Large-scale Assessments in Education

REVIEW **Open Access**

The potential of international large-scale assessments for meta-analyses in education

Ronny Scherer[1,2*], Fazilat Siddiq[3] and Trude Nilsen[2,4]

*Correspondence:
ronny.scherer@cemo.uio.no

[1] Faculty of Educational Sciences, Centre for Educational Measurement at the University of Oslo (CEMO), University of Oslo, Forskningsparken, PO Box 1161, 0318 Oslo, Norway
[2] Centre for Research on Equality in Education (CREATE), Faculty of Educational Sciences, University of Oslo, Oslo, Norway
[3] Department of Education and Quality in Learning, Unit for Digitalisation and Education, University of South-Eastern Norway, Drammen, Norway
[4] Faculty of Educational Sciences, Department of Teacher Education and School Research, University of Oslo, Oslo, Norway

Abstract

Meta-analyses and international large-scale assessments (ILSA) are key sources for informing educational policy, research, and practice. While many critical research questions could be addressed by drawing evidence from both of these sources, meta-analysts seldom integrate ILSAs, and current integration practices lack methodological guidance. The aim of this methodological review is therefore to synthesize and illustrate the principles and practices of including ILSA data in meta-analyses. Specifically, we (a) review four ILSA data inclusion approaches (analytic steps, potential, challenges); (b) examine whether and how existing meta-analyses included ILSA data; and (c) provide a hands-on illustrative example of how to implement the four approaches. Seeing the need for meta-analyses on educational inequalities, we situated the review and illustration in the context of gender differences and socioeconomic gaps in student achievement. Ultimately, we outline the steps meta-analysts could take to utilize the potential and address the challenges of ILSA data for meta-analyses in education.

Keywords: Gender differences, Digital literacy, International large-scale assessments, Meta-analysis, Socioeconomic status

Evidence-based decision-making is key to educational policy and practice. To facilitate this, researchers synthesize the body of evidence on, for instance, the effectiveness of educational programs, the factors related to desirable educational outcomes, and possible sources of variation or inequalities in education via meta-analyses (Hattie et al., 2014; Oh, 2020). These quantitative research syntheses must provide reliable, meaningful, and unbiased evidence so that valid inferences can be drawn by researchers, practitioners, and policymakers (Slavin, 2008). However, meta-analyses in education and other disciplines face several challenges compromising their validity (e.g., Ahn et al., 2012; Rios et al., 2020; Sharpe, 1997): small-sample primary studies (e.g., low power to detect practically relevant effect sizes, high uncertainty, risk of invalid generalizations to student populations), study characteristics that may affect the quality and magnitude of effects (e.g., convenience samples, lack of stratification, matching, or control groups), and insufficient psychometric quality of the outcome measures (e.g., low reliability, limited construct coverage)—just to name a few. International large-scale assessments (ILSAs), such as ICILS (International Computer and Information Literacy Study), TIMSS (Trends in International Mathematics and Science Study), and PISA (Programme for International

Springer Open

Student Assessment), address many of these issues (Braun & Singer, 2019; Klieme, 2020; Rutkowski et al., 2010; Wagemaker, 2016).

Despite this potential, it is not a common practice to include ILSA data in meta-analyses on key educational research questions. For instance, in their meta-analysis of the relation between socioeconomic status and student achievement, Kim et al. (2019) and Scherer and Siddiq (2019) included ILSA and non-ILSA data side-by-side, while Sirin (2005) and Harwell et al. (2017) based their meta-analyses solely on non-ILSA data, although ILSA data would have been eligible for inclusion. Similarly, some meta-analyses of the gender differences in student achievement included both ILSA and non-ILSA data (Lietz, 2006; Siddiq & Scherer, 2019), while some focused only on non-ILSA (Lindberg et al., 2010) or ILSA data (Else-Quest et al., 2010). The complexities of analyzing primary ILSA data and the resultant meta-analytic data may provide some reasoning for these varying practices. These complexities include the multi-stage cluster sampling designs that need to be represented when estimating effect sizes, the availability of multiple effect sizes per ILSA, ILSA cycle, or country, and the lack of analytic approaches guiding the integration of ILSAs in meta-analyses (e.g., Hedges, 2007; Rutkowski et al., 2010).

To this end, the inclusion of ILSA data in meta-analyses has faced two key challenges: *Varying inclusion practices*, likely due the lack of methodological guidance, and the *complex structures of ILSA and meta-analytic data* that demand non-standard effect size computation and advanced meta-analyses. Our methodological review addresses these challenges by (a) describing an analytic framework that comprises four inclusion approaches; (b) reviewing systematically whether and how existing meta-analyses included ILSA data; and (c) illustrating these approaches with an example meta-analysis. Drawing from the results of our review, we offer recommendations for researchers on how to include ILSA data in their meta-analyses to inform evidence-based practice and policymaking.

International large-scale assessments (ILSAs) informing meta-analyses in education

Purposes and contribution

Meta-analyses and ILSAs have similar purposes. Oh, (2020) identified three evidence-based uses of *meta-analyses*: (a) Informing the design of empirical studies; (b) informing the interpretation of the effect sizes resulting from primary studies by creating context and providing benchmarks; (c) informing educational practice and the development of professional guidelines for research. Besides, meta-analyses have several theoretical uses, such as providing information about population effect sizes, quantifying heterogeneity, and identifying the extent to which sample, study, and measurement characteristics could explain this heterogeneity (Borenstein et al., 2009). Ultimately, meta-analyses are aimed at supporting research, practice, and policy in drawing robust conclusions about key educational issues and explaining how and why specific findings may fit together or deviate (Glass, 1976; Siddaway et al., 2019).

Similarly, *ILSAs* provide large-scale, representative, and international data to (a) increase the understanding of key factors influencing teaching and learning, including contextual factors; (b) identify key educational issues; (c) inform national strategies for monitoring and improvement, including evaluating the effectiveness of curricula, instruction, and policies; (d) contribute to the research community to facilitate educational evaluation and document progress in research; (e) create de facto benchmarking, providing context for small-scale research and tracing student achievement across nations and over time (Braun & Singer, 2019; Hopfenbeck et al., 2018; Wagemaker, 2016). The next sections demonstrate that ILSAs have a value in their own right for meta-analyses in education and how they may address some of the challenges meta-analyses are facing.

Potential, challenges, and limitations of including ILSA data in meta-analyses in education

Key educational issues and constructs

ILSAs contain rich indicators of educational achievement, oftentimes in several domains and sub-domains, motivational and affective constructs, background characteristics, and contextual factors, which are measured across the different levels of educational systems and over time. These indicators are documented transparently and allow researchers to assess and monitor key educational issues, such as equity and equality, trends and profiles of student achievement, and the link between school practices and educational outcomes (e.g., Klieme, 2020; Lenkeit et al., 2015). Despite the rich set of constructs and indicators, the feasibility and time constraints in which ILSAs operate allow for including only a selection of constructs, types of tasks, and scales (Gustafsson, 2018; Kuger & Klieme, 2016), so that selection of educationally relevant constructs is by no means exhaustive. Hence, ILSA data do not qualify for inclusion in any meta-analysis in education but need to undergo a rigorous eligibility check.

Country selection

ILSAs follow a rigorous sampling designs with multiple stages of quality assurance (Musu et al., 2020; Wagemaker, 2020). While much emphasis has been placed on the random sampling within countries (e.g., random sampling of schools and students or teachers within schools in PISA and, respectively, TALIS), the sampling of countries participating in ILSAs is not random. In fact, countries decide to participate in ILSAs and ILSA cycles and, essentially, self-select depending on their needs and capabilities. This self-selection has several consequences, such as the varying participation across ILSAs and ILSA cycles with countries remaining, dropping out, or joining in, and the possible underrepresentation of cultures or World regions (e.g., Rutkowski & Rutkowski, 2021). The varying participation of countries challenges the study of educational trends due to the lack of consistent longitudinal data at the level of countries (e.g., Lohmann et al., 2022).

Given that ILSAs include a broad range of countries, cultures, and educational systems, including ILSA data in meta-analyses can balance the representation of cultural and language groups—in fact, possible cultural and language bias may be reduced in meta-analyses (Morrison et al., 2012). While meta-analysts oftentimes exclude studies and reports that were not published in English, the information on the various ILSA samples, assessments, and results are made available in English, irrespective of the language of origin in the countries. For instance, in their meta-analysis of the relation between academic achievement and self-concept, Möller et al. (2020) included multiple PISA samples from around the world—cultural balance was of particular importance in this study, due to the cultural differences in students' self-concept.

Availability and comparability of ILSA data

Finally, we would like to highlight the availability of the primary data along with rich documentation as another strength of ILSAs—specifically, for the most part, ILSA data are freely available to meta-analysts through open-access platforms of the respective organizations (primarily the IEA and OECD). Given the availability of these so-called individual-participant data (IPD), meta-analysts may not have to rely on the reporting from secondary sources, but can extract or estimate the relevant effect sizes themselves (Riley et al., 2021). For instance, if researchers are interested in the achievement differences between private and public schools after controlling for schools' socioeconomic composition and individual differences in socioeconomic status, they can specify and estimate a multilevel regression model with the variables of interest, utilizing multiple ILSA data sets. Across these data sets, the model generating the effect sizes is the same, and comparability of the type of effects is given. However, if the researchers extracted the achievement differences from secondary reports which were based on different multilevel regression models (e.g., with different predictors), the resultant effect sizes would no longer be comparable, and the validity of the meta-analytic results would be in question (Becker & Wu, 2007). This is a key issue in meta-analyses that are based on aggregated data (AD) and effect sizes generated from different analytic models (e.g., Polanin et al., 2020; Riley et al., 2021). In this sense, ILSA data allow meta-analysts to control the specification and estimation of the statistical models used to generate the effect sizes (Cheung & Jak, 2016).

Measurement invariance

Besides the comparability of the data setup, many ILSAs, ILSA cycles, and samples are based on the same or linked measures of constructs. However, although this design ensures some degree of comparability or, more precisely, a similar exposure to items and tasks, it does not ensure measurement invariance—the comparability of the measurement models underlying reflectively defined constructs—per se. Researchers and large-scale data analysts still have to provide evidence that the measurement models representing specific constructs are sufficiently invariant (van de Vijver et al., 2019). However, extending the range of participating countries and educational systems, population heterogeneity in ILSAs can be problematic, because deficits in invariance may undermine the comparability of measures (Rutkowski et al., 2019). This issue does not only concern the measurement of student achievement, which has received most of the attention in ILSA-related research, but also the measures taken via the accompanying questionnaires (Rutkowski & Rutkowski, 2018). For instance, in some ILSAs, the constructs measured via the background questionnaires are not fully aligned with the core achievement domains, so that obtaining evidence on convergent validity is hardly possible.

Correlational nature of the data

The correlational nature of the ILSA data, resulting from the cross-sectional study design, may be another issue that could exclude these data from meta-analyses in education, especially when effectiveness questions are addressed that require (quasi-)experimental designs (Klieme, 2013). In fact, ILSAs offer only limited opportunities to draw causal inferences (Rutkowski & Delandshere, 2016), and may inform meta-analysis primarily by group differences (e.g., gender differences in student achievement) or relations among constructs (e.g., relation between self-concept and student achievement). For instance, research questions on the effectiveness of instruction can hardly be addressed directly, and randomized-controlled trials would obviously be the gold standard to inform such questions. Given their design, ILSA data would not be eligible for inclusion in meta-analyses of the effectiveness of instruction. Yet, ILSAs could still provide information about the distribution of relevant variables and their relations to educational achievement reported (Braun & Singer, 2019; Klieme, 2020).

Complex survey designs and large samples

Another challenge associated with the use of ILSA data in meta-analysis is the extraction of the correct effect sizes. ILSA data follow a complex survey design with multiple stages of sampling that require advanced methods to estimate effects (Rust, 2014)—among others, the key elements include the multilevel data structure (e.g., students hierarchically nested in schools), the use of sampling weights (e.g., student- and school-level weights), the correct variance estimation (e.g., via jackknifing techniques and replicate weights), and the achievement estimation (e.g., via plausible value techniques). For instance, if meta-analysts are interested in the relation between measures of instructional quality in classrooms and student achievement, multilevel modeling is required to account for the nested structure of the primary ILSA data ("primary clustering") and obtain the contextual effect. While these elements have been discussed and presented in the extant literature extensively (Rutkowski et al., 2010), we suspect that addressing them to extract the relevant effect sizes from secondary data analyses can pose barriers for meta-analysts. Associated with the complex survey design are the large sample sizes within ILSA data. Large-sample studies may well increase precision and reduce sampling error, yet may also influence the effect size estimate and its variance components substantially due to large weights in the meta-analytic data set (Turner et al., 2013).

Complex meta-analytic data structures

Besides the primary clustering of the ILSA data, including them in meta-analyses can create a nested structure of the meta-analytic data ("secondary clustering"), with multiple effect sizes extracted from the ILSAs, ILSA cycles, or countries (e.g., Pigott & Polanin, 2020). Such structures can violate the independence assumption in meta-analysis and require meta-analysts to address them, for instance, via multilevel meta-analysis, robust variance estimation, or pooling approaches (Cheung, 2019; Pustejovsky & Tipton, 2021; Scammacca et al., 2014). Notice that the primary clustering represents a different analytic problem than the secondary clustering: While the former describes the structure of the *primary study data* with, for instance, students nested in classrooms, the latter describes the structure of the *meta-analytic data* with multiple effect sizes nested in, for instance, ILSAs. Addressing the secondary clustering still requires from meta-analysts the knowledge and skills to engage in advanced meta-analytic techniques.

The present study

Given the diversity of the ways in which meta-analysts include or ignore ILSA data and, at the same time, the current lack of guidelines informing this inclusion, our methodological review describes and illustrates the principles and practices of including ILSA data in meta-analyses. Specifically, we address the following three research questions:

1. Which analytic approaches can meta-analysts take to include ILSA alongside non-ILSA data, and what are their advantages and disadvantages? (*Inclusion approaches*)
2. To what extent have ILSA data been included in existing meta-analyses in education, and how? (*Inclusion status*)
3. How can the different inclusion approaches be implemented in meta-analyses? (*Inclusion implementation*)

We constrained our review of existing meta-analyses and the presentation of an illustrative example to the context of equality in education, given the otherwise unmanageably large body of meta-analyses in education (Ahn et al., 2012) and given the need for meta-analyses of issues related to educational equality (Broer et al., 2019).

Approaches of including ILSA data in meta-analyses in education

The meta-analytic literature on multilevel meta-analysis (Fernández-Castilla et al., 2020), meta-analysis with individual participant data (Burke et al., 2017), and Bayesian meta-analysis (Röver, 2020) offers a plethora of approaches to synthesize effect sizes from small- and large-scale studies, with or without complex data structures, including one- and two-stage procedures. On the basis of these approaches and the knowledge gained from the systematic reviews addressing our first research question, we propose an analytic framework that contains four approaches to include ILSA data in existing meta-analyses at the level of effect sizes:

1. Separate meta-analyses: ILSA and non-ILSA data are meta-analyzed separately.
2. Indirect inclusion via Bayesian meta-analysis: In a first step, the multiple effect sizes per ILSA are meta-analyzed, yielding estimates of the weighted average effect size and heterogeneity. In a second step, one or more of these estimates inform the prior distribution of the weighted average effect size and the heterogeneity for the non-ILSA data.
3. One-stage direct inclusion: ILSA and non-ILSA data (i.e., effect sizes) are included in a meta-analysis side-by-side and at the level of the effect sizes. For ILSA data, multiple effect sizes (e.g., for multiple countries or domains) are extracted.
4. Two-stage direct inclusion: In the first stage, the multiple effect sizes per ILSA are meta-analyzed or aggregated following some aggregation rules (e.g., Borenstein et al., 2009). In the second stage, the resultant, aggregated effect sizes for each ILSA are included in the meta-analysis next to the non-ILSA data.

In the following, we review the analytic steps, advantages, and challenges associated with each of these four approaches (see also Table 1).

Separate meta-analyses

Separate meta-analyses of ILSA and non-ILSA data do not directly integrate the two data sources. Ultimately, they result in separate estimates of the weighted average effect sizes and variance components (Table 1), which could inform alternative approaches, such as the two-stage direction and indirect inclusion approaches, or serve the purpose of benchmarking (e.g., ILSA effect sizes as benchmarks for non-ILSA effect sizes, or vice versa). Nevertheless, if the same meta-analytic models are specified for the two data sources, direct comparisons of the overall effect sizes are possible utilizing mixed-effects models and Wald tests even under heteroscedasticity (Rubio-Aparicio et al., 2020). Meta-analysts can examine moderator effects separately and compare the results qualitatively. If comparisons of effect sizes are not the main focus, conducting separate meta-analyses further allows researchers to specify different meta-analytic models for the ILSA and non-ILSA data, addressing their individual complexities (e.g., non-nested structure of the non-ILSA data, nested structure of the ILSA data; Table 1).

Indirect inclusion via Bayesian meta-analysis

If the primary interest of the meta-analysts lies in the meta-analysis of non-ILSA data, information from the meta-analysis of ILSA data could be incorporated indirectly via Bayesian meta-analysis. In this approach, the weighted average effect size and/or variance components derived from ILSA data can inform the distributions of the respective estimates for non-ILSA data (see Table 1 and Additional file 5: S5). While a general discussion of Bayesian meta-analysis is beyond the scope of this study, one key advantage lies in the possibilities for researchers to incorporate some prior knowledge in their meta-analysis, even when only few effect sizes are available (Röver, 2020). This inclusion approach is similar to Bayesian (historical) borrowing, in which prior information about distributions or effect sizes from previous ILSAs or ILSA cycles is used to inform the data analysis of new ILSAs, ILSA cycles, or other studies (Kaplan et al., 2023). However, specifying informative priors and random-effects models in the Bayesian framework requires some understanding of the possible parameter distributions and may thus not be easily accessible to meta-analysts. Moreover, the meta-analytic outcomes for the non-ILSA data may depend on the choice of priors, thus necessitating additional sensitivity analyses.

One-stage direct inclusion

The one-stage direct inclusion approach combines the ILSA and non-ILSA data directly at the level of effect sizes. For each ILSA study or wave (e.g., PISA 2006, PISA 2015), each country or cohort sample contributes an effect size (see Fig. 1a). This inclusion is comparable to the one-stage meta-analysis of individual participant data, in which multiple data sets are combined directly (Burke et al., 2017). If meta-analysts allow for including multiple countries or cohort samples, this direct inclusion ultimately results in a complex meta-analytic structure with multiple effect sizes per ILSA. Such a structure violates the basic assumption of the independence of effect sizes, because effect sizes

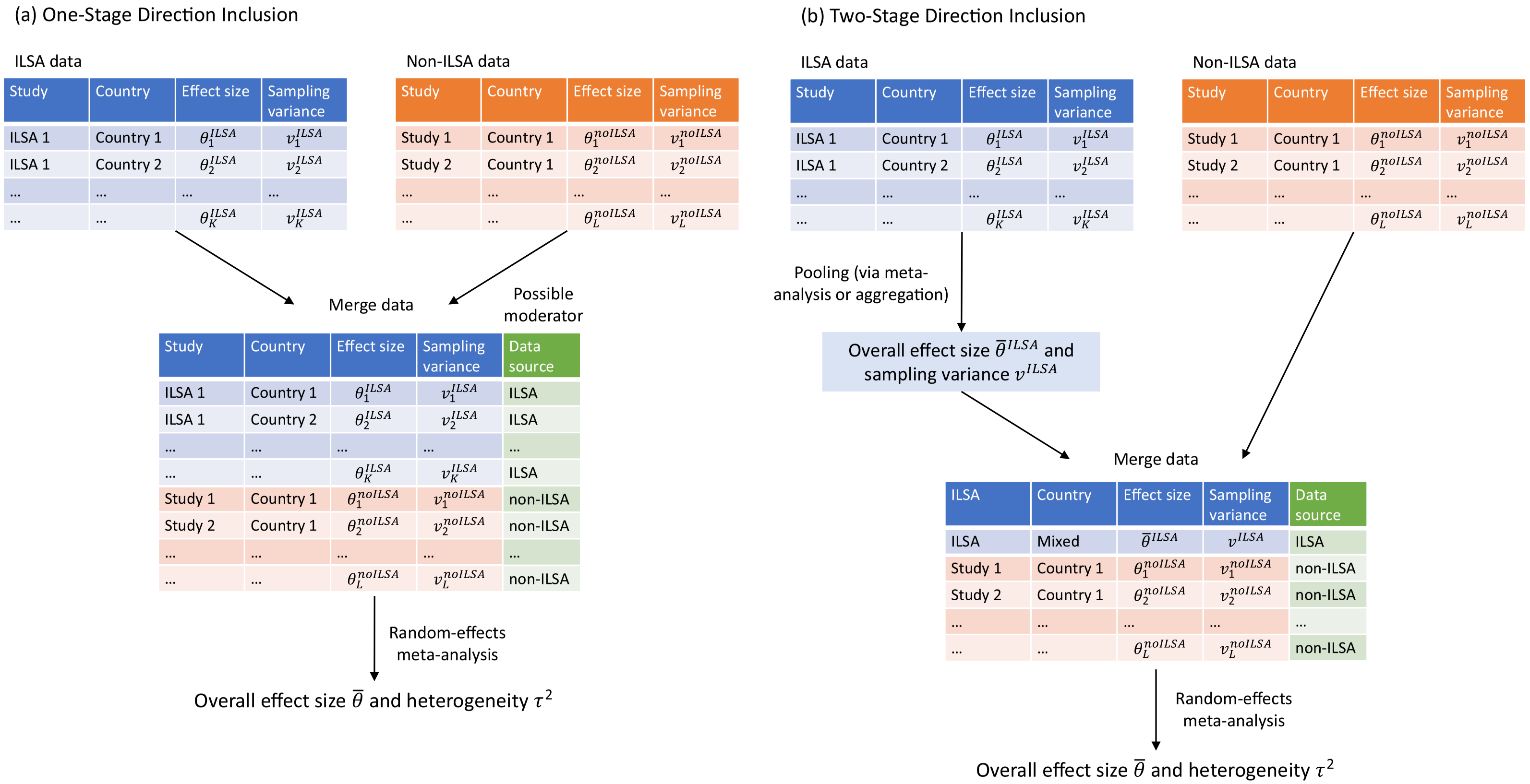
Table 1 Overview of the Different Inclusion Approaches

	Separate meta-analyses	Direct inclusion		Indirect inclusion
		One-stage inclusion	Two-stage inclusion	
Description	Primary effect sizes obtained from ILSA and non-ILSA data are meta-analyzed separately and considered as different meta-analytic data sets	Primary effect sizes obtained from ILSA and non-ILSA data are directly included in the meta-analyses, allowing for multiple effect sizes per study, cycle, or country	*Stage 1:* Primary effect sizes obtained from ILSA data are pooled *Stage 2:* The pooled ILSA effect size(s) and the primary effect sizes of the non-ILSA data are meta-analyzed	Primary effect sizes obtained from ILSA data are meta-analyzed in the first step, and the resultant overall effect size and/or variance estimates inform the meta-analysis of non-ILSA data as priors
Analytic steps	1. Compute the primary effect sizes for ILSA and non-ILSA data, taking into account the (complex) study designs 2. Identify the meta-analytic data structures for the ILSA and non-ILSA data sets 3. Synthesize the primary effect sizes of the ILSA data via (multilevel) random-effects meta-analysis and obtain the overall effect size(s) and variance component(s) for the level of studies or cycles (e.g., PISA or PISA 2015) 4. Synthesize the primary effect sizes of the non-ILSA data via (multilevel) random-effects meta-analysis and obtain the overall effect size(s) and variance component(s) 5. Conduct moderator analyses to identify possible sources of heterogeneity for ILSA and non-ILSA data separately 6. Examine publication bias and file-drawer issues for ILSA and non-ILSA data separately	1. Compute the primary effect sizes for ILSA and non-ILSA data, taking into account the (complex) study designs 2. Combine the primary effect-size ILSA and non-ILSA data 3. Identify the meta-analytic data structure 4. Quantify the overall effect size and the respective variance component(s) via (multilevel) random-effects meta-analysis 5. Examine the possible differences in effect sizes between ILSA and non-ILSA data 6. Conduct moderator analyses to identify other possible sources of heterogeneity 7. Examine publication bias and file-drawer issues 8. Conduct sensitivity analyses with respect to ignoring the hierarchical data structure vs. modeling this structure	1. Compute the primary effect sizes for ILSA and non-ILSA data, taking into account the (complex) study designs 2. Identify the meta-analytic data structures for the ILSA and non-ILSA data sets 3. Synthesize the primary effect sizes of the ILSA data via (multilevel) random-effects meta-analysis or other aggregation approaches and obtain the overall effect size(s) and variance component(s) for the level of studies or cycles (e.g., PISA or PISA 2015) 4. Combine the synthesized ILSA effect sizes with the primary effect-size non-ILSA data 5. Quantify the overall effect size and the respective variance component(s) via random-effects meta-analysis 6. Conduct moderator analyses to identify possible sources of heterogeneity 7. Examine publication bias and file-drawer issues 8. Conduct sensitivity analyses with respect to excluding vs. including the synthesized effect size of the ILSA data	1. Compute the primary effect sizes for ILSA and non-ILSA data, taking into account the (complex) study designs 2. Identify the meta-analytic data structures for the ILSA and non-ILSA data sets 3. Synthesize the primary effect sizes of the ILSA data via (multilevel) random-effects meta-analysis and obtain the overall effect size(s) and variance component(s) for the level of studies or cycles (e.g., PISA or PISA 2015) 4. Meta-analyze the primary non-ILSA effect sizes via Bayesian random-effects meta-analysis using priors for the overall effect size and/or the variance component(s) that are informed by the overall effect size and variance component(s) of the ILSA data 5. Conduct moderator analyses to identify possible sources of heterogeneity 6. Examine publication bias and file-drawer issues 7. Conduct sensitivity analyses with respect to the choice of priors
Primary analytic approaches	Random-effects meta-analysis, meta-analysis with robust variance estimation	Multilevel random-effects meta-analysis, meta-analysis with robust variance estimation	Random-effects meta-analysis, meta-analysis with robust variance estimation	Bayesian meta-analysis

Table 1 (continued)

	Separate meta-analyses	Direct inclusion		Indirect inclusion
		One-stage inclusion	**Two-stage inclusion**	
Key advantages	■ Less complex meta-analytic models are required to synthesize the data (e.g., within-ILSA variation is no longer estimated)—standard meta-analytic models available	■ Full information on the ILSA characteristics at the level of effect sizes can be incorporated ■ Variation between and within studies or other entities can be quantified ■ Moderator analyses at different levels of analysis are possible ■ Comparisons of overall effect sizes between ILSA and non-ILSA data possible via mixed-effects models	■ Less complex meta-analytic models are required to synthesize the data (e.g., within-ILSA variation is no longer estimated)—standard meta-analytic models available	■ Less complex meta-analytic models are required to synthesize the data (e.g., within-ILSA variation is no longer estimated) ■ Overall effect size and variance component(s) based on non-ILSA data are informed by ILSA data without their direct inclusion
Key challenges	■ Separate meta-analyses of ILSA and non-ILSA data do not inform each other ■ Fewer studies available for each of the separate meta-analyses	■ Advanced meta-analytic methods are needed (e.g., multilevel random- and mixed-effects meta-analysis)	■ Sensitivity analyses are required to examine the influence of including one or more synthesized ILSA effect sizes based on very large primary samples ■ Fewer studies available for meta-analysis ■ Limited possibilities to examine variation within ILSAs or ILSA cycles	■ Sensitivity analyses are required to examine the influence of the priors ■ Only non-ILSA data are meta-analyzed in the main analyses

ILSA International large-scale assessment


(a) One-Stage Direction Inclusion
ILSA data
Non-ILSA data
Merge data
Possible moderator
Random-effects meta-analysis
Overall effect size $\overline{\theta}$ and heterogeneity τ^2
(b) Two-Stage Direction Inclusion
ILSA data
Non-ILSA data
Pooling (via meta-analysis or aggregation)
Overall effect size $\overline{\theta}^{ILSA}$ and sampling variance v^{ILSA}
Merge data
Random-effects meta-analysis
Overall effect size $\overline{\theta}$ and heterogeneity τ^2

(a) One-Stage Direction Inclusion

ILSA data

Study	Country	Effect size	Sampling variance
ILSA 1	Country 1	θ_1^{ILSA}	v_1^{ILSA}
ILSA 1	Country 2	θ_2^{ILSA}	v_2^{ILSA}
…	…	…	…
…	…	θ_K^{ILSA}	v_K^{ILSA}

Non-ILSA data

Study	Country	Effect size	Sampling variance
Study 1	Country 1	θ_1^{noILSA}	v_1^{noILSA}
Study 2	Country 1	θ_2^{noILSA}	v_2^{noILSA}
…	…	…	…
…	…	θ_L^{noILSA}	v_L^{noILSA}

Merge data

Study	Country	Effect size	Sampling variance	Data source
ILSA 1	Country 1	θ_1^{ILSA}	v_1^{ILSA}	ILSA
ILSA 1	Country 2	θ_2^{ILSA}	v_2^{ILSA}	ILSA
…	…	…	…	…
…	…	θ_K^{ILSA}	v_K^{ILSA}	ILSA
Study 1	Country 1	θ_1^{noILSA}	v_1^{noILSA}	non-ILSA
Study 2	Country 1	θ_2^{noILSA}	v_2^{noILSA}	non-ILSA
…	…	…	…	…
…	…	θ_L^{noILSA}	v_L^{noILSA}	non-ILSA

(b) Two-Stage Direction Inclusion

ILSA data

Study	Country	Effect size	Sampling variance
ILSA 1	Country 1	θ_1^{ILSA}	v_1^{ILSA}
ILSA 1	Country 2	θ_2^{ILSA}	v_2^{ILSA}
…	…	…	…
…	…	θ_K^{ILSA}	v_K^{ILSA}

Non-ILSA data

Study	Country	Effect size	Sampling variance
Study 1	Country 1	θ_1^{noILSA}	v_1^{noILSA}
Study 2	Country 1	θ_2^{noILSA}	v_2^{noILSA}
…	…	…	…
…	…	θ_L^{noILSA}	v_L^{noILSA}

Merge data

ILSA	Country	Effect size	Sampling variance	Data source
ILSA	Mixed	$\overline{\theta}^{ILSA}$	v^{ILSA}	ILSA
Study 1	Country 1	θ_1^{noILSA}	v_1^{noILSA}	non-ILSA
Study 2	Country 1	θ_2^{noILSA}	v_2^{noILSA}	non-ILSA
…	…	…	…	…
…	…	θ_L^{noILSA}	v_L^{noILSA}	non-ILSA

Fig. 1 Overview of the **a** one-stage and **b** two-stage direct inclusion approaches. Note. ILSA = International large-scale assessment, K = Number of effect sizes extracted from ILSAs, L = Number of effect sizes extracted from the non-ILSA studies

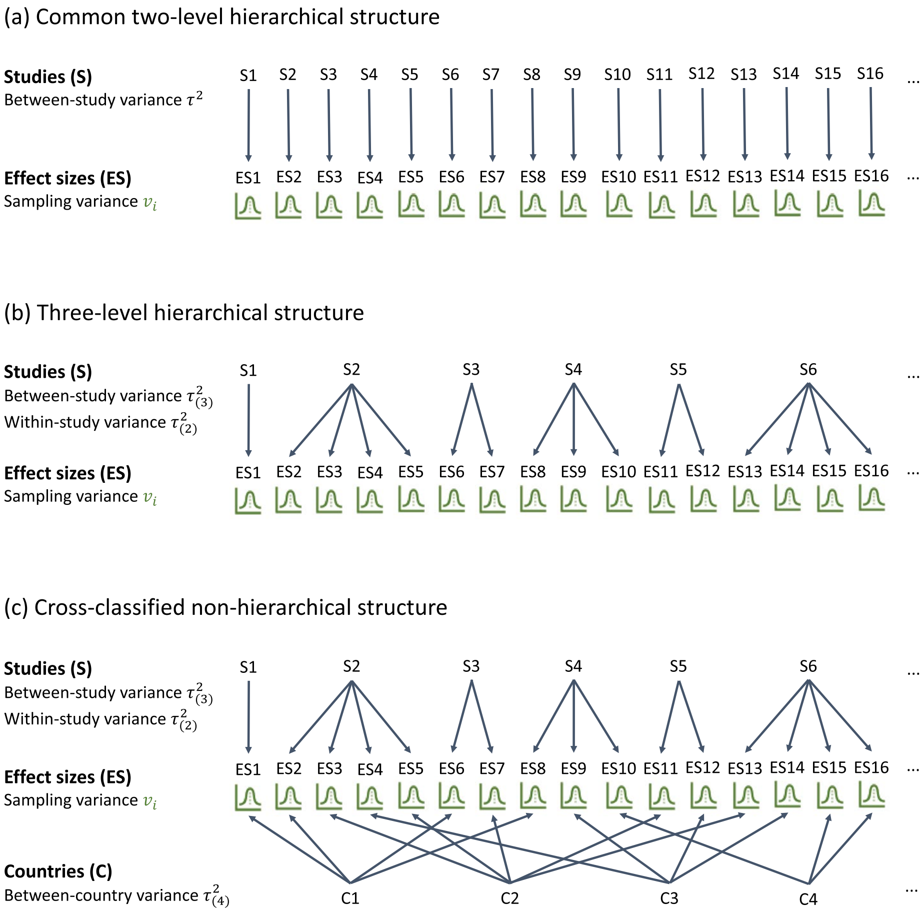

(a) Common two-level hierarchical structure
Studies (S)
Between-study variance τ^2
S1 S2 S3 S4 S5 S6 S7 S8 S9 S10 S11 S12 S13 S14 S15 S16 ...
Effect sizes (ES)
Sampling variance v_i
ES1 ES2 ES3 ES4 ES5 ES6 ES7 ES8 ES9 ES10 ES11 ES12 ES13 ES14 ES15 ES16 ...
(b) Three-level hierarchical structure
Studies (S)
Between-study variance $\tau^2_{(3)}$
Within-study variance $\tau^2_{(2)}$
S1 S2 S3 S4 S5 S6 ...
Effect sizes (ES)
Sampling variance v_i
ES1 ES2 ES3 ES4 ES5 ES6 ES7 ES8 ES9 ES10 ES11 ES12 ES13 ES14 ES15 ES16 ...
(c) Cross-classified non-hierarchical structure
Studies (S)
Between-study variance $\tau^2_{(3)}$
Within-study variance $\tau^2_{(2)}$
S1 S2 S3 S4 S5 S6 ...
Effect sizes (ES)
Sampling variance v_i
ES1 ES2 ES3 ES4 ES5 ES6 ES7 ES8 ES9 ES10 ES11 ES12 ES13 ES14 ES15 ES16 ...
Countries (C)
Between-country variance $\tau^2_{(4)}$
C1 C2 C3 C4 ...

Fig. 2 Meta-analytic data structures: **a** Common two-level hierarchical structure; **b** Three-level hierarchical structure; and **c** Cross-classified non-hierarchical structure

from the same ILSA may be more homogeneous than effect sizes from different ILSAs (Borenstein et al., 2009). As a consequence, meta-analysts must determine the structure of the meta-analytic data set and choose among suitable approaches to estimate the overall effect sizes and/or moderation effects that represent this structure (Cheung, 2019). Figure 2 illustrates two of the possible structures meta-analysts may encounter in this situation: Given the availability of multiple effect sizes per ILSA, the “ideal” structure with one effect size per ILSA only does no longer apply (Fig. 2a). Instead, a hierarchical structure with multiple effect sizes nested in ILSAs (Fig. 2b) or a non-hierarchical cross-classified structure with multiple effect sizes nested in ILSAs and countries (Fig. 2c) may better represent the meta-analytic data. The latter may be especially useful when including multiple ILSAs or ILSA cycles. If however only one ILSA or ILSA cycle is included, the country-specific effect sizes are considered independent, and the non-ILSA data contribute one effect per study, the structure may be simplified to Fig. 2a.

Having identified the data structure, meta-analysts can then choose how to handle such dependencies (see Table 1). While the described structures can be modelled explicitly via multilevel meta-analysis, a random-effects modeling approach that quantifies variation at the respective levels of analysis (e.g., within and between studies), or implicitly considered via robust variance estimation (RVE; e.g., Fernández-Castilla et al., 2020;

Hedges et al., 2010). Given the variety of approaches to handling multiple effect sizes, meta-analysts may consider conducting sensitivity analyses, varying these approaches and examining the possible differences in the resultant estimates (Table 1). Later in the data-analytic process, the possible differences between the effect sizes extracted from ILSA and non-ILSA studies can be examined and the effects of including ILSA data quantified. While the direct inclusion may require advanced meta-analytic models, meta-analysts can obtain information on different variance components, examine moderator effects at different levels of analysis, and gain precision in the effect size and variance estimates due the increased sample size (see Table 1).

In situations where ILSA studies provide the individual-participant data, and non-ILSA studies provide aggregated data, the possible differences in effect sizes between them may point to "availability bias"—a form of bias that occurs when the availability of IPD is associated with the quality of the primary study or its effect size (Riley et al., 2021). Although incorporating IPD from ILSAs can reduce publication bias due to the possible inclusion of unpublished data sets and studies, IPD may not be available for every primary study, for instance, due to issues related to data protection or accessibility. Hence, we consider the sensitivity analyses and testing for possible differences between ILSA and non-ILSA data to be important for meta-analyses combining these data.

Two-stage direct inclusion

Unlike the one-stage approach, the two-stage approach handles the multiple effect sizes per ILSA or ILSA cycle by pooling them first and submitting the resultant, pooled effect size and sampling variance to the meta-analysis with non-ILSA data (see Fig. 1b). This approach is similar to that two-stage meta-analysis of individual participant data (Burke et al., 2017). To perform the first stage, meta-analysts may rely on, for instance, Borenstein et al., (2009) formula to pool the effect sizes (to the average effect size) and the respective sampling variances (to a pooled variance which includes correlations between the effect sizes within a study). Alternatively, the pooled effect size may also be derived via separate meta-analyses for each of the ILSAs or ILSA cycles (see Table 1). This first stage can simplify the meta-analytic data structure in the second stage, because only one effect size per ILSA or ILSA cycle is included—ultimately, this may result in more robust variance estimates (Declercq et al., 2020). At the same time, the first pooling stage discards the within-ILSA variation (e.g., across countries within ILSA cycles; Fig. 1b)—an important source of variation and heterogeneity (Van den Noortgate et al., 2013). Moreover, meta-analysts may face the challenge of including effect sizes that are based on very large ILSA samples which ultimately receive larger weights (Borenstein et al., 2009). Examining the sensitivity of the meta-analytic results with respect to including such effect sizes and diagnosing influential effect sizes become key steps in this approach (see Table 1; e.g., Scherer & Siddiq, 2019).

Effect size measures

The four presented approaches are all based on the assumption that the correct effect sizes have been extracted from the ILSA and the non-ILSA data. In this context, "correct" refers to effect size and sampling (co-)variance estimates in which the complex survey design was accounted for, especially the hierarchical structure of the ILSA data (Lai

& Kwok, 2016; Tymms, 2004). For instance, when meta-analysts are interested in deriving the correct effect size measures for gender differences in achievement, the standardized mean difference (*SMD*) may be the effect size of their choice (Borenstein et al., 2009). When computing *SMD*, the pooled standard deviation can incorporate information about the nesting of the primary study data (e.g., students nested in classrooms or schools; Brunner et al., 2022). Hedges (2007) proposed several ways to incorporate the intraclass correlation ICC_1 into the estimate of the pooled standard deviation. While such adjustments are available, they depend on the authors' reporting of the relevant statistics, especially the intraclass correlation.

Besides the accounting for the nesting of the primary data, further elements may inform the estimation of the effect sizes, such as the use of sampling weights or performance assessment designs that draw from a set of plausible values (Rutkowski et al., 2010). Given that the raw primary data are oftentimes not available, meta-analysts may have to trust the estimation and reporting of the effect sizes in the publication and have hardly any chance to perform further adjustments. However, such adjustments are possible for most ILSA data—in fact, if the raw data of primary studies are available, the meta-analysts are in full control of the effect size estimation and can estimate them and the respective sampling (co-)variances from analytic models that incorporate the complex survey design features of ILSAs, such as multilevel models with sampling weights, stratifying variables, plausible values, and multi-group structures (Campos et al., 2023). Overall, meta-analysts have at least two options to address the complex survey design, especially the nested data structure, in primary studies: (a) Adjust the reported effect sizes by the ICC_1 (for details, please see Hedges, 2007); or (b) analyze the raw data (if available) via multilevel modeling (Kim et al., 2012).

The status of including ILSA data in meta-analyses of gender differences and SES gaps in student achievement

Substantive background

Educational research has long been concerned with examining and ultimately reducing gaps in educational outcomes between groups of students. Much of the discussion has centered around equity and equality in general (Espinoza, 2007), and the educational gaps associated with gender and socioeconomic status in particular (Berkowitz et al., 2017; Else-Quest et al., 2010). For instance, describing the SES-achievement relation in the domain of reading, PISA 2018 identified substantial variation in this relation across more than 70 educational systems (OECD, 2019). This ILSA also revealed cross-country variation in the gender gaps in reading achievement, yet with girls consistently outperforming boys. Similarly, other PISA cycles and ILSAs have mapped such gaps in student achievement across educational systems, age groups, subject domains, and over time and thus provide a rich data source for exploring their effect sizes, heterogeneity, and possible explanatory mechanisms (Broer et al., 2019; Gray et al., 2019).

To examine the extent to which ILSA data have been utilized to inform the meta-analytic body of knowledge and which approaches to including these data meta-analysts have taken, we systematically reviewed existing meta-analyses of the gender differences and SES gaps in student achievement. In this sense, the following two systematic reviews showcase the status of inclusion and inclusion approaches.

Methods

We used the systematic review methodology to identify the relevant studies within the scope of this paper, and followed the recommended steps, including predefining research questions, development of the search strategy, defining inclusion and exclusion criteria, screening, data extraction, appraisal, and synthesis (Higgins et al., 2019). In the following sections, we describe the application of these steps.

Search strategy

To retrieve the relevant meta-analyses, we developed a search strategy by first identifying the key terms for answering the aims of this study and identified the most commonly used synonymous for each term. We then performed two independent searches in the databases ERIC (Education Resources Information Center) and PsycINFO, combining search terms related to (a) the *study design*: meta-analysis or meta-analytic; (b) the *outcome variable*: achievement or performance or literacy or numeracy or reading or math* or science; and (c) the *independent variable*. For the latter, we used the search terms "gender difference* or sex difference* or gender gap" for meta-analyses of gender differences and "SES or socioeconomic status or socio-economic status or number of books or parent* education or parent* occupation or income or ESCS or HISEI or ISEI or possession* or capital" for meta-analyses reporting the relation between SES and student achievement. We extended these searches by hand-searching publications in key journals in the field (Educational Research Review, Review of Educational Research, Psychological Bulletin, Journal of Educational Psychology, Large-scale Assessments in Education) and the database PsyArXiv to identify possible preprint publications eligible for inclusion. Additional file 6: S6 contains the full search strategies, including the specific search terms. After removing duplicates, these searches yielded 318 publications for the gender meta-analyses and 271 publications for the SES meta-analyses (see Fig. 3).

Screening and coding

The retrieved publications were then screened in two steps: First, we reviewed the abstracts for their topic fit, considering meta-analyses that were published in English between 1995 and 2020. Besides, the full texts of these publications must have been made available, the topic must have related to the designated content areas (i.e., gender differences in student achievement or relations between SES and student achievement), and the authors must have performed a meta-analysis—theoretical reviews, comments, methodological papers, and errata were excluded. This first step resulted in 19 published meta-analyses eligible for further screening for the gender meta-analyses and 36 publications for the SES meta-analyses (see Fig. 3). Second, we reviewed the full texts according to the following criteria:

- *Type of research question and data:* The research question concerning gender differences in student achievement or the relation to SES are of correlational nature,

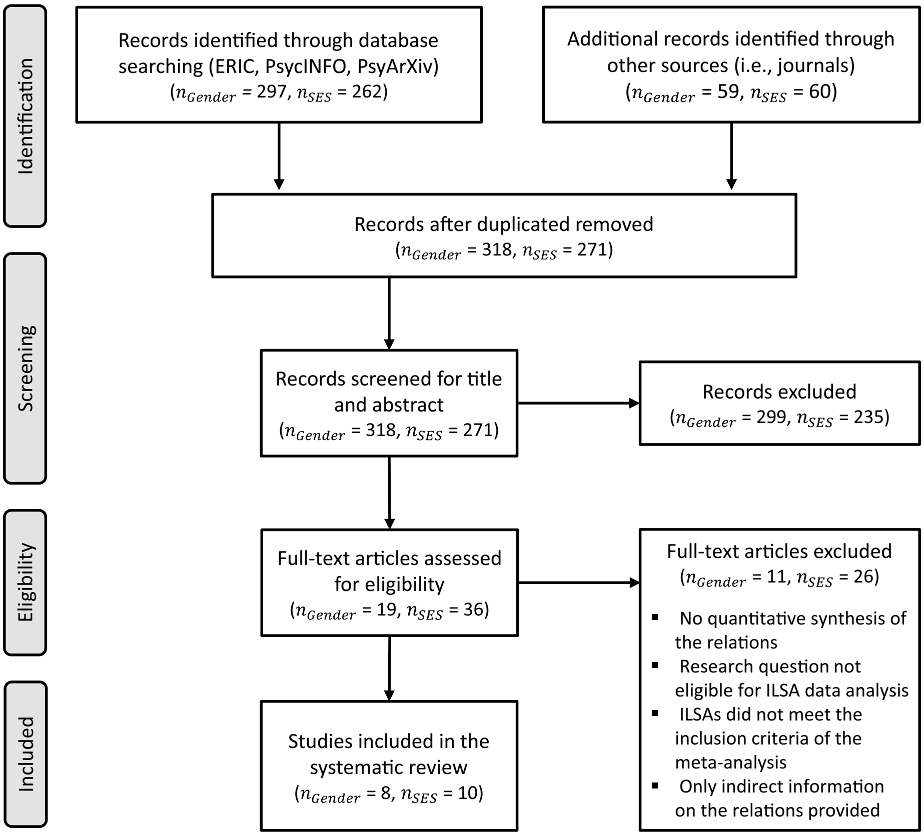

Identification
Screening
Eligibility
Included
Records identified through database searching (ERIC, PsycINFO, PsyArXiv) (n_{Gender} = 297, n_{SES} = 262)
Additional records identified through other sources (i.e., journals) (n_{Gender} = 59, n_{SES} = 60)
Records after duplicated removed (n_{Gender} = 318, n_{SES} = 271)
Records screened for title and abstract (n_{Gender} = 318, n_{SES} = 271)
Records excluded (n_{Gender} = 299, n_{SES} = 235)
Full-text articles assessed for eligibility (n_{Gender} = 19, n_{SES} = 36)
Full-text articles excluded (n_{Gender} = 11, n_{SES} = 26)
▪ No quantitative synthesis of the relations
▪ Research question not eligible for ILSA data analysis
▪ ILSAs did not meet the inclusion criteria of the meta-analysis
▪ Only indirect information on the relations provided
Studies included in the systematic review (n_{Gender} = 8, n_{SES} = 10)

Fig. 3 PRISMA Flow Diagram of the Search, Screening, and Inclusion Processes of the Meta-Analyses. *ILSA* International large-scale assessment, *SES* Socioeconomic status

and the data were observational. Exclude: Meta-analyses on the effectiveness of interventions.

- *Sample:* ILSAs contain the student samples the meta-analyses focused on. Exclude: Meta-analyses focusing on children younger than primary school students, children with medical conditions or disorders, and children that were selected according to some criterion that could not be found in ILSAs (e.g., executive function scores).
- *Content and constructs:* The constructs and contents of the meta-analyses were included in ILSAs. Exclude: Meta-analyses utilizing achievement or SES measures that were not assessed in ILSAs (e.g., working memory measures, school grades, parents' income).
- *Direct relations:* Direct relations between the constructs (i.e., gender or SES and student achievement) were reported. Exclude: Meta-analyses that use the key constructs as moderators (e.g., Peng et al., 2019).
- *Reported statistics:* Independent of their inclusion, ILSA data could provide the statistics and effect sizes needed for the meta-analysis.
- *Inclusion criteria:* Irrespective of their inclusion, ILSA studies fulfilled the inclusion criteria of the meta-analysis (i.e., would be eligible for inclusion). Exclude: Meta-analyses that focused on national large-scale assessments (e.g., Petersen, 2018).

These screening steps yielded eight gender meta-analyses and ten SES meta-analyses which set their inclusion and exclusion criteria so that ILSA studies had been eligible for inclusion (Fig. 3). A flowchart describing the screening decisions is shown in Additional file 6: S6.

The coding of these meta-analyses included key characteristics of the studies (i.e., publication year and status, number of studies and effect sizes, context), the measures (i.e., achievement domain, SES dimension(s), SES source(s), SES metric), the meta-analytic models (i.e., type of model(s), addressing the dependence structure, pooled effect size(s)), and the extent to which ILSA data were included (i.e., inclusion of ILSA data [yes/no], data sources, type(s) of ILSAs, cycle(s), inclusion approach, sensitivity analyses). Additional file 1: S1 and Additional file 2: S2 contain the detailed coding of the gender and SES meta-analyses, along with their pooled effects.

Results

Meta-analyses of gender differences in student achievement

Overall, the $m = 8$ meta-analyses examining gender differences in student achievement included 448 studies, oftentimes operationalized as independent study samples, and yielded 6428 effect sizes in total (see Table 2). These meta-analyses covered the domains of reading ($m = 5$), mathematics ($m = 6$), science ($m = 3$), and digital literacy ($m = 1$). One meta-analysis was based only on non-ILSA primary studies to avoid redundancies with other meta-analyses (Lindberg et al., 2010), four only on ILSA data (Baye & Monseur, 2016; Else-Quest et al., 2010; Gray et al., 2019; Keller et al., 2022), and the remaining two meta-analyses included both ILSA and non-ILSA data (Lietz, 2006; Siddiq & Scherer, 2019). To a large extent, PISA and TIMSS data were included in the six meta-analyses that extracted information from ILSA data, followed by PIRLS, SACMEQ, and ICILS data. The two meta-analyses that included ILSA and non-ILSA data side-by-side took a one-stage direct inclusion approach, that is, the authors considered the participating countries and/or ILSAs to be studies yielding multiple effect sizes. None of these meta-analyses considered meta-analytic models with dependency structures—in fact, only one of the eight meta-analyses addressed such structures explicitly via multilevel meta-analysis (Keller et al., 2022). Only Siddiq and Scherer (2019) performed sensitivity analyses comparing the one-stage inclusion approach with a two-stage inclusion approach. The latter was based on two steps of meta-analysis: First, ILSA data were meta-analyzed, and the resultant weighted average effect size was extracted as a representative of the effects from ILSA studies. Second, this effect size was combined with the non-ILSA data and then meta-analyzed. Finally, the gender meta-analyses reported mainly standardized mean differences as effect sizes ($m = 7$), along with variance ratios ($m = 2$). In sum, six of the eight meta-analyses utilized ILSA data, only two of which directly included ILSA and non-ILSA data.

Meta-analyses of the relation between SES and student achievement

The sample of $m = 10$ meta-analyses describing the relation between SES and student achievement yielded 1631 effect sizes based on 556 studies (see Table 3). These effect sizes were mainly reported as correlations ($m = 9$) and in only one meta-analysis as a standardized mean difference. The meta-analyses covered a broad range of achievement

Table 2 Characteristics of the meta-analyses on the gender differences in student achievement

Reference	n	k	Achievement domain(s)	Data	Type(s) of ILSA	Inclusion approach	Addressing dependence	Sensitivity analyses	Pooled effect size(s)
Baye and Monseur, (2016)	36	1654	Mathematics, Science, Reading	Only ILSA data	PISA 2000, PISA 2003, PISA 2006, PISA 2009, PISA 2012, PIRLS 2001, PIRLS 2006, PIRLS 2011, TIMSS 1995, TIMSS 1999, TIMSS 2003, TIMSS 2007	Only ILSA data	No	No	$d = -0.06$–0.34; $VR = 1.12$–1.15
Else-Quest et al., (2010)	2	476	Mathematics	Only ILSA data	TIMSS 2003, PISA 2003	Only ILSA data	No	No	$d = -0.42$–0.40
Gray et al., (2019)	102	2609	Mathematics, Science, Reading	Only ILSA data	PISA 2000, PISA 2003, PISA 2006, PISA 2009, PISA 2012, PISA 2015, PIRLS 2001, PIRLS 2006, PIRLS 2011, TIMSS 1995, TIMSS 1999, TIMSS 2003, TIMSS 2007, TIMSS 2011, TIMSS 2015	Only ILSA data	No	No	$VR = 1.118$
Keller et al., (2022)[#]	6	1028	Mathematics, Science, Reading	Only ILSA data	PISA 2000, PISA 2003, PISA 2006, PISA 2009, PISA 2012, PISA 2015	Only ILSA data	Yes (three-level RE Model)	No	$d = -0.23$–0.05
Lietz, (2006)	31	139	Reading	ILSA and Non-ILSA	IEA RC 1970/71, IEA RL 1990/91, PISA 2000	One-stage direct inclusion	No	No	$d = 0.19$
Lindberg et al., (2010)	242	441	Mathematics	Only Non-ILSA data	-	None	No	–	$d = -0.15$–0.22
Ouma & Nam, (2015)	6	35	Mathematics, Reading	Only Non-ILSA data	SACMEQ I 2003, SACMEQ I 2006, SACMEQ I 2008; SACMEQ II 2005, SACMEQ II 2007, SACMEQ II 2009	One-stage direct inclusion	No	–	$d = -0.06$–0.08
Siddiq and Scherer, (2019)	23	46	Digital literacy	ILSA and Non-ILSA	ICILS 2013	One-stage direct inclusion	No	Yes (two-stage direction inclusion)	$g = 0.12$

n = Number of primary studies, k = Number of effect sizes, ILSA = International large-scale assessment, d = Cohen's d, g = Hedges' g, VR = Variance ratio, PISA = Programme for International Student Assessment, TIMSS = Trends in International Mathematics and Science Study, ICILS = International Computer and Information Literacy Study, PIRLS = Progress in International Reading Literacy Study, SACMEQ = Southern and Eastern Africa Consortium for Monitoring Educational Quality, IEA = International Association for the Evaluation of Educational Achievement, RC = Reading Comprehension Study, RL = Reading Literacy Study, RE Model = Random-effects model. [#] This study was originally published as a preprint in 2020 (https://doi.org/10.31234/osf.io/73wap)

Table 3 Characteristics of the meta-analyses on the relation between socioeconomic status and student achievement

Reference	n	k	Achievement domain(s)	SES measure(s)	Data	Type(s) of ILSA	Inclusion approach	Addressing dependence	Sensitivity analyses	Pooled effect size(s)
Harwell et al., (2017)	143	297	Mathematics, Science, Literacy, General disciplines	Parent education, parent occupation, home resources, free/reduced-price lunches, composite SES indices	Non-ILSA data	–	–	No	No	$r = 0.12–0.26$
Kim et al., (2019)	49	49	General achievement, Educational attainment	Parent education, parent occupation, home resources and wealth indices, composite SES indices	ILSA and non-ILSA data	PISA, TIMSS, SACMEQ, MICS, Young Lives International Longitudinal Study	One-stage direct inclusion	No	No	$r = 0.08–0.20$
Letourneau et al., (2011)	7	7	Literacy, General cognitive skills	Parent education, parent occupation, home resources and wealth indices, composite SES indices	Non-ILSA data	–	–	No	No	$g = 0.35$
Liu et al., (2020)	78	480	Mathematics, Science, Chinese, English, General achievement	Parent education, parent occupation, home resources and wealth indices, income, composite SES indices	ILSA and non-ILSA data	PISA 2009, PISA 2012, PISA 2015	One-stage direct inclusion	Yes (robust variance estimation)	No	$r = 0.18–0.29$
Rodríguez-Hernández et al., (2020)	23	23	General achievement, Persistence	Parent education, parent occupation, parent income, home resources, neighborhood resources	Non-ILSA data	–	–	No	No	$r = 0.06$

Table 3 (continued)

Reference	n	k	Achievement domain(s)	SES measure(s)	Data	Type(s) of ILSA	Inclusion approach	Addressing dependence	Sensitivity analyses	Pooled effect size(s)
Scherer and Siddiq, (2019)	11	75	Digital literacy	Parent education, parent occupation, home resources and wealth indices, income, composite SES indices	ILSA and non-ILSA data	ICILS 2013	One-stage direct inclusion	Yes (three-level random-effects modeling)	Yes (two-stage direct inclusion, exclusion of ILSA data)	*r* = 0.18–0.25
Sirin, (2005)	58	167	Mathematics, Science, Literacy, General disciplines	Parent education, parent occupation, home resources and wealth indices, income, neighborhood, free or reduced-price lunches, home	Non-ILSA data	–	–	No	–	*r* = 0.25–0.51
Tan, (2017)	52	155	Mathematics, Science, Literacy	Home educational resources, cultural participation, reading at home or outside school, parent–child cultural discussions, educational expectations, parent home and school involvement, parent educational attainment	ILSA and non-ILSA data	TIMSS 1999	One-stage direct inclusion	No	No	*r* = -0.02–0.38

Table 3 (continued)

Reference	n	k	Achievement domain(s)	SES measure(s)	Data	Type(s) of ILSA	Inclusion approach	Addressing dependence	Sensitivity analyses	Pooled effect size(s)
Tan et al. (2019)	105	190	Mathematics, Science, Literacy	Home educational resources, cultural participation, reading at home or outside school, parent–child cultural discussions, educational expectations, parent home and school involvement, parent educational attainment, composite SES indices	ILSA and non-ILSA data	TIMSS 1999	One-stage direct inclusion	No	No	$r=0.00–0.37$
Van Ewijk et al., (2010)	30	188	Mathematics, Science, Literacy, Social Sciences, Economics, General achievement	Parent education, parent occupation, home resources, composite SES indices	ILSA and non-ILSA data	PISA 2000, PISA 2003	One-stage direct inclusion	Yes (weighted meta-regression model)	Yes (exclusion of ILSA data)	$r=0.32$

n = Number of primary studies, *k* = Number of effect sizes, ILSA = International large-scale assessment, *g* = Hedges' *g*, *r* = Pearson correlation, SES = Socioeconomic status domains, including literacy ($m = 7$), mathematics ($m = 6$), science ($m = 6$), general cognitive skills ($m = 6$), social sciences ($m = 1$), and digital literacy ($m = 1$), some of which were assessed not only by achievement tests but also school grades. The SES measures covered multiple dimensions, including parents' income, occupation, and education, in all meta-analyses. Four meta-analyses were based on non-ILSA data and did not provide any reason for this exclusion (Harwell et al., 2017; Letourneau et al., 2013; Rodríguez-Hernández et al., 2020; Sirin, 2005), while six included both ILSA and non-ILSA data (Kim et al., 2019; Liu et al., 2020; Scherer & Siddiq, 2019; Tan, 2017; Tan et al., 2019; van Ewijk & Sleegers, 2010). None of the meta-analyses were based only on ILSA data. Primarily, the meta-analysts chose the PISA, TIMSS, ICILS, and SACMEQ data to inform their meta-analyses and consistently took a one-stage direct inclusion approach, considering the countries or ILSA cycles as separate studies. Two meta-analyses reported sensitivity analyses: Scherer and Siddiq (2019) compared the one-stage direct inclusion with the two-stage direct inclusion and examined the effects of excluding ILSA data; van Ewijk & Sleegers (2010) also examined the effects of excluding ILSA data. Accounting for the dependencies among multiple effect sizes per study, Liu et al. (2020) performed robust variance estimation, Scherer and Siddiq (2019) and Keller et al. (2022) conducted three-level meta-analysis, and van Ewijk and Sleegers (2010) modified the weights in a meta-regression model similar to the robust variance estimation. In sum, six of the ten SES meta-analyses included ILSA data next to non-ILSA data utilizing mainly the one-stage direct inclusion.

Summary of key findings

Our systematic review of meta-analyses on gender differences in student achievement and the relation between SES and achievement indicated that (a) ILSA data were not eligible for all meta-analyses on these topics, for instance, due to misfit of the target samples, types of achievement measures, or the focus on national rather than international assessment data; (b) several meta-analyses included ILSA data, yet to different degrees (i.e., ILSA data only, ILSA and non-ILSA data side-by-side); (c) meta-analysts mostly took the one-stage direct inclusion approach, yet hardly considered alternative approaches and sensitivity analyses; (d) the structure of the meta-analytic data sets with multiple effect sizes per study was hardly considered.

Illustrative example: Gender differences in digital literacy

In the following, we illustrate the application of the inclusion approaches and show how to implement them. Additional file 4: S4 and Additional file 5: S5 contain the R code, the detailed analytic steps (see also Table 1), and the respective results.

Meta-analytic data set and aims

Siddiq's & Scherer's (2019) original meta-analysis contained 23 primary studies yielding 46 standardized mean effect sizes and included the data from ICILS (International Computer and Information Literacy Study) 2013. We updated this meta-analysis by adding the openly available data from ICILS 2018 (Fraillon et al., 2020). We performed this update for several reasons: First, it increased the number of effect sizes and, ultimately, the statistical power to detect gender differences and possible moderator effects. Second, ICILS 2018 contained a different set of participating countries, and, by including it, we extended the range of educational systems, cultures, and languages to test some hypotheses on the moderating effects of cultural orientation (i.e., power distance index) and innovation (i.e., global innovation index). Third, meta-analysts may be able to include several ILSAs or ILSA cycles rather than only one. In this sense, our illustrative example mimics, to some extent, a typical inclusion scenario, and we use it to showcase the resultant complexities of the meta-analytic data.

Ultimately, this data set contained 24 primary studies and 59 effect sizes. Hedges' g represented the standardized mean differences between girls and boys with positive effect sizes indicating higher performance scores for girls. In the original study, the authors aimed to quantify an overall effect size ($\overline{g}$), the between-study heterogeneity (τ^2), and the moderator effects of, for instance, test fairness (*0 = test fairness was not examined, 1 = test fairness was examined*) and publication type (*0 = published, 1 = grey literature*). Illustrating the inclusion approaches, we addressed these aims and further examined whether two country-level variables, Power Distance Index (PDI; see Hofstede, 2001) and the Global Innovation Index were additional moderators (GII; see Cornell University et al., 2020). Given that we relied on an updated data set with ICILS 2013 and 2018 data included, the meta-analytic models we used to estimate the weighted average effect size were more complex than the ones used in the original publication. Specifically, we used multilevel meta-analytic models and quantified multiple sources of heterogeneity—hence, we report multiple variance estimates at different levels of analysis (e.g., within studies $\tau^2_{(2)}$, between studies $\tau^2_{(3)}$, between countries $\tau^2_{(4)}$). We represented the proportion of non-random variance that is due to heterogeneity by the I^2 value and the degree of inconsistency by Cochrane's Q statistic (Borenstein et al., 2009). Additional file 1: S1 contains the data and describes how these variables were derived.

Separate meta-analysis

Performing separate meta-analyses via random-effects modeling, we obtained estimates of the weighted average effect sizes for the ICILS 2013, ICILS 2018, the combined ICILS 2013 and 2018, and the non-ILSA data sets. Table 4 shows these estimates, which ranged between $\overline{g} = 0.12$ and 0.21 and exhibited heterogeneity between the samples within these data sets. Notably, the overall effect size of the ICILS 2013 data was comparable to that of the non-ILSA data ($z = -0.3$, $p = 0.76$); yet, the ICILS 2018 data showed a significantly higher overall effect ($z = -1.8$, $p = 0.07$). The degree of heterogeneity varied between these data sets (see also Fig. 4): While the non-ILSA effect sizes varied substantially ($\tau^2_{(2)} = 0.033$, $I^2 = 95.4\%$), the effect sizes for the ICILS 2018 varied less ($\tau^2_{(2)} = 0.012$, $I^2 = 91.2\%$), and varied the least for the ICILS 2013 data ($\tau^2_{(2)} = 0.005$, $I^2 = 78.2\%$). We extended the random-effects model by adding the variables test fairness, publication status, power distance, and global innovation to the non-ILSA data. For the non-ILSA data, publication status negatively moderated the gender differences, with grey literature exhibiting smaller effects, and test fairness positively moderated these differences, with larger effects for studies examining test fairness (Table 5). For the ICILS 2018 and the combined ICILS data, more innovative countries exhibited significantly larger gender effects; this moderation effect was not apparent for ICILS 2013.

Table 4 Results of the random-effects meta-analyses of the gender differences in digital literacy

	Separate meta-analyses				**Direct inclusion**			**Indirect inclusion**
	ICILS 2013 data	**ICILS 2018 data**	**ICILS 2013 & 2018 data**	**Non-ILSA data**	**One-stage inclusion**	**One-stage inclusion**	**Two-stage inclusion**	**Bayesian meta-analysis**
Weighted average effect size								
$\overline{g}$	0.13	0.21	0.16	0.12	0.13	0.10	0.12	0.12
95% CI	[0.10, 0.17]	[0.15, 0.27]	[0.12, 0.19]	[0.04, 0.20]	[0.05, 0.21]	[0.01, 0.18]	[0.05, 0.19]	[0.03, 0.20]
k	21	13	34	25	59	59	27	25
m	21	13	28	25	24	24	27	25
Heterogeneity								
$\tau^2_{(2)}$	0.005	0.012	0.003	0.033	0.007	0.016	0.030	0.036
95% CI	[0.002, 0.012]	[0.005, 0.033]	[0.000, 0.013]	[0.016, 0.074]	[0.004, 0.013]	[0.000, 0.079]	[0.014, 0.065]	[0.017, 0.078]
$\tau^2_{(3)}$	–	–	–	–	0.027	0.031	–	–
95% CI	–	–	–	–	[0.007, 0.071]	[0.011, 0.077]	–	–
$\tau^2_{(4)}$	–	–	0.006	–	–	0.006	–	–
95% CI	–	–	[0.000, 0.015]	–	–	[0.000, 0.012]	–	–
Baseline model	REM2	REM2	REM3	REM2	REM3	CCREM4	REM2	REM2

ILSA = International large-scale assessment, $\overline{g}$ = Weighted average effect size of gender differences (Hedges' *g*), *k* = Number of effect sizes (samples), *m* = Number of studies, $\tau^2_{(2)}$ = Within-study (between-sample) heterogeneity, $\tau^2_{(3)}$ = Between-study heterogeneity, $\tau^2_{(4)}$ = Between-country heterogeneity, REM2 = Two-level random-effects model, REM3 = Three-level random-effects model, CCREM4 = Four-level cross-classified random-effects model (with 31 countries). (2)-(4) refer to the level of analysis

Indirect inclusion via *Bayesian meta-analysis*

Utilizing the information from the separate meta-analyses, we conducted Bayesian meta-analysis for the non-ILSA data with informative priors on the weighted average effect and the heterogeneity estimates—these priors were based on the effect size and variance estimate of the combined ICILS 2013 and 2018 data (for the detailed specification of the priors, see Additional file 5: S5). The overall effect size was $\overline{g} = 0.12$, with a 95% confidence interval similar to the effect for the non-ILSA data and a between-sample variance of $\tau^2_{(2)} = 0.036$ (Table 4). The Potential Scale Reduction Factors $\widehat{R}^2$ of the model parameters were all below 1.01, and the simulated distributions were similar to the observed distributions in the posterior predictive checks (see Additional file 5: S5). Moreover, the Monte Carlo Markov Chains showed a stable pattern without any clear trends or systematic changes over time and scattered around the model parameter estimates (see the trace plots in Additional file 5: S5). These observations supported that the meta-analytic model had converged and that stable estimates were obtained (Harrer et al., 2022). Moreover, varying the prior distributions did not show substantial sensitivity of the Bayesian effect size and variance estimates. Similar to the separate meta-analysis of the non-ILSA, the publication status moderated the gender differences in digital literacy; yet not the test fairness (Table 5).

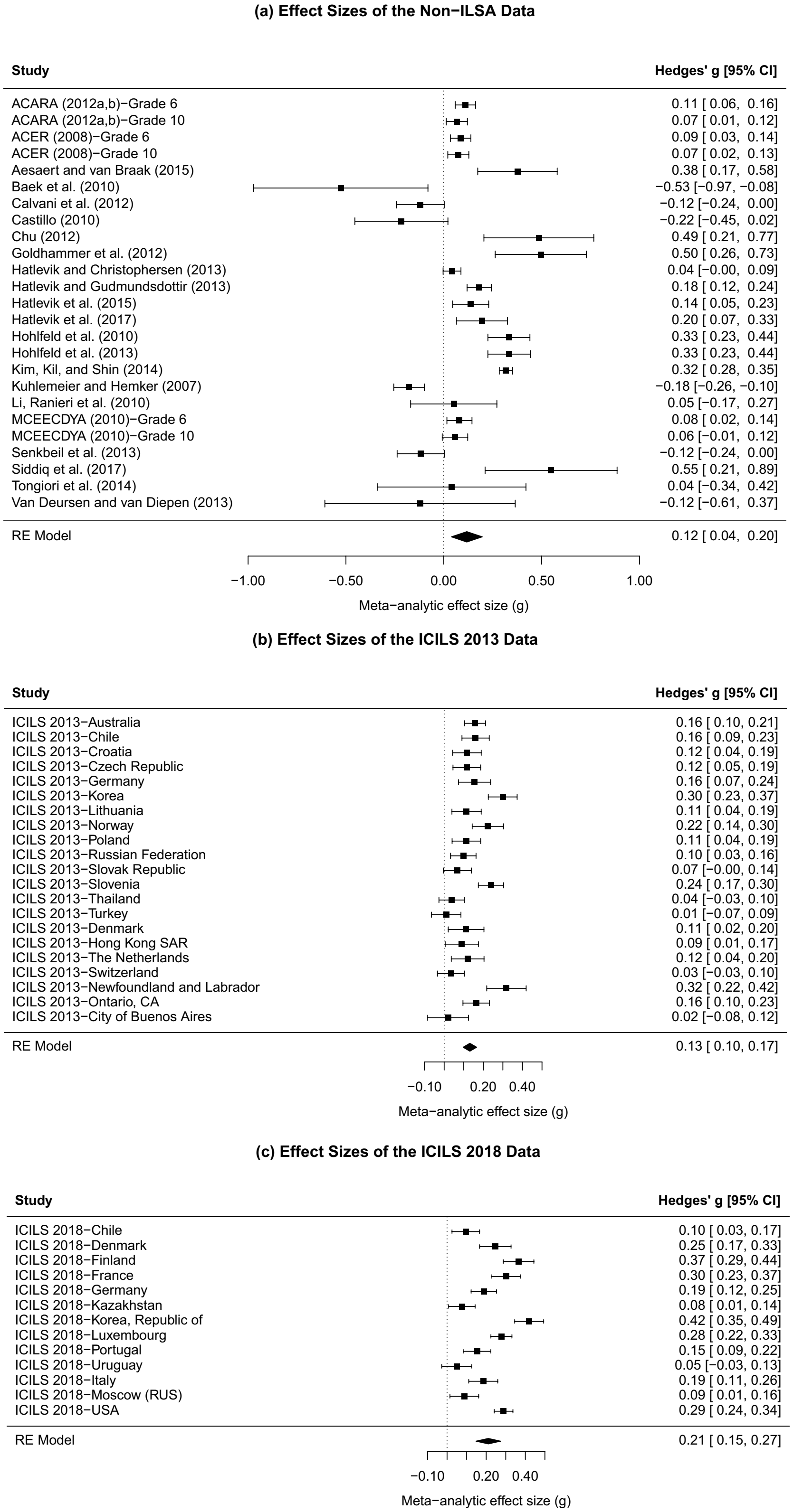


Fig. 4 Forest plots of the effect sizes for the ILSA and non-ILSA data. Note. The weighted average effect sizes were based on common (two-level) random-effects (RE) models. Positive standardized mean differences (Hedges' g) suggested that girls performed better than boys

Table 5 Results of the mixed-effects meta-regression analyses of the standardized mean differences across gender including study- and country-level moderators

	Separate meta-analyses				Direct inclusion			Indirect inclusion
	ICILS 2013 data	**ICILS 2018 data**	**ICILS 2013 & 2018 data**	**Non-ILSA data**	**One-stage inclusion**	**One-stage inclusion**	**Two-stage inclusion**	**Bayesian meta-analysis**
Regression coefficients *B* (95% CI)								
Intercept	− 0.03 [− 0.37, 0.30]	− 0.54[− 0.90, − .018]	− 0.17 [− 0.49, 0.14]	− 0.22 [− 1.49, 1.05]	− 0.30 [− 0.58, -0.02]	− 0.28 [− 0.57, 0.02]	0.12 [0.01, 0.22]	− 0.21 [− 1.57, 1.15]
Test fairness	–	–	–	0.20* [0.01, 0.39]	0.17 [− 0.01, 0.36]	0.18* [0.00, 0.36]	0.17* [0.00, 0.34]	0.20 [− 0.01, 0.41]
Publication type	–	–	–	− 0.24* [− 0.43, − 0.05]	− 0.21* [− 0.40, -0.01]	− 0.21* [− 0.41, -0.01]	− 0.22* [− 0.39, -0.05]	− 0.24* [− 0.45, − 0.04]
Power distance index (PDI)	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	0.00 [0.00, 0.00]	0.00 [0.00, 0.01]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	–	0.00 [0.00, 0.01]
Global innovation index (GII)	0.00 [0.00, 0.01]	0.01* [0.01, 0.02]	0.01* [0.00, 0.01]	0.00 [− 0.02, 0.02]	0.01* [0.00, 0.01]	0.01* [0.00, 0.01]	–	0.00 [− 0.02, 0.02]
Baseline model	MEM2	MEM2	MEM3	MEM2	MEM3	CCMEM4	MEM2	MEM2

ILSA = International large-scale assessment. The results of the two-stage direct inclusion are based on the pooling of the ICILS 2013 and 2018 data via random-effects meta-analyses in the first stage. The GII was derived as the country average of the global innovation indices across publication years. PDI and GII are country-level variables, while test fairness and publication type are study- level variables. MEM2 = Two-level mixed-effects meta-regression model, MEM3 = Three-level mixed-effects meta-regression model, CCMEM4 = Four-level cross-classified mixed-effects meta-regression model (with 31 countries). Positive standardized mean differences (Hedges' *g*) suggested that girls performed better than boys. * $p < .05$

One-stage direct inclusion

Directly combining the effect sizes obtained from ILSA and non-ILSA data resulted in a nested structure with multiple effect sizes per study. We therefore specified a three-level random-effects model addressing this structure (see Fig. 2b)—this model exhibited a significantly better fit to the meta-analytic data than a model ignoring the nesting (see Fig. 2a), $\chi^2(1) = 10.0$, $p = 0.002$. Moreover, the three-level model exhibited substantial within-study variation in addition to the between-study variation (see Table 4). The respective overall effect size was $\overline{g} = 0.13$ (95% CI [0.05, 0.21]) and showed significant heterogeneity ($Q_E[58] = 592.5$, $p < 0.001$). Adding the potential moderator variables resulted in a significant effect of publication status ($B = -\ 0.21$, $SE = 0.10$, $p = 0.04$) and global innovation ($B = 0.01$, $SE = 0.00$, $p < 0.001$; see Table 5). Overall, about 46% of the between-sample and 2% of the between-study variation could be explained. Moreover, the difference in gender effects between ILSA and non-ILSA data was insignificant, $B = 0.05$, $SE = 0.13$, $p = 0.72$. In our example, random-effect models with RVE only identified the moderating effect of the publication type (Additional file 5: S5).

Given that some countries in the samples contributed multiple effect sizes (e.g., to the ICILS 2013, ICILS 2018, and non-ILSA data), an additional level of nesting may exist. To examine the degree of possible between-country variation in the effect sizes, we extended the three-level model to a four-level cross-classified random-effects model (see Fig. 2c). This model exhibited a better fit than the three-level model ($\chi^2[1] = 4.2$, $p = 0.04$) and showed that between-country variation existed, in addition to within- and between-study variation (see Table 4). The corresponding effect size was $\overline{g} = 0.10$, 95% CI [0.01, 0.18]. Similar to the three-level model, the effects of publication status ($B = -\ 0.21$, $SE = 0.10$, $p = 0.04$) and global innovation existed ($B = 0.01$, $SE = 0.00$, $p < 0.001$; see Table 5). However, this model showed that most variance could be explained at the country level (49.1%), yet not the study level (2.9%).

Two-stage direct inclusion

Utilizing the weighted average effect size and variance estimates of the separate meta-analyses, we combined the non-ILSA effect sizes with one overall ICILS 2013 and one overall ICILS 2018 effect size. Estimating the random-effects model without a nested structure, we obtained an overall gender effect of $\overline{g} = 0.12$ (95% CI [0.05, 0.19]; see Table 4), and the moderation effect of publication status ($B = -\ 0.22$, $SE = 0.09$, $p = 0.01$; see Table 5). The effect of test fairness was statistically significant, $B = 0.17$, $SE = 0.08$, $p = 0.05$ (see Table 5). This additional moderation effect suggested that larger effects were exhibited for studies that examined test fairness, after controlling for the interactivity of the assessment tasks and the publication status. Finally, pooling the ILSA effect sizes via Borenstein et al.'s (2009) procedure in the first stage did not show any different results: The weighted average effect size was $\overline{g} = 0.12$ (95% CI [0.05, 0.19]), and the two moderator effects persisted (publication status: $B = -0.22$, $SE = 0.08$, $p = 0.01$; test fairness: $B = 0.17$, $SE = 0.08$, $p = 0.05$). Further analyses neither flagged the large ILSA-data effect sizes as influential (see Additional file 5: S5).

Summary of key findings

Across the direct inclusion approaches, the overall effect sizes were consistently small and positive. Notably, these gender differences favored girls and tended to be smaller than in more curricular oriented domains such as mathematics, science, and reading (for specific ranges of effect sizes, please see Additional file 1: S1). All approaches revealed the heterogeneity of the gender effects. The cross-classified model represented the data best for the one-stage direct inclusion and highlighted three additional sources of heterogeneity (next to sampling variation): samples within studies, studies, and countries. Next to the consistency of the fixed effects, the moderator effects of publication status were almost identical in direction and magnitude. Some differences however existed for test fairness and global innovation: The one-stage inclusion approach identified the GII moderation effect and located it to the country level—these effects did not exist when synthesizing only the non-ILSA or ICILS 2013 data. The two-stage inclusion approach and the separate meta-analysis of the non-ILSA data further indicated moderation by test fairness.

Discussion

Including ILSA data in meta-analyses in education

Our systematic review of the extent to which ILSA data were included in existing meta-analyses of gender differences or SES gaps in student achievement showed that ILSA data were not eligible for all meta-analyses. This may have been the main reason why their inclusion was limited. For instance, the seminal meta-analysis of gender differences in student achievement by Voyer & Voyer, (2014) focused solely on teacher-assigned grades as achievement measures and thus excluded ILSA data. Evaluating the eligibility of studies for inclusion also applies to ILSA data, and meta-analysts should carefully evaluate whether the ILSA samples, constructs, and study designs fit to their inclusion criteria and, ultimately, research purposes. Irrespective of the outcome of this evaluation, communicating the reasons for excluding ILSA data should be an integral part of the methodological rigor of meta-analyses in education (Pigott & Polanin, 2020). Moreover, given our review of the potential and the analytic opportunities associated with the inclusion of ILSA data in meta-analyses, we argue that searching the existing ILSA databases should become part of the meta-analytic standard procedures in education.

As noted earlier, one key issue of including ILSA data in meta-analyses lies in the methodological complexities these large-scale data may impose. As we have showcased while presenting the one-stage direct inclusion approach, the meta-analytic structure of the data that include ILSA and non-ILSA effect sizes can become complex, with hierarchical or even cross-classified structures. While modeling such structures may shed light on the possible sources of variation and the level at which moderators operate, the underlying meta-analytic models are advanced (Fernández-Castilla et al., 2020)—this may have been one reason why most meta-analysts refrained from addressing such complex data structures in their meta-analyses of the gender differences and SES gaps in student achievement. Our extension of the meta-analysis of the gender differences in digital literacy included two ILSAs and thus required meta-analytic models accounting for the multiple effect sizes per study and country. Meta-analysts should be aware which structure their meta-analytic data set including ILSA and non-ILSA data exhibits to obtain accurate estimates of fixed and random effects (Fernández-Castilla et al., 2020).

Another complexity is associated with the decision of which type of ILSA data are included, primary or secondary data? Given the availability of most ILSA data, meta-analysts do not need to rely on the results reported in secondary ILSA data analyses, yet can compute the effect sizes themselves. Although appealing, this opportunity requires that meta-analysts must be aware of the methodological complexities of the primary ILSA data and that they can address them analytically (Rutkowski et al., 2010). Hence, we see the need for training meta-analysts in both the analysis of primary ILSA data to derive the correct effect size estimates and the inclusion approaches for meta-analyses.

Concerning the four inclusion approaches, notably, our illustrative example showed consistently small estimates of the weighted average gender effect size. With the exception of the separate meta-analysis of the ICILS 2018 data, the estimates were comparable and did not lead to another conclusion. Nonetheless, we refrain from generalizing this result—in other context, with other measures and effect size, and for a different set of ILSA or non-ILSA data, the fixed effects may indeed vary considerably, especially when meta-analyzed separately (Gray et al., 2019). At the same time, some specifications within the inclusion approaches were homogeneous. For instance, the overall gender effects were identical for the separate meta-analysis of the non-ILSA data and the indirect inclusion in our study—in fact, both approaches focused on the non-ILSA data and differed only in the extent to which information from the ILSA data was incorporated (e.g., Röver, 2020). Similarly, the different direct inclusion approaches agreed on the size of the pooled effect.

Recommendations for including ILSA data in meta-analyses

Considering the marginal differences in meta-analytic findings in our illustrative example, meta-analysts may well argue that the choice of the specific approaches may not matter for the reporting of the overall effects. However, some of these approaches are more useful than others, especially for quantifying the sources of variation and the moderator effects (Fernández-Castilla et al., 2020), and we recommend that meta-analysts choose an approach in light of the goals of their study.

First, we recommend to meta-analysts who wish to compare the effects obtained from non-ILSA studies to ILSA data to conduct separate meta-analyses of these two types of data. This approach facilitates the benchmarking and interpreting of effect sizes from non-ILSA data (Wagemaker, 2016). Moreover, we argue that conducting separate meta-analyses could also provide initial insights into the potential similarities and differences of effects across data sources and may, at the least, serve as form of robustness check for the other approaches.

Second, if the purpose of a meta-analysis is to synthesize evidence from non-ILSA data sources (e.g., due to some substantively motivated inclusion criterion), we recommend considering an indirect inclusion approach. Without influencing the core meta-analytic findings or choices of data, such an approach can inform and potentially improve the estimates of the heterogeneity estimates by incorporating the knowledge about such parameters in ILSAs (Brunner et al., 2018).

Third, in situations where the heterogeneity and possible moderator effects for ILSA and non-ILSA data are the primary interest, we recommend taking a *direct inclusion* approach. Both the one- and two-stage direct inclusion can shed light on between-study heterogeneity and moderation by study-level features. The one-stage approach can further include between-country heterogeneity and country-level moderation effects (see also Cheng et al., 2018). Via direct inclusion, meta-analysts can test specific hypotheses on which factors at which levels of analyses may explain the heterogeneity of the effects. Moreover, they can compare directly via subgroup or moderator analyses to what extent the type of data (i.e., ILSA vs. non-ILSA data) also explains heterogeneity. In this sense, the direct inclusion approaches offer several analytic possibilities to quantify and explore heterogeneity, which is why we considered them to be the preferred choice in meta-analyses in education.

Each of the steps within the inclusion approaches should be documented, and the analytic decisions within justified (Pigott & Polanin, 2020). Once the eligible primary studies and ILSA data sets have been identified, the following analytic aspects are key when meta-analyzing non-ILSA and ILSA data side-by-side:

- *Generate the effect sizes from the primary data incorporating the complex sampling survey design features.* As noted earlier, the correct effect size and sampling variance estimates must be derived from both the non-ILSA and ILSA data. For the latter, both adjustments of effect sizes and the re-analysis of the raw data are largely available—the ILSA official reports already contain some effect sizes that are based on the complex survey design (e.g., gender differences, relations between SES and achievement). Meta-analysts should clearly communicate the ways in which they derived the effect size measures, their sampling variances and covariances, and how they dealt with the complex survey design features of ILSAs, such as weighting, multi-stage and cluster sampling, rotated questionnaire designs, and stratification. Moreover, if IPD sets are analyzed and model-based effect sizes are estimated, the analytic modeling procedures should be mimicked across ILSAs or ILSA cycles, so that effect sizes are comparable and have the same meaning.
- *Indicate the structure of the meta-analytic data.* Despite the nested structure of the primary data (e.g., students nested in classrooms or schools), meta-analytic data can also follow complex structures (e.g., multiple effect sizes nested in studies or ILSAs; see Fig. 2). To derive overall estimates of a weighted average effect, meta-analytic models that account for this structure are needed (Fernández-Castilla et al., 2020). Meta-analysts should identify the structure of their data and select the respective meta-analytic models (e.g., multilevel meta-analysis, robust variance estimation). Selecting one effect size per ILSA is not recommended.
- *Choose an inclusion approach based on the research questions and goals.* As we reviewed the inclusion approaches in our framework, we identify both their strengths and weaknesses. Meta-analysts should carefully consider them and decide for an approach in light of their research questions and purposes. For instance, if only small-scale primary studies are in the focus, ILSA data may only inform the meta-analysis via an indirect inclusion approach. If differences between studies with

random versus convenience samples are in the focus, both non-ILSA and ILSA data may inform the meta-analysis via a direct inclusion approach.

- *Conduct sensitivity analyses.* Sensitivity analyses can shed light on the impact the inclusion of ILSA data in the meta-analysis of non-ILSA data may have on the substantive findings and estimates. Moreover, they indicate the robustness of the specific inclusion approach researchers have taken.
- *Report the analytic steps and decisions transparently.* We encourage meta-analysts to document each of the analytic steps and decisions and share their analytic code to facilitate transparency and possible updates of their meta-analyses. This is especially relevant for replicating the model-based generation of effect sizes accounting for the complexities of the primary ILSA data (IPD) and the meta-analytic models accounting for the complexities of the secondary (meta-analytic) data.

Limitations and future directions

The present study has several limitations: First, the two systematic reviews provide information about the inclusion of ILSA data in meta-analyses for the two selected topics (i.e., gender differences and the relation between SES and achievement). Although these topics concern key issues in education (e.g., OECD, 2016), especially in the context of equity and equality, the respective findings may not be fully generalizable. In this sense, we encourage researchers to consider extending these reviews into other, educationally relevant topics.

Second, our study reviewed the advantages and challenges associated with the application of four inclusion approaches, yet did not examine their performance in large-scale meta-analyses and simulations. Knowledge about their performance, especially their efficiency, bias, and the precision of the meta-analytic estimates, could further guide the decisions for one or the other approach.

Third, our review focused on situations in which ILSA and non-ILSA data are combined. However, in practice, meta-analysts may also face situations in which only ILSA data are combined meta-analytically, for example, from multiple ILSAs and ILSA cycles. Such situations offer the possibility to generate effect sizes and sampling (co-)variances from the same kind of analytic model. Recently, some ways to meta-analyze only ILSA data have been proposed (Brunner et al., 2022; Campos et al., 2023) with respective examples (e.g., Blömeke et al., 2021; Keller et al., 2022).

Conclusions

Overall, we argue that ILSA data hold great potential for informing meta-analyses in education, especially due to their rigorous study and sampling designs, the availability of indicators describing educational systems at multiple levels, and their focus on key issues and constructs in education. This potential may not only assist meta-analysts in expanding their data sets and ultimately improve the precision of the meta-analytic estimates, but also reduce possible publication, cultural, and methodological bias. Another key advantage is that the primary ILSA data are almost entirely available to meta-analysts, who can define and implement the analytic models themselves, yielding effect sizes based on complex survey design directly. At the same time, including ILSA data requires a careful choice of an appropriate methodological approach and may extend the analytic steps involved in a meta-analysis by further sensitivity and moderator analyses. Moreover, the complex structure of both the primary ILSA and the resultant meta-analytic ILSA and non-ILSA data must be addressed.

Our paper describes four ILSA data inclusion approaches, outlines the steps meta-analysts may take to examine the possible effects of including ILSA data in their meta-analyses, and provides information on their potential, challenges, and fit to the specific research purposes. We believe that this framework of approaches informs and stimulates the inclusion of ILSA data in meta-analyses on key issues in education to ultimately improve the quality, precision, and informativeness of research evidence.

Abbreviations

ERIC	Education resources information center
ESCS	Economic, social and cultural status
GII	Global innovation index
HISEI	Highest international socio-economic index of occupational status
ICC1	Intraclass correlation coefficient
ICILS	International computer and information literacy study
IEA	International association for the evaluation of educational achievement
ILSA	International large-scale assessment
IPD	Individual-participant data
Non-ILSA	Studies other than international large-scale assessments
OECD	Organisation for economic co-operation and development
PDI	Power distance index
PIRLS	Progress in international reading literacy study
PISA	Programme for international student assessment
SACMEQ	The Southern and Eastern Africa Consortium for Monitoring Educational Quality
SES	Socioeconomic status
SMD	Standardized mean difference
TALIS	Teaching and learning international survey
TIMSS	Trends in international mathematics and science study
UNESCO	United Nations Educational, Scientific and Cultural Organization

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-024-00191-1.

Additional file 1. Gender Differences in Academic Achievement.

Additional file 2. The Relationship between Socioeconomic Status and Academic Achievement .

Additional file 3. Primary Study Data.

Additional file 4. Meta-Analysis of Gender Differences in Digital Skills: Separate Meta-Analyses.

Additional file 5. Meta-Analysis of Gender Differences in Digital Skills: Direct and Indirect Inclusion Approaches.

Additional file 6. Search Strategies, Screening, and Included References.

Acknowledgements
Not applicable.

Author contributions
RS: Conceptualization, methodology, software, formal analysis, investigation, resources, data curation, writing-original draft, writing-review & editing, visualization; TN: Writing-original draft, writing-review & editing; FS: Writing-original draft, writing-review & editing.

Funding
This work was supported by the Centres of Excellence scheme, funded by the Research Council of Norway, project number 331640.

Availability of data and materials
The IPD analysed during the current study are available in the IEA ICILS data repository, https://www.iea.nl/data-tools/repository/icils. The meta-analytic dataset can be accessed via the Open Science Framework (OSF) at https://doi.org/https://doi.org/10.17605/OSF.IO/NT5QJ and is made available as supplementary material. The manuscript has been made available as a preprint via MetaArXiv at https://doi.org/https://doi.org/10.31222/osf.io/hcjqm.

Declarations

Ethics approval and consent to participate

This review article utilizes the primary study data of the IEA ICILS and the secondary study data published in the eligible articles. Ethics approval and consent to participate concerning ICILS were organized and given by the national ICILS centres in the participating countries and conformed to the IEA ethical standards. For more details, please refer to the respective ICILS documentation (e.g., the ICILS 2013 and 2018 technical reports). The secondary study data were based summary data of the primary study data sets, so that no additional ethics approval or consent were required.

Consent for publication

We provide our consent to publish this manuscript upon acceptance in the Springer open-access journal "Large-scale Assessments in Education". No further consent is required.

Competing interests

The authors declare that they have no competing interests.

Received: 29 October 2022 Accepted: 25 January 2024
Published online: 14 February 2024

References

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A Review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research, 82*(4), 436–476. https://doi.org/10.3102/0034654312458162

Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-Scale Assessments in Education, 4*(1), 1–16. https://doi.org/10.1186/s40536-015-0015-x

Becker, B. J., & Wu, M.-J. (2007). The synthesis of regression slopes in meta-analysis. *Statistics Science, 22*(3), 414–429. https://doi.org/10.1214/07-STS243

Berkowitz, R., Moore, H., Astor, R. A., & Benbenishty, R. (2017). A research synthesis of the associations between socioeconomic background, inequality, school climate, and academic achievement. *Review of Educational Research, 87*(2), 425–469. https://doi.org/10.3102/0034654316669821

Blömeke, S., Nilsen, T., & Scherer, R. (2021). School innovativeness is associated with enhanced teacher collaboration, innovative classroom practices, and job satisfaction. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000668

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken: Wiley.

Braun, H. I., & Singer, J. D. (2019). Assessment for monitoring of education systems: international comparisons. *The ANNALS of the American Academy of Political and Social Science, 683*(1), 75–92. https://doi.org/10.1177/0002716219843804

Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic inequality and educational outcomes*. Springer.

Brunner, M., Keller, L., Stallasch, S. E., Kretschmann, J., Hasl, A., Preckel, F., Lüdtke, O., & Hedges, L. V. (2022). Meta-analyzing individual participant data from studies with complex survey designs: A tutorial on using the two-stage approach for data from educational large-scale assessments. *Research Synthesis Methods*. https://doi.org/10.1002/jrsm.1584

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

Burke, D. L., Ensor, J., & Riley, R. D. (2017). Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ. *Statistics in Medicine, 36*(5), 855–875. https://doi.org/10.1002/sim.7141

Campos, D. G., Cheung, M.W.-L., & Scherer, R. (2023). A primer on synthesizing individual participant data obtained from complex sampling surveys: A two-stage IPD meta-analysis approach. *Psychological Methods*. https://doi.org/10.1037/met0000539

Cheng, C., Cheung, M.W.-L., & Wang, H.-Y. (2018). Multinational comparison of internet gaming disorder and psychosocial problems versus well-being: Meta-analysis of 20 countries. *Computers in Human Behavior, 88*, 153–167. https://doi.org/10.1016/j.chb.2018.06.033

Cheung, M.W.-L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review, 29*(4), 387–396. https://doi.org/10.1007/s11065-019-09415-6

Cheung, M.W.-L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.00738

Cornell University, INSEAD, & WIPO. (2020). Global Innovation Index 2020: Who will finance innovation? Cornell University, INSEAD, and the World Intellectual Property Organization. https://www.globalinnovationindex.org/Home

Declercq, L., Jamshidi, L., Fernández Castilla, B., Moeyaert, M., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2020). Multilevel meta-analysis of individual participant data of single-case experimental designs: one-stage versus two-stage methods. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2020.1822148

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103–127. https://doi.org/10.1037/a0018053

Espinoza, O. (2007). Solving the equity–equality conceptual dilemma: A new model for analysis of the educational process. *Educational Research, 49*(4), 343–363. https://doi.org/10.1080/00131880701717198

Fernández-Castilla, B., Jamshidi, L., Declercq, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). The application of meta-analytic (multi-level) models with multiple random effects: A systematic review. *Behavior Research Methods, 52*, 2031–2052. https://doi.org/10.3758/s13428-020-01373-9

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). IEA International Computer and Information Literacy Study 2018 Technical Report. IEA. https://www.iea.nl/sites/default/files/2020-05/ICILS%202018%20Technical%20Report-FINAL_0.pdf

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

Gray, H., Lyth, A., McKenna, C., Stothard, S., Tymms, P., & Copping, L. (2019). Sex differences in variability across nations in reading, mathematics and science. *Large-Scale Assessments in Education, 7*(1), 1–29. https://doi.org/10.1186/s40536-019-0070-9

Gustafsson, J.-E. (2018). International large scale assessments: Current status and ways forward. *Scandinavian Journal of Educational Research, 62*(3), 328–332. https://doi.org/10.1080/00313831.2018.1443573

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2022). Doing Meta-Analysis in R: A Hands-on Guide. PROTECT Lab. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

Harwell, M., Maeda, Y., Bishop, K., & Xie, A. (2017). The surprisingly modest relationship between SES and educational achievement. *Journal of Experimental Education, 85*(2), 197–214. https://doi.org/10.1080/00220973.2015.1123668

Hattie, J., Rogers, H. J., & Swaminathan, H. (2014). The role of meta-analysis in educational research. In: A. D. Reid, E. P. Hart, & M. A. Peters (Eds). A companion to research in education. Springer Netherlands. 197-207. https://doi.org/10.1007/978-94-007-6809-3_26

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). Cochrane handbook for systematic reviews of interventions. *Wiley*. https://doi.org/10.1002/9781119536604

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, Institutions and Organizations Across Nations. Thousand Oaks.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research, 62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

Kaplan, D., Chen, J., Yavuz, S., & Lyu, W. (2023). Bayesian dynamic borrowing of historical information with applications to the analysis of large-scale assessments. *Psychometrika, 88*(1), 1–30. https://doi.org/10.1007/s11336-022-09869-3

Keller, L., Preckel, F., Eccles, J., & Brunner, M. (2022). Top-performing math students in 82 countries: A meta-analysis of gender differences in achievement, achievement profiles, and achievement motivation. *Journal of Educational Psychology, 114*(5), 966–991. https://doi.org/10.1037/edu0000685

Kim, J.-S., Anderson, C. J., & Keller, B. (2012). Multilevel analysis of assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 390–425). CRC Press.

Kim, S. W., Cho, H., & Kim, L. Y. (2019). Socioeconomic status and academic outcomes in developing countries: A meta-analysis. *Review of Educational Research, 89*(6), 875–916. https://doi.org/10.3102/0034654319877155

Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 115–147). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-4629-9_7

Klieme, E. (2020). Policies and practices of assessment: A showcase for the use (and Misuse) of international large scale assessments in educational effectiveness research. In J. Hall, A. Lindorff, & P. Sammons (Eds.), *International perspectives in educational effectiveness research* (pp. 147–181). Cham: Springer. https://doi.org/10.1007/978-3-030-44810-3_7

Kuger, S., & Klieme, E. (2016). Dimensions of context assessment. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: an international perspective* (pp. 3–37). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-45357-6_1

Lai, M. H. C., & Kwok, O.-M. (2016). Estimating standardized effect sizes for two- and three-level partially nested data. *Multivariate Behavioral Research, 51*(6), 740–756. https://doi.org/10.1080/00273171.2016.1231606

Lenkeit, J., Chan, J., Hopfenbeck, T. N., & Baird, J.-A. (2015). A review of the representation of PIRLS related research in scientific journals. *Educational Research Review, 16*, 102–115. https://doi.org/10.1016/j.edurev.2015.10.002

Letourneau, N. L., Duffett-Leger, L., Levac, L., Watson, B., & Young-Morris, C. (2011). Socioeconomic status and child development: A meta-analysis. *Journal of Emotional and Behavioral Disorders, 21*(3), 211–224. https://doi.org/10.1177/1063426611421007

Letourneau, N. L., Duffett-Leger, L., Levac, L., Watson, B., & Young-Morris, C. (2011). Socioeconomic status andchild development: A meta-analysis. Journal of Emotional and Behavioral Disorders, *21*(3), 211–224.https://doi.org/10.1177/1063426611421007

Lietz, P. (2006). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation, 32*(4), 317–344. https://doi.org/10.1016/j.stueduc.2006.10.002

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123–1135. https://doi.org/10.1037/a0021276

Liu, J., Peng, P., & Luo, L. (2020). The relation between family socioeconomic status and academic achievement in China: A meta-analysis. *Educational Psychology Review, 32*, 49–76. https://doi.org/10.1007/s10648-019-09494-0

Lohmann, J. F., Zitzmann, S., Voelkle, M. C., & Hecht, M. (2022). A primer on continuous-time modeling in educational research: An exemplary application of a continuous-time latent curve model with structured residuals (CT-LCM-SR) to PISA Data. *Large-Scale Assessments in Education, 10*(1), 5. https://doi.org/10.1186/s40536-022-00126-8

Möller, J., Zitzmann, S., Helm, F., Machts, N., & Wolff, F. (2020). A meta-analysis of relations between achievement and self-concept. *Review of Educational Research, 90*(3), 376–419. https://doi.org/10.3102/0034654320919354

Morrison, A., Polisena, J., Husereau, D., Moulton, K., Clark, M., Fiander, M., Mierzwinski-Urban, M., Clifford, T., Hutton, B., & Rabb, D. (2012). The effect of English-language restriction on systematic review-based meta-analyses: A systematic review of empirical studies. *International Journal of Technology Assessment in Health Care, 28*(2), 138–144. https://doi.org/10.1017/s0266462312000086

Musu, L., Dohr, S., & Netten, A. (2020). Quality control during data collection: Refining for rigor. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (pp. 131–150). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_8

OECD. (2016). PISA 2015 Results (Volume I): excellence and equity in education. *OECD Publishing*. https://doi.org/10.1787/9789264266490-en

OECD. (2019). PISA 2018 Results (Volume II): where all students can succeed. *OECD Publishing*. https://doi.org/10.1787/b5fd1b8f-en

Oh, I.-S. (2020). Beyond meta-analysis: Secondary uses of meta-analytic data. *Annual Review of Organizational Psychology and Organizational Behavior, 7*(1), 125–153. https://doi.org/10.1146/annurev-orgpsych-012119-045006

Ouma, C., & Nam, J. (2015). A meta-analysis of gender gap in student achievement in African countries.International Review of Public Administration, *20*(1), 70–83. https://doi.org/10.1080/12294659.2014.967372

Peng, P., Wang, T., Wang, C., & Lin, X. (2019). A meta-analysis on the relation between fluid intelligence and reading/mathematics: Effects of tasks, age, and social economics status. *Psychological Bulletin, 145*(2), 189–236. https://doi.org/10.1037/bul0000182

Petersen, J. (2018). Gender difference in verbal performance: A meta-analysis of united states state performance assessments. *Educational Psychology Review, 30*(4), 1269–1281. https://doi.org/10.1007/s10648-018-9450-x

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research, 90*(1), 24–46. https://doi.org/10.3102/0034654319877153

Polanin, J. R., Espelage, D. L., Grotpeter, J. K., Spinney, E., Ingram, K. M., Valido, A., El Sheikh, A., Torgal, C., & Robinson, L. (2020). A meta-analysis of longitudinal partial correlations between school violence and mental health, school performance, and criminal or delinquent acts. *Psychological Bulletin*. https://doi.org/10.1037/bul0000314

Pustejovsky, J. E., & Tipton, E. (2021). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*. https://doi.org/10.1007/s11121-021-01246-3

Riley, R. D., Tierney, J. F., & Stewart, L. A. (2021). *Examining the potential for Bias in IPD meta-analysis. Results individual participant data meta analysis* (pp. 237–251). Hoboken: Wiley.

Rios, J. A., Ihlenfeldt, S. D., Dosedel, M., & Riegelman, A. (2020). A topical and methodological systematic review of meta-analyses published in the educational measurement literature. *Educational Measurement Issues and Practice, 39*(1), 71–81. https://doi.org/10.1111/emip.12282

Rodríguez-Hernández, C. F., Cascallar, E., & Kyndt, E. (2020). Socio-economic status and academic performance in higher education: A systematic review. *Educational Research Review, 29*, 100305. https://doi.org/10.1016/j.edurev.2019.100305

Röver, C. (2020). Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software, 1*(6), 1–51. https://doi.org/10.18637/jss.v093.i06

Rubio-Aparicio, M., López-López, J. A., Viechtbauer, W., Marín-Martínez, F., Botella, J., & Sánchez-Meca, J. (2020). Testing categorical moderators in mixed-effects meta-analysis in the presence of heteroscedasticity. *The Journal of Experimental Education, 88*(2), 288–310. https://doi.org/10.1080/00220973.2018.1561404

Rust, K. (2014). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 117–154). Boca Raton: CRC Taylor & Francis.

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education, 4*(1), 6. https://doi.org/10.1186/s40536-016-0019-1

Rutkowski, D., & Rutkowski, L. (2021). Running the wrong race? The case of PISA for development. *Comparative Education Review, 65*(1), 147–165. https://doi.org/10.1086/712409

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Rutkowski, L., & Rutkowski, D. (2018). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research, 62*(3), 354–367.

Rutkowski, L., Rutkowski, D., & Liaw, Y.-L. (2019). The existence and impact of floor effects for low-performing PISA participants. *Assessment in Education Principles Policy and Practice, 26*(6), 643–664. https://doi.org/10.1080/0969594X.2019.1577219

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research, 84*(3), 328–364. https://doi.org/10.3102/0034654313500826

Scherer, R., & Siddiq, F. (2019). The relation between students' socioeconomic status and ICT literacy: Findings from a meta-analysis. *Computers and Education, 138*, 13–32. https://doi.org/10.1016/j.compedu.2019.04.011

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review, 17*(8), 881–901. https://doi.org/10.1016/S0272-7358(97)00056-1

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology, 70*(1), 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

Siddiq, F., & Scherer, R. (2019). Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. *Educational Research Review, 27*, 205–217. https://doi.org/10.1016/j.edurev.2019.03.007

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Slavin, R. E. (2008). Perspectives on evidence-based research in education—what works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5–14. https://doi.org/10.3102/0013189X08314117

Tan, C. Y. (2017). Examining cultural capital and student achievement: Results of a meta-analytic review. *Alberta Journal of Educational Research, 63*(2), 139–159.

Tan, C. Y., Peng, B., & Lyu, M. (2019). What types of cultural capital benefit students' academic achievement at different educational stages? Interrogating the meta-analytic evidence. *Educational Research Review, 28*, 100289. https://doi.org/10.1016/j.edurev.2019.100289

Turner, R. M., Bird, S. M., & Higgins, J. P. T. (2013). The impact of study size on meta-analyses: Examination of underpowered studies in Cochrane reviews. *PLoS ONE, 8*(3), e59202. https://doi.org/10.1371/journal.pone.0059202

Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (pp. 55–66). National Foundation for Educational Research.

van de Vijver, F. J. R., Jude, N., & Kuger, S. (2019). Challenges in international large-scale educational Surveys. In L. Suter, E. Smith, & B. Denman (Eds.), *The SAGE handbook of comparative studies in education* (pp. 83–102). Sage Publications.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review, 5*(2), 134–150. https://doi.org/10.1016/j.edurev.2010.02.001

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Wagemaker, H. (2016). International large-scale assessments: from research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: background, technical issues, and methods of data analysis* (pp. 11–36). CRC Press.

Wagemaker, H. (2020). Study design and evolution, and the imperatives of reliability and validity. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment : understanding IEA's comparative studies of student achievement* (pp. 7–21). Cham: Springer International Publishing.

Publisher's Note

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