J Comput High Educ (2014) 26:183–209
DOI 10.1007/s12528-014-9084-z

An exploration of bias in meta-analysis: the case of technology integration research in higher education

Robert M. Bernard · Eugene Borokhovski · Richard F. Schmid · Rana M. Tamim

Published online: 12 November 2014

Abstract This article contains a second-order meta-analysis and an exploration of bias in the technology integration literature in higher education. Thirteen meta-analyses, dated from 2000 to 2014 were selected to be included based on the questions asked and the presence of adequate statistical information to conduct a quantitative synthesis. The weighted random effects average was $g^{++} = 0.393$, $p < .000$. The article goes on to report an assessment of the methodological quality of the thirteen studies based on Cooper's (Research synthesis and meta-analysis: a step-by-step approach. Sage, Thousand Oaks, 2010) seven stages in the development of a meta-analysis. Two meta-analyses were found to have five out of seven stages where methodological flaws could potentially create biased results. Five meta-analyses contained two flawed stages and one contained one flawed stage. Four of the stages where methodological flaws can create bias are described in detail. The final section attempts to determine how much influence the methodological flaws exerted on the results of the second-order meta-analysis.

Keywords Technology · Computers · Meta-analysis · Bias · Higher education

Introduction

In April of 2014 our research team published an article (Bernard et al. 2014) in the *Journal of Computing in Higher Education* entitled "A Meta-analysis of Blended Learning and Technology use in Higher Education: From the General to the

R. M. Bernard (✉) · E. Borokhovski · R. F. Schmid
Center for the Study of Learning and Performance, Concordia University, Montreal, QC, Canada
e-mail: bernard@education.concordia.ca

R. M. Tamim
College of Education, Zayed University, Dubai, UAE

Applied" in which we simultaneously described the steps in doing a meta-analysis (i.e., the general) while presenting the results of a meta-analysis of the quantitative research literature of blended learning in higher education (i.e., the applied). This article might be judged as largely pedagogical in nature using a new meta-analysis as an example, or it might be thought of as the presentation of a new meta-analysis in an area of emerging concern in higher education supplemented by an extended description of the methodological issues involved in conducting it with the potential to be used for pedagogical purposes. Either way, the article barely scratches the surface of issues relating to bias and the reliability and validity of meta-analysis as a whole and the large number of technology integration meta-analyses that have been conducted over the last 15 years.

In this article we probe deeper into this literature and the practices that can make or break a meta-analysis in an attempt to arm consumers with the information needed to detect bias in the meta-analyses they read and producers with strategies to avoid bias in the first place. As a lead up to this, we conducted a second-order meta-analysis of 13 meta-analyses that have appeared in the technology integration literature in the last 15 years. Selected studies from this collection will be used to provide examples of both good and poor practices in conducting meta-analyses that can lead to biased and possibly misrepresentative results.

As a special form of systematic review, meta-analysis has several major goals that go beyond simply reviewing the literature around a given question (Jackson 1980). As a study of the population of primary research studies, one goal is to assess the state of the research literature—how big it is, how much commonality exists among research questions, what is the quality of the primary studies, what demographics it covers, etc. Another goal is to estimate the average effect size in the population (i.e., how well/poorly does the treatment work compared to a control condition). A related goal is to estimate the variability (i.e., between-study variance) of a distribution of effect sizes in order to judge whether it exceeds what would be expected by chance (i.e., sampling error). These steps would be analogous to calculating the mean and standard deviation of individual outcomes in a primary research study. In cases where excess between-study variability exists, a meta-analyst usually goes deeper by examining the effects of coded moderator variables. Moderator variable analysis lends texture, qualification and detail to the overall assessment.

Meta-analyses on technology integration conducted since 2000

The main purpose of this part of the paper is to identify and review meta-analyses that have been done in technology integration in higher education since 2000. We searched the literature for meta-analyses that addressed either higher education alone or included separate average effect sizes broken down by categories of grade level (i.e., effect sizes reported separately for higher education and K-12). Also only meta-analyses that contained a comparison between a technology treatment and a no-technology control group were considered.

Meta-analyses are conducted for a number of reasons and therefore can differ somewhat in focus and scope. Table 1 shows the meta-analyses on technology

Table 1 Thirteen recent meta-analyses of technology use in higher education and their characteristics ordered by year of publication

Meta-analyses	k	ES^{+}	Type of technology	Subject matter	Publication type	Inclusive dates
Bayraktar (2000)	108	0.27	CAI	Science	Dissertation	1970–1999
Christmann and Badgett (2000)	26	0.13	Various	Various	Article	1983–1996
Hsu (2003)	31	0.43	Various	Statistics	Dissertation	1985–2002
Zhao (2003)	9	1.12	ITT	Language Learning	Article	1997–2001
Koufogiannakis and Wiebe (2006)	8	−0.09	CAI	Information Literacy	Article	1967–2005
Timmerman and Kruepke (2006)	118	0.24	CAI	Various	Article	1994–2005
Michko (2007)	123	0.43	Various	Engineering	Dissertation	1996–2005
Schenker (2007)	46	0.24	Various	Statistics	Dissertation	1975–2005
Tekbiyik and Akdeniz (2010)	65	1.12	Various	Science	Article	2001–2007
Larwin and Larwin (2011)	219	0.57	CAI	Statistics	Article	1960–2010
Sitzmann (2011)	39	0.28	Simulations	Various	Article	1976–2009
Sosa et al. (2011)	45	0.33	CAI	Statistics	Article	1974–2005
Schmid et al. (2014)	479	0.25	Various	Various	Article	1990–2010

integration conducted in higher education from 2000 to the present that we included, along with average effect sizes and some specific characteristics of each. Some, like the Zhao (2003) meta-analysis, were performed to inform consumers in a particular specialized area, specifically technology use in language learning, about the effects based on a relatively few studies (9) conducted in a fairly constrained timeframe (1997–2001) and based on extremely limited sample of only published studies. The Koufogiannakis and Wiebe (2006) meta-analysis on information literacy is also of this type—a relatively small number of studies (8) and a very specific target audience.

In the collection, we found (amazingly) that four out of the 13 meta-analyses focus on the teaching of statistics using technology. Statistics is a somewhat more general content area than the previous two, owing largely to the fact that statistics is taught widely in departments of psychology, education, commerce, sociology, and in many other social science disciplines. It is often perceived as a difficult subject matter (remember the aphorism "my students call it sadistics") and that technology might play a role in making it more palatable and thus easier to learn. The four meta-analyses of statistics education are Hsu (2003), Schenker (2007), Larwin and Larwin (2011) and Sosa et al. (2011). Two of these are unpublished dissertations and the other two appear as journal articles.

Then, there are meta-analyses targeting a more general scope of content (e.g., engineering, science, mathematics) that are intended for a wider audience of practitioners. The meta-analyses by Bayraktar (2000), Michko (2007), Tekbiyik and Akdeniz (2010) addressed technology issues in these content areas. One meta-analysis (Sitzmann 2011) dealt with a training audience (including higher education) and focused specifically on educational simulations.

Finally, there is the category of meta-analysis that aims to look at technology integration in higher education in a very general way. There are three such meta-analyses. Two looked at all forms of technology and a variety of subject matters (Christmann and Badgett 2000; Schmid et al. 2014) and the other, Timmerman and Kruepke (2006) looked only at CAI (e.g., computer-assisted learning).

By far the largest and most wide-ranging meta-analysis is the one by Schmid et al. (2014). In fact, the effect sizes shown in Table 1 are only half of those that were published by Schmid et al. The other half of the meta-analysis included technology in the treatment condition compared to a control condition that also received some form of technology and thus would not align with the other meta-analyses included here that considered only technology-free control conditions. We will return to a discussion of this side of the Schmid et al. meta-analysis later in this paper.

There were two meta-analysis that were not included in the final collection, one that purported to address gaming and simulation (Vogel et al. 2006) and the other that dealt with virtual reality at various educational levels (Merchant et al. 2014), but the results of K-12 grade levels were not separated from higher education results. In addition, the Vogel et al. study does not include average effect sizes, just confidence intervals. Another excluded meta-analysis is one conducted by Rolfe and Gray (2011) on the effects of technology in life sciences courses (e.g., physiotherapy, medicine, dentistry). In many ways this study was carefully conducted but the

authors chose to use unstandardized effect sizes (i.e., raw mean differences not divided by a standard deviation) and so the average effect sizes do not match any of the other meta-analyses reported here. Borenstein et al. (2009) present this as an option, but only if all of the studies in a meta-analysis are measured on the same common metric (e.g., blood pressure measurement). Here, the primary studies clearly differed in terms of instrumentation rendering this approach untenable and the results uninterpretable.

The final collection of 13 meta-analyses encompassed 1,316 primary studies, some that were undoubtedly overlapping, especially in the more general meta-analyses. Overlap in research participants can be a problem in first-order meta-analyses, especially when a single control condition was used repeatedly for multiple treatment comparisons. These overlaps lead to dependencies in the data that tend to inflate α and increase the possibility of making a Type I error (i.e., rejecting the null hypothesis when it should be accepted). Various remedies are proposed by (Scammacca et al. 2013) for first-order meta-analyses. In a second-order meta-analysis the possibility of overlapping studies is high, but because the standard error (i.e., the denominator of a *z*-ratio) is based on the number of studies, instead of participants, the problem is less severe.

Second-order meta-analysis

A second-order meta-analysis is a meta-analysis of meta-analyses that is intended to convey the highest-level information about the relationship between a treatment and a control condition in the population. John Hattie (2009) has synthesized more than 800 meta-analyses around 138 educational variables that are related to student achievement. Unlike a regular meta-analysis, a second-order meta-analysis typically does not include much if any moderator variable analysis, mainly because it addresses pre-existing meta-analyses that usually have unique moderator structures that cannot be easily reconciled. Therefore, as Cooper and Koenka (2012) state: "moderating and mediating variables must exist at the level of the research syntheses that are the constituent elements, not at the level of the individual studies" (p. 458). Another point is that second-order meta-analysts may observe a change in methodological practices over time, since the methodology of meta-analysis itself has undergone significant changes since it was first introduced by Glass (1976). This includes not only advances in statistical methodology (e.g., Hedges and Olkin 1985; Borenstein et al. 2009, 2010), but also the accessibility and richness of literature sources. With this in mind we have chosen 2000 as the cut-off date for this second-order meta-analysis.

The entire literature of technology integration in education (K-higher education) was addressed in a second-order meta-analysis (Tamim et al. 2011). They found a weighted average random effect size of 0.35, $k = 25$, $p < .001$. Tamim et al. did two things before synthesizing 25 meta-analyses dating from 1985 to 2008 that we will not do. They selected 25 meta-analyses out of nearly 75 that had an overlap no greater that 25 % of primary studies and coded them for methodological quality. The former was done to reduce the dependency problem that occurs when samples

Table 2 Summary statistics for the second-order meta-analysis

Models and heterogeneity	Effect size and 95 % confidence interval				Test of null (2-tail)	
	K[a]	g^{++}[b]	*SE*	95 % CI	*z*-value	*p* value
Unweighted average	13	0.409	No statistics can be computed			
Fixed effect	13	0.362	0.04	0.28/0.44	9.14	.00
Random effects	13	0.393	0.07	0.25/0.53	5.48	.00
Heterogeneity	Q-total = 28.37	$df = 12$	$p < .01$		$I^2 = 57.70$	$\tau^2 = 0.03$

[a] K means number of meta-analyses

[b] g^{++} refers to the Hedges' g for weighted average effect sizes in the second-order meta-analysis

are used repeatedly. Our purpose is different. We want to retain as many meta-analyses as possible so that later on in this article the methodological strengths and weaknesses, potentially leading to bias, can be discussed. The dependency problem is somewhat reduced here because only four out of 13 studies examine technology in general. The rest are in specific content areas like mathematics, statistics and language learning which draw from the literatures of different content areas.

We conducted a search for studies in much the same way that is done for first-order meta-analyses. To be included, meta-analyses had to be published from the year 2000 onward, had to deal with higher education or some defined adult population and in the case where K-12 studies were also addressed had to contain average effect sizes with higher education separated from other grade levels. Studies addressing any subject matter or type of technology were admitted and technology used in experimental conditions had to be compared to technology-free control conditions in terms of student achievement outcomes. There were no minimum standards for methodological quality.

All effect sizes that were originally calculated as Cohen's d, or some other metric, were converted to Hedges' g for the sake of consistency. Hedges' g is considered an unbiased estimator because it adjusts for the tendency for small samples to over-estimate the true effect size. In samples that are larger than about $k = 30$, Cohen's d and Hedges' g converge. We used the reported average adjusted weighted effect sizes but produced separately calculated standard errors. This was because the unit of analysis in a second-order meta-analysis is the average effect size of each meta-analysis, constructed from samples, rather than participants of the individual research studies. Thus, we used K (i.e., the number of effect sizes in each included meta-analysis) as the unit of analysis, and since each primary study contained both a treatment and control condition, we doubled K. The same principle is used when conducting a meta-analysis; both treatment and control participants are counted in the overall sample size. As a result, the standard error for each meta-analysis is more conservative (i.e., larger) than was reported in the original first-order meta-analyses.

The results of this analysis are shown in Table 2. The weighted average fixed and random effects are close but not identical ($g^{++} = 0.362$ and 0.393, respectively) and they are both significantly greater than zero. Interestingly, the weighted random effect size that Tamim et al. (2011) found in their second-order meta-analysis was

slightly lower but in the same range as these findings. The unweighted average was slightly higher than that produced by the fixed effect model ($g^{++} = 0.409$, $k = 13$), but basically the same as the random effects weighted average effect size in this study. Since all studies were weighted the same, the relatively large effect size produced by Zhao (2003, $ES^{+} = 1.12$), for example, brings the overall unweighted average up, even though its sample size was small. Like in Tamim et al. the distribution in this second-order meta-analysis was heterogeneous ($Q_{\text{Total}} = 28.37$, $p < .01$), suggesting that these data are *not* a good fit to the fixed effect model, and that between-study variability that exceeds chance expectation is present. The I^2, or the percentage of true heterogeneity exceeding chance expectations, was 57.70 % indicating moderate variability among meta-analyses. The value $\tau^2 = 0.03$, the average between-study variability, was added to within-study variability to produce the weights for the random effects model.

There are several ways to interpret this average effect size. One way is to think of the average effect size in standard deviation terms; the average treatment condition exceeded the average control condition by 0.393*sd*. Another way is to rate it according to Cohen's (1988) effect size criteria. An average effect size of 0.393 is considered to be on the lower boundary of a medium effect size. Also it can be described in terms of the percentile difference between the mean of the control and treatment conditions, under the normal distribution. In this case, a person at the median of the no technology control condition would be presumed to increase from the 50th to the 65.29$^{\text{th}}$ percentile in achievement if they had received the treatment (i.e., a 15.29 % increase).

Publication bias analysis is the process of estimating the number of studies that may be missing from a distribution of effect sizes (i.e., still in the "file drawers of researchers") that might change the conclusions of the meta-analysis. The presumption is that a lack of non-published studies (e.g., conference papers, technical reports) that may contain lower or even negative effect sizes can create a positive bias that is misleading. There are a number of statistical techniques, each using a different approach to estimation, which can be used to assess publication bias. In addition, an effect size by sample size (represented by the standard error) plot, called a "funnel plot" can help a meta-analyst visualize the shape and symmetry of the distribution. In the absence of publication bias we would expect the studies to be distributed symmetrically around the combined effect size. By contrast, in the presence of bias, we would expect that the bottom of the plot would show a higher concentration of studies on one side of the mean than the other. This would reflect the fact that smaller studies (which appear toward the bottom) are more likely to be published if they have larger than average effects, which makes them more likely to meet the criterion for statistical significance. The distribution shown in Fig. 1 is symmetrical around the weighted average effect size. There is only one effect size (i.e., Tekbiyik and Akdeniz 2010) that is outside of the funnel on the right side of the graphic.

We also ran a publication bias analysis on the second-order data and found that, according to the Classic Fail-Safe *N* test, it would take 233 additional zero-effect meta-analyses to nullify the effect of $z = 8.51$ and bring it below $z = 1.96$ (the 2-tailed z value at $p = .05$). Orwin's Fail-Safe *N* indicates that an additional 19

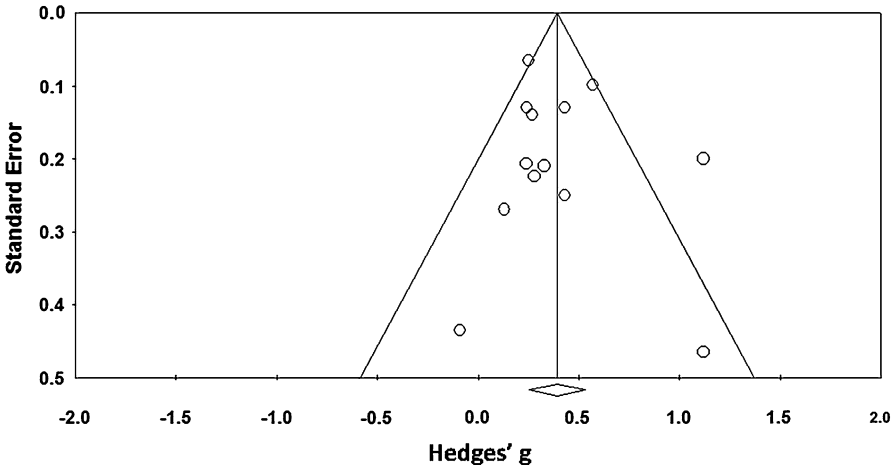

Fig. 1 Funnel plot (random effects model) of the 13 meta-analyses

meta-analyses would be necessary to bring the g^{++} of 0.393 to a trivial level of $g^{++} = 0.15$ (this value was chosen arbitrarily; 81 are need for a $g^{++} = 0.05$). Duval and Tweedie's (2000) Trim and Fill approach found that no studies needed to be trimmed or filled to reach homogeneity of effect size. As stated previously, a funnel plot (i.e., effect size by standard error) of these effect sizes produced symmetry around the weighted average of random effects mean of 0.393 indicating that no additional studies needed to be imputed. The conclusion is that the data are a reasonable fit to the assumptions of the fixed and random effects model and that it can be interpreted safely.

These findings, as positive as they seem statistically, do not mean that the meta-analyses that make up this distribution are methodologically perfect. It simply means that as a collection, they work well together. But this is not the only way of examining the question of the validity of the results. Meta-analyses are intended to produce top-level information about a research question that is of interest to a particular segment of practitioners, researchers, or policy makers.

We also performed a one-study-removed analysis in *Comprehensive Meta-Analysis*™ version 2.2 (Borenstein et al. 2005) to determine if any of the average effect sizes had a undue influence on the averaged weighted random effect size of $g^{++} = 0.393$. The results are shown in Table 3. The second column from the left (referred to as "Point") is the average weighted effect size with each study removed and the overall average re-calculated. The maximum increase is +0.025 (Schmid et al. 2014) and the maximum decrease is −0.056 (Larwin and Larwin 2011) and all of the re-calculated effect sizes fall within the 95th confidence interval of the weighted average ($CI = +0.227/+0.440$). None of these average effect sizes were considered to be outliers.

These results are slightly higher than the second-order meta-analysis results reported by Tamim et al. ($ES^{++} = 0.35$) for 25 meta-analyses (i.e., six of which overlapped the present study) from all disciplines and all age groups. In addition, the average effect size here ($g^{++} = 0.393$) falls within the confidence interval of the

Table 3 One-study-removed analysis of the average effect sizes of 13 meta-analyses (replica of comprehensive meta-analysis table)

Study name	Statistics with study removed						
	Point	Standard error	Variance	Lower limit	Upper limit	z-value	*p* value
Bayraktar (2000)	0.41	0.08	0.01	0.25	0.56	5.15	0.00
Christmann et al. (Christmann and Badgett 2000)	0.41	0.07	0.01	0.26	0.55	5.45	0.00
Hsu (2003)	0.39	0.08	0.01	0.24	0.54	5.14	0.00
Zhao (2003)	0.38	0.07	0.00	0.24	0.52	5.35	0.00
Koufogiannakis and Wiebe (2006)	0.40	0.07	0.01	0.26	0.55	5.55	0.00
Timmerman and Kruepke (2006)	0.41	0.08	0.01	0.26	0.57	5.21	0.00
Michko (2007)	0.39	0.08	0.01	0.23	0.55	4.86	0.00
Schenker (2007)	0.40	0.08	0.01	0.25	0.55	5.29	0.00
Tekbiyik and Akdeniz (2010)	0.34	0.05	0.00	0.24	0.43	6.93	0.00
Larwin and Larwin (2011)	0.37	0.08	0.01	0.22	0.51	4.86	0.00
Sitzmann (2011)	0.40	0.08	0.01	0.25	0.55	5.24	0.00
Sosa et al. (2011)	0.40	0.08	0.01	0.25	0.55	5.17	0.00
Schmid et al. (2014)	0.42	0.08	0.01	0.26	0.57	5.16	0.00
Random	0.39	0.07	0.01	0.25	0.53	5.47	0.00

Tamim et al. results ($CI = 0.30/0.413$) rendering the results essentially identical. Both studies produced heterogeneous results.

Bias in meta-analysis

Methodological bias has been defined as *systematic inaccuracy in data due to characteristics of the processes employed in its collection, manipulation, analysis, interpretation and/or presentation of research findings* (Bernard 2014). Meta-analyses are, by definition, time-bound and always retrospective—in some sense a meta-analyst is a prisoner of the past, with little latitude to go beyond the efforts of previous primary researchers. This means that the meta-analyst can do little to improve the quality of the evidence that is included. However, a meta-analyst can present either an accurate or a biased picture of the past, depending upon a host of large and small decisions they must make. There are many sources of information about how to conduct a meta-analysis (e.g., Lipsey and Wilson 2001; Cooper 2010; Bernard et al. 2014) including discussions on how to avoid bias at various stages. However, it is sometimes the case and for a variety of reasons, that researchers conducting actual meta-analyses do not always follow best research practices. As a result there have been a number of attempts to develop assessment tools for meta-analyses that have already entered the literature (e.g., Higgins et al. 2012; Schlosser et al. 2005, 2008; Shea et al. 2007), in much the same way that Valentine and Cooper (2008) have addressed the validity of primary research by developing *The Study Design and Implementation Assessment Device* (Study DIAD).

Our research team has been working on an instrument for assessing meta-analyses in the social sciences, including education (Tamim et al. 2011). The instrument, tentatively called the *Methodological Quality Instrument for Meta-Analysis* (MQIM), consists of 22 items, expressed as questions and broken into sections that roughly map onto Cooper's (2010) steps for conducting a meta-analysis: Step 1—Formulating the problem; Step 2—Searching the literature; Step 3—Gathering information from studies; Step 4—Evaluating the quality of studies; Step 5—Analyzing and integrating the outcomes of research; Step 6—Interpreting the evidence; and Step 7—Presenting the results. The seventh step is not included, as quality of reporting (e.g., amount of details may depend on limitations in space) does not necessarily reflects the quality of the conducted meta-analysis directly. See "Appendix" for an outline of these criteria.

In the section that follows we explain how the MQIM was used to evaluate the 13 meta-analyses and in Table 1, the results of its application and subsequent analyses are presented. We chose four of the sections where bias might be present to discuss in detail. These are illustrated by problematic areas that we identified in each of the meta-analyses marked with an **X** in Table 4.

Rating the meta-analyses

Two expert reviewers who have participated in conducting many previous meta-analyses rated the 13 meta-analyses (Table 1) on each of the 22 items using a three-

Table 4 Rating of meta-analyses across Cooper's categories for the development and presentation of a meta-analysis (see "Appendix" for the entire instrument)

Meta-analyses	Formulating the problem	Searching for literature	Gathering information from studies	Evaluating the quality of studies	Analyzing research outcomes	Interpreting evidence	Interrater reliability[a]
Christmann and Badgett (2000)	✔	**X**	**X**	**X**	**X**	**X**	0.82
Bayraktar (2000)	✔	✔	✔	✔	✔	✔	0.90
Zhao (2003)	✔	**X**	**X**	**X**	**X**	**X**	0.82
Hsu (2003)	✔	✔	**X**	✔	✔	✔	0.82
Koufogiannakis and Wiebe (2006)	✔	✔	**X**	**X**	✔	✔	0.90
Timmerman and Kruepke (2006)	✔	**X**	✔	**X**	✔	✔	0.91
Michko (2007)	✔	✔	✔	**X**	✔	✔	0.90
Schenker (2007)	✔	✔	✔	**X**	✔	✔	0.90
Tekbiyik and Akdeniz (2010)	✔	✔	✔	**X**	**X**	✔	0.64
Larwin and Larwin (2011)	✔	✔	**X**	✔	**X**	✔	0.82
Sosa et al. (2011)	✔	✔	✔	✔	✔	✔	0.64
Sitzmann (2011)	✔	✔	✔	✔	✔	✔	1.00
Schmid et al. (2014)	✔	✔	✔	✔	✔	✔	0.91

Seems OK (✔), Possible bias (**X**)

[a] Cohen's κ = percentage of agreement (number of agreements/number of times) –0.50/0.5

point scale: (1) fully meets methodological standards; (2) does not meet methodological standards fully; and (3) does not meet methodological standards at all. These 13 meta-analyses were part of a collection of 48 meta-analyses being rated at the same time as a part of the larger project of developing and validating the MQIM. The 13 meta-analyses were rated in different orders and the coders were unaware which would be included here and which would not. These 22 ratings for each meta-analysis were collapsed into six methodological categories and averaged so that each category was rated on the three-point scale (✔ $\geq$ 1.51, **X** $\leq$ 1.50). The inter-rater reliabilities are shown in the far right column of Table 4. In all but two cases, the inter-rater reliability was above $\kappa = 0.80$. Where disagreements did arise, they were settled through discussion between the reviewers.

The **X**-marks do not necessarily mean that bias is present. Instead, it is an indication of concern and a way of signaling to consumers and producers of meta-analyses that these areas should be given special attention. Since we worked from published accounts of meta-analyses or dissertations that apply meta-analytic processes, it is entirely possible that authors simply failed to report details within some of the sections. This is in itself problematic, since the published work (or unpublished dissertation) is all that a consumer or evaluator has to work from in judging the quality of the effort. An accepted principle of all research, including systematic reviews, is that its methods and results should be fully transparent to its audience, and in the best cases, replicable by other researchers.

Searching for literature

Searching the literature to locate relevant primary experimental studies that inform the research question is one of the key components of conducting a meta-analysis based upon a systematic review. As we have emphasized throughout this paper, bias may be introduced at any stage, but this failure to thoroughly search for studies lays the foundation for many of the others that follow. If the studies identified through literature searches are not representative of the population with respect to the research question, bias is inevitable no matter how flawlessly all the following stages are conducted.

To be comprehensive and relatively protected from so-called publication bias (a tendency to over-represent significant findings or large effect sizes in the published literature), searches should target unpublished sources often referred to as grey literature (e.g., conference presentations, unpublished dissertations and theses, public and private research reports). To create a representative picture of the relevant research literature, a broad variety of electronic databases should be searched with necessary adjustments in search terminology and strategies for each. Thorough and systematic literature searches would also include reviews of the bibliographies of previously published reviews and major primary studies in the field (also called branching), web-searches for various grey literature and manual searches of the tables of contents of the most relevant journals and conference proceedings. In addition, the search strategy (i.e., targeted literature, sources of data

and key words in specific combinations) should be made transparent (i.e., well documented and incorporated into the report of the study) to enable the possibility of replication. For more detail on the concept of properly examining research literature for a meta-analysis and on specifics for search strategies see Hammerstrøm et al. (2010) and Lefebvre et al. (2011).

There is a controversy in the literature of systematic review regarding the advisability of contacting individual researchers in a field to solicit studies that might not appear in the published literature (i.e., true file drawer studies). While it is an established practice in the health sciences literature, we see two objections. First, if the field of study is large and diverse, as the literature of technology integration is, the list of potential authors to be contacted is likely to be very large and the "hit rate" very low, resulting in considerable inefficiency. In smaller literatures, this inefficiency might be tolerable. Second, and probably of greater importance is that the procedure violates the important principle of replicability. Since studies found in this way are, by definition, not present in the public record, another reviewer attempting a replication cannot examine the original studies without contacting the same authors. Creating and maintaining a "trials register" (i.e., a very large database of individual intervention studies) is one way of providing reviewers with difficult to access studies. The EPPI-Centre at the University of London provides one called CERUK that holds randomized and quasi-experimental studies on all aspects of educational research in the United Kingdom (http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=185).

Three meta-analyses in our collection appear not to meet requirements for sufficient methodological quality when it comes to the stage of searching the literature. Christmann and Badgett (2000) limited their literature searches to three electronic databases and reported only a short list of key-terms without denoting how they were applied to individual databases. Similarly, Timmerman and Kruepke (2006) reported only "principal" key-terms, and more importantly, indicated no efforts to locate unpublished literature sources.

Finally, in our view the most problematic case is the meta-analysis by Zhao (2003). Technically, this meta-analysis does not belong to the category of systematic review. What was done, more or less systematically, is the reduction of literature sources (this process is well described) from the original collection of journals to five titles that the author deemed to be more informative (i.e., with higher frequencies of publications on the topic of his meta-analyses). However, the result of this process could not be considered a representative sample of relevant experimental research. It is not only limited to articles published in peer-reviewed journals (identified through search in a single electronic database using an extremely restrictive set of key-words), but the list of these journals is abridged from twenty-two to just five. That leaves the final collection of studies wide open to publication bias. The Zhao review appears to us to be more like a "brief review," as described by Abrami et al. (2010), as an attempt to summarize the literature in an abbreviated form without including everything. In all cases where weaknesses were detected, the corresponding meta-analysis received an "**X**".

Gathering information from studies

This step begins with the assessment of manuscripts identified through searches and the decision to retain them for further consideration based on previously established inclusion/exclusion criteria. Setting these criteria is for the most part linked to the purpose of the meta-analysis and the availability and relevance of the data within studies, with a view towards answering the research question. The process of gathering information from studies continues with coding of study characteristics, including effect size data, methodology and demographic data as well as coding for substantive study features. Accuracy in these processes is critical to maintaining the quality of a meta-analysis and minimizing bias.

Cooper (2010) devotes considerable attention to describing the development of a codebook, coding sheets, training coders, establishing inter-rater reliability, dealing with missing information, etc. One of the principles that he espouses is absolute necessity for multiple independent coders who are trained in extracting various kinds of information from studies. He distinguishes between *low-inference coding* and *high-inference coding* (essentially the difference between locating and extracting factual information and using judgment in rating study characteristics on a scale). Cooper emphasizes the requirement for multiple trained coders by saying "There is simply too much room for bias (conscious or unconscious), for idiosyncratic interpretation of coding questions and responses, and for simple mechanical error for the unverified codes of a single person to be considered part of a scientific synthesis of research" (p. 101). He also advocates for extensive training (less for *low-inference coding*, and more for *high-inference coding*) and that inter-coder reliability be calculated both during training and throughout the coding process. Missing or unclear information needs to be dealt with uniformly across a set of studies and coding needs to be examined for potential bias that can be introduced through the predisposition of coders' to favor one interpretation over another when judging ambiguous reporting.

In our collection, if a meta-analysis described all aspects of the study retention and coding processes and was in compliance with Cooper's suggestions, it was given a ✔. Four out of the thirteen meta-analyses (all marked with an **X**) either explicitly used only one coder or failed to discuss the coding process, including how many coders were used, whether coders were trained or how they performed (i.e., inter-coder reliability). Christmann and Badgett (2000), Zhao (2003) and Koufogiannakis and Wiebe (2006) did not mention coding or coders at all. Hsu's (2003) meta-analysis was a dissertation and explicitly states that the author alone made all extraction and coding decisions. Larwin and Larwin (2011) used the word "we" in describing their coding process but give no further information or details of how the coding was carried out and how reliable its results were. Likewise, they gave no information about inclusion/exclusion decisions or effect size extraction.

Evaluating the quality of studies

The data that comprise a meta-analysis are results derived from the primary literature in a chosen field of study. Allowing unreliable or questionable data to

enter into a meta-analysis can contaminate the analytical stage of the review, potentially rendering its findings biased, or in the worse case, completely untrustworthy. The imperative at this stage, then, is to minimize risk by allowing only the highest quality evidence possible into the review. However, in the social sciences, and particularly in education, several questions arise that should be considered before the process of evaluating quality is undertaken. First, what is the best quality evidence around a given question? Often it is randomized control trials and high-quality quasi-experiments, but sometimes, random assignment and pretesting are not possible so that lesser quality studies must be included. A second question involves decisions about the quality (i.e., validity and reliability) of measurement. Sometimes high quality standardized instruments are the norm, but usually they are not. Ultimately, the meta-analyst must decide whether it is better to offer some evidence to consumers or if evidence is of such poor quality that it is better not to conduct a meta-analysis at all. In the former case, it is incumbent of meta-analysts to be clear about the choices they make.

Among the meta-analyses we reviewed, too many lacked attention to these and other aspects of methodological quality and rigor. Below we provide examples of the included meta-analyses that were judged to be susceptible to bias at the stage of evaluating quality of research evidence as it is depicted (**X**) in Table 4.

There are four main issues that should be addressed to minimize bias in a meta-analysis, or at least to categorize the findings in terms of the methodological quality. First and foremost, it is the validity of individual primary research. One of the most recognized and thorough instruments for determining this is the Study DIAD: *Study Design and Implementation Assessment Device* (Valentine and Cooper 2008), mentioned earlier. At its top level, the instrument provides the meta-analysts with the means of assessing: (1) internal validity; (2) measurement and construct validity; (3) statistical validity; and (4) external validity of the studies that are considered for inclusion in a meta-analysis. Many researchers tend to focus almost exclusively on the first category by classifying research designs used in individual studies and often excluding pre-experiments or even quasi-experiments. Measurement, statistical, and external validity may be totally overlooked. We argue that a balanced approach should be used by meta-analysts; one that considers the validity of the design, the psychometric quality of assessment tools, the appropriateness and precision of statistical procedures, as well as the representativeness and generalizability of the findings. Even if compromises to best quality evidence are required, the issue should be addressed openly by the meta-analyst so that it is clear how the results are qualified or how the overall methodological quality estimate (i.e., some kind of a composite score of all four validity measures) is used in moderator variable analyses or as an adjustment term for the average effect size (e.g., Bernard et al. 2009).

Concerning the meta-analyses being addressed here, Timmerman and Kruepke (2006) provide no evidence that they assessed the quality of primary research. None of the threats to study validity were considered at the inclusion/exclusion stage of their review, nor were any aspects of study validity coded and analyzed in moderator variable analysis. With the exception of excluding studies that used "… additional non-equivalent manipulations across the CAI and non-CAI groups without reporting performance scores for each manipulation …" (p. 80), we know

practically nothing about the quality of primary empirical studies included in this meta-analysis. By contrast, Koufogiannakis and Wiebe (2006) documented in detail all types of research design in the studies admitted to their meta-analysis. However, since pre-experiments were not excluded and no differential treatment or any adjustment (e.g., through weighting or moderator variable analysis) for them is reported, we judged this stage of evaluating the quality of the evidence in this study to be insufficient.

A no less important aspect of a reputable meta-analysis is its attention to the issue of data independence. In primary empirical studies, practically no one would use data collected from the same participant several times, unless of course it is a repeated measures design. It would also be poor research practice to average scores for several incompatible measures—this is why multivariate analysis of variance (*MANOVA*) was created. Unfortunately, even well trained and experienced researchers, when conducting a meta-analysis, far too often violate these basic principles of primary research. It is a mistake to average, say, achievement, attitudes and various behavioral measures into one aggregate effect size, if for no other reason because the meaning of such composite of various dependent measures is obscured. Likewise, repeatedly using the same control group in comparisons with several experimental groups in different treatment conditions causes a rise in the Type I error rate just like it does in primary research. When reading a meta-analysis, one needs to be mindful about a possibility of these and similar missteps as some of them may result in shortcomings well beyond a potential bias.

One example from this collection of meta-analyses where violating the principle of aggregating measures was used allegedly to reduce issues of data dependency appears in the study by Zhao (2003). The author reports averaging effect sizes when more than one was extracted from the same study (i.e., a legitimate procedure if the measures are of the same type). However, since the aggregated effect size included "… measures of listening, reading, writing, cultural knowledge, and student attitudes …" (p. 18), the meaning of the average effect size was lost. This course of action without clearly separating outcomes into internally coherent types (i.e., distinguishing between listening comprehension and writing, not to mention attitudes and cultural knowledge) appears to be completely invalid, even if it is done to avoid issues of dependency.

The only difference between averaging at the level of an individual study or at the level of aggregating effects across studies is that in the latter case the undesirable consequences could be aggravated as dependency of repeatedly using the same participants is added to dependency/incompatibility among outcomes. In many of the 13 meta-analyses in this collection we do not really know for sure whether averaged outcomes are of exactly the same nature or not, but it is obvious that in some, data from the same participants were used more than once. Consider, for example, data from Larwin and Larwin (2011) or Christmann and Badgett (2000) studies (in both cases data were presented in a table format summarizing the outcome source/measure and the corresponding number of participants for each included effect size). It is clear that participant groups have been used repeatedly to

increase the number of effect sizes in the meta-analyses, thus potentially introducing bias associated with the data dependence.

The last two issues at this stage of conducting a meta-analytical review, publication bias (e.g., Rothstein et al. 2005) and detection and treatment of outliers (e.g., Viechbauer and Cheung 2010), are typically dealt with by applying special analytical procedures. As previously noted, publication bias is rooted in a tendency of research journals to primarily publish significant findings and refers to the possibility that a large portion of research from other sources (often denoted as the grey literature), that may have produced less positive results, have been excluded from the analysis. Publication bias is partially addressed through exhaustive literature searches that should target both published (e.g., in peer-reviewed journals) and unpublished (e.g., dissertations, conference presentations) empirical research. In addition, meta-analysts employ a combination of other statistical techniques, some of which have been addressed previously in the section that describes our own second-order meta-analysis.

Attention to the issue of publication bias was not very common among the reviewed meta-analyses. Eight of 13 of them either did not acknowledge it at all or attempt to detect its presence and assess its magnitude. When truly exhaustive systematic literature searches are conducted, publication bias is somewhat of a lesser concern. However, when the literature in a meta-analysis is limited to published studies, a set of selected journals (as in the case of Zhao 2003, a single one per each relevant thematic area) the issue of publication bias can represent a serious methodological flaw.

The presence of outliers in a meta-analysis creates a problem similar to that in primary research, and can substantially affect the mean and variability of the distribution of effect sizes. A statistical technique called "one study removed" is often used to detect outliers (high leverage effects, either in terms of their own magnitude or associated with it anomalously large sample size, or both). It repeatedly recalculates the average effect size with each study removed from the distribution, in turn, to estimate the relative affect of each effect size on the mean. Outliers may also be detected by examining the magnitude of the residuals in simple linear meta-regression. Both of these procedures are available in *Comprehensive Meta-Analysis*™ (Borenstein et al. 2005). Outliers detected in this fashion need to be given particular consideration, especially if some underlying aspect of the study is a contributory factor (e.g., very low variability). Effect sizes that are judged to be out of a reasonable range, generally $\pm$ 3.0*sd*, can either be removed or Winsorized (i.e., reduced to the next highest value on either side of the distribution). The latter applies to both the magnitude of the effect sizes themselves and to the associated sample sizes, since the average effect size is weighted by sample size, both under the fixed effect and the random effects models.

None of the meta-analyses marked with **X** in Table 4 dealt with outliers (at least, not explicitly judging from the corresponding reports). Of course, a possibility exists that outliers (either in effect size magnitude or in sample size) were not present there, but the failure to share this information with the readers, was nevertheless classified in our review as inadequate attention to the issue.

Analyzing research outcomes

There are several important issues that arise when effect sizes are synthesized and moderator variable analysis is conducted. The first is the choice of the analytical model that is to be used to synthesize the results. There are three approaches that can be taken to arrive at an average effect size: (1) average the distribution of effect sizes without weighting (i.e., a simple average); (2) average the distribution using the inverse of the variance of each study as the weighting factor (i.e., fixed effect model; $W_{\mathrm{i}} = 1/SE_i^2$); or (3) average the distribution using the inverse variance of each study plus average between-study variance (i.e., random effects model, $W_i = 1/SE_i^2 + \tau^2$). The average effect size ES^+, then, is $ES^+_{Weighted} = \sum(W_i)(ES_i)/\sum W_i$. Due to the presence of average between-study variance in weighting under the random effects model (τ^2), the distribution of weighting across studies tends to be smoother than weighting under the fixed effect model (Borenstein et al. 2010).

The first approach gives equal weight to each study, regardless of sample size. If the sample size of each study is approximately equal, this is not an unreasonable approach to take, but this is rarely the case in the social sciences. Seven of the meta-analyses reviewed here combined studies using an unweighted means approach to analysis (all except one before 2007). This is generally not recommended in the literature of meta-analysis (e.g., Hedges and Olkin 1985; Hunter and Schmidt 2004) because synthesizing in this way gives equal weight to each study in the distribution, whether it has a very large or very small sample size.

Four of the 13 meta-analyses used the fixed effect model. According to Borenstein et al. (2010), using this model for synthesis in studies of this type is problematic for two reasons, one conceptual and the other statistical. First, the fixed effect model assumes that a single weighted average effect size can describe a distribution of effect sizes in the population. This condition is only true when the studies being synthesized are very much alike in sample and treatment definition, procedures, outcome metric, etc. Clearly, none of the studies in these meta-analyses meet this condition.

Second, the fixed effect model tends to produce skewed (or biased) results when a very large-sample study lies on either the positive or negative margin of the distribution. This is because the very large sample size is given more weight than smaller studies and its outlying effect size tends to be leveraged or exaggerated. This effect can be seen in this second-order meta-analysis. The Schmid et al. (2014) has by far the largest number of studies and receives the most weight ($k = 479$). Its average effect size is below the average weighted fixed effect size ($g^+ = 0.250$ compared to $g^{++} = 0.362$). Hence, the average weighted fixed effect size is depressed compared to the average unweighted effect size and the average random effect size ($g^{++} = 0.362$ compared to $g^{++} = 0.393$). When the Schmid et al. is removed, so that $k = 12$, the recalculated differential becomes less dramatic ($g^{++} = 0.427$ for the fixed vs. $g^{++} = 0.417$ for the random).

It is difficult to judge the biasing effect in those studies that used an unweighted or a fixed effect approach to synthesizing effect sizes. It is safe to say, however, that any bias in individual meta-analyses would have a negligible affect upon the results of this second-order meta-analysis.

Uneven effects of bias

In the foregoing sections, we have examined only four types of bias in detail. However, as previously stated, bias can reside in any of Cooper's stages in conducting a meta-analysis. But the effect of bias at different stages is not necessarily the same. Here are several examples. Publication bias can affect a review in at least two ways. It can affect the overall outcome of the study if mostly positively signed studies are included, thus potentially leading to an overestimation of the treatment effect in the population. Moreover, regardless of the magnitude of the resulting effect size, the presence of publication bias impedes generalizability of the meta-analysis findings—what is not based on a representative (or exhaustive) dataset cannot be confidently projected beyond the study limits. For interpretation of the results of a meta-analysis, the latter is even more treacherous, as there are statistical means for detecting publication bias and adjusting the magnitude of the average effect size accordingly (i.e., previously mentioned Trim and Fill procedure), while little help is available to consumers of a meta-analysis to judge how representative and generalizable its findings are.

Much more damage could result from bias associated with improper treatment of outcomes included in a meta-analysis. When a meta-analyst decides to average incompatible outcomes (e.g., achievement with attitudes, standardized reading scores with self-reports of engagement in collaborative activities), it is no longer a question of magnitude or representativeness. This kind of misstep renders the results of a meta-analysis completely uninterpretable, providing readers with little insight into the research question.

The bottom line

Bias in this second-order meta-analysis

There are two issues considered here. The first has to do with the average effect size over the meta-analyses that we selected to include in the second-order meta-analysis. This is a standard question that meta-analysts ask and the analysis is the standard approach to analysis, albeit with some modifications because these are meta-analyses and not primary studies. The second issue deals with potential bias that might qualify or even nullify the results of each of the meta-analyses. Based on our fairly rigorous assessment methodology, the results shown in Table 4 indicate that two meta-analyses have a high potential (five **X**s each) to misrepresent the results and three other studies have a moderate potential (three **X**s each) for inaccuracy. As we have previously stated, these studies are not necessarily flawed, but they contain potential flaws, either through the omission of crucial information at various stages in the review or through actual questionable or even unacceptable practices.

In this final section, we would like to attempt to join these two issues, since it is generally recognized that the quality of studies included in a meta-analysis—or a second-order meta-analysis in this case—can impinge upon the conclusions that can be derived from its outcomes (Cooper 2010).

Table 5 Summary statistics with studies of lower methodological quality removed

Studies removed	Effect size and 95 % confidence interval				Test of null (2-tail)	
	K	g^{++}	*SE*	95 % CI	*z*-value	*p* value
Two studies removed[a]						
Random model results	11	0.391	0.07	0.25/0.53	5.32	.00
Heterogeneity	$Q_{\text{Total}} = 24.97$	$df = 10$	$p < .01$		$I^2 = 59.94$	$\tau^2 = 0.03$
Five studies removed[b]						
Random model results	8	0.283	0.05	0.20/0.37	6.26	.00
Heterogeneity	$Q_{\text{Total}} = 2.09$	$df = 7$	$p < .96$		$I^2 = 00.00$	$\tau^2 = 0.00$

[a] Christman et al. and Zhao

[b] Plus Tekbiyik et al., Koufogiannakis et al. and Larwin et al.

To answer this question, we reanalyzed the collection by removing studies that contained more than one methodological flaw (**X**). We discounted the first category (i.e., formulating the problem), because we decided that this stage usually does not have a direct effect on the statistical analysis. In the first round, we removed two studies that had four **X**s each and recalculated the random effects model. The results are shown in the top half of Table 5. Notice that the average effect size does not change greatly ($g^{++} = 0.393$–0.391) and the distribution is still heterogeneous. In the second round, we also removed meta-analyses that received two **X**s. When five studies, with either two or four **X**s were removed, the results did diminish ($g^{++} = 0.393$–0.283), and the distribution became homogeneous with virtually no between-study variability (the median of the entire distribution is 0.28). This is partially the result of the reduced number of studies (*Q*-Total tends to go down with a fewer number of studies—Type II error) but it is also because the two highest effect sizes ($g^{+} = +1.12$) and the two lowest effect sizes ($g^{+} = -0.09$ and $+0.13$) were removed. One could speculate from these results that bias can both raise and lower effect sizes accruing from individual studies.

In spite of these changes, both of the recalculated average effect sizes remain within the confidence interval of the overall analysis, so the difference is actually negligible. This suggests that methodological quality had little effect on the original statistical outcomes of the second-order meta-analysis. In addition, when effect size was regressed on the interval-level MQIM (using the mixed effects method of moments approach) a slightly negative but not significant slope resulted ($\beta = -0.004$, $p = .62$).

There is a suggestion in these data that the estimated quality of these meta-analyses improved over the years 2000–2014. The two-tailed correlation coefficient between the total methodological quality score for each meta-analysis and the year of publication was significant, but not overly large ($r = 0.561$, $df = 11$, $p = .046$). This was a positive finding as it suggests that meta-analysts are likely paying attention to the growing methodological literature on meta-analysis, and concurrently, journal editors and reviewers are becoming more knowledgeable, asking more from authors.

One issue that implicates both primary researchers and meta-analysts is the nature of the question asked in the educational technology literature and related areas such as distance and online learning. In this second-order meta-analysis, we included only studies that compared a technology treatment group to a no-technology control condition. This was done for the sake of consistency. However, this form of question is rapidly becoming "old school" for several reasons. First, starkly worded either/or questions do not respond to the rising trend in schools and universities of offering some form of technology in every classroom. Technology-free classrooms are becoming the exception rather than the rule. Second, there is a sense that this form of question does not move the field forward so that we better understand how different technology features, for example, contribute to successful learning and achievement. Cook (2009) compares this form of question to the research developments in the automotive industry. Researchers quickly realized that comparisons to horses would never produce better "horseless carriages."

Potentially, meta-analysts may find that answering subtle questions is more challenging than answering either/or questions. Meta-analysis works best when there is a clear distinction between the treatment and the control—when the coding is of the *low-inference* variety described earlier. However, it is often necessary to engage in *high inference* coding when exploring the more nuanced differences in one treatment versus an alternative treatment. For instance, in the area of distance education, Bernard et al. (2009) investigated how three different forms of interaction treatments affect achievement. Distance education courses in both the treatment and control conditions were rated for the presence of the stronger and weaker forms of interaction treatments. The stronger form was considered the treatment and the other condition the control. Effect sizes, then, represented the difference between the means of the stronger and weaker interaction treatments.

Likewise, in the other half of the Schmid et al. (2014) meta-analysis, that was not used in this paper ($k = 400$), technology resided in both classroom conditions and so three major criteria were used to judge the condition with the most technology: (1) higher frequency and/or intensity of use; (2) more advanced forms of technology that contained more features; and 3) larger number of technology tools available to students. The condition that contained more technology served as the treatment and the other condition was the control. In both of these cases, the rating scheme served to produce a common relationship among all studies, a relationship that did not exist uniformly before the coding began. The results were similar to the collection of technology/no technology studies reported here ($g^{+} = 0.31$, $k = 400$, $p < .01$).

There are examples of meta-analyses that synthesize studies that contain technology in both conditions by capitalizing on a common "present/absent" relationship. For instance, in a meta-analysis by Karich et al. (2014) the issue of learner and program control (e.g., pacing, sequencing) in educational technology applications (i.e., mostly computer-based instruction) was examined across all contents and grade levels. The overall result for 25 effect sizes was negligible, with a median of 0.05, $p > .05$. None of the learner or program control study features was significant. Typically studies of this type are relatively small-scale (compared to Schmid et al., for instance) because of the narrow definition of the independent variable. On the other hand, they provide the kind of detailed design advice that is

required when educators are working on a particular instructional problem. In this example, the Karich et al. study might well represent a form of *low-inference* coding, while the more complex treatment definition in Schmid et al. is decidedly *high-inference* coding, according to Cooper's description. As a result, the former might contain less potential for coding bias while the latter might contain more. This form of bias, however, would probably not be systematic and therefore would be irregular across an entire study.

Ultimately, however, it is up to consumers, designers and educators, interested in the application of technology to instruction, to decide what research questions could improve their practice. But it is the responsibility of primary researchers to provide the grist for the meta-analysts mill in order to inform all kinds of research questions.

Acknowledgments The development of this article was supported in part by a grant to Bernard and Schmid from the Social Sciences and Humanities Research Council of Canada.

Appendix: Evaluation criteria within Cooper's (2010) categories

1. Formulating the Problem
   - 1.1 Research Question
     Are the research objectives and/or questions clearly stated?
   - 1.2 Context of the M-A
     Are the purposes of the M-A described within the context of prior work and current practice?
   - 1.3 Time Frame
     Is the time frame defined and adequately justified in the context of the research question and prior M-As?
   - 1.4 Contextual Positioning of the Research Problem
     Is the rationale for the M-A adequate, conceptually relevant and supported by empirical evidence?
   - 1.5 Experimental and Control Groups
     Are the experimental and control group clearly defined and described in detail?
   - 1.6 Outcome Measures
     Are outcome measures relevant to the research question and representative of the outcomes found in real classrooms?

2. Searching the Literature
   - 2.1 Inclusion Criteria
     Are the inclusion criteria clearly and operationally stated and described in detail?
   - 2.2 Resources Used
     Are the resources used to identify relevant literature representative of the field and exhaustive?

2.3 Literature Included
Is the included literature exhaustive and includes all types of published and unpublished literature?
2.4 Search Strategy
Is the list of search terms provided and appropriate for each individual source (e.g. modifying key words for specific databases)?

3. Extracting Effect Sizes and Coding Study Features

3.1 Effect Size Extraction
Is effect size extraction implemented by at least two raters with a reasonable level of inter-rater reliability?
3.2 Study Feature Coding
Is study feature coding implemented by at least two raters with reasonable inter-rater reliability?

4. Methodological Quality of the Data

4.1 Validity of Included Studies
Are all aspects of validity explicitly and operationally defined and consistently applied across studies?
4.2 Publication Bias
Are procedures for addressing publication bias adequately substantiated and reported in detail?
4.3 Independence of Data
Is the issue of dependency addressed in detail with methods for assuring data independence being appropriate and adequately described?
4.4 Effect Size Metrics and Extraction Procedures
Are the used ES metrics and extraction procedures appropriate and fully described including necessary transformations?
4.4 Treatment of Outliers
Are criteria and procedures for identifying and treating outliers adequately substantiated and reported in detail?

5. Synthesizing effect sizes

5.1 Overall Analyses
Is the overall analysis performed according to standard procedures (e.g., correct model use, homogeneity assessed, standard errors reported, confidence intervals reported)?
5.2 Moderator Variable Analyses
Are moderator variable analyses performed according to the proper analytical model and is appropriate information reported (e.g., Q_{Between}, test statistics provided)?

5.3 Post hoc Analysis
If appropriate to the analysis are post hoc test conducted using appropriate measures for controlling Type I error?

6. Interpreting Evidence

6.1 Reporting Statistical Results
Are the appropriate statistics supplied for all analyses and explained in enough detail that the reader will understand the findings?

6.2 Appropriate Interpretation
Are the results interpreted appropriately and correctly?

7. Presenting the Results

7.1 Discussing Results
Does the discussion relate the results to previous research?

7.2 Emphasis
Does the interpretation place emphasis on the main findings?

7.3 Limitations to the Results
Does the discussion expose and explain limitations to the M-A?

7.4 Application to Practice
Does the discussion provide advice to other researchers, practitioners, policy makers, etc.?

References

Asterisks (*) are meta-analyses in the second-order meta-analysis. Double asterisks (**) are rejects

Abrami, P. C., Borokhovski, E., Bernard, R. M., Wade, C. A., Tamim, R., Persson, T., et al. (2010). Issues in conducting and disseminating brief reviews. *Evidence and Policy: A Journal of Research, Debate and Practice, 6*(3), 371–389. doi:10.1332/174426410X524866.

*Bayraktar, S. (2000). A meta-analysis study of the effectiveness of computer assisted instruction in science education (Unpublished doctoral dissertation). Ohio State University, Columbus, OH (UMI Number: 9980398).

Bernard, R. M. (2014). Things I have learned about meta-analysis since 1990: Reducing bias in search of "The Big Picture." *Canadian Journal of Learning and Instruction, 40*(3). http://www.cjlt.ca/index.php/cjlt/article/view/870

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., & Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: From the general to the applied. *Journal of Computing in Higher Education., 26*(1), 87–122. doi:10.1007/s12528-013-9077-3.

Bernard, R. M., Abrami, P. C., Borokhovski, E., Wade, A., Tamim, R., Surkes, M., et al. (2009). A meta-analysis of three interaction treatments in distance education. *Review of Educational Research, 79*(3), 1243–1289. doi:10.3102/0034654309333844.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2005). *Comprehensive meta-analysis version 2*. Englewood, NJ: Biostat.

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed effect and random effects models for meta-analysis. *Research Synthesis Methodology, 1*, 97–111. doi:10.1002/jrsm.12.

*Christmann, E. P., & Badgett, J. L. (2000). The comparative effectiveness of CAI on collegiate performance. *Journal of Computing in Higher Education, 11*(2), 91–103.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, D. A. (2009). The failure of e-learning research to inform educational practice, and what we can do about it. *Medical Teacher, 31*(2), 158–162. doi:10.1080/01421590802691393.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Cooper, H. M., & Koenka, A. C. (2012). The overview of review: Unique challenges and opportunities when research syntheses are the principal elements of new integrative scholarship. *American Psychologist, 67*(6), 446–462. doi:10.1037/a0027119.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463. doi:10.1111/j.0006-341X.2000.00455.x.

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

Hammerstrøm, K., Wade, A, & Jørgensen A. M. K. (2010) *Searching for studies: A guide to information retrieval for Campbell Systematic Reviews,* Supplement 1. Oslo, Norway: The Campbell Collaboration. doi:10.4073/csrs.2010.1. http://www.campbellcollaboration.org/resources/research/new_information_retrieval_guide.php)

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Hedges, L. V., & Olkin, I. (1985). *Statistical aspects of meta-analysis*. New York, NY: Academic Press.

Higgins, J. P. T., Land, P. W., Anagnostelis, J. A.-C., Baker, N. F., Cappelleri, S. H., Hollis, S., et al. (2012). A tool to assess the quality of a meta-analysis. *Research Synthesis Methods*, *4*, 351–366. doi:10.1002/jrsm.1092.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-Analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

*Hsu, Y.-c., (2003). The effectiveness of computer-assisted instruction in statistics education: A meta-analysis. (Unpublished doctoral dissertation). The University of Arizona, Tucson, AZ (UMI Number: 3089963).

Jackson, G. B. (1980). Methods for integrative reviews. *Review of Educational Research, 50*, 438–460. doi:10.3102/00346543050003438.

Karich, A. C., Burns, M. K., & Maki, K. E. (2014). Updated meta-analysis of learner control within educational technology. *Review of Educational Research.* OnlineFirst, March 10, 2014. doi:10.3102/0034654314526064

*Koufogiannakis, D., & Wiebe, N. (2006). Effective methods for teaching information literacy skills to undergraduate students: A systematic review and meta-analysis. *Evidence-Based Library and Information Practice, 1*(3), 3–43.

*Larwin, K., & Larwin, D. (2011). A meta-analysis examining the impact of computer-assisted instruction on postsecondary statistics education: 40 years of research, *Journal of Research on Technology in Education, 43*(3), 253–278. http://www.editlib.org/p/54098/

Lefebvre, C., Manheimer, E., & Glanville J. (2011). Chapter 6: Searching for studies. In J. P. T. Higgins & Green, S. (Eds.), *Cochrane handbook for systematic reviews of interventions version 5.1.0* (updated March 2011). http://www.cochrane-handbook.org

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

**Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. R. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education, 70*, 29–40.

*Michko, G. M. (2007). A meta-analysis of the effects of teaching and learning with technology in undergraduate engineering education (Unpublished doctoral dissertation). University of Huston, Huston, TX (UMI Number: 3289807).

**Rolfe, V., & Gray, D. (2011). Are multimedia resources effective in life science education? A meta-analysis. *Bioscience Education, 18*(December). www.bioscience.heacademy.ac.uk/journal/vol18/beej-18-3.pdf

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis—Prevention, assessment and adjustments*. Chichester: Wiley.

Scammacca, N., Roberts, G., & Stuebing, K. K. (2013). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research, 84*(3), 328–364. doi:10.3102/0034654313500826.

*Schenker, J. D. (2007). The effectiveness of technology use in statistics instruction in higher education: A meta-analysis using hierarchical linear modeling (Unpublished doctoral dissertation). Kent State University, Kent, OH.

Schlosser, R. W., Wendt, O., Angermeier, K., & Shetty, M. (2005). Searching for and finding evidence in augmentative and alternative communication: Navigating a scattered literature. *Augmentative and Alternative Communication, 21*(4), 233–255. doi:10.1080/07434610500194813.

Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R. M., Abrami, P. C., Surkes, M. A., et al. (2014). The effects of technology use in postsecondary education: A meta-analysis of classroom applications. *Computers & Education, 72*, 271–291. doi:10.1016/j.compedu.2013.11.002.

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., et al. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7*(10). doi:10.1186/1471-2288-7-10.

*Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology, 64*, 489–528. doi:10.1111/j.1744-6570.2011.01190.x.

*Sosa, G. W., Berger, D. E., Shaw, A. T., & Mary, J. C. (2011). Effectiveness of computer-assisted instruction in statistics: A meta-analysis. *Review of Educational Research, 81*(1), 97–128. doi:10.3102/0034654310378174.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research, 81*(3), 4–28. doi:10.3102/0034654310393361.

*Tekbiyik, A., & Akdeniz, A. R. (2010). A meta-analytical investigation of the influence of computer assisted instruction on achievement in science, *Asia-Pacific Forum on Science Learning and Teaching, 11*(2). http://www.ied.edu.hk/apfslt/v11_issue2/tekbiyik/index.htm

*Timmerman, C. E., & Kruepke, A. (2006). Computer-assisted instruction, media richness and college student performance. *Communication Education, 55*(1), 73–104. doi:10.1080/03634520500489666.

Valentine, J. C., & Cooper, H. M. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods, 13*(2), 130–149. doi:10.1037/1082-989X.13.2.130.

**Vogel, J. J., Vogel, D. S., Cannon-Bower, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*(3), 229–243.

Viechbauer, W., & Cheung, M.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*, 112–125. doi:10.1002/jrsm.11.

*Zhao, Y. (2003). Recent developments in technology and language learning: A literature review and meta-analysis. *CALICO Journal, 21*(1), 7–27.

Robert M. Bernard Ph.D. is Professor of Education and the Systematic Review Team Leader for the Centre for the Study of Learning and Performance at Concordia University. His research interests include distance and online and blended learning and instructional technology. His methodological expertise is in the areas of research design, statistics and meta-analysis.

Eugene Borokhovski Ph.D. is the Systematic Review Manager for the Centre for the Study of Learning and Performance at Concordia University. His areas of expertise and interests include cognitive and educational psychology, language acquisition, and methodology and practices of systematic review in education, meta-analysis in particular.

Richard F. Schmid Ph.D. is Professor of Education at Concordia University and Chair of the Department of Education. He is also the Educational Technology Team Leader for the Centre for the Study of

Learning and Performance. His research interests include examining pedagogical strategies supported by technologies, and the cognitive/affective factors they influence.

Rana M. Tamim Ph.D. is Associate Professor at Zayed University, Dubai, United Arab Emirates. She is a collaborator with the Centre for the Study of Learning and Performance at Concordia University. Her research interests include online and blended learning, learner-centered instructional design, and science education. Her research expertise includes quantitative and qualitative research methods in addition to systematic review and meta-analysis.