilevel regression) model of reading achievement contained gender, SES, the interaction between gender and SES, as well as mathematics achievement as predictors. Mathematics achievement was used as an auxiliary variable[‡] because of (a) the small, albeit consistent gender differences found in this achievement domain in many countries⁹ and (b) the consistent relationships found between students' SES and achievement.⁸ Finally, to account for the fact that students were nested within schools, we added a random intercept for schools to each conditional model.⁹¹ Of note, we did not impute missing values for gender because the mechanism underlying missing values on gender may be MNAR, thus reflecting that students' gender identity was not represented well by the binary response format used in the PISA student questionnaire.[§] In summary, the applied imputation model was compatible with all regression model sets that we used to estimate the target effect sizes.

Model sets to estimate effect sizes and their sampling variances

To estimate the various effect sizes and their sampling variances, we applied three sets of linear regression

models—Model Sets 1, 2, and 3—to the IPD of each student sample (see, e.g., Figure 4). Further details on the specification and interpretation of these models are presented in OSM.6. We highlight the most important aspects of these analyses here. First, in all model sets, the regression coefficients were estimated by using the final weights, which implied that students with larger weights had a larger influence on the values of the regression parameters.^{27,28} Further, the sampling variances of all effect sizes were estimated by using Fay BRR.⁶⁶ Second, we z -standardized the outcome variable (i.e., reading achievement) and the continuous predictor variable (i.e., the SES measure *HISEI*) with $M = 0$ and $SD = 1$ for each multiply imputed data set as obtained for a certain student sample.⁹² We did not z -standardize the indicator variable for gender. The standardized regression coefficient for gender (β_{Gender} as obtained from Model Set 1) therefore depicts gender differences as a standardized mean difference (equivalent to d), with positive values indicating that female students outperformed male students. Third, using this z -standardization procedure yielded a bivariate regression coefficient (β_{SES} as obtained from Model Set 2) that was identical in size and interpretation to the correlation between reading achievement and the SES index.⁹² Fourth, to examine how gender differences in reading achievement were moderated by students' SES, we specified Model Set 3, which included gender, the z -standardized SES index, and a multiplicative term between gender and the z -standardized SES index to represent their interaction. The standardized regression coefficient $\beta_{\text{Gender} \times \text{SES}}$ indicates how gender differences in reading achievement (in terms of d) change when students' SES increases by 1 (cycle- and country-specific) SD (see Figure 4d). Finally, in accordance with Burgess et al.'s⁸⁸ recommendations, we computed the (pooled) average effect size (e.g., $\bar{\beta}_{\text{Gender}}$) and its sampling variance (e.g., $v(\bar{\beta}_{\text{Gender}})$) for each model set across the multiply imputed data sets as obtained for a certain student sample (using the nested multiple imputation formula; see OSM.5). These average effect sizes were then used in Stage 2 of the IPD meta-analysis.

Software

There are several software packages that can be used to take the methodological features of ELSAs into account when estimating effect sizes and their sampling variances.⁹³ We used the "survey" package (version 4.1.1) that is implemented in the free software environment R (version 4.1.1)⁹⁴ because it is well-documented.²⁸ Further, it allows researchers to specify many different generalized linear models (e.g., linear, logistic, Poisson, or quantile regression) to estimate effect sizes and their sampling errors for a large variety of stratified multistage sampling designs. Finally, using R allowed us to repeat the data

analytic steps as required in Stage 1 for the IPD of each sample from an ELSA, and it provided a versatile interface to the applied meta-analytic models in Stage 2 of the IPD meta-analysis.

5.2 | Stage 2: Integrating effect sizes and analyzing their heterogeneity with meta-analytic models

5.2.1 | Choosing a meta-analytic model

When integrating effect sizes resulting from descriptive analyses with ELSAs in an IPD meta-analysis, researchers should choose a meta-analytic model that provides a good fit to (a) the inference population and (b) the dependent structure and distributional form of the observed effect sizes.

Inference populations in the fixed-effect model and random-effects model

There are two popular statistical models—the fixed-effect model (also known as the common-effect model) and the random-effects model—which differ in their respective inference populations.^{36,44} If the goal is to make inferences about the effect size parameters only in the observed set of ELSA samples, the fixed-effect model can be applied. However, in this case, the meta-analyst assumes that all observed effect sizes share a common population value for the true effect size.^{36,44} Observed variations between effect sizes are only expected due to sampling errors as estimated by the sampling variances of the observed effect sizes.^{36,44} As for ELSAs, this set of assumptions seems most plausible when the effect sizes to be integrated stem from direct or very close replication studies using IPD from the same target population.⁴⁴

Given these restrictions of the fixed-effect model, the random-effects model will be a more plausible match for most meta-analyses of descriptive analyses with IPD from ELSAs.⁴⁴ Specifically, the random-effects model allows researchers to make inferences about a population of studies from which the observed ELSA samples are considered to be a random (representative) sample.^{36,44} The random-effects model therefore assumes that there is a distribution of true effect sizes θ rather than a single true common effect size. This assumption seems plausible, for example, when the effect sizes to be integrated stem from several cycles of an ELSA that draw on independent student samples that were observed at different points in historical time (e.g., several PISA cycles). The random-effects model can also be applied to integrate the results as obtained from related ELSAs (e.g., PISA and PIRLS)

that draw on different independent student samples and that also differ in the materials that were used.^{44(p107)}

Dependencies among observed effect sizes and their distributional form

Statistical inferences from parameters in meta-analytic models are based on some assumptions about the dependencies among observed effect sizes and their distributional form.^{95,96} Standard meta-analytic models (e.g., the two-level random-effects model) assume independent effect sizes. Is this a reasonable assumption for IPD meta-analyses of ELSAs? The two-level model can be used, for example, to meta-analyze the descriptive analyses for a certain outcome as obtained from several independent samples in the same country⁷ or from independent student samples in a single ELSA cycle,⁵ for example, when a certain target variable is assessed in this cycle only. In these examples, the effect sizes are based on independent samples, and hence, the assumptions of the two-level random-effects model apply. OSM.7 presents an application of the two-level random-effects model for independent student samples participating in a single PISA cycle (i.e., PISA 2018).

However, in many applications of IPD meta-analyses of ELSAs, the assumption of independent effect sizes may be unrealistic. In particular, in international ELSAs (e.g., PISA, TIMSS, PIRLS), independent student samples participated in a certain cycle of an ELSA. Several such samples were nested within countries (see Figure 2a). The effect sizes within the same country can be expected to be more similar to each other (e.g., because individuals are educated in the same system) than to effect sizes from other countries. Modeling these dependencies requires researchers to account for the hierarchical structure of the observed effect sizes in the meta-analytic model. To this end, a model-based approach drawing on a three-level meta-analytic model can be used.^{35,96,97}

Importantly, following Pustejovsky and Tipton,⁹⁸ we recommend combining the model-based approach with robust variance estimation (RVE)⁹⁵ to adjust the sampling variances and the 95% confidence intervals (95% CIs) for the average true effect size and meta-regression coefficients.** Doing so safeguards statistical inferences against (a) violations of distributional assumptions, (b) small sample bias when integrating effect sizes from a small number of countries, (c) the misspecification of the structure of the dependence of observed effect sizes, and (d) the misspecification of the applied meta-analytic model, for example, when using a mixed-effects meta-regression model rather than a random-slope meta-regression model.^{36,98} Moreover, using a model-based approach in combination with RVE is a reasonable

strategy when a meta-analyst cannot fully accommodate the complex structure of the dependence of observed effect sizes in a model-based approach (e.g., when combining an IPD meta-analysis of ELSAs with an AD meta-analysis). Finally, the sampling distribution for some effect sizes (e.g., correlations or standardized regression coefficients, such as β_{SES}) might not be well approximated by a normal distribution unless the sample size is very large.¹⁰³ However, the standard assumption in meta-analytic models is that the observed effect sizes are normally distributed (see Sections 5.2.2 and 5.2.4). In such situations, RVE helps to reduce potential bias in the SEs and confidence limits of the average true effect size and meta-regression coefficients because it makes no assumptions about the specific form of the sampling distributions of the observed effect sizes.^{95,104}

5.2.2 | Three-level random-effects model

Model specification

Accounting for the hierarchical dependency of effect sizes as found in many ELSAs requires a three-level random-effects model³⁵ that involves three stages of sampling (see Figure 2a). Given that sporadically missing data are the rule (and not the exception) in ELSAs, we illustrate this model for the observed, pooled effect sizes \bar{T}_{ct} and their sampling variances $v(\bar{T}_{ct})$ that were obtained (e.g., by means of [nested] multiple imputation) for a certain sample in country c in cycle t . The first stage (Level 1) of the sampling process assumes that the observed effect sizes \bar{T}_{ct} estimate a true effect size ϑ_{ct} with some estimation error e_{ct} .

$$\bar{T}_{ct} = \vartheta_{ct} + e_{ct} \quad (1)$$

The second stage (Level 2) of the sampling process assumes that the true effect sizes ϑ_{ct} may vary around the country-specific true mean effect size β_{0c} . The deviation of a certain true effect size ϑ_{ct} is depicted by the random effect u_{ct} .

$$\vartheta_{ct} = \beta_{0c} + u_{ct} \quad (2)$$

The third stage (Level 3) of the sampling process assumes that the true country-specific mean effect sizes β_{0c} may vary around the (grand) true mean effect size γ_{00} . The deviation for a certain true country-specific mean effect size from γ_{00} is depicted by the random effect u_{0c} .

$$\beta_{0c} = \gamma_{00} + u_{0c} \quad (3)$$

When we combine Equations (1)–(3), the three-level random-effects model can be written as³⁵:

$$\bar{T}_{ct} = \gamma_{00} + u_{0c} + u_{ct} + e_{ct}. \quad (4)$$

In the three-level random-effects model, the sampling errors e_{ct} and the random effects u_{ct} and u_{0c} are assumed to be uncorrelated within and across levels and to follow normal distributions with means of zero and variances $v(\bar{T}_{ct})$, $\tau_{\text{Level } 2}^2$, and $\tau_{\text{Level } 3}^2$, respectively. Thus, three sources of variation affect the distribution of observed effect sizes: the variance of the sampling errors $v(\bar{T}_{ct})$, the variance of the true sample-specific effect sizes around the mean true effect size within countries $\tau_{\text{Level } 2}^2$ (which is assumed to be identical for all countries), and the variance of the country-specific true mean effect sizes around the average true effect size $\tau_{\text{Level } 3}^2$. The total variance of the true effect sizes is therefore $\tau^2 = \tau_{\text{Level } 2}^2 + \tau_{\text{Level } 3}^2$. When estimating the meta-analytic parameters for the true effect sizes, an inverse-variance weighting scheme is applied. The weights are based on all three sources of variance, which also implies that effect sizes that were estimated more precisely (as reflected by $v(\bar{T}_{ct})$) receive greater weight.³⁵ The computation of the weights requires estimates of the variances $\tau_{\text{Level } 2}^2$ and $\tau_{\text{Level } 3}^2$.³⁵ Veroniki et al.¹⁰⁵ recommended the restricted maximum likelihood estimator (REML) to estimate these variances. Reliable 95% CIs can then be obtained (a) by using the profile likelihood method for variances or SDs¹⁰⁵ and (b) by drawing on the estimates of $\tau_{\text{Level } 2}^2$ and $\tau_{\text{Level } 3}^2$ with RVE for the average true effect.⁹⁸ Model convergence can be assessed with profile likelihood plots (see OSM.8).¹⁰⁶

Evaluation of the heterogeneity of effect sizes

The three-level random-effects model provides several key meta-analytic parameters: the average true effect size and the variance estimates for describing the heterogeneity of the true effect sizes in total as well as at Levels 2 and 3. There are several approaches that can be used to assess the heterogeneity of effect sizes.¹⁰⁷ First, the Q test statistic is computed by summing the squared deviations of each individual effect size estimate from the average true effect estimate where individual effect sizes are weighted by their sampling variance. A statistically significant value of Q is typically taken to indicate effect size heterogeneity.¹⁰⁷ Second, the I^2 statistic provides information about the proportion of observed heterogeneity that is real and not due to random noise and has a range of 0%–100%.¹⁰⁷ The Cochrane handbook for research syntheses¹⁰⁸ offers guidelines that characterize I^2 values

falling in the intervals $30\% \leq I^2 \leq 60\%$, $50\% \leq I^2 \leq 90\%$, and $75\% \leq I^2 \leq 100\%$ as representing moderate, substantial, and considerable heterogeneity. Notably, any evaluation of the size of I^2 should also take into account the sign and size of the effect sizes as well as the precision with which the heterogeneity could be estimated (e.g., the p value from the Q test statistic). Moreover, when using three-level random-effects models, the I^2 statistic can be computed for each level separately with $I^2_{\text{Level } 2}$ and $I^2_{\text{Level } 3}$ providing information about the proportions of the total variance of the effect sizes that can be explained by true effect size heterogeneity at Level 2 (e.g., between samples within countries) or Level 3 (e.g., between countries).⁹⁷ Third, the 95% prediction interval (95% PI) takes into account τ^2 as well as the uncertainty (i.e., the sampling variance) in estimating the average true effect size.¹⁰⁷ In doing so, the 95% PI gives a hint about the variation in true effect sizes by providing a plausible range of values in which the true effect sizes of about 95% of all relevant populations will fall.

5.2.3 | Applications of the three-level random-effects model

We applied three three-level random-effects models to meta-analyze the (pooled) effect sizes $\bar{\beta}_{\text{Gender}}$, $\bar{\beta}_{\text{SES}}$, and $\bar{\beta}_{\text{Gender} \times \text{SES}}$ as obtained from Model Sets 1, 2, and 3 (see Figure 5) by using the R package “metafor” (version 3.0.2).¹⁰⁶ We applied “metafor” because of its comprehensive modeling capabilities.¹⁰⁹ Further, we implemented RVE by using the “clubSandwich” package (version 0.5.3).¹¹⁰

Meta-analyzing gender differences in reading achievement

Our results (Table 1) indicated that, on average, female students outperformed male students in reading with an average estimated (true) effect size $\gamma_{00}(\bar{\beta}_{\text{Gender}}) = 0.37$ SD units across 92 countries (95% CI [0.35, 0.39]). This gender difference in favor of 15-year-old female students is slightly larger than the gender gap ($d = 0.30$) reported for eighth-grade students in the United States.⁷ Assessing the heterogeneity of effect sizes showed that gender differences in reading achievement varied substantially, $Q(df = 423) = 5878.4$, $p < 0.001$; $I^2 = 93.2\%$; 95% PI [0.13, 0.62], with larger variation observed between ($\tau_{\text{Level } 3} = 0.10$) than within countries ($\tau_{\text{Level } 2} = 0.08$). According to the benchmark ranges that Hyde¹¹¹ proposed for evaluating effect sizes for gender differences (negligible: $0.00 < |d| \leq 0.10$, small: $0.10 < |d| \leq 0.35$, moderate: $0.35 < |d| \leq 0.65$, large: $0.65 < |d| \leq 1.00$, and very large:

$|d| > 1.00$), the average gender difference in reading achievement can be considered moderate in size. The 95% PI (Figure 5a) indicates that about half of the gender differences will likely be small, whereas the other half will probably be moderate in size.

Meta-analyzing the relationship between reading achievement and students' SES

Our results (Table 1) showed that the average (true) standardized regression coefficient between reading achievement and students' SES was $\gamma_{00}(\bar{\beta}_{\text{SES}}) = 0.30$ (95% CI [0.29, 0.31]) across 96 countries. That is, on average, an increase of 1 SD in the SES index was associated with an increase of about 0.30 SD units in reading achievement. A similar, yet slightly lower, relationship between reading achievement and student SES was reported in Sirin's meta-analysis⁸ for the United States ($r = 0.27$). Assessing the heterogeneity of effect sizes showed that the relationship between reading achievement and students' SES varied substantially, $Q(df = 422) = 4917.3$, $p < 0.001$; $I^2 = 91.0\%$, with considerably larger variation observed between ($\tau_{\text{Level } 3} = 0.06$) than within countries ($\tau_{\text{Level } 2} = 0.03$). The 95% PI ranged from 0.17 to 0.43 (Figure 5b), which suggests that social inequality in reading achievement that is related to students' SES can be expected in most countries and PISA cycles.

Meta-analyzing the interaction between gender differences and students' SES

Our results (Table 1) showed that, on average, gender differences in reading achievement were slightly moderated by students' SES with $\gamma_{00}(\bar{\beta}_{\text{Gender} \times \text{SES}}) = -0.018$ (95% CI [-0.023, -0.013]). The interaction implies, for example, that if students' SES is 2 SD units above the mean, we would expect the gender gap in reading achievement to be, on average, 0.036 SD units (i.e., about 10%) smaller than the average gender difference. Assessing the heterogeneity of effect sizes showed that the interaction terms $\bar{\beta}_{\text{Gender} \times \text{SES}}$ varied significantly but only to a moderate degree, $Q(df = 422) = 481$, $p = 0.023$; $I^2 = 41.1\%$. All variation in the true effect sizes was observed between countries ($\tau_{\text{Level } 3} = 0.02$). The 95% PI ranged from -0.05 to 0.02, emphasizing that true effect sizes can be expected to be heterogeneous and not to have the same sign or size in all student populations (Figure 5c; see also, e.g., Figure 4d). Taken together, these findings provide some support for an intersectional effect¹⁴ by showing that students' gender and their SES interact in explaining achievement differences in reading achievement. However, our results also suggest that the intersectional effect varies across countries. Of note, further meta-regression analyses (see OSM.11) suggested

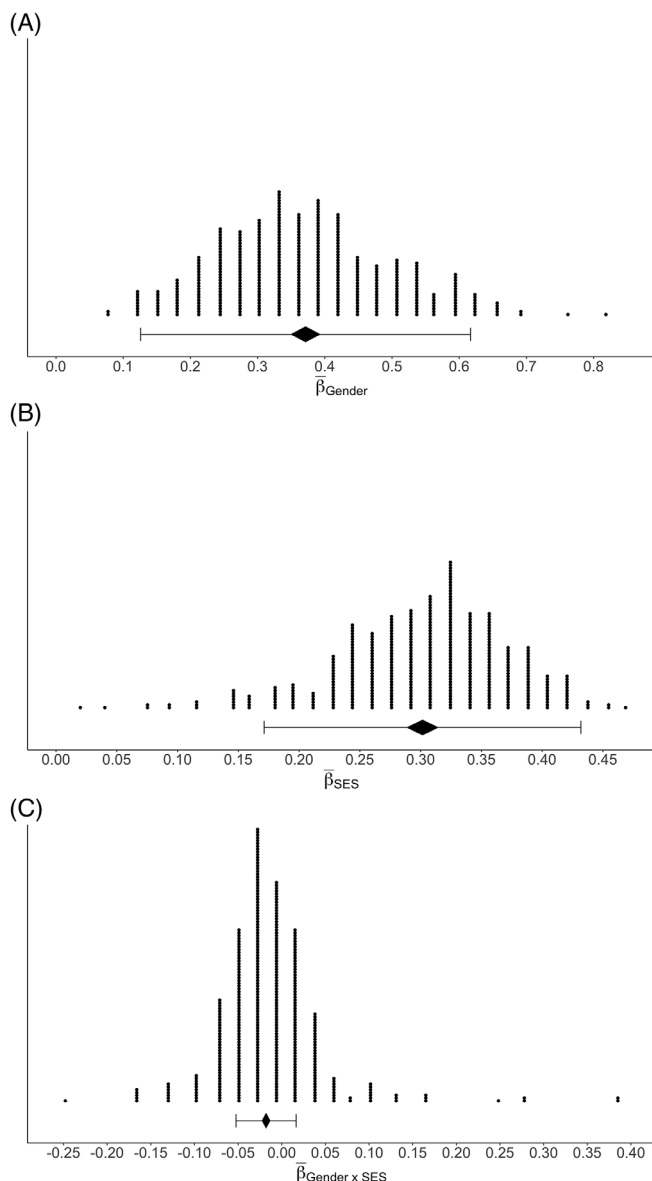


FIGURE 5 Distribution of pooled effect sizes: (a) Gender differences in reading achievement ($\bar{\beta}_{\text{Gender}}$ as obtained from Model Set 1), (b) relationship between reading achievement and students' socioeconomic status (SES) ($\bar{\beta}_{\text{SES}}$ as obtained from Model Set 2), and (c) relationship between reading achievement and the interaction between gender and SES ($\bar{\beta}_{\text{Gender} \times \text{SES}}$ as obtained from Model Set 3). The dots represent the observed, pooled effect sizes; the diamond the true average effect size (vertical line through the vertical points of the diamond) and the 95% CIs (horizontal points of the diamond); and the error bars the 95% PIs as obtained from three-level random-effects models

that intersectionality effects may depend on the test mode. In particular, intersectionality effects were (on average) about -0.02 when a paper-pencil test of reading achievement was used and about zero when a computer-based assessment was used.

5.2.4 | Meta-regression

In the previous sections, we showed how to achieve one key goal of IPD meta-analyses of ELSAs: assessing the mean and the heterogeneity of the distribution of (true) effect sizes. For example, we observed substantial heterogeneity in gender differences in reading achievement within countries over time (i.e., across PISA cycles) and between countries (Table 1 and Figures 5a and 6). A second key goal of IPD meta-analyses of ELSAs is therefore to better understand and explain this heterogeneity. To this end, we illustrate how the three-level random-effects model can be expanded into a mixed-effects meta-regression model that is typically applied to examine moderating effects.^{36,45} In particular, we used this model to examine moderator variables that had the potential to explain the heterogeneity in gender differences in reading achievement (in terms of $\bar{\beta}_{\text{Gender}}$; Model Set 1). OSM.10 and OSM.11 provide results on meta-regression models for $\bar{\beta}_{\text{SES}}$ and $\bar{\beta}_{\text{Gender} \times \text{SES}}$ (i.e., Model Sets 2 and 3).

Expanding the three-level random-effects model into a meta-regression model

Drawing on IPD from ELSAs from the United States, previous meta-analytic research found no significant change in gender differences in reading achievement between 1960 and 1994⁶ or between 1988 and 2015.⁷ To investigate the generalizability of these results to 92 countries, we analyzed the change in gender differences in reading achievement over time (Time_{ct} was specified as a continuous variable at Level 2 and coded 0/3/6 etc. indicating the 2000/2003/2006 etc. cycles of PISA). Further, reading achievement was measured with paper-pencil tests in PISA 2000 to 2012, whereas computer-based assessments were applied in PISA 2015 and 2018. Meta-analytic research showed that the test mode (paper-pencil test vs. computer-based assessment) has no significant effect on reading achievement test performance.¹¹³ Accordingly, gender differences in reading achievement are not expected to change when the test mode changes. To test this hypothesis, we examined the extent to which gender differences were moderated by test administration mode (Mode_{ct} was located at Level 2 and coded 0 = paper-pencil test if $\text{Time}_{ct} \leq 12$, 1 = computer-based assessment if $\text{Time}_{ct} > 12$). Finally, a key goal of educational policies in OECD countries is to reduce educational disparities between male and female students.^{112(p142)} We therefore examined the extent to which (a) the magnitude of gender differences and (b) the development of gender differences over time were moderated by a country's OECD membership (OECD_c was located at Level 3 and coded 0 = other countries, 1 = OECD country). Notably, the assumption that the development of gender differences

TABLE 1 Three-level meta-analytic models to integrate reading achievement’s relationships with gender (Model Set 1) and students’ SES (Model Set 2) and their interaction (Model Set 3)

Meta-analytic statistics	Model Set 1 $\bar{\beta}_{\text{Gender}}$	Model Set 2 $\bar{\beta}_{\text{SES}}$	Model Set 3 $\bar{\beta}_{\text{Gender} \times \text{SES}}$
$N_{\text{countries}}$	92	92	92
k	424	422	422
k_{country}			
<i>Min</i>	1	1	1
<i>Mdn</i>	5	5	5
<i>Max</i>	7	7	7
γ_{00}	0.37	0.30	−0.018
[95% CI]	[0.35, 0.39]	[0.29, 0.31]	[−0.023, −0.013]
95% PI	[0.13, 0.62]	[0.17, 0.43]	[−0.052, 0.017]
Q			
value	5878.4	4917.3	481.0
<i>df</i>	423	421	421
<i>p</i>	<0.001	<0.001	0.023
τ^2	0.016	0.004	0.0003
[95% CI]	[0.012, 0.020]	[0.0033, 0.006]	[0.0002, 0.0005]
$\tau^2_{\text{Level 2}}$	0.006	0.0007	0.0000
[95% CI]	[0.005, 0.007]	[0.0005, 0.0009]	[0.0000, 0.0001]
$\tau^2_{\text{Level 3}}$	0.010	0.004	0.0003
[95% CI]	[0.007, 0.014]	[0.003, 0.005]	[0.0001, 0.0005]
τ	0.125	0.066	0.017
[95% CI]	[0.112, 0.141]	[0.058, 0.077]	[0.016, 0.023]
$\tau_{\text{Level 2}}$	0.075	0.026	0.000
[95% CI]	[0.068, 0.082]	[0.023, 0.030]	[0.000, 0.012]
$\tau_{\text{Level 3}}$	0.100	0.061	0.017
[95% CI]	[0.084, 0.120]	[0.052, 0.072]	[0.011, 0.023]
I^2	93.2%	91.0%	41.1%
$I^2_{\text{Level 2}}$	33.5%	13.9%	0.0%
$I^2_{\text{Level 3}}$	59.8%	77.1%	41.1%

Note: The table shows the meta-analytic results from integrating the standardized effect sizes that were obtained from using Model Sets 1, 2, and 3. The specification of the model sets is shown in Figure 4. $N_{\text{countries}}$ = number of countries (i.e., Level 3 units). k = total number of effect sizes. k_{country} = number of effect sizes per country (i.e., Level 2 units). γ_{00} = estimated average true effect size with 95% confidence interval. $\tau^2, \tau^2_{\text{Level 2}}, \tau^2_{\text{Level 3}}$ = estimated variances of true effect sizes with 95% confidence intervals: total, within countries, between countries. $\tau, \tau_{\text{Level 2}}, \tau_{\text{Level 3}}$ = estimated SDs of true effect sizes with 95% confidence intervals: total, within countries, between countries. $I^2, I^2_{\text{Level 2}}, I^2_{\text{Level 3}}$ = proportion of variance due to true effect sizes to total variance: total, within countries, between countries.

may depend on OECD membership is depicted by a so-called cross-level interaction between $Time_{ct}$ (a Level 2 variable) and $OECD_c$ (a Level 3 variable).
The present moderator analyses exemplify a typical challenge that many meta-analysts have to face. Specifically, time and mode of test administration were confounded (i.e., $Mode_{ct} = 0$ if $Time_{ct} \leq 12$ and $Mode_{ct} = 1$ if $Time_{ct} > 12$). To account for the (potential)

confounding of moderator variables, it is recommended that meta-regression models that include multiple moderator variables be run.^{23,45,114} Following this advice, we specified a mixed-effects meta-regression model that included $Time_{ct}$ and $Mode_{ct}$ (both of which may vary within and between countries) as moderator variables to explain the heterogeneity in observed effect sizes at Level 2.³⁵

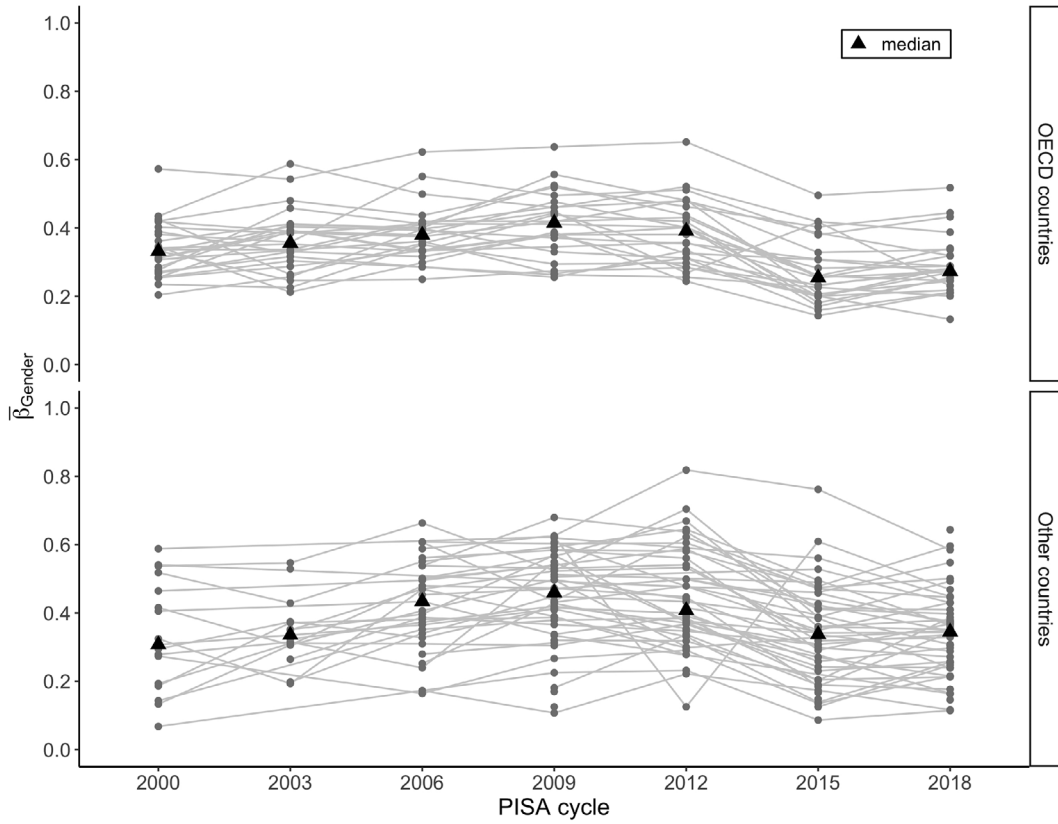


FIGURE 6 Country-specific development of gender differences in reading achievement ($\bar{\beta}_{\text{Gender}}$) across Programme for International Student Assessment (PISA) cycles for OECD and other countries. Each dot represents a pooled effect size ($\bar{\beta}_{\text{Gender},ct}$). The lines connect effect sizes as obtained for independent student samples participating in a certain cycle t within the same country c . OECD countries were defined as those countries that were members of the OECD as of the year 2000.^{112(p13)} The triangles indicate the median value of the observed effect sizes as obtained for a certain PISA cycle and the group of OECD or other countries.

$$\bar{\beta}_{\text{Gender},ct} = \beta_{0c} + \beta_{1c} \cdot \text{Time}_{ct} + \beta_{2c} \cdot \text{Mode}_{ct} + u_{ct} + e_{ct} \quad (5)$$

Further, $OECD_c$ (which may vary between but not within countries) was expected to moderate the magnitude of gender differences in reading achievement. To probe this prediction, $OECD_c$ was used as a moderator variable to explain the variability in β_{0c} among true effect sizes between countries.

$$\beta_{0c} = \gamma_{00} + \gamma_{01} \cdot OECD_c + u_{0c} \quad (6)$$

Of note, Equation (6) contains a random effect u_{0c} to allow true (average) gender differences in reading achievement to vary between countries while controlling for countries' OECD membership. Further, $OECD_c$ was also expected to moderate the development of gender differences in reading achievement across time as depicted by β_{1c} .

$$\beta_{1c} = \gamma_{10} + \gamma_{11} \cdot OECD_c \quad (7)$$

Because we specified a mixed-effects meta-regression model (rather than a random-slope model), Equation (7) contains no random effects when explaining β_{1c} .³⁵ This implies that β_{1c} is assumed to be either identical for all countries (i.e., $\gamma_{11} = 0$) or that β_{1c} can be fully explained by countries' OECD membership. Finally, we had no expectations that the relationship between $Mode_{ct}$ and the observed effect size would depend on moderator variables or that it might vary between countries. As we specified a mixed-effects model, β_{2c} was therefore assumed to be the same for all countries.

$$\beta_{2c} = \gamma_{20} \quad (8)$$

Combining Equations (5)–(8) yields a three-level mixed-effects meta-regression model.

$$\begin{aligned} \bar{\beta}_{\text{Gender},ct} = & \gamma_{00} + \gamma_{01} \cdot OECD_c + u_{0c} \\ & + (\gamma_{10} + \gamma_{11} \cdot OECD_c) \cdot \text{Time}_{ct} + \gamma_{20} \cdot \text{Mode}_{ct} + u_{ct} \\ & + e_{ct} \end{aligned} \quad (9)$$

This meta-regression model draws on the same set of assumptions about the error terms (e_{ct}) and the residual random effects (u_{0c} and u_{ct}) as the three-level random-effects model with the variance $\tau^2_{\text{Res.Level 2}}$ depicting (residual) heterogeneity within countries and $\tau^2_{\text{Res.Level 3}}$ between countries.³⁵ Pseudo R^2 s can be computed to quantify the proportion of variance accounted for at each level by the meta-regression model: $P^2_{\text{Level 2}}$ and $P^2_{\text{Level 3}}$ depict the proportion of the total variance in true effect sizes (as estimated by the three-level random-effects model without moderator variables, see Table 1) explained by the moderator variables within and between countries.³⁶

$$P^2_{\text{Level 2}} = (\hat{\tau}^2_{\text{Level 2}} - \hat{\tau}^2_{\text{Res.Level 2}}) / \hat{\tau}^2_{\text{Level 2}} \quad (10)$$

$$P^2_{\text{Level 3}} = (\hat{\tau}^2_{\text{Level 3}} - \hat{\tau}^2_{\text{Res.Level 3}}) / \hat{\tau}^2_{\text{Level 3}} \quad (11)$$

To estimate the parameters in the meta-regression model, we used the same methods as for the three-level random-effects model (i.e., REML in combination with RVE). The interpretation of the parameters in the meta-regression is equivalent to the regression parameters in multilevel models.⁵⁷ Thus, the intercept $\gamma_{00}(\bar{\beta}_{\text{Gender}}) = 0.37$ (95% CI [0.33, 0.42]) represents the estimated average (true) gender differences in reading achievement for the year 2000 cycle as measured for a paper-pencil test in countries that were not members of the OECD. Further, $\gamma_{10} = 0.007$ indicates that, on average, gender differences in reading achievement have widened in favor of girls at 0.007 SD per year (95% CI [0.004, 0.010]) in non-OECD countries. This implies, for example, that gender differences in these countries have increased (on average) by about 0.07 SD per decade (i.e., 0.007 SD/year · 10 years). Moreover, $\gamma_{20} = -0.16$ (95% CI [-0.19, -0.13]) indicates that the gap in reading achievement between male and female students narrowed, on average, by 0.16 SD when a computer-based assessment was used instead of a paper-pencil test. Hence, this result suggests that the test mode may affect female and male students' performance on reading achievement tests differently. Finally, γ_{01} represents the mean differences in the magnitude of gender differences in reading achievement between OECD and other countries, whereas γ_{11} depicts how changes in gender differences over time depend on a country's OECD membership. Thus, $\gamma_{01} = -0.03$ (95% CI [-0.08, 0.02]) and $\gamma_{11} = -0.001$ (95% CI [-0.003, 0.001]) indicate that gender differences in reading achievement as well as their development over time did not differ substantially (or significantly) between OECD and non-OECD countries.

Further analyses of the residual variances showed that the meta-regression model did not fully account for the heterogeneity of effect sizes between countries ($\tau^2_{\text{Res.Level 3}} = 0.011$; 95% CI [0.008, 0.015]) or within countries ($\tau^2_{\text{Res.Level 2}} = 0.003$; 95% CI [0.002, 0.003]). This conclusion was corroborated by the Q statistic, which indicated a significant amount of residual heterogeneity, $Q(df = 419) = 4777.1$, $p < 0.001$. Finally, $P^2_{\text{Level 2}} = 0.55$ indicates that *Time*, *Mode*, and *OECD* jointly explained 55% of the variance in true effects sizes within countries, whereas $P^2_{\text{Level 3}} = -0.09$ (truncated to 0) indicates that OECD membership explained no variance in effect sizes between countries.

Potential issues in meta-regression analyses and suggested solutions

When using the meta-regression approach with hierarchically dependent effect sizes, several issues may occur. First, all problems that are well-documented for regression analyses in single or multilevel contexts may also occur in meta-regression models. These include outliers in the effect sizes that may exert a strong influence on the results or the multicollinearity of the moderator variables that may lead to unreliable sampling variances for the regression coefficients.^{36,115} Viechtbauer and Cheung¹¹⁶ developed diagnostic tools that can be used to examine outliers and influential effect sizes. Further, inspecting the correlations among the moderator variables and excluding some highly correlated predictors may help to mitigate the problem of multicollinearity.³⁶

Second, using a mixed-effects model to perform a meta-regression has the limitation that moderating effects cannot be specified to vary between countries. This limitation can be addressed by adding random effects to the corresponding meta-regression coefficients. For example, one could add a random effect to Equations (7)–(9) to depict the assumption that OECD membership does not fully explain between-country variation in how gender differences change over time. Doing so changes the mixed-effects model into a random-slope model.³⁵ Importantly, some scholars⁵⁷ recommend that random-slope models be used when investigating cross-level interactions (e.g., the cross-level interaction between *OECD* and *Time*) to allow for valid statistical inferences. When we introduced a random slope for *Time*, however, the corresponding variance component was estimated to be very close to zero (see OSM.9). This result suggests that the rate of change in gender differences in reading achievement did not differ between countries when controlling for a country's OECD membership. We therefore followed Raudenbush and Bryk's^{58(p28)} advice and based our conclusions on the more parsimonious mixed-effects model. Importantly, we embedded our analyses in

the RVE framework, which safeguards the statistical inferences for the meta-regression coefficients against model misspecification, for example, when using a mixed-effects rather than a random-slope meta-regression model.⁹⁸

Third, when conducting meta-regression analyses, (unplanned) systematically missing data may occur for the effect sizes (e.g., when a country did not participate in a certain PISA cycle) or the moderator variables (e.g., when information about a certain gender equality factor is not available for a certain country or cycle). In our empirical examples, the number of observed effect sizes varied between countries (see Table 1), which implies that some countries had systematically missing data on effect sizes in one or more PISA cycles. Notably, the meta-regression extension of the three-level random-effects model estimates the meta-regression coefficients and their sampling variances under the assumption that the underlying missingness mechanism is MAR or MCAR.⁵⁷ This is a plausible assumption for ELSAs when data cannot be suppressed, particularly for political reasons. For example, in PISA, countries can generally only withhold certain variables (but not all their data) when the information contained in these variables would threaten the confidentiality and anonymity of individuals (e.g., students or teachers).¹⁵ However, the technical standards in PISA do not allow variables to be withheld to avoid “inconvenient” results (e.g., large gender differences in reading achievement). Further, we observed no missing data on the moderator variables. When moderator variables are missing systematically (assuming MCAR or MAR) at the cycle or country level, (multilevel) multiple imputation approaches can be applied.^{60,77} To this end, the available information on (a) effect sizes and (b) moderator variables at the cycle and country level can be used. Notably, following Pigott and Polanin²³ we do not recommend imputing systematically missing effect sizes by using moderator variables at the cycle or country level (e.g., imputing missing effect sizes for gender differences in reading achievement when a country participated in some but not all PISA cycles or for countries that never participated in PISA).

6 | GENERAL DISCUSSION

6.1 | Research opportunities with IPD meta-analyses of ELSAs

IPD meta-analyses of studies with complex survey designs offer a powerful way to study the consistency, replicability, and generalizability of socially important or theoretically interesting phenomena and trends. More specifically, we see two major research opportunities for how IPD meta-analyses of ELSAs may substantially contribute to cumulative knowledge, particularly

in the behavioral and social sciences. First, one of the founding fathers of the meta-analytic enterprise, Gene Glass, envisioned that “[m]eta-analysis needs to be replaced by archives of raw data that permit the construction of complex data landscapes that depict the relationships among independent, dependent, and mediating variables”.^{117(p230)} The present paper directly responds to Glass’ call. IPD meta-analyses of ELSAs allow researchers to draw such “landscapes” for important policy-relevant subgroups, for example, by focusing descriptive analyses on socially disadvantaged minority students or students at the lowest or highest points on the achievement distribution.⁹ Further, by using a standardized protocol to manage and analyze the data, we illustrated one key strength of IPD meta-analyses, namely, the capability to investigate how individual characteristics at the participant level may moderate the magnitude of effect sizes (which is usually not possible with AD meta-analyses, e.g., due to a lack of reported information). Questions about how outcome relationships are moderated by individuals’ characteristics are highly relevant in the behavioral and social sciences, for example, in quantitative research on intersectionality¹⁴ where one of our empirical examples was located, in the context of aptitude treatment interactions to adapt teaching to learners’ characteristics,¹¹⁸ or research on individual differences in environmental sensitivity (e.g., parenting).¹¹⁹ To address such questions, the two-stage approach we presented can be expanded to meta-analyze effect sizes that are estimated with more complex models, such as generalized linear models,^{27,28} structural equation models,¹²⁰ or multilevel models.⁵⁶ However, given their nonexperimental design, the results of IPD meta-analyses of ELSAs, do not allow conclusions to be drawn about causality.¹²¹ Nevertheless, they may help improve our understanding of causal relationships (e.g., when a postulated cause does not show the expected effect) or point to important moderating effects that can be tested in subsequent experimental studies.¹

Second, the evidence base and consequently the reach of systematic reviews on many key research topics can be substantially enriched by combining traditional AD meta-analyses with IPD meta-analyses of ELSAs.¹²² Currently, researchers in the behavioral and social sciences use AD meta-analyses to statistically integrate findings from systematic reviews.¹²³ However, the way such systematic reviews are currently conducted, it seems unrealistic to expect them to cover the complete evidence base. For example, most systematic reviews are based on publications in English and are thereby missing publications in other languages.¹²⁴ Further, the search for unpublished (gray) literature is labor-intensive and must be stopped at some point in

light of limited time and financial resources.¹²⁵ Thus, it is highly plausible that there is empirical evidence that cannot be identified via standard search practices. For instance, the literature search for the systematic review of the AD meta-analysis by Barroso et al.¹²⁶ on the relationship between anxiety and achievement in mathematics identified effect sizes from 332 samples (total $N > 385,000$ individuals). Yet, the search missed 59 out of 65 samples from the PISA 2012 cycle and 2 samples from the PISA 2003 cycle, which contained data from about 465,000 students.¹²⁶ To sum up, adding evidence from IPD meta-analyses of ELSAs to AD meta-analyses will likely help to substantially improve the reliability of meta-analytic knowledge (because parameters can be estimated more precisely) and to draw more nuanced conclusions in systematic reviews in the behavioral and social sciences. For example, integrating the results obtained from IPD meta-analyses of ELSAs and AD meta-analyses creates the opportunity to probe whether the results are moderated by quality characteristics of the samples (e.g., convenience samples vs. probability samples in ELSAs).

6.2 | Where to look for further methodological guidance

To tailor the two-stage approach to the complexities of ELSAs, we focused on typical challenges that meta-analysts have to face in the parts of Stages 1 and 2 devoted to analyses. Setting this focus, our tutorial does not elaborate on the challenges that exist before and after an IPD meta-analysis of an ELSA is carried out. In particular, important further steps that need to be considered to carry out an IPD meta-analysis involve planning and preregistration,¹²⁷ computing a statistical power analysis,¹²⁸ and locating the studies from which the IPD are to be retrieved.^{9,125} Further, we did not discuss how to set the inclusion or exclusion criteria for selecting among the identified studies,⁴ to harmonize measures across studies,^{4,40} to assess the problem of publication bias,¹²⁹ or to apply the relevant reporting standards (PRISMA-IPD).¹³⁰ Moreover, further guidance on meta-analyzing a smaller number of effect sizes or countries in addition to the ones discussed here (RVE and model-based approaches)⁹⁸ can be found in Konstantopoulos³⁵ and Bender et al.⁹⁹

6.3 | Conclusion

The present tutorial introduced the two-stage IPD meta-analysis approach, which we tailored to the methodological challenges of studies with complex survey designs,

such as ELSAs. A two-stage IPD meta-analysis can be flexibly applied to tackle (typical) meta-analytic research objectives when synthesizing empirical evidence from descriptive analyses. The guidance offered in this paper can be helpful for synthesizing research evidence from complex surveys and panel studies in the behavioral, social, economic, educational and health sciences. IPD meta-analyses of studies with complex survey designs have (a) the potential to significantly enrich the extant body of knowledge with reliable and widely generalizable evidence for socially important or theoretically interesting phenomena and trends, and (b) open up new and unique research opportunities to synthesize evidence.

AUTHOR CONTRIBUTIONS

Martin Brunner was the project leader and conceptualized the tutorial. Martin Brunner, Lena Keller, and Oliver Lüdtke designed and conducted the statistical analyses. All authors contributed substantially to the final draft and have approved its submission.

ACKNOWLEDGMENT

Open Access funding enabled and organized by Projekt DEAL.

CONFLICTS OF INTEREST

The authors declare no conflicts of interests to disclose.


DATA AVAILABILITY STATEMENT


The R syntax for reproducing all results, figures, and tables as well as the data with effect sizes used in the present paper can be accessed via the Open Science Framework at <https://osf.io/wfd6p/>. We made a preprint of our tutorial available on psyarxiv at <https://psyarxiv.com/t79vk/>. We did not pre-register the analyses presented in this tutorial. The raw PISA data that we used in Stage 1 of the two-stage approach to IPD meta-analysis can be downloaded here: PISA 2000 (https://www.oecd.org/pisa/pisaproducts/intstud_math.zip; https://www.oecd.org/pisa/pisaproducts/intstud_read.zip), PISA 2003 (https://www.oecd.org/pisa/pisaproducts/INT_stui_2003_v2.zip), PISA 2006 (https://www.oecd.org/pisa/pisaproducts/INT_Stu06_Dec07.zip), PISA 2009 (https://www.oecd.org/pisa/pisaproducts/INT_STQ09_DEC11.zip), PISA 2012 (https://www.oecd.org/pisa/pisaproducts/INT_STU12_DEC03.zip), PISA 2015 (https://webfs.oecd.org/pisa/PUF_SPSS_COMBINED_CMB_STU_QQQ.zip), and PISA 2018 (https://webfs.oecd.org/pisa2018/SPSS_STU_QQQ.zip). Of note, the PISA 2000 to 2012 files are in an ASCII format for which SAS or SPSS control files are available at the cycle-specific data repositories (see <https://www.oecd.org/pisa/data/>). The effect sizes that we meta-analytically integrated in Stage 2 can be accessed via the Open Science Framework at <https://osf.io/wfd6p/>.

ORCID

Martin Brunner  <https://orcid.org/0000-0001-7182-5622>

Lena Keller  <https://orcid.org/0000-0002-3242-0208>

Sophie E. Stallasch  <https://orcid.org/0000-0002-4433-2600>

Julia Kretschmann  <https://orcid.org/0000-0002-2884-2061>

Andrea Hasl  <https://orcid.org/0000-0002-4945-2388>

Larry V. Hedges  <https://orcid.org/0000-0002-7531-0631>

ENDNOTES

* This also became evident from a literature review in which we searched the ERIC and PsycINFO databases for records that were relevant to four major ELSA programs: NAEP, PISA, TIMSS, and PIRLS. Our review showed that only 59 out of 15,923 identified records (0.4%) relevant to these ELSAs were classified as meta-analyses (for details, see OSM.1). Given that we focused on four major ELSA programs, this result provides a plausible estimate of the application of meta-analytic models to results from ELSAs.

† For some types of analyses, the final weights are linearly transformed to obtain different sets of weights.⁶⁷ Normalized weights (also called house weights) are transformed such that the sum of the weights is equal to the sample size. Normalized weights are often used in structural equation modeling.⁶⁸ Further, for some research questions, it is of interest to obtain pooled results by using a single data set encompassing the IPD from all samples (e.g., all country-specific samples) where each sample contributes equally to the pooled results. To this end, so-called Senate weights are computed by using a linear transformation of the final weights to ensure that the sum of the Senate weights equals a constant (say 500) in each sample.⁶⁷ Notably, these linear transformations of the final weights do not affect the values obtained for effect sizes as applied in the two-stage IPD meta-analyses when the IPD for each sample are analyzed separately.

‡ Mathematics achievement was used as an auxiliary variable for multiple imputations in PISA cycles 2003–2018. In PISA 2000, the proportion of missing values on the mathematics achievement variable was too high for it to serve as an auxiliary variable.

§ There is no consensus on how best to handle missing data on binary measures that are aimed at capturing multidimensional constructs, such as gender. Further, in PISA, the amount of missing data on gender was very small (the largest sample-specific percentage was 2.6% for Canada in PISA 2003). When there are so few missing data, the applied procedure for missing data on gender (i.e., listwise deletion) may be a reasonable option for handling missing data.^{62(p554)} An alternative approach is to use a multiple imputation strategy to impute missing data on gender, assuming that these missing data are MCAR or MAR. We compared the results obtained from using both missing data approaches in the three countries with the largest amount of missing data on gender. The two approaches yielded very similar results for the effect sizes and their SEs.

** RVE shares vital statistical characteristics with the widely recommended^{99,100} Hartung–Knapp–Sidik–Jonkman method^{101,102} for two-level random-effects meta-analysis.⁹⁵ However, RVE makes

less stringent assumptions about the weights for estimating the SEs for meta-analytic averages or meta-regression coefficients⁹⁵ and can easily be extended to safeguard statistical inferences for meta-analytic models with three levels or even more complex random-effects structures.⁹⁸

REFERENCES

- Loeb S, Dynarski S, McFarland D, Morris P, Reardon S, Reber S. *Descriptive Analysis in Education: A Guide for Researchers*. (NCEE 2017–4023). Institute of Education Sciences; 2017.
- Naemi B, Gonzales E, Bertling J, et al. Large-scale group score assessments: past, present, and future. In: Saklofske DH, Reynolds CR, Schwan VL, eds. *The Oxford Handbook of Child Psychological Assessment*. Oxford Library of Psychology, Oxford University Press; 2013. doi:10.1093/oxfordhb/9780199796304.013.0006
- Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;340:c221. doi:10.1136/bmj.c221
- Rao SR, Graubard BI, Schmid CH, et al. Meta-analysis of survey data: application to health services research. *Health Serv Outcomes Res Methodol*. 2008;8(2):98–114. doi:10.1007/s10742-008-0032-0
- Else-Quest NM, Hyde JS, Linn MC. Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychol Bull*. 2010;136:103–127.
- Nowell A, Hedges LV. Trends in gender differences in academic achievement from 1960 to 1994: an analysis of differences in mean, variance, and extreme scores. *Sex Roles*. 1998;39:21–43.
- Reilly D, Neumann DL, Andrews G. Gender differences in reading and writing achievement: evidence from the National Assessment of Educational Progress (NAEP). *Am Psychol*. 2019;74(4):445–458. doi:10.1037/amp0000356
- Sirin SR. Socioeconomic status and academic achievement: a meta-analytic review of research. *Rev Educ Res*. 2005;75:417–453.
- Keller LK, Preckel F, Eccles JS, Brunner M. Top-performing math students in 82 countries: an integrative data analysis of gender differences in achievement, achievement profiles, and achievement motivation. *J Educ Psychol*. 2022;114(5):966–991. doi:10.1037/edu0000685
- Keller LK, Preckel F, Brunner M. Nonlinear relations between achievement and academic self-concepts in elementary and secondary school: an integrative data analysis across 13 countries. *J Educ Psychol*. 2021;113(3):585–604. doi:10.1037/edu0000533
- Blömeke S, Nilsen T, Scherer R. School innovativeness is associated with enhanced teacher collaboration, innovative classroom practices, and job satisfaction. *J Educ Psychol*. 2021;113(8):1645–1667. doi:10.1037/edu0000668
- Guo J, Hu X, Marsh HW, Pekrun R. Relations of epistemic beliefs with motivation, achievement, and aspirations in science: generalizability across 72 societies. *J Educ Psychol*. 2022;114(4):734–751. doi:10.1037/edu0000660
- Marsh HW. Cross-cultural generalizability of year in school effects: negative effects of acceleration and positive effects of retention on academic self-concept. *J Educ Psychol*. 2016;108(2):256–273. doi:10.1037/edu0000059

14. Else-Quest NM, Hyde JS. Intersectionality in quantitative psychological research: II. Methods and techniques. *Psychol Women Q*. 2016;40(3):319-336. doi:[10.1177/0361684316647953](https://doi.org/10.1177/0361684316647953)
15. OECD. *PISA 2015*. Technical Report. OECD; 2017.
16. Rutkowski L, von Davier M, Rutkowski D, eds. *Handbook of International Large-Scale Assessment Background, Technical Issues, and Methods of Data Analysis*. CRC Press; 2014.
17. Findley MG, Kikuta K, Denly M. External validity. *Annu Rev Polit Sci*. 2021;24(1):365-393. doi:[10.1146/annurev-polisci-041719-102556](https://doi.org/10.1146/annurev-polisci-041719-102556)
18. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company; 2002.
19. Kelley K, Preacher KJ. On effect size. *Psychol Methods*. 2012;17(2):137-152. doi:[10.1037/a0028086](https://doi.org/10.1037/a0028086)
20. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. Vol 25303. National Academies Press; 2019. doi:[10.17226/25303](https://doi.org/10.17226/25303)
21. Simmonds M, Stewart G, Stewart L. A decade of individual participant data meta-analyses: a review of current practice. *Contemp Clin Trials*. 2015;45:76-83. doi:[10.1016/j.cct.2015.06.012](https://doi.org/10.1016/j.cct.2015.06.012)
22. Ahn S, Ames AJ, Myers ND. A review of meta-analyses in education: methodological strengths and weaknesses. *Rev Educ Res*. 2012;82(4):436-476. doi:[10.3102/0034654312458162](https://doi.org/10.3102/0034654312458162)
23. Pigott TD, Polanin JR. Methodological guidance paper: high-quality meta-analysis in a systematic review. *Rev Educ Res*. 2020;90(1):24-46. doi:[10.3102/0034654319877153](https://doi.org/10.3102/0034654319877153)
24. Tierney JF, Stewart LA, Clarke M. Chapter 26: Individual participant data. In: Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane; 2021 www.training.cochrane.org/handbook
25. Debray TPA, Moons KGM, van Valkenhoef G, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Res Synth Methods*. 2015;6(4):293-309. doi:[10.1002/jrsm.1160](https://doi.org/10.1002/jrsm.1160)
26. Gray H, Lyth A, McKenna C, Stothard S, Tymms P, Copping L. Sex differences in variability across nations in reading, mathematics and science: a meta-analytic extension of Baye and Monseur (2016). *Large-Scale Assess Educ*. 2019;7(1):2. doi:[10.1186/s40536-019-0070-9](https://doi.org/10.1186/s40536-019-0070-9)
27. Heeringa S, West BT, Berglund PA. *Applied Survey Data Analysis*. Taylor & Francis; 2010.
28. Lumley T. *Complex Surveys: A Guide to Analysis Using R*. John Wiley; 2010.
29. Valliant R, Dever JA, Kreuter F. *Practical Tools for Designing and Weighting Survey Samples*. 2nd ed. Springer; 2018.
30. Riley RD, Tierney JF, Stewart LA, eds. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Wiley; 2021.
31. Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Stat Med*. 2017;36(5):855-875. doi:[10.1002/sim.7141](https://doi.org/10.1002/sim.7141)
32. Morris TP, Fisher DJ, Kenward MG, Carpenter JR. Meta-analysis of Gaussian individual patient data: two-stage or not two-stage? *Stat Med*. 2018;37(9):1419-1438. doi:[10.1002/sim.7589](https://doi.org/10.1002/sim.7589)
33. Cooper HM, Hedges LV, Valentine JC. *The Handbook of Research Synthesis and Meta-Analysis*. 3rd ed. Russel Sage Foundation; 2019.
34. Siddaway AP, Wood AM, Hedges LV. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annu Rev Psychol*. 2019;70(1):747-770. doi:[10.1146/annurev-psych-010418-102803](https://doi.org/10.1146/annurev-psych-010418-102803)
35. Konstantopoulos S. Fixed effects and variance components estimation in three-level meta-analysis. *Res Synth Methods*. 2011;2(1):61-76. doi:[10.1002/jrsm.35](https://doi.org/10.1002/jrsm.35)
36. Konstantopoulos S, Hedges LV. Statistically analyzing effect sizes: fixed- and random-effects models. In: Cooper HM, Hedges LV, Valentine JC, eds. *Handbook of Research Synthesis and Meta-Analysis*. 3rd ed. Russell Sage Foundation; 2019: 245-280.
37. Riley RD, Burke DL, Morris T. One-stage versus two-stage approach to IPD meta-analysis: differences and recommendations. In: Riley RD, Tierney JF, Stewart LA, eds. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Wiley Series in Statistics in Practice. Wiley; 2021: 199-217.
38. Belias M, Rovers MM, Reitsma JB, Debray TPA, Int'Hout J. Statistical approaches to identify subgroups in meta-analysis of individual participant data: a simulation study. *BMC Med Res Methodol*. 2019;19(1):183. doi:[10.1186/s12874-019-0817-6](https://doi.org/10.1186/s12874-019-0817-6)
39. Nevitt SJ, Marson AG, Davie B, Reynolds S, Williams L, Smith CT. Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: systematic review. *BMJ*. 2017;357:j1390. doi:[10.1136/bmj.j1390](https://doi.org/10.1136/bmj.j1390)
40. Hussong AM, Curran PJ, Bauer DJ. Integrative data analysis in clinical psychology research. *Annu Rev Clin Psychol*. 2013;9(1):61-89. doi:[10.1146/annurev-clinpsy-050212-185522](https://doi.org/10.1146/annurev-clinpsy-050212-185522)
41. Siddique J, de Chavez PJ, Howe G, Cruden G, Brown CH. Limitations in using multiple imputation to harmonize individual participant data for meta-analysis. *Prev Sci*. 2018;19(1): 95-108. doi:[10.1007/s11121-017-0760-x](https://doi.org/10.1007/s11121-017-0760-x)
42. Cheung MWL, Jak S. Analyzing big data in psychology: a split/analyze/meta-analyze approach. *Front Psychol*. 2016;7. doi:[10.3389/fpsyg.2016.00738](https://doi.org/10.3389/fpsyg.2016.00738)
43. Hedges LV. What are effect sizes and why do we need them. *Child Dev Perspect*. 2008;2(3):167-171. doi:[10.1111/j.1750-8606.2008.00060.x](https://doi.org/10.1111/j.1750-8606.2008.00060.x)
44. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*. 2010;1(2):97-111. doi:[10.1002/jrsm.12](https://doi.org/10.1002/jrsm.12)
45. Tipton E, Pustejovsky JE, Ahmadi H. A history of meta-regression: technical, conceptual, and practical developments between 1974 and 2018. *Res Synth Methods*. 2019;10(2):161-179. doi:[10.1002/jrsm.1338](https://doi.org/10.1002/jrsm.1338)
46. Robinson WS. Ecological correlations and the behavior of individuals. *Am Sociol Rev*. 1950;15:351-357.
47. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11): 1559-1573. doi:[10.1002/sim.1187](https://doi.org/10.1002/sim.1187)
48. Riley RD, Debray TPA, Fisher D, et al. Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates: statistical recommendations for conduct and planning. *Stat Med*. 2020;39(15):2115-2137. doi:[10.1002/sim.8516](https://doi.org/10.1002/sim.8516)

49. Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *BMJ*. 2017;356: j573. doi:10.1136/bmj.j573
50. Enders CK, Tofighi D. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychol Methods*. 2007;12:121-138.
51. Cohen J, Cohen P, Aiken LS, West SG. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd ed. Lawrence Erlbaum Associates; 2003.
52. Dalal DK, Zickar MJ. Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organ Res Methods*. 2012;15(3):339-362. doi:10.1177/1094428111430540
53. Fisher DJ, Copas AJ, Tierney JF, Parmar MKB. A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *J Clin Epidemiol*. 2011;64(9):949-967. doi:10.1016/j.jclinepi.2010.11.016
54. Simmonds MC, Higgins JPT. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. *Stat Med*. 2007;26(15):2982-2999. doi:10.1002/sim.2768
55. Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Stat Med*. 2012;31(29):3821-3839. doi:10.1002/sim.5471
56. Stapleton LM. Incorporating sampling weights into single- and multilevel analyses. In: Rutkowski L, von Davier M, Rutkowski D, eds. *Handbook of International Large-Scale Assessment Background, Technical Issues, and Methods of Data Analysis*. CRC Press; 2014:363-388.
57. Snijders TAB, Bosker RJ. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. Sage; 2012.
58. Raudenbush SW, Bryk AS. *Hierarchical Linear Models*. 2nd ed. Sage; 2002.
59. Tanner-Smith EE, Tipton E. Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS. *Res Synth Methods*. 2014; 5(1):13-30. doi:10.1002/jrsm.1091
60. Audigier V, White IR, Jolani S, et al. Multiple imputation for multilevel data with continuous and binary variables. *Stat Sci*. 2018;33(2):160-183. doi:10.1214/18-STS646
61. OECD. *PISA 2006*. Technical Report. OECD; 2009.
62. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol*. 2009;60:549-576.
63. Peugh JL, Enders CK. Missing data in educational research: a review of reporting practices and suggestions for improvement. *Rev Educ Res*. 2004;74(4):525-556. doi:10.3102/00346543074004525
64. Van Buuren S. *Flexible Imputation of Missing Data*. 2nd ed. CRC Press; 2018 <https://stefvanbuuren.name/fim/>
65. Skinner C, Wakefield J. Introduction to the design and analysis of complex survey data. *Stat Sci*. 2017;32(2):165-175. doi:10.1214/17-STS614
66. Organisation for Economic Co-operation and Development. *PISA Data Analysis Manual*. SPSS. 2nd ed. OECD; 2009.
67. Rutkowski L, Gonzalez E, Joncas M, von Davier M. International large-scale assessment data: issues in secondary analysis and reporting. *Educ Res*. 2010;39(2):142-151. doi:10.3102/0013189X10363170
68. Stapleton LM. An assessment of practical solutions for structural equation modeling with complex sample data. *Struct Equ Model Multidiscip J*. 2006;13(1):28-58. doi:10.1207/s15328007sem1301_2
69. Brunner M, Keller U, Wenger M, Fischbach A, Lüdtke O. Between-school variation in students' achievement, motivation, affect, and learning strategies: results from 81 countries for planning group-randomized trials in education. *J Res Educ Eff*. 2018;11(3):452-478. doi:10.1080/19345747.2017.1375584
70. Rust K, Rao J. Variance estimation for complex surveys using replication techniques. *Stat Methods Med Res*. 1996;5(3):283-310. doi:10.1177/096228029600500305
71. Rust K. Sampling, weighting, and variance estimation in international large-scale assessments. In: Rutkowski L, von Davier M, Rutkowski D, eds. *Handbook of International Large-Scale Assessment Background, Technical Issues, and Methods of Data Analysis*. CRC Press; 2014:117-153.
72. Borenstein M, Hedges LV. Effect sizes for meta-analysis. In: Cooper HM, Hedges LV, Valentine JC, eds. *Handbook of Research Synthesis and Meta-Analysis*. 3rd ed. Russell Sage Foundation; 2019:207-243.
73. Pek J, Flora DB. Reporting effect sizes in original psychological research: a discussion and tutorial. *Psychol Methods*. 2018; 23(2):208-225. doi:10.1037/met0000126
74. Cumming G, Calin-Jageman R. *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. Routledge; 2016.
75. Hill CJ, Bloom HS, Black AR, Lipsey MW. Empirical benchmarks for interpreting effect sizes in research. *Child Dev Perspect*. 2008;2:172-177.
76. von Davier M, Sinharay S. Analytics in international large-scale assessments: item response theory and population models. In: Rutkowski L, von Davier M, Rutkowski D, eds. *Handbook of International Large-Scale Assessment Background, Technical Issues, and Methods of Data Analysis*. CRC Press; 2014:155-201.
77. Debray TPA, Snell KIE, Quartagno M, Jolani S, Moons KGM, Riley RD. Dealing with missing data in an IPD meta-analysis. In: Riley RD, Tierney JF, Stewart LA, eds. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Wiley Series in Statistics in Practice. Wiley; 2021: 499-524.
78. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res*. 2018;27(6):1634-1649. doi:10.1177/0962280216666564
79. Lüdtke O, Robitzsch A, West SG. Regression models involving nonlinear effects with missing data: a sequential modeling approach using Bayesian estimation. *Psychol Methods*. 2020; 25(2):157-181. doi:10.1037/met0000233
80. Resche-Rigon M, White IR, Bartlett WJ, SAE P, Thompson SG. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Stat Med*. 2013;32(28):4890-4905. doi:10.1002/sim.5894
81. Jolani S, Debray TPA, Koffijberg H, Van Buuren S, Moons KGM. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Stat Med*. 2015;34(11):1841-1863. doi:10.1002/sim.6451
82. Kunkel D, Kaizar EE. A comparison of existing methods for multiple imputation in individual participant data meta-analysis. *Stat Med*. 2017;36(22):3507-3532. doi:10.1002/sim.7388

83. Enders CK, Du H, Keller BT. A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychol Methods*. 2020;25(1):88-112. doi:10.1037/met0000228
84. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. doi:10.1136/bmj.b2393
85. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*. John Wiley & Sons, Ltd; 2013. doi:10.1002/9781119942283
86. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377-399. doi:10.1002/sim.4067
87. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6(4):330-351.
88. Burgess S, White IR, Resche-Rigon M, Wood AM. Combining multiple imputation and meta-analysis with individual participant data. *Stat Med*. 2013;32(26):4499-4514. doi:10.1002/sim.5844
89. Rubin DB. Nested multiple imputation of NMES via partially incompatible MCMC. *Stat Neerlandica*. 2003;57(1):3-18. doi:10.1111/1467-9574.00217
90. Robitzsch A, Lüdtke O. *Mdmb: Model Based Treatment of Missing Data*. R package version 1.5-8; 2021. <https://cran.r-project.org/web/packages/mdmb/index.html>
91. Grund S, Lüdtke O, Robitzsch A. Multiple imputation of missing data in multilevel models with the R package mdmb: a flexible sequential modeling approach. *Behav Res Methods*. 2021;53(6):2631-2649. doi:10.3758/s13428-020-01530-0
92. van Ginkel JR. Standardized regression coefficients and newly proposed estimators for R2 in multiply imputed data. *Psychometrika*. 2020;85(1):185-205. doi:10.1007/s11336-020-09696-4
93. West BT, Sakshaug JW, Aurelien GAS. Accounting for complex sampling in survey estimation: a review of current software tools. *J Off Stat*. 2018;34(3):721-752. doi:10.2478/jos-2018-0034
94. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021 <https://www.R-project.org/>
95. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Res Synth Methods*. 2010;1(1):39-65. doi:10.1002/jrsm.5
96. Scammacca N, Roberts G, Stuebing KK. Meta-analysis with complex research designs: dealing with dependence from multiple measures and multiple group comparisons. *Rev Educ Res*. 2014;84(3):328-364. doi:10.3102/0034654313500826
97. Cheung MWL. *Meta-Analysis: A Structural Equation Modeling Approach*. Wiley; 2015.
98. Pustejovsky JE, Tipton E. Meta-analysis with robust variance estimation: expanding the range of working models. *Prev Sci Off J Soc Prev Res*. 2022;23(3):425-438. doi:10.1007/s11121-021-01246-3
99. Bender R, Friede T, Koch A, et al. Methods for evidence synthesis in the case of very few studies. *Res Synth Methods*. 2018; 9(3):382-392. doi:10.1002/jrsm.1297
100. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol*. 2014; 14(1):25. doi:10.1186/1471-2288-14-25
101. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med*. 2003;22(17): 2693-2710. doi:10.1002/sim.1482
102. Sidik K, Jonkman JN. On constructing confidence intervals for a standardized mean difference in meta-analysis. *Commun Stat - Simul Comput*. 2003;32(4):1191-1203. doi:10.1081/SAC-120023885
103. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Academic Press; 1985.
104. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. *Psychol Methods*. 2015;20(3): 375-393. doi:10.1037/met0000011
105. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods*. 2016;7(1):55-79. doi:10.1002/jrsm.1164
106. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3). doi:10.18637/jss.v036.i03
107. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Res Synth Methods*. 2017;8(1):5-18. doi:10.1002/jrsm.1230
108. Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane; 2021 www.training.cochrane.org/handbook
109. Polanin JR, Hennessy EA, Tanner-Smith EE. A review of meta-analysis packages in R. *J Educ Behav Stat*. 2017;42(2): 206-242. doi:10.3102/1076998616674315
110. Pustejovsky JE. *ClubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections*. R package version 0.5.3. 2021. <https://CRAN.R-project.org/package=clubSandwich>
111. Hyde JS. The gender similarity hypothesis. *Am Psychol*. 2005; 60:581-592.
112. OECD. *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*. OECD; 2003.
113. Wang S, Jiao H, Young MJ, Brooks T, Olson J. Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: a meta-analysis of testing mode effects. *Educ Psychol Meas*. 2008;68(1):5-24. doi:10.1177/0013164407305592
114. Tipton E, Pustejovsky JE, Ahmadi H. Current practices in meta-regression in psychology, education, and medicine. *Res Synth Methods*. 2019;10(2):180-194. doi:10.1002/jrsm.1339
115. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, Inc; 1980. doi:10.1002/0471725153
116. Viechtbauer W, Cheung MWL. Outlier and influence diagnostics for meta-analysis. *Res Synth Methods*. 2010;1(2):112-125. doi:10.1002/jrsm.11
117. Glass GV. Meta-analysis at middle age: a personal history. *Res Synth Methods*. 2015;6(3):221-231. doi:10.1002/jrsm.1133
118. Pellegrino JW, Baxter GP, Glaser R. Addressing the “two disciplines” problem: linking theories of cognition and learning with assessment and instructional practice. *Rev Res Educ*. 1999;24:307-354.
119. Slagt M, Dubas JS, Deković M, van Aken MAG. Differences in sensitivity to parenting depending on child temperament: a meta-analysis. *Psychol Bull*. 2016;142(10):1068-1110. doi:10.1037/bul0000061
120. Oberski D. *lavaan.survey: an R package for complex survey analysis of structural equation models*. *J Stat Softw*. 2014; 57(1):1-27. doi:10.18637/jss.v057.i01

121. Singer JD, Braun HI. Testing international education assessments. *Science*. 2018;360(6384):38-40. doi:[10.1126/science.aar4952](https://doi.org/10.1126/science.aar4952)
122. Pigott TD, Williams R, Polanin J. Combining individual participant and aggregated data in a meta-analysis with correlational studies. *Res Synth Methods*. 2012;3(4):257-268. doi:[10.1002/jrsm.1051](https://doi.org/10.1002/jrsm.1051)
123. Johnson BT. Toward a more transparent, rigorous, and generative psychology. *Psychol Bull*. 2021;147(1):1-15. doi:[10.1037/bul0000317](https://doi.org/10.1037/bul0000317)
124. Neimann Rasmussen L, Montgomery P. The prevalence of and factors associated with inclusion of non-English language studies in Campbell systematic reviews: a survey and meta-epidemiological study. *Syst Rev*. 2018;7(1):129. doi:[10.1186/s13643-018-0786-6](https://doi.org/10.1186/s13643-018-0786-6)
125. Giustini D. Retrieving grey literature, information, and data in the digital age. In: Cooper HM, Hedges LV, Valentine JC, eds. *Handbook of Research Synthesis and Meta-Analysis*. 3rd ed. Russell Sage Foundation; 2019:101-126.
126. Barroso C, Ganley CM, McGraw AL, Geer EA, Hart SA, Daucourt MC. A meta-analysis of the relation between math anxiety and math achievement. *Psychol Bull*. 2021;147(2):134-168. doi:[10.1037/bul0000307](https://doi.org/10.1037/bul0000307)
127. Stewart LA, Riley RD, Tierney JF. Planning and initiating an IPD meta-analysis project. In: Riley RD, Tierney JF, Stewart LA, eds. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Wiley Series in Statistics in Practice. Wiley; 2021:21-43.
128. Ensor J, Burke DL, Snell KIE, Hemming K, Riley RD. Simulation-based power calculations for planning a two-stage individual participant data meta-analysis. *BMC Med Res Methodol*. 2018;18(1):41. doi:[10.1186/s12874-018-0492-z](https://doi.org/10.1186/s12874-018-0492-z)
129. Ahmed I, Sutton AJ, Riley RD. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ*. 2012;344:d7762. doi:[10.1136/bmj.d7762](https://doi.org/10.1136/bmj.d7762)
130. Stewart LA, Clarke M, Rovers M, et al. Preferred reporting items for a systematic review and meta-analysis of individual participant data: the PRISMA-IPD statement. *Jama*. 2015;313(16):1657-1665. doi:[10.1001/jama.2015.3656](https://doi.org/10.1001/jama.2015.3656)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Brunner M, Keller L, Stallasch SE, et al. 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. *Res Syn Meth*. 2022;1-31. doi:[10.1002/jrsm.1584](https://doi.org/10.1002/jrsm.1584)