



A Topical and Methodological Systematic Review of Meta-Analyses Published in the Educational Measurement Literature

Joseph A. Rios , Samuel D. Ihlenfeldt, Michael Dosedel, and Amy Riegelman ,
University of Minnesota

This systematic review investigated the topics studied and reporting practices of published meta-analyses in educational measurement. Our findings indicated that meta-analysis is not a highly utilized methodological tool in educational measurement; on average, less than one meta-analysis has been published per year over the past 30 years (28 meta-analyses were published between 1986 and 2016). Within the field, researchers have utilized meta-analysis to study three primary subject areas: test format effects, test accommodations, and predictive validity of operational testing programs. In regard to reporting practices, authors often failed to provide descriptive details of both their search strategy and sample characteristics limiting reproducibility and generalizability of findings, respectively. Furthermore, diagnostic analyses of outliers, publication bias, and statistical power were not provided for the majority of studies, putting into question the validity of inferences made from the meta-analyses sampled. The lack of transparent and replicable practices of meta-analyses in educational measurement is a concern for generating credible research syntheses that can assist the field in improving evidence-based practices. Recommendations are provided for improving training and editorial standards of meta-analytic research.

Keywords: educational measurement, evidence syntheses, meta-analyses, systematic review

There has been a recent public outcry to end the use of large-scale educational assessments in the U.S. public education system as they have been perceived to be both unfair and educationally irrelevant (i.e., the opt-out movement; Bennett, 2016). To combat this anti-testing movement, it is imperative for educational measurement specialists to demonstrate that tests are fair, valid, authentic, and provide actionable feedback for practitioners. Though this is a difficult task, by incorporating replicable findings from past rigorous experiments into current test development, test administration, analytic, and score reporting efforts, we may be able to promote and establish evidence-based practices in educational measurement and put the field into a constant state of innovation, evaluation, and progressive improvement. However, for evidence-based practices to have the same impact in educational measurement as they have had in the fields of medicine, agriculture, and technology, *trusted* quantitative reviews of research (meta-analyses) are needed to

guide practice (Slavin, 2017). A trusted meta-analysis is one that is methodologically sound in terms of its sampling, variable coding, analysis, and transparency of reporting, which is needed to adequately evaluate its quality and ensure its replicability (see Cooper, Hedges, & Valentine, 2009). However, numerous studies have found that in other fields appropriate meta-analytic methodology and reporting practices are often lacking (e.g., Harwell & Maeda, 2008). To evaluate whether this is the case in educational measurement, the objective of this study is to conduct a systematic review of published meta-analyses in the field to identify the prevalence of this methodology, understand the topics that have been studied, and critically evaluate the methodology and reporting practices that have been employed. Below we describe currently accepted meta-analysis reporting guidelines and prior systematic reviews of educational meta-analyses.

Meta-Analysis Reporting Guidelines

Since Glass's (1976) seminal work introducing the meta-analytic methodology, there have been thousands of quantitative syntheses in the education field (Ahn, Ames, & Myers, 2012). To address quality issues noted in early syntheses (Slavin, 1984), numerous individual researchers (see Cooper et al., 2009) and research organizations (e.g., American Psychological Association) have put forward best practice guidelines to improve problem formulation, sampling,

Joseph A. Rios, Department of Educational Psychology, University of Minnesota, Minneapolis, MN 55455; jrrios@umn.edu; Samuel D. Ihlenfeldt, Department of Educational Psychology, University of Minnesota, Minneapolis, MN 55455; ihlen010@umn.edu; Michael Dosedel, Department of Educational Psychology, University of Minnesota, Minneapolis, MN 55455; dose0018@umn.edu; Amy Riegelman, University Libraries, University of Minnesota, Minneapolis, MN 55455; aspringe@umn.edu.

variable coding, analytic approaches, and reporting guidelines [see Quality of Reporting of Meta-analyses (QUOROM; Moher et al., 2000); Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher, Liberati, Tetzlaff, Altman, & Prisma Group, 2009); and Meta-Analysis Reporting Standards (MARS; American Psychological Association, 2008)]. Though each of these guidelines has somewhat different foci, they all relate to the goals of quality, transparency, and replicability. Below we briefly summarize the key points for each major stage of a meta-analysis.

Problem formulation. Problem formulation is an important aspect of any meta-analysis as it impacts many decisions, such as search strategies, criteria for inclusion/exclusion of study selection, and choice of variables for data extraction, to name a few. However, unlike primary researchers who have the ability to generate novel research questions on topics that have garnered little attention, meta-analysts are dependent on primary research existing on a topic prior to a synthesis being conducted. Therefore, meta-analysts must have a strong grasp of the existing literature and propose research questions that address conceptually broad topics for two reasons. First, a conceptually narrow topic may have an insufficient number of studies to provide adequately powered estimation of average effect sizes, effect size heterogeneity, and moderator coefficients, thus making study conclusions less definitive and robust (for a discussion on the necessary number of studies for adequate statistical power, see Valentine, Pigott, & Rothstein, 2010). Second, synthesists that identify a research topic in which primary studies differ in terms of populations studied, data collection methods, operational variations that are relevant to the concepts being studied, and effect size direction and magnitude will capitalize on naturally occurring differences within the literature. In doing so, it will allow the synthesist to provide an opportunity to test theoretical hypotheses concerning moderators and mediators that have never been tested in primary research when variability between studies is present, thereby fulfilling the full potential of the meta-analytic methodology (Cooper, 2009).

Search strategy. The need for rigorous, relevant, primary research is critical for the validity of inferences made from a meta-analysis, and identifying a comprehensive sample requires a thorough search strategy (Harwell & Maeda, 2008). The PRISMA reporting guidelines (Moher et al., 2009) state that a good description of a search strategy will describe all of the sources of data, the keywords used alongside connecting Boolean operators, limits constrained on the search (e.g., date limits, language, and geography), and the dates that these searches were employed. The goals of these guidelines are to permit readers to evaluate the comprehensiveness of the search strategy and facilitate the reproducibility of meta-analytic samples for replicative purposes. High-quality searches have been found to be generated when a librarian is part of the meta-analytic team and when the search has been peer-reviewed before being conducted (Koffel, 2015).

Study selection and variable coding. A detailed description of the coding process and evaluation of rater agreement is critical to assessing the transparency and potential replicability of results. Meta-analysts are often presented with a number of ambiguities when coding qualitative information into a quantitative form, which thus requires a concrete cod-

ing plan to reduce subjectivity and increase rating reliability. A good coding plan describes both the eligibility criteria for inclusion of primary studies into the meta-analysis and documents the procedures used to extract information from studies. The development of the eligibility criteria is directly influenced by the study objective and defined universe of generalization. As such, it should explicitly address requirements of the following: (a) defining features of the studies of interest, (b) the eligible research designs, (c) sample restrictions, (d) required statistical data, and (e) geographical, linguistic, and time frame restrictions. In regard to data extraction, a quality coding protocol should ensure that clear operationalizations of variables are provided that minimize rater inferences, describe how the variable is coded (e.g., categorical or open-ended), explain whether assumptions are made to impute missing data, as well as delineate whether variables are coded for descriptive or moderator purposes. Furthermore, to allow for the reader to assess the validity of the coding process, meta-analysts should report rater training as well as raters' coding consistency.

Analytic approaches. Meta-analysts are presented with a number of validity threats, such as publication bias, outliers, effect size dependencies, missing data, and obtaining adequate statistical power, that must be addressed to ensure that important measures, such as the central tendency, heterogeneity of the effect size distribution, and statistical inferences are accurate (see Matt & Cook, 2009). Of these validity threats, two are unique to meta-analysis: publication bias and effect size dependencies. Publication bias occurs when the sample differs significantly from the population of studies that should be included leading to artificial inflation of study results. To remedy this issue, it is vital that meta-analysts conduct thorough literature searches and include articles not published in peer-reviewed journals (i.e., gray literature), which are more likely to possess nonstatistically significant findings and/or negligible effect sizes (Polanin, Tanner-Smith, & Hennessy, 2016). The second unique validity threat, effect size dependencies, occurs when multiple effect sizes are based on overlapping samples. Failure to deal with this issue can lead to inaccuracies in standard error estimates, and, thus, incorrect inferences concerning null and moderator analyses (e.g., Tipton, 2015). A number of approaches, each possessing advantages and limitations, have been used in practice to deal with this threat (see Scammacca, Roberts, & Stuebing, 2014). Although not unique to meta-analysis, best practice guidelines stipulate that researchers must provide a clear and transparent reporting of how validity threats are handled to allow readers to evaluate the validity of inferences made.

Reporting. Transparency of findings and implications requires clear reporting practices. To begin with, as noted in both the PRISMA and MARS guidelines, it is recommended that authors disclose their funding sources so that the audience can infer the role of funders in the development of the study and inferences made. Additionally, meta-analysts should provide detailed information on the sample characteristics (e.g., demographics, age, nationality, sample size [number of studies and effect sizes]) to allow the reader to evaluate the external validity of the findings. In terms of reporting results, both standard error estimates and effect size heterogeneity ought to be reported to permit the audience

to judge the relationship between uncertainty in estimates and inferences. Additionally, to allow for replication, there has been a recent push for authors to provide their raw data, which can be done either in-text, as supplementary online information, or through open-source platforms such as the Open Science Framework. Finally, it is vital that authors note general limitations of the meta-analysis (e.g., related to sampling or analyses) as well as implications of study results to theory, policy, or practice. Overall, meta-analytic research can only be effective if decision-makers feel that they are trustworthy, and clear reporting practices permit decision-makers to make such an evaluation.

Previous Methodological Systematic Reviews of Meta-Analyses in Education

To date, a number of systematic reviews summarizing and critiquing the methods and reporting practices of previous educational meta-analyses have been performed. Across reviews, effective reporting of search strategies has been found to be lacking as many studies have failed to include sufficient details (e.g., databases, keywords, search limits) to reproduce search results (e.g., Ahn et al., 2012; Harwell & Maeda, 2008; Polanin, Maynard & Dell, 2017). In terms of study inclusion, only 4% to 8.6% of studies that have been evaluated in the literature fully described their eligibility criteria (e.g., Harwell & Maeda, 2008; Polanin et al., 2017). Although rater consistency is a critical source of meta-analytic validity (Harwell & Maeda, 2008), a limited percentage of studies have been found to report the number of coders included or coder training employed, and one study found that interrater reliability was assessed in only 40% of analyses (e.g., Ahn et al., 2012; Harwell & Maeda, 2008). Combined, these results suggest that search processes and reliability of data extraction may be questionable.

Although statistical procedures for estimating an average effect size and conducting moderator analyses was explicitly stated in 82% of studies reviewed by Ahn et al. (2012), details on the underlying assumptions required for those procedures were missing in a large number of studies reviewed (e.g., Ahn et al., 2012; Harwell & Maeda, 2008). As an example, the presence of effect size dependencies, how dependencies were handled, sensitivity to outliers, and the presence of publication bias were not discussed (e.g., Ahn et al., 2012; Lin, Chen & Liou, 2017). There have not been reviews on power in educational meta-analyses, but previous research in the social sciences has found that a large number of analyses lacked statistical power, thus limiting the validity of inferences made from these meta-analytic studies (Cafri, Kromrey, & Brannick, 2010). One area in which reviews were consistently positive was that educational meta-analysts summarized their major findings (Ahn et al., 2012; Lin et al., 2017). However, some researchers have noted that other reporting practices, such as clearly describing the sample size, sample makeup, and quality of the sample, have been sparse, limiting the ability of readers to judge the generalizability of the meta-analytic findings (Harwell & Maeda, 2008). Combined, the results from prior systematic reviews call into question the validity of meta-analyses in the educational research literature and demonstrate the need to determine if there is a disconnect between practice and guidelines in the educational measurement field.

Need for Current Study

To date, there is no clear indication of how many meta-analyses have been conducted in educational measurement, the topics that have been studied, nor the rigor of methodological and reporting practices. To address this limitation, this study seeks to comprehensively synthesize published meta-analyses conducted in educational measurement. Based on this synthesis, we look to understand the breadth and depth of research topics that have been investigated using meta-analytic methodology, which can inform the field of potential areas of future meta-analytic research. Further, by critically evaluating reporting practices we look to identify strengths and areas in need of improvement when conducting and publishing educational measurement meta-analyses. These objectives are addressed via the following research questions:

1. How prevalent is the meta-analytic methodology in the educational measurement literature? What topics in educational measurement have been studied using meta-analytic procedures?
2. Do researchers describe their search strategies to allow for reproducibility? As the validity of inferences from meta-analyses is contingent on the reliability of data extraction, do researchers clearly operationalize variable definitions and their coding process? How do researchers deal with meta-analytic assumptions and issues of power? Are results reported in a manner that is in-line with best practice reporting guidelines?

Method

Search Strategy

As the objective of this study was to evaluate *published* meta-analyses in journals related to educational measurement, four distinctive search strategies were employed: (a) manual, (b) database, (c) backward citation, and (d) forward citation searches. This literature search was conducted by the fourth author, who is a social sciences librarian, between September 21, 2018 and November 19, 2018. Below is a full description of each strategy, presented in the order conducted.

Database search. The first search strategy consisted of conducting bibliographic and manual searches for 26 journals publishing research related to educational measurement (the full list of twenty-six journals can be found in the online Supporting Information as Appendix A). These journals were identified from a list of 78 journals in educational measurement, statistics, research, and psychometrics compiled at the University of Massachusetts (Khademi, 2013; available upon request from Joseph Rios). The first three authors evaluated the aims and scope of each journal to determine whether the journal published both research related to the field of educational measurement as well as research utilizing meta-analytic methodologies. Any disagreements were settled based on consensus amongst authors.

Upon deciding on which journals to include, four of the selected educational measurement journals were manually searched due to irregular indexing (*Assessment and Evaluation in Higher Education, Assessment for Effective Intervention, Language Testing in Asia, Papers in Language Testing and Assessment*). This manual searching resulted in discovering only one relevant study. The remaining literature

was searched for in the journals identified via PsycINFO via Ovid, ERIC via EBSCO, Education Source, and Scopus. Adhering to PRISMA guidelines, one full search string is included in Appendix B (found in the online Supporting Information). The systematic searching targeted title fields that included “*meta-analy**” OR “*meta analy**” as well as the *meta analysis* subject heading when applicable. Publication titles were located by searching with both current and past journal titles as well as corresponding print and electronic ISSNs provided in Ulrich’s Periodicals Directory.

Backward and forward citation search. To conduct the backward (i.e., examining the reference list of every included study) and forward (i.e., examining papers that cited every included study) citation searches, the *Social Sciences Citation Index* was utilized as it provides resource efficiency and rapid linking to full texts for the references via the databases outlined previously. The first step in this process was to search the reference lists of all the studies meeting the eligibility criteria (discussed below) found via the methods listed above. Specifically, studies found in reference lists underwent title and abstract review, and if necessary, full-text review. Forward citation searching was then implemented by examining any papers that had cited the articles found to meet the eligibility criteria from the hand, database, and backward citation searches. If articles from this search strategy met the eligibility criteria, they then underwent backward citation searching. Articles found to meet the eligibility criteria from the backward citation search were then subjected to forward citation searching. This process was repeated until no new studies met the eligibility criteria.¹

Eligibility Criteria

To be included in the systematic review, each study had to conduct a meta-analysis consisting of a literature review (i.e., many authors refer to a study aggregating data sets collected from a testing program to be a meta-analysis and these were not included; e.g., Talento-Miller, 2008) as well as meet the eligibility criteria set forth along three dimensions: (a) study, (b) assessment, and (c) participant characteristics.

Study characteristics. Eligible studies were published in both peer-reviewed journals and the English language. Studies were excluded if they were narrative reviews, primary study research, or papers related to methodological issues in meta-analysis.

Participant characteristics. Eligible studies were those analyzing K-12 and higher education student populations. In regard to the latter, studies could include either student or general populations taking assessments for entrance to higher education programs (either undergraduate or graduate studies). Furthermore, studies were included if the sample consisted of a mixture of K-12, higher education, and/or adult populations with no restrictions placed on participants’ nationality. In cases where the study authors did not specify the participant characteristics, a review of the included primary studies in the meta-analysis was conducted to ascertain the population type. Ineligible studies comprised empirical research solely consisting of participants from the general

population not taking a higher education admissions test, individuals taking certification or licensure tests, or students in preschool.

Assessment characteristics. To be included, meta-analyses had to focus on test development, administration, score reporting, and/or validity evidence of assessments group-administered in educational settings for formative, summative, or admissions purposes (analyses that included a mixture of these and other forms of assessment such as licensure and certification examinations were included so long as they focused on one or more of the above test components; e.g., Rodriguez, 2005). Ineligible studies evaluated achievement differences between subgroups using an educational test (e.g., Hyde, Fennema, & Lamon, 1990), used an educational test as a learning event as opposed to assessing learning (e.g., Rowland, 2014), analyzed assessments administered in an individual setting (e.g., psychoeducational), administered an assessment exclusively for purposes of licensure and/or certification, or implemented peer-grading as it has limited application to large-scale testing programs (e.g., Sanchez, Atkinson, Koenka, Moshontz, & Cooper, 2017).

Variable Coding

A total of 76 variables were coded, which can be categorized into the following themes: (a) background, (b) search strategy, (c) study selection and variable coding, (d) data analyses, and (e) reporting of findings. These variables reflect those suggested in the PRISMA reporting guidelines (Moher et al., 2009) as well as those found in previous systematic reviews of meta-analyses (e.g., Harwell & Maeda, 2008; Koffel, 2015; Koffel & Rethlefsen, 2016). Below we present a description and rationale for the inclusion of these variables. The coding protocol for this study, which includes an operationalization and coding strategy for each variable, can be found online as Supporting Information in Appendix C.

Background variables. For each article, background information, such as authors’ last names, year of publication, and journal name, were coded for. Furthermore, we included variables on study characteristics that included the topic of the meta-analysis, student population type (general, English language learner [ELL], or special education), and student grade level. This information was coded to provide a general description of the topical and general characteristics of meta-analyses published in the field.

Search strategy variables. Description of the following variables by study authors was coded for as they are essential in reproducing meta-analytic search results: search method employed, identification of databases, search terms, actual search dates, Boolean operators, and search limits (e.g., language). Additionally, we coded for librarian involvement and search strategy peer review, as these variables have been found to be associated with quality meta-analytic searches (Koffel, 2015).

Study selection and variable coding variables. A well-defined coding plan contains both detailed information on (a) which studies are to be included in the analysis in an effort to reduce ambiguities that may arise during the study

selection process and (b) the process of data extraction from primary studies. Thus, we analyzed whether the authors included an operationalized description of what was required of primary research for its inclusion in the meta-analysis. In addition, we were interested in whether meta-analysts in educational measurement used study quality measures or proxies of quality (e.g., excluding based on high attrition, low reliability, or study design) for exclusion. In regard to data extraction, the following variables were analyzed in terms of whether each was described in the primary study: (a) the number of coders, (b) description of coder training, (c) inclusion of explicit operational definitions of key variables (i.e., variables included as moderators and outcomes), and (d) description of how variables were coded (e.g., categorical vs. open-ended). Three variables were analyzed for rater agreement, which included the procedure for resolving rater disagreement, whether rater agreement was evaluated, and the method for assessing rater agreement.

Data analyses variables. The appropriateness of data analytic procedures that are chosen by research synthesists has a clear impact on the interpretation of results. To evaluate these decisions, a number of variables were included, which can be categorized into the following subclasses: (a) evaluating assumptions, (b) handling missing data, (c) conducting power analyses, (d) calculating effect sizes, and (e) performing moderator analyses. In reference to assumptions, we evaluated whether publication bias was assessed, the method used, whether publication bias was identified, and if so, how it was handled for subsequent analyses. Additionally, we investigated whether meta-analysts explicitly noted the presence of influential outliers as a validity threat, and if that was the case, how they were examined prior to conducting their analyses. Similarly, we were interested in knowing if meta-analysts conducted power analyses, and if so, whether they were prospective or retrospective analyses, the degree of heterogeneity assumed, and the criterion level for acceptable power. Finally, we coded for whether missing data were noted as part of the analysis, and in such a case, the methodology employed for handling them. In regard to the actual analyses, information related to both the calculation of effect sizes and moderator analyses was evaluated. In terms of the former, the following variables were included: (a) number of studies (n), (b) number of effect sizes (k), (c) total number of participants (N), (d) effect size type, (e) effect size correction (i.e., correcting for measurement error, restriction of range, and study design), and (f) heterogeneity of effect sizes examined. Lastly, the approach to modeling average effect sizes and moderators was coded for, with particular focus on the weighting approach taken. For those analyses that took a regression approach we coded for whether random or fixed effects were modeled and if the authors gave a methodological justification for their choice of model.

Reporting results. Reporting should be transparent regarding all aspects of the search results, sample, analysis, and conclusions of the study. The following variables were identified: (a) inclusion of a PRISMA flow diagram, (b) the number of citations that were included in the final analysis and excluded after the initial search, (c) the inclusion of gray literature, and (d) if the author provided a way to view the raw data. Whether the author provided a breakdown of the time frame, nationality, language, and gender of the sample

was also analyzed. As for the quantitative synthesis, much like in primary research it may be tempting for meta-analysts to omit unfavorable results. Thus, the following variables were coded for (a) whether the average effect size was reported, (b) whether the error associated with the average effect size was reported, and (c) if a figure displaying the distribution of effect sizes was included. Lastly, based on previous syntheses, we looked for explicit statements of support or nonsupport for primary and secondary hypotheses, discussion of the funding, implications and limitations of the study, and an explicit statement describing the universe of generalization of the results.

Interrater Agreement

Interrater reliability was assessed for three distinctive stages of the coding process, which included title and abstract inclusion, full-text inclusion, and variable coding. Rayyan (rayyan.qcri.org), a collaborative cloud-based tool, was used for blind screening. At each stage, the first author provided training to the second and third authors who served as the coders. These individuals were graduate students in educational measurement and had taken a formal course on meta-analysis providing them adequate substantive and methodological knowledge. Upon completion of training, the second and third authors then double-coded 20% of articles at each stage, which were evaluated for agreement accounting for chance using Cohen's kappa (κ). Any disagreements in coding were resolved by the first author. The criterion established a priori for adequate interrater agreement was a κ value of .80 (Landis & Koch, 1977). The κ values observed between the two raters for title and abstract inclusion, full-text inclusion, and variable coding were 1, .857, and .815, respectively.

Results

The combined database and citation searches produced 1,411 unique publications, which underwent title and abstract exclusion. Of these, 52 articles were found to warrant a full-text review as the content of their titles and abstracts communicated that they were original meta-analytic research conducted in educational measurement. However, upon conducting full-text reviews, we found that 28 fully met the inclusion and exclusion criteria established a priori (Figure 1; see online Supplementary Information for the reference list of these studies [Appendix D]). These 28 studies were published between 1986 and 2016 with approximately 36% being published since 2010. The median number of studies and effect sizes in each analysis were 15 and 42, respectively, with as few as studies and 1 effect size per analysis, and as many as 1,521 studies and 6,589 effect sizes. A large number of journals ($n = 20$) were found to publish educational measurement meta-analyses with the following journals publishing more than two studies in our sample: *Educational and Psychological Measurement* ($n = 4$), *Educational Measurement: Issues and Practice* ($n = 3$), *Applied Measurement in Education* ($n = 2$), and *Psychological Bulletin* ($n = 2$).

Topical and Population Characteristics

In our sample, the meta-analytic methodology was applied to the study of predominantly three topics: (a) predictive validity of operational testing programs ($n = 8$; 28.57%),

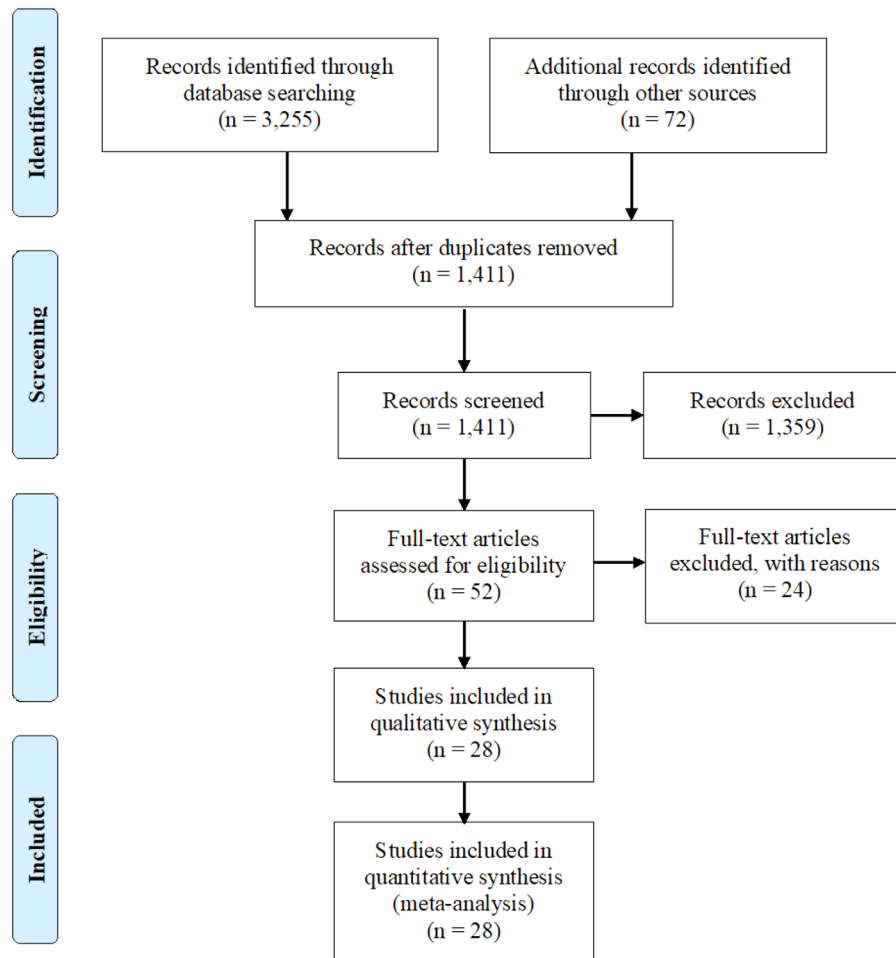


FIGURE 1. Flow diagram of search strategy. Note that the database search consisted of finding published meta-analyses in 31 unique peer-reviewed journals. [Color figure can be viewed at wileyonlinelibrary.com]

(b) test accommodations ($n = 8$; 28.57%), and (c) test presentation ($n = 7$; 25%). Of the eight predictive validity meta-analyses, four studied the Graduate Record Examination (GRE) (14.29%), while the remaining studies examined the Graduate Management Admission Test (GMAT), Pharmacy College Admission Test (PCAT), Medical College Admission Test (MCAT), and a combination of multiple college admissions tests. Equally as representative in our sample was the study of testing accommodations for both English language learners ($n = 4$; 14.29%) and students with disabilities ($n = 4$; 14.29%). The third most popular topic studied was test presentation, which included studies on test mode effects ($n = 4$; 14.29%) and item formatting ($n = 3$; 10.71%). The former area of study included multiple papers comparing computer and paper-based test presentation, while the latter examined issues around item types and number of response options. In addition to these three main topics, our sample also included studies investigating standard setting procedures ($n = 1$; 3.57%), effectiveness of formative assessments ($n = 1$; 3.57%), practice effects ($n = 1$; 3.57%), feedback on improving test scores ($n = 1$; 3.57%), and the impact of testing on academic achievement ($n = 1$; 3.57%). Of the 26 studies that explicitly noted their sample characteristics, there was nearly an equal breakdown of samples consisting of K-12 (34.62%), higher education (seven were related to

graduate school; 30.77%), and a mixture of both populations (34.62%).

Search Strategy

Of the 10 search strategies we coded for, eight were present in the sample of 28 studies, while none of the studies employed forward citation searching or contacting professional organization listservs (Table 1). Of the search strategies utilized, the most common was research databases (96.43%). For those that employed this strategy, 82.14% of studies stipulated the use of at least two unique databases. Beyond employing databases, the second most popular search utilized was the backward citation search (75%). Other strategies employed in less than half of sampled studies included making use of hand searches (39.29%), web browsing (28.57%), searching citation indices (25%), expert consultations (21.43%), browsing publishers' websites (21.43%), and searching conference abstracts (7.14%). On average, each study made use of approximately three of the coded strategies. In terms of providing details of database searches, less than 10% of studies explicitly noted employing search limits related to language (7.14%), geography (3.57%), age (7.14%), or other (0%) variables; though, more than half of the studies documented study date range limits. However, very few of the studies divulged the specific search terms employed (28.57%) nor their Boolean

Table 1. Search Strategy, Study Selection, and Variable Coding Results

Stage	Strategy	No (Percent)	Yes (Percent)	
Search Strategy	Database	3.57	96.43	
	Journals Handsearched	60.71	39.29	
	Citation Indices Searched	75	25	
	Backward Citation Search	25	75	
	Forward Citation Search	100	0	
	Conference Abstracts Reviewed	92.86	7.14	
	Expert Consultation	78.57	21.43	
	Listservs	100	0	
	Internet Browsing	71.43	28.57	
	Publisher Website	78.57	21.43	
	Database Search Information			
	Databases Included	14.29	85.71	
	Two or More Databases	10.71	82.14	
	Dates Ranges for Database Search	92.86	7.14	
	Search Variables			
	Search Terms Included	71.43	28.57	
	Boolean Phrases Included	92.86	7.14	
	Inclusion of Timeframe (Years) Search For	42.86	57.14	
	Language Limits	92.86	7.14	
	Geography Limits	96.43	3.57	
	Age Range	92.86	7.14	
	Other Limits	100	0	
	Peer Review			
	Librarian Involved	100	0	
	Search Strategy Peer Reviewed	92.86	7.14	
	Inclusion and Exclusion of Studies			
	Selection	Inclusion/Exclusion Criteria	21.43	78.57
		Use of Quality Scale for Exclusion	100	0
		Exclusion Based on Nonrandomization	78.57	21.43
		Exclusion Based on Reliability	96.43	3.57
		Exclusion Based on Attrition	96.43	3.57
	Coding			
	Variable Coding	Number of Coders	42.86	57.14
Coder Training		89.29	10.71	
Method of Resolving Rater Disagreement		50	50	
Rater Agreement Evaluated		64.29	35.71	
Independent Variables Defined Operationally		14.29	85.71	
Independent Variables Coding Description		71.43	28.57	

operators (7.14%), and only two studies (7.14%) documented the actual dates that the databases were searched. The lack of search description provided may be attributed to none of the studies mentioning librarian involvement and only two (7.14%) studies utilizing peer review of search strategies.

Study Selection and Variable Coding

Although operationalizing the inclusion and exclusion criteria is crucial for minimizing uncertainties between raters and enabling replications, only 79% of studies explicitly specified their eligibility criteria (Table 1). For those that did, exclusion was based on nonrandomization ($n = 6$), low reliability ($n = 1$), and high participant attrition ($n = 1$). None of the evaluated meta-analyses used a quality scale for inclusion/exclusion decisions. In terms of variable coding, 82% of studies included explicit definitions of key independent variables, and 29% explicitly noted if these variables were categorical or continuous. Sixteen of 28 (57.14%) studies specified the number of coders used to code primary studies (most often two raters); however, only 10.71% of studies described how coders were trained. In regard to rater consistency, 14 studies specified how disagreements were settled

when they arose (50%), and 10 out of 28 studies formally evaluated rater agreement. Of these studies, 75% reported percent agreement, 25% reported both percent agreement and Cohen's kappa, and one study used a correlation between the aspects of the primary studies that they coded on a rating scale.

Data Analysis

Diagnostic analyses are imperative to ensure the validity of average effect size, heterogeneity, and moderator estimates (see Table 2). In our sample, 42.86% of papers discussed this form of bias, but it was assessed in a much smaller number of studies (17.86%). Of the studies that did evaluate publication bias, three methods were utilized: visual inspection of a funnel plot ($n = 2$), Rosenthal's fail safe N ($n = 2$), and Egger's test ($n = 1$); though none of the studies identified or adjusted for publication bias. Missing data were factored into the analysis in a little over half the studies (57.14%) and addressed using solely data imputation in seven papers, listwise deletion in three cases, by contacting the primary study author in two cases, or by some combination of those noted. For those using data imputation, mean imputation ($n = 3$) and

Table 2. Data Analysis and Reporting Results

Stage	Diagnostic Analyses	No (Percent)	Yes (Percent)	
Data Analysis	Publication Bias Discussed	57.14	42.86	
	Publication Bias Assessed	82.14	17.86	
	Missing Data Examined	42.86	57.14	
	Outliers Examined	67.86	32.14	
	Data Corrections			
	Measurement Error Correction	82.14	17.86	
	Restriction of Range Correction	85.71	14.29	
	Study Design Correction	89.29	10.71	
	Main and Moderator Analyses			
	Main Analysis	3.57	96.43	
	Justification of Fixed vs. Random-Effects	17.65	82.35	
	Studies Weighted	21.43	78.57	
	Dependent Effect Sizes Discussed	42.86	57.14	
	Heterogeneity of Effect Size Examined	32.14	67.86	
	Moderator Analyses	25	75	
	Power Analyses	100	0	
	Description of Sample			
	Reporting	PRISMA Flow Diagram	100	0
		Number of Included Citations	3.57	96.43
Number of Excluded Citations		42.86	57.14	
Gray Literature		7.14	92.86	
Raw Data Provided to Readers		89.29	10.71	
Sample Nationality		89.29	10.71	
Date Range of Studies		46.43	53.57	
Languages of Studies		89.29	10.71	
Gender Breakdown of Studies		96.43	3.57	
Analysis				
Average Effect Size Reported		10.71	89.29	
Error Associated With Average Effect Size Reported		17.86	82.14	
Figure Displaying Distribution of Effect Sizes		67.86	32.14	
Post Hoc Analysis		89.29	10.71	
Discussion and Conclusion				
Support for Primary Hypothesis		0	100	
Support for Secondary Hypothesis		0	100	
Universe of Generalization		46.43	53.57	
Implications		3.57	96.43	
Limitations	21.43	78.57		

regression-based imputation ($n = 2$) were employed, while four studies did not stipulate the method used. Of the meta-analyses sampled, fewer than half (32.14%) noted that outliers were a potential threat to the validity of inferences made from the analyses. Specific information on how outliers were handled was limited, as five studies provided no additional details beyond stating that an outlier analysis was performed. Finally, as many studies possess artifacts that can distort results, a minority of researchers in our sample applied effect sizes corrections for measurement error (17.86%), range restriction (14.29%), and study design (10.71%).

In terms of analyses, of the 28 studies sampled, 57.14% noted the potential impact of dependencies on their results; however, three studies noted that dependencies impacted estimates, but did nothing to mitigate their effect. The most common approaches to dealing with dependencies in our sample were to average across effect sizes from the same primary study or select one of many effect sizes that was most representative of the findings. The approach to modeling the average effect size varied across studies. A traditional approach, such as a mean or a weighted mean, was the most common method for determining the average effect size for both the main analysis and the analysis of moderators ($n = 16$). Others relied on metaregression ($n = 13$), and a small number of

studies utilized hierarchical linear modeling (HLM; $n = 4$). In the latter two methods, random effects ($n = 8$), fixed effects ($n = 2$), and a combination ($n = 3$) were used for the main analyses. Of the 17 studies that employed regression or HLM for either the main or moderator analysis, 14 contained a justification for their model choice. In modeling the average effect size, primary studies were most commonly weighted by standard error ($n = 11$) or sample size ($n = 10$). Three studies did not conduct a main analysis and focused only on the moderators. The heterogeneity of effect sizes was analyzed in about two-thirds of the studies, and all but seven studies conducted a moderator analysis. Moderators were modeled in separate analyses ($n = 8$), simultaneously ($n = 6$), and in three cases both approaches were used. None of the studies included any discussion of acceptable power, nor a power analysis.

Reporting Results

Considering measures of transparency, six studies reported external funding (21.43%; Table 2). Further, three meta-analyses provided more transparency by providing or indicating how a reader could access their full data (10.71%). Although most of the meta-analyses reviewed included some of the information that would be presented in a PRISMA

diagram (96.43%), none we reviewed included the recommended diagram. This is perhaps because the diagram was introduced in 2009, after 17 of the studies were published. All but one meta-analysis reported the number of studies included (96.43%), and slightly more than half also reported the number of studies reviewed and excluded (57.14%). Nearly all of our studies reported an average effect size (89.29%) and effect size error (82.14%), and some included a figure displaying the distribution of the effect sizes from primary studies (32.14%). Studies that did not report an average effect size tended to have research questions more focused on moderator analyses. Of the nine studies using figures to display effect sizes, forest, stem-and-leaf, and histogram plots were the most common ($n = 3$, $n = 2$, and $n = 2$, respectively). All studies in our sample provided support for or against primary and secondary hypotheses, noting implications for research and practice (96.43%), as well as study limitations (78.57%). However, fewer studies directly discussed the appropriate universe of generalization (53.57%). In lieu of direct discussion of appropriate generalization, information regarding the final sample of included studies can provide some insight. Unfortunately, many of the reviewed meta-analyses did not provide much information. Specifically, slightly more than half provided the date range of studies included (53.57%), only three reported the countries and languages represented in the sample (10.71%), and only one provided the gender breakdown of the final sample (3.57%).

Discussion

The primary objectives of this study were to survey educational measurement journals to determine which topics in the field have been studied using meta-analytic methods and evaluate their reporting practices. Overall, we found that meta-analysis is not a highly utilized methodological tool in the field as, on average, less than one meta-analysis was published per year; though, we observed an increase in the last decade. In terms of topical areas of study, most of the published meta-analyses came from three primary subject areas: test format effects, test accommodations, and the predictive validity of operational testing programs. Though the use of simulation techniques in educational measurement applications has become an important and popular tool (Feinberg & Rubright, 2016), we found no meta-analyses of such techniques in our sample. We speculate that the lack of utilization of the meta-analytic methodology may largely be due to an unawareness of this approach to conducting research, and/or to a dearth of primary studies on similar topics necessary to conduct an adequately powered meta-analysis. The latter hypothesis is supported by the fact that the median number of primary studies included in the meta-analyses sampled was 15, suggesting that many of the research questions investigated using the meta-analytic methodology may be too narrow to maintain a large sample of primary studies for analyses. However, to remedy this uncertainty a priori, a survey of educational measurement literature focusing on the topics studied, methodologies employed, and populations sampled would be extremely helpful to identify potential areas for future meta-analytic research.

In terms of surveying reporting practices of search strategies, our results demonstrated that meta-analyses in the field of educational measurement did an exceptional job in using multiple methodologies in their literature searches in an

attempt to obtain a comprehensive sample, but did a poor job describing the details of their searches, such as the specific search terms used, search limits employed, and dates that the search took place. One potential reason for the poor documentation is that meta-analysts in our sample failed to mention the involvement of librarians, experts in information retrieval, in their study, nor did they conduct peer review of their search strategies, which innately requires detailing of the search process. Thus, to answer our second research question, it would be impossible to reproduce the samples in the reviewed studies calling into question the validity of the results as replicative efforts could not take place. As direct involvement of a librarian or information specialist improves the quality and reproducibility of a search strategy (Koffel & Rethlefsen, 2016), we recommend that meta-analysts closely collaborate with librarians during both the sampling and writing phases. Ultimately, the concern of transparency of the study selection and variable coding process is one of validity much like every other step of the meta-analytic process. Authors within our sample were careful to define their eligibility criteria and variables operationally but did not describe how they were coded (e.g., binary or multiple categories), which could lead to inaccuracies in replication efforts. This inconsistency is compounded by a lack of description of the coder training process (and in many cases even the number of coders was excluded) making the reliability, and thus, the validity of data extraction questionable. Much like with other aspects of the study, we presume the lack of detail is due to a lack of knowledge concerning best practice reporting guidelines and/or a consequence of page length requirements enforced by journals. If the latter is true, we encourage authors to preregister their study and publish their coding strategies as supplemental material on websites such as the Open Science Framework (www.osf.io). Not only does this allow for improved replication in the future, it adds evidence to the validity of the result by demonstrating that the coding process was operationalized a priori.

Previous reviews of meta-analysis in education found that there was typically a lack of information about the underlying assumptions that were necessary for the validity and generalizability of results (e.g., Ahn et al., 2012; Harwell & Maeda, 2008; Slavin, 1984). Overwhelmingly this is true within our sample. Specifically, some types of bias were more broadly discussed (and assessed) than others, but nearly every type of biasing feature, such as missing data and publication bias, were not described in the majority of studies. Surprisingly, statistical power was never assessed in any of the 28 papers included in the sample. Although recommended by the MARS reporting guidelines, neither the PRISMA nor QUOROM guidelines make reference to it. Nonetheless, we conducted a retroactive power analysis following a guideline presented in the Cochrane Handbook (9.6.5.1; Higgins & Green, 2011) recommending that no more than 1 moderator is included for every 10 primary studies, which was met by only 67.86% of the included studies, suggesting that a large percentage of studies in our sample are underpowered. We recommend that journal editors and reviewers require power analyses to be included in published meta-analytic research so that readers can judge the quality of inferences being made.

Finally, we observed that meta-analyses in educational measurement tended to do an excellent job of reporting their findings, including an explicit statement of support or non-support for their hypotheses, and noting the limitations and

implications of their study. However, there is a clear need to improve the reporting of sample characteristics in order to be in line with reporting guidelines. The generalizability of inferences made from meta-analyses is based on the alignment between the stipulated universe of generalization and the characteristics of the sampled studies. However, only slightly more than half of the studies had explicit statements about this universe. Additionally, many details necessary for determining generalizability were missing (perhaps due to poor reporting in primary research), such as the nationality of the sample, the language of the selected studies, and the gender breakdown of the sample. To provide readers the ability to make inferences concerning the external validity of inferences made from educational measurement meta-analyses, improved reporting is needed as suggested in best practice guidelines, such as PRISMA, QUOROM, or MARS.

Limitations

Two limitations associated with our study should be noted. First, there may have been studies that were not included in our sample due both to the eligibility criteria and search strategy implemented. One potential consequence associated with our eligibility criteria is that we only included studies published in peer-reviewed journals. This was done to explicitly investigate editorial practices in the field. However, in taking this approach, it is possible that a number of meta-analyses published as gray literature were not included. Consequently, there may be excellent examples of meta-analytic studies that provided transparent and valid practices that were missed. In terms of our search strategy, one potential limitation is that our results may have been susceptible to language search bias as we only included studies published in the English language. Therefore, future systematic reviews of meta-analyses in the field should include non-English languages to evaluate whether these studies provide more transparent reporting.

Second, a number of the reviewed studies may have implemented rigorous meta-analytic procedures, but did not provide sufficient details. One potential challenge for meta-analysts in supplying adequate detail to allow for both transparency and replicability is page length requirements put into effect by journals. Thus, it is unclear whether the limited specifications of the search strategies, variable coding, data analyses, and reporting observed in our sample was due to a lack of awareness of best practice guidelines from authors or an inability to provide detail due to page length requirements stipulated by journals. A post hoc analysis demonstrated that none of the sampled studies cited PRISMA, MARS, or QUOROM guidelines, suggesting that many authors may not have been aware of these best practice guidelines at the time of their writing; though, page limitations may have also served as a barrier for detailing practices.

Conclusions and Recommendations

This systematic review shows that over the past 30 years researchers have attempted to employ meta-analytic methodologies to study topics in educational measurement. However, our analysis revealed that there are significant deficiencies in nearly all of the studies performing meta-analyses in the field, particularly in evaluating analytical assumptions. Thus, the internal validity of the inferences drawn from the results are uncertain, leading to a concern of negative consequences associated with potentially faulty findings (e.g., adopting test

accommodation practices that may not be effective) from these meta-analyses. In addition, deficient reporting practices observed for many of the sampled studies puts into question the external validity of this literature as the lack of detail provided makes it impossible to both reproduce samples and replicate results. These conclusions imply that enhancing the quality of meta-analyses in our field will require both improvements in training and editorial standards.

To improve training of educational measurement graduate students and experts in meta-analytic methodology, we must increase the number of graduate-level courses and workshops on the topic. In a post hoc analysis of curricula taught at 37 U.S. graduate programs in quantitative methodology in the 2017–2018 academic year, 10 programs were found to offer a course on research synthesis/meta-analysis, and only three made this course mandatory for graduation. This result clearly demonstrates that knowledge of meta-analytic methodology is currently not valued as a necessary component of training in educational measurement. One solution to remedy this issue is to fund intensive multiple day workshops on meta-analysis, such as the Meta-Analysis Training Institute (<https://www.meta-analysis-training-institute.com/>), that current faculty members in quantitative methodology can attend. By doing so, these faculty members can take their newfound knowledge of meta-analysis and offer a graduate course on the topic, thereby improving training for the next generation of measurement specialists.

Additionally, it is crucial that we raise the editorial standards for publishing meta-analyses in peer-reviewed journals of educational measurement by endorsing and upholding the PRISMA and MARS best practice guidelines. As of the time of this writing, over 170 journals in the health sciences as well as editorial organizations, such as the Cochrane and Campbell Collaborations, have formally endorsed the PRISMA guidelines (<http://www.prisma-statement.org/Endorsement/PRISMAEndorsers>). However, a post hoc analysis of the journals sampled in this systematic review found that 0% of journals made reference to the PRISMA statement as part of their instructions to authors. By explicitly endorsing the PRISMA or MARS guidelines, journals will communicate to authors the importance of adhering to best practice guidelines as part of the review process. Journal editors will also need to be aware of the potential barriers associated with page length requirements on providing transparency of the meta-analytic process and allow for sensible solutions (e.g., printing aspects of manuscripts as online supplementary information for the reader).

Both of these recommendations will take time to implement across the field; however, we expect to see immediate improvements in meta-analytic research if, at minimum, synthesists: (a) identify research questions with a broad enough scope to support adequately powered estimation of average effect sizes, effect size heterogeneity, and moderator coefficients; (b) involve a librarian early in the research planning stages so that they can assist in strengthening the search strategy to improve the representativeness of meta-analytic samples; (c) clearly operationalize search strategies, eligibility criteria for study inclusion, and study variables to advance replication efforts; (d) evaluate meta-analytic assumptions and include power analyses to improve internal validity; and (e) adequately describe their universe of generalization and sample characteristics to allow readers to evaluate external validity. However, as noted earlier, meta-analysts within the

educational measurement field should make every effort to adhere to best practice guidelines, such as PRISMA. Regardless, it is our hope that this systematic review has brought to light some of the concerning practices related to conducting and reporting meta-analyses in the field of educational measurement, and as a result, sparks conversations about how we can improve transparent and accurate reporting of meta-analyses in an effort to improve evidence-based practices in our field.

Author Contribution Statement

The first author conceived of the presented idea. All authors were involved in data collection with the fourth author conducting the search, authors two and three selecting studies and coding variables, and the first author supervising all work. The first three authors conducted data analyses and interpreted the findings. Although the majority of writing was done by the first two authors, every author wrote different aspects of the manuscript. Authors three and four provided feedback on the draft, and the first and second authors conducted critical revisions of the article throughout the review process. Final approval of the version to be published was made by all authors.

Note

¹Based on one reviewer's suggestion, we conducted an additional database search of the journals which published the five included studies produced by our citation search. These journals included: *Academy of Management Learning and Education*, *American Journal of Pharmaceutical Education*, *Journal of Learning Disabilities*, *Journal of Educational Psychology*, and *Academic Medicine*. This search, which was completed on May 10, 2019, produced a total of 79 articles (four of which underwent a full-text review); however, none were found to meet the eligibility criteria of this study.

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*, 436–476.

American Psychological Association. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*, 839–851.

Bennett, R. E. (2016). *Opt out: An examination of issues* (ETS RR-16-13). Princeton, NJ: Educational Testing Service.

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research, 45*, 239–270.

Cooper, H. (2009). Hypotheses and problems in research synthesis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 19–36). New York, NY: Russell Sage Foundation.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36–49.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3–8.

Harwell, M., & Maeda, Y. (2008). Deficiencies of reporting in meta-analyses and some remedies. *The Journal of Experimental Education, 76*, 403–430.

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions*. London, England: Cochrane Collaboration.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139–155.

Khademi, A. (2013). *Journals and publications in educational measurement, statistics, research and psychometrics*. Unpublished manuscript, University of Massachusetts, Amherst, MA.

Koffel, J. B. (2015). Use of recommended search strategies in systematic reviews and the impact of librarian involvement: A cross-sectional survey of recent authors. *PLoS One, 10*(5), e0125931.

Koffel, J. B., & Rethlefsen, M. L. (2016). Reproducibility of search strategies is poor in systematic reviews published in high-impact pediatrics, cardiology and surgery journals: A cross-sectional study. *PLoS One, 11*(9), e0163309.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lin, H., Chen, T., & Liou, H. (2017). Transparency of reporting in CALL meta-analyses between 2003 and 2015. *Recall, 30*, 253–277.

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 537–560). New York, NY: Russell Sage Foundation.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. (2000). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *Oncology Research and Treatment, 23*, 597–602.

Moher D., Liberati A., Tetzlaff J., Altman D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097.

Polanin, J. R., Maynard, B. R., & Dell, N. A. (2017). Overviews in educational research: A systematic review and analysis. *Review of Educational Research, 87*, 172–203.

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research, 86*, 207–236.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432–1463.

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*, 1049–1066.

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*, 328–364.

Slavin, R. E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher, 13*, 6–15.

Slavin, R. E. (2017). Evidence-based reform in education. *Journal of Education for Students Placed at Risk, 22*, 178–184.

Talento-Miller, E. (2008). Generalizability of GMAT® validity to programs outside the US. *International Journal of Testing, 8*, 127–142.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*, 375–393.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*, 215–247.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

- Appendix A
- Appendix B
- Appendix C
- Appendix D