J Comput High Educ
DOI 10.1007/s12528-013-9077-3

A meta-analysis of blended learning and technology use in higher education: from the general to the applied

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

Abstract This paper serves several purposes. First and foremost, it is devoted to developing a better understanding of the effectiveness of blended learning (BL) in higher education. This is achieved through a meta-analysis of a sub-collection of comparative studies of BL and classroom instruction (CI) from a larger systematic review of technology integration (Schmid et al. in Comput Educ 72:271–291, 2014). In addition, the methodology of meta-analysis is described and illustrated by examples from the current study. The paper begins with a summary of the experimental research on distance education (DE) and online learning (OL), encapsulated in meta-analyses that have been conducted since 1990. Then it introduces the Bernard et al. (Rev Educ Res 74(3):379–439, 2009) meta-analysis, which attempted to alter the DE research culture of always comparing DE/OL with CI by examining three forms of *interaction treatments* (i.e., student–student, student–teacher, student–content) within DE, using the theoretical framework of Moore (Am J Distance Educ 3(2):1–6, 1989) and Anderson (Rev Res Open Distance Learn 4(2):9–14, 2003). The rest of the paper revolves around the general steps and procedures (Cooper in Research synthesis and meta-analysis: a step-by-step approach, 4th edn, SAGE, Los Angeles, CA, 2010) involved in conducting a meta-analysis. This section is included to provide researchers with an overview of precisely how meta-analyses can be used to respond to more nuanced questions that speak to underlying theory and inform practice—in other words, not just answers to the "big questions." In this instance, we know that technology has an overall positive impact on learning ($g^{+} = +0.35$, $p < .01$, Tamim et al. in Rev Educ Res 81(3):4–28, 2011), but the sub-questions addressed here concern BL interacting with technology in higher

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

R. M. Tamim
Zayed University, Dubai, United Arab Emirates

Published online: 09 February 2014

education. The results indicate that, in terms of achievement outcomes, BL conditions exceed CI conditions by about one-third of a standard deviation ($g^+ = 0.334$, $k = 117$, $p < .001$) and that the kind of computer support used (i.e., cognitive support vs. content/presentational support) and the presence of one or more interaction treatments (e.g., student–student/–teacher/–content interaction) serve to enhance student achievement. We examine the empirical studies that yielded these outcomes, work through the methodology that enables evidence-based decision-making, and explore how this line of research can improve pedagogy and student achievement.

Keywords Bended learning · Technology use · Higher education · Meta-analysis

Introduction

Quantitative research on distance education and online learning

Few would argue these days that quantitative studies of nearly any aspect of educational practice represent the final word, but there has been a tendency to relegate to the periphery these kinds of evidence when questions of the effectiveness of distance education (DE) and online learning (OL) arise. DE, and the more recently OL, refers to instructional conditions where learners are for the most part physically separated from their teachers and where at least two-way communication connects them (Keegan 1996). DE/OL may be conducted either synchronously or asynchronously, although Keegan considers the former to be a special case of classroom instruction (CI). Literally thousands of comparative primary studies, where DE/OL conditions conform to these instructional specifications, have pitted DE/OL against CI and since 2000, sixteen major meta-analyses have been mounted to assess the differences between CI and DE/OL.[1] These are among the important things that we have learned from all of this primary research and synthesis activity:

1. there is general consensus of the effectiveness of all forms of DE (including OL) compared with CI (i.e., the average effect sizes range from $d^+ \approx 0.00$ for conventional DE and correspondence education to $d^+ \approx 0.05$–0.15 for OL)—in other words there is little difference in these two instructional patterns;
2. there is wide variability among studies, from those strongly favoring DE to those favoring CI, thereby bringing into question the value of point one;
3. there is a tendency for researchers to describe the DE/OL condition in great detail while characterizing the CI condition as "traditional classroom instruction," thereby diminishing the opportunity to describe and compare salient study features;

[1] Means et al. (2013), Cook et al. (2008), Jahng et al. (2007), Lou et al. (2006), Allen et al. (2006), Sitzmann et al. (2006), Williams (2006), Zhao et al. (2005), Allen et al. (2004), Bernard et al. (2004), Cavanaugh et al. (2004), Shachar and Neumann (2003), Ungerleider and Burns (2003), Cavanaugh (2001), Allen et al. (2002), Machtmes and Asher (2000).

4. comparative primary research is plagued with a variety of methodological problems and confounds that make them very hard to interpret (Is it the distance, the media, the instructional strategies, etc., or combinations of these?); and
5. only a few substantive moderator variables have yielded any information about what makes DE/OL and CI different.

While these findings are useful, and methodologically heuristic, we still know little about how to design and implement DE/OL. Beyond experimental research, the literature is limited to articles on praxis, "one-off" quantitative and qualitative studies (often survey or institutionally-bounded research) and theory-based applications of instructional design. Furthermore, it has been argued (Bernard et al. 2004, 2009; Cook 2009) that there is little left to learn about DE/OL from studies that compare them with CI. These studies answer the "big" questions (e.g., Is DE/OL more effective than CI?), but they generally fail to establish an alignment of evidence that addresses the "do's" and "don'ts" of instruction via rigorous research.

More useful information can be extracted from studies that include DE/OL in both conditions. A 2009 meta-analysis by Bernard et al. examined comparisons of greater (i.e., treatment condition) versus lesser-strength (i.e., control condition) interaction treatments (IT). ITs were defined as media and instructional conditions, implemented within DE contexts, intended to increase student–student, student–teacher and/or student–content interaction. A selection of the results of this meta-analysis is shown in Table 1. All greater-strength IT effects were positive and significant. In addition, the both student–student and student–content ITs outperformed student–teacher treatments ($z = 2.69$, $p = .004$ and $z = 3.09$, $p = .001$, respectively). Moreover, post hoc analysis revealed that combinations of student–content ITs and student–student ITs outperformed student–teacher ITs on achievement outcomes ($z = 2.62$, $p = .004$). The implication is that when students are given the means to communicate and interact with one another or online with content, an increase in achievement may result. These findings may have particular relevance for the design of BL.

Beyond DE and OL, there is a large and growing literature of studies investigating blended learning, which involves a combination of elements of

Table 1 Average effect sizes for categories of interaction

Categories of interaction treatments	Effect sizes and standard errors			Confidence interval	
	k	g^+	SE	Lower 95th	Upper 95th
Student–student	10	0.49*	0.08	0.33	0.65
Student–teacher	44	0.32*	0.04	0.24	0.40
Student–content	20	0.46*	0.05	0.36	0.56
Total	74	0.38*	0.03	0.33	0.44
Between-classes: $Q = 7.05$, $p = .03$					

* $p < .01$

face-to-face CI and OL outside of class time, and is increasingly becoming a substitute for CI. It is sometimes argued that BL is the "best of both worlds" because it is a marriage of the best elements of the two practices (Bele and Rugelj 2007). But experience amply defies this broad generalization because, as always, the devil is in the detail. The study described here is an attempt to fill in some of those details.

Blended learning

In an early review of BL research, Bliuc et al. (2007) pegged the first use of the term BL to the year 2000. By 2006, there was a handbook devoted almost solely to its educational implementation and issues related to BL (i.e., *The Handbook of Blended Learning: Global Perspectives, Local Designs*; Bonk and Graham 2006). In 2008, an influential book appeared, focusing on effective use of BL in higher education from the perspective of "communities of inquiry" (i.e., *Blended Learning in Higher Education: Framework, Principles, and Guidelines*, Garrison and Vaughan 2008). The growing popularity of BL has been documented in several surveys of instructors (e.g., Arabasz and Baker 2003) and students (Albrecht 2006) in higher education institutions, and in 2004, Marquis found that 90 % of university instructors believed BL to be "more effective than CI."

This enthusiasm for BL has not been matched by a large literature of primary research studies. In an unpublished vote count literature review of blended/hybrid learning studies, Zhoa and Breslow (2013) found only 25 studies of blended/hybrid learning dating from 1999 that met their criteria. Eleven (11) showed a "significant difference between the treatment and the control" in favor of the BL condition, five (5) found "mixed results" and nine (9) found "no significant difference." Most of the studies were conducted in higher education contexts. To date, there has been only one meta-analysis devoted to BL (Means et al. 2013). This article was based on a US Department of Education meta-analysis, originally published in 2009 and updated in 2010 (Means et al. 2010), which also included purely online learning in a separate section. BL conditions were found to significantly outperform fully face-to-face CI where no blending occurred ($g^{+} = 0.35, k = 23, p = .001$). The majority of the studies here were also from research in higher education and the authors acknowledged the difficulty with controlling for time spent outside of class and the somewhat unequal distribution of materials used.

Similar to the Bernard et al. (2004) meta-analysis of the DE literature, Means et al. (2013) found that pedagogical setups for BL do make a difference. Namely, students in both collaborative interactive learning and teacher-directed expository instructional conditions significantly outperformed those engaged in active self-study. No other moderator variables revealed significant differential effects on learning outcomes, though some findings were suggestive. For example, in computer-mediated communications with instructor and among students, the asynchronous mode only was more effective than it was in combination with the synchronous mode. Also, learners in undergraduate courses seemed to benefit more from BL than graduate students. Interestingly, a balance of course time in favor of OL instruction (as compared to time spent face-to-face) produced a relatively higher

weighted average effect size, as well as the opportunity to interact face-to-face with the instructor during the class time rather than before or after instruction. The BL treatments of longer duration were more effective compared to shorter ones. As promising as they appear to be, these results were not statistically significant.

The issue of BL is a complicated one; there has been considerable discussion even of the meaning of the term itself. Driscoll and Carliner (2005) describe four patterns, each of which they call a form of BL: (1) a mix of Web-based technologies; (2) a mix of various pedagogical approaches (e.g., constructivism, behaviorism, cognitivism); (3) a combination of any form of instructional technology with face-to-face instructor-led conditions; or (4) a combination of instructional technology with actual job tasks to form an effective mix of learning and working. Our operational definition is closer to number three, above, and the one espoused by Graham (2005) where BL is defined as "the combination of instruction from two historically separate models of teaching and learning: traditional face-to-face learning systems and distributed learning systems" (p. 5), emphasizing distributed learning as the use of computer-based technologies outside of class time. We go further by focusing on the proportion of time associated with the classroom/online mixture. BL is thus defined as instructional conditions in which at least 50 % of total course time is face-to-face CI and students working online outside of the classroom spend the remainder of time, up to the additional 50 %, online. In some cases this definition produces an equal blend of CI and OL (i.e., 50–50 %). In most other cases, BL could accrue from as little as 25 % online work and 75 % face-to-face work. We argue that this a conservative test of BL that explores the lower limits of the addition of OL components. The classroom use of computers or other educational technologies in the treatment and/or control groups does not count as BL in this study.

The meta-analysis presented here is part of a recently completed project that examines all forms and configurations of technology use in higher education from 1990 through 2010 (Schmid et al. 2014). The results of 674 studies, yielding 879 effect sizes of achievement outcomes, revealed a moderate average effect size, $g^{+} = 0.334$, $p < .001$), that was significantly heterogeneous.

A portion of these studies was designated as BL because of their mix of CI (i.e., face-to-face) and out-of-class OL where the online work substituted for class time. It is those studies that are addressed in this meta-analysis. So in essence, the corpus of studies examined here are defined by their "pedagogical pattern" rather than the technology used—they by definition use technology, but in a way that supports this increasingly popular mixture.

The goals of this paper are twofold. One is to describe the characteristics of meta-analysis as an analytical tool, with some commentary on the various aspects of this methodology. Basic information about major steps and procedures of meta-analytical research appear in sections labeled "General." They are then illustrated by the decisions and findings of the study itself in sections labeled "Application." To accomplish these two goals, the descriptions in the method section, in particular, are truncated from their original form in Schmid et al. (2014).

Research questions in the literature of technology integration

Since 1990, numerous meta-analyses have been published each intending to capture the difference between technology-enhanced classrooms and "traditional classrooms" that contain no technology. Tamim et al. (2011) summarized 25 of the most important of these in a second-order meta-analyses. Many are specific to particular grade levels and subject matters, and most deal with specific forms of technology (e.g., computer-assisted instruction), yet all ask the starkly worded question: *Is some technology better than no technology?* The exceptions to this (i.e., Rosen and Salomon 2007; Lou et al. 2001; Azevedo and Bernard 1995) have attempted to discern among different instructional strategies within technology use (i.e., constructivist use of technology in K-12 mathematics, the use of technology with small groups and feedback in computer-based learning, respectively). It is arguable that since about 1990 few classrooms contain no technology at all, so it makes more sense to cast the question in the form of: *What is the difference between this technology application (e.g., type, amount) and another?* Schmid et al. (2014) attempted to capture the answers that might accrue from both questions, and by extension so will this study. We also classified the use of technology for teaching and learning purposes to see if there is a difference between, for instance, technologies used for cognitive support (e.g., simulations, serious games) and technologies used to present information and subject matter content (i.e., content/presentational support).

The remainder of this article is organized around Cooper's (2010) seven steps for conducting a systematic review/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. These stages in conducting a systematic review are neither mutually exclusive nor entirely distinct; rather, they should be viewed as key steps in a continuous and iterative process. Since every systematic review is somewhat different, the subheadings under each stage reflect the actual nature of this project.

Step 1: formulating the problem (research questions, definitions, inclusion/exclusion criteria)

Research questions—general

The questions posed in a systematic review help to focus attention on the goals of the research endeavor, the important variables that will be addressed and their relationship to one another. In the case of a meta-analysis of standardized differences between a treatment condition or intervention and a control condition, the questions are often expressed as "the impact of" or "the effect of" the difference in treatments on an outcome or dependent measure. Sometimes this step is quite straightforward and sometimes it can be quite complex. It is appropriate here to search for and examine previous reviews of all kinds. There are three important reasons for not excluding this step. It helps to determine how the problem

has been dealt with in the past and how recently. It helps to gather others' views on the nature of the problem and the terms that have been used and how they were operationalized. It helps to determine if there are questions left unanswered or if there are other unexplored ways of looking at the body of research. Complexity is introduced when there are synonymous terms and/or fuzzy descriptions of the treatments and dependent measures. For instance, Abrami et al. (2008) expended considerable effort examining the relationships between the terms "critical thinking," "creative thinking," "problem-solving," "higher-order thinking" and the like. The researchers also searched the literature for the range of standardized measures that were available and their psychometric properties. At one point, a limited factor analytical study was conducted (Bernard et al. 2008) to determine if the subscales of the *Watson–Glaser Critical Thinking Appraisal* (See Watson and Glaser 1980) were empirically distinguishable from one another, or whether the test should be considered as a global measure. These are the kinds of issues that often must be dealt with and resolved before the main work of meta-analysis can begin.

Research questions—application

BL is a special case of CI and OL, because it contains elements of both. There are many previous meta-analyses in the technology integration literature, over and above the ones from the DE/OL literature previously described. Tamim et al. (2011) summarized 25 of them in a second-order meta-analysis and validation study of this literature. The researchers determined that virtually all of the previous meta-analyses, and their underlying primary studies, had addressed the *no technology in the control condition* question. Schmid et al. (2014) went beyond this by adding studies where there was *some technology in both the treatment and the control condition*, paying special attention to the purpose of technology use. So in that meta-analysis, the total collection of 879 effect sizes were divided by the form of the research question—*no technology* ($k = 479$) or *some technology* ($k = 400$) in the control condition. Originally, our intention was to make the same distinction here, but only $k = 13$ comparisons out of a total of $k = 117$ (11 %) contained some technology in the control condition, so this is not a primary focus in our research questions. It is important to note that in none of these studies did control participants use technology for BL, so that all of the comparisons reflect the distinction between BL and CI.

The following questions formed the basis for the current meta-analysis:

- What is the impact of blended learning (i.e., courses that contain components of both face-to-face and OL) on the achievement of higher education students in formal educational settings?
- How do course demographic study features (e.g., course subject matter) moderate the overall average effect size?
- How do various pedagogical factors, like the amount of time spent online outside of class and the purpose of technology use in the treatment condition, moderate this effect?

- How do various interaction treatments (i.e., defined as in Bernard et al. 2009) modify the overall treatment effect?
- Finally, is there a difference between studies that have *no technology in the control condition* and those that contain *some technology in the control condition*?

Definitions—general

This involves establishing working or operational definitions of terms and concepts related to the purposes of the meta-analysis. This is done to help further clarify the research questions and to inform the process of devising information search strategies. Definitions also convey what the researchers mean by particular terms, especially when the terms have multiple definitions in the literature. This was the case in the critical thinking project just alluded to in the previous section. This step is important because a well-defined and clearly articulated review question will have an impact on subsequent steps in the process, especially the step of searching the literature and making inclusion/exclusion decisions.

Definitions—application

The key terms that frame the research questions above are defined as follows:

- *Educational technology use* is any use of technology for teaching and learning as opposed to technology that may serve administrative and/or managerial purposes. This following quotation from Ross et al. (2010) explains the term educational technology as it is used here: "a broad variety of modalities, tools, and strategies for learning, [whose] effectiveness… depends on how well [they] help teachers and students achieve the desired instructional goals" (p. 19).
- *Learning achievement,* in this study, is the primary educational goal and is operationalized to include any measure of academic performance.
- *Pedagogical factors* refer to elements of instructional design that can be manipulated by a teacher/instructor in an attempt to provide the best conditions/support for learning. These might or might not include adaptations of technology use.
- *Blended Learning* is the combination of face-to-face and online learning outside of class, where the latter does not exceed 50 % of the course time. Face-to-face classroom time therefore can be greater than 50 %.
- *Formal educational settings* include instructional interventions of any duration for CI in accredited institutions of higher education.

Inclusion/exclusion criteria—general

Inclusion/exclusion criteria are primarily a set of rules that are used both by information specialists to tailor the literature searches to the literature implicated in the research question, and by reviewers to choose which studies to retain (or

exclude) in/from the meta-analysis. They also determine the inclusivity or exclusivity of the meta-analysis as a whole. For instance, if the researchers have decided to include only studies of the highest methodological quality (e.g., randomized control trials only), the inclusion criteria will specify this. Likewise, if the review is to have a particular beginning date (e.g., 1990) or include only particular contents or populations, the inclusion/exclusion criteria will indicate this.

Inclusion/exclusion criteria—application

Review for selecting studies for the meta-analysis was conducted in two stages. First, studies identified through literature searches were screened at the abstract level. Then, the review of full-text documents identified at the first stage led to decisions about whether or not to retain each individual study for further analyses. To be included, a study had to have the following characteristics:

- Be published no earlier than 1990.
- Be publicly available or archived.
- Address the impact of computer technology (including CBI, CMC, CAI, simulations, e-learning) on students' achievements or academic performance.
- Be conducted in formal higher education settings (i.e., a course or a program unit leading to a certificate, diploma, or degree).
- Represent BL in the experimental condition and CI in the control condition, excluding DE and purely OL courses. However, the control condition is allowed to have technology but not for the purposes of BL.
- Contain sufficient statistical information for effect size extraction.
- Contain at least two independent samples. This includes true experiments and quasi-experiments and excludes two-group pre-experiments and one-group pretest–posttest designs. All studies must somehow control for selection bias (Campbell and Stanley 1963).

Failure to meet any of these criteria led to exclusion of the study with the reason for rejection documented for further summary reporting. Two researchers working independently rated studies on a scale from 1 (definite exclusion) to 5 (definite inclusion), discussed all disagreements until they were resolved, and documented initial agreement rates expressed both as Cohen's Kappa (κ) and as Pearson's r between two sets of ratings.

Step 2: searching the literature

This step involves identifying sources of information, specifying search terms and developing and implementing a search strategy.

Searching the literature—general

This is arguably one of the most important aspects of conducting a systematic review/meta-analysis, as it may be compared to the data collection phase of a primary study. To meet the criterion of comprehensiveness and minimize what is

known as the "publication bias" phenomenon, it is necessary to look beyond the published literature to the "grey literature" found in conference presentations, dissertations, theses, reports of research to granting agencies, government agencies, the archives of organizations, etc. For a complete picture of the literature, a diversity of bibliographic and full-text databases must be searched, including those in related fields and geographic regions. Since different fields (and cultures) use somewhat different terminologies, strategies for each database must be individually constructed. In addition to the database searches, web searches for grey literature, manual searches through the tables of contents of the most pertinent journals and conference proceedings, and branching from previous review articles or selected manuscripts should also be conducted. In some cases researchers will contact prominent and knowledgeable individuals in the field to determine if they know of additional works that fit the inclusion/exclusion criteria. Literature searches may continue even as other stages in the review are proceeding, so that the process of information search and retrieval is best described as iterative. Naïve information search and retrieval (e.g., not using a trained professional reference librarian or information specialist) will result in a systematic review that has limited generalizability or, even worse, biased results. Guidelines from information search and retrieval for systematic reviews can be found in various publications. The Campbell Collaboration publishes such a document (Hammerstrøm et al. 2010) on its website that is especially useful for the social sciences: http://www.campbellcollaboration.org/resources/research/new_information_retrieval_guide.php

Searching the literature—application

Extensive literature searches were designed to identify and retrieve primary empirical studies relevant to the major research question. Key terms used in search strategies, with some variations (to account for specific retrieval sources), primarily included: "technolog*," "comput*" "web-based instruction," "online," "Internet," "blended learning," "hybrid course*," "simulation," "electronic," "multimedia" OR "PDAs" etc.) AND ("college*," "university," "higher education," "postsecondary," "continuing education," OR "adult learn*") AND ("learn,*" "achievement*," "attitude*," "satisfaction," "perception*," OR "motivation," etc.), but excluding "distance education" or "distance learning" in the subject field. To review the original search strategies, please visit http://doe.concordia.ca/cslp/cslp_cms/SR.

The following electronic databases were among those sources examined: ERIC (WebSpirs), ABI InformGlobal (ProQuest), Academic Search Premier (EBSCO), CBCA Education (ProQuest), Communication Abstracts (CSA), EdLib, Education Abstracts (WilsonLine), Education: A SAGE Full-text Collection, Francis (CSA), Medline (PubMed), ProQuest Dissertation and Theses, PsycINFO (EBSCO), Australian Policy Online, British Education Index, and Social Science Information Gateway.

In addition, a Google Web search was performed for grey literature, including a search for conference proceedings. Review articles and previous meta-analyses were used for branching, as well as the table of contents of major journals in the field of educational technology (e.g., *Educational Technology Research and Development*).

Step 3: gathering information from studies (select studies for inclusion, assign treatment and control conditions, extract effect sizes, identify the number of effect sizes, code study features and moderator variables)

Select studies for inclusion—general

In this step raters apply the inclusion/exclusion criteria to the studies that have been obtained through searches. The first step normally involves an examination of abstracts, so as to avoid the cost of retrieving full text articles on the first step. Since many abstracts do not contain elaborate information, raters should err on the side of inclusion at this stage. The next step is to retrieve full text documents for further examination. Again, raters apply the inclusion/exclusion criteria as they examine the entire document for relevance. Normally, two raters are used to accomplish these selection tasks and inter-rater reliability is calculated to indicate the degree of agreement between them.

Select studies for inclusion—application

This meta-analysis is a subset of a larger meta-analysis (Schmid et al. 2014). Therefore, the statistics and other quantitative information presented here is from the larger study. Overall, more than 9,000 abstracts were identified and reviewed, resulting in full-text retrieval of about 3,700 primary research studies potentially suitable for the analysis. Out of this number, through a thorough review of full-text documents, 674 studies were retained for further analysis. They yielded 879 effect sizes in the *Achievement* category. Inter-rater agreements at different stages of the review were as follows:

- Screening abstracts—86.89 % (Cohen's $\kappa = 0.74$) or $r = 0.75$, $p < .001$; and
- Full-text manuscript inclusion/exclusion—85.57 % ($\kappa = 0.72$) or $r = 0.84$, $p < .001$.

For the purposes of this meta-analysis of BL, there were 96 studies and $k = 117$ effect sizes ($N = 10{,}800$ students) selected for inclusion from the larger study. There are no statistics for inclusion at this stage because selection was based on previously coded study features.

Assign treatment and control conditions—general

Before effect sizes can be extracted the researchers must determine which condition will be designated as the treatment group and which will be designated as the control group. It is very important to get this step right as the valence of the effect size depends on it—getting some studies wrong will greatly affect the veracity of the findings. In most meta-analyses, designation of the treatment or intervention group and the control group is clear. Usually, the treatment group receives the intervention in question and the control group does not. A good example of this is the meta-analysis by Bernard et al. (2004) in which DE conditions (the treatment) were compared to CI conditions (the control). However, there are some circumstances,

especially when two treatments are being compared, when this designation is not clear. When the question being asked is "*which of these two treatments is most effective?"* it is necessary to have some framework or rational basis for establishing what characteristics of the intervention will define the "treatment" and what characteristics will define the "control." In Bernard et al. (2009), the meta-analysis described in the introduction, different interaction treatments in DE environments were compared. The intervention or treatment condition was determined to be the condition that likely evoked the most interaction between: (1) students and other students, (2) students and the teacher; and (3) students and the content to be learned. The lesser condition was deemed the control even if it also contained elements of ITs.

Assign treatment and control conditions—application

The current meta-analysis contains studies that have a clear designation of treatment and control (i.e., studies with no technology in the one condition), and studies in which this distinction is less clear (i.e., studies with technology in both conditions). We handled these two types of studies differently.

No technology in one group Since this was a meta-analysis of the effects of technology implementation in higher education, some form of educational technology was required in at least one condition. When this was the case, the group that received some form of educational technology was deemed the treatment group and its no technology companion was treated as the control condition.

Some technology in both groups When studies were found with some form of technology in both conditions, it became necessary to develop and use a set of standards that could be applied uniformly across all studies of this type. The degree of technology used was the way we determined the distinction between the experimental and control conditions. Conditions that contained more technology use were designated as the experimental conditions while the conditions with less use were considered the control. The degree of technology use in each condition was determined as follows:

- Intensity of use (frequency of technology use, and/or length of time used);
- Nature of the technology used (number of options and/or number of different functions); and/or
- Use of more advanced tools, devices, software programs, etc.

The summative use of educational technology was gauged for each group independently and rated on a 3-point scale: minimal, moderate, and high. Experimental and control group designations, based on these rules, were determined by independent coders using procedures similar to those described in other stages of the review. The condition with the greatest total among coders was deemed the experimental condition and the other the control condition. Inter-rater statistics for the attribution of dimensions and estimation of the magnitude of the difference between conditions was 75.44 % ($\kappa = 0.51$), based on sample coding of 100 studies at the beginning of the project.

Identify the number of effect sizes—general

Many studies contain multiple conditions, usually multiple variations of a given treatment. Since in these circumstances there is usually only one control condition, multiple comparisons may involve the same participants, resulting in dependency among some effect sizes. A dependency here is defined more formally as two or more effect sizes that are correlated by virtue of sharing research participants. While dependency in this case is related to multicollinearity in multiple regression, where different predictors are correlated, it is more akin to the issue of dependency encountered in controlling Type I error rate after performing one-way *ANOVA*. Type I error rises as pairs of conditions containing the same study participants are used repeatedly. There are a number of approaches in the meta-analysis literature that have been devised to handle dependencies of this sort. Scammacca et al. (2013) compared three practical solutions for dealing with dependency to the results achieved by treating all comparisons as if they were independent. These were: (1) selecting the single highest effect size from a group of dependent studies; (2) selecting a single group at random; and (3) selecting the single most representative comparison.

To our minds, all of these approaches involve an unfortunate loss of information and the potential for lower power to find differences in moderator analysis. Ideally, a method should preserve all comparisons while controlling for the inflation of within-study variability associated with treating each comparison as if it were independent. Our preferred approach to controlling for dependencies is to reduce sample by a multiple of the number of comparisons per study. For instance, if the two treatments are compared to a single control condition, the sample size of the control condition is reduced by half. This procedure increases the standard error of each comparison, and hence each variance, so that when studies are synthesized any bias due to dependent samples is minimized.

Identify the number of effect sizes—application

There were no dependent effect sizes in this study of BL, even though in some cases multiple effect sizes were extracted from a single study. In this rare instance, each treatment group had its own independent control group (e.g., comparisons within different semesters).

Extract effect sizes—general

Effect size extraction is defined as the process of locating and coding information contained in research reports that allows for the calculation of an effect size. There are three forms of this metric: (1) *d*-type, or standardized mean differences; (2) *r*-type, or correlations; and (3) *OR*-type, or odds-ratios. Each is calculated differently, but they can be mixed, as there are conversion equations for them all. In many education meta-analyses the *d*-type is used, because of the experimental nature of much of the literature, where some form of intervention is pitted against a non-intervention condition or some alternative. In the best instance, this information

is in the form of means, standard deviations and sample sizes for the experimental and the control conditions. Since there is little in the way of standardized reporting in the experimental literature of education, it is sometimes necessary to extract effect sizes from test statistics (e.g., *t*-ratios), exact probabilities (e.g., $p = .023$) or even inexact hypothesis-test outcomes (e.g., $p < .05$) (See Glass et al. 1981; Hedges et al. 1989 for more on this). There is also a modification of the basic effect size equation for studies reporting pretest and posttest data for both experimental and control groups (Borenstein et al. 2009).

In constructing *d*-type effect sizes, Glass originally used the standard deviation of the control group as the denominator of the effect size equation (i.e., $\Delta = \bar{X}_E - \bar{X}_C / SD_C$), because the untreated control group was considered to be "unbiased by the treatment." Cohen (1988) modified this equation to represent the joint variation in the treatment and the control groups by producing an effect size metric (called Cohen's *d*; Table 2, Eq. 1) based on division of the mean difference by the pooled standard deviations of both groups (Table 2, Eq. 2). Cohen's *d* has become the accepted standardized difference effect size. The equations for additional study-level statistics are shown and described in Table 2.

For a general qualitative assessment of the magnitude of an effect size there is the set of benchmarks established by Cohen (1988), where: (1) $d \geq 0.20 \leq 0.50$ is referred to as a small average effect; (2) $d > 0.50 \leq 0.80$ is referred to as a medium effect) and (3) $d > 0.80$ is called a large effect. Valentine and Cooper (2003) warn that these qualitative descriptors may be misleading in fields like education where smaller effect sizes tend to be the norm.

Another descriptor used in interpreting effect sizes is referred to as U_3 (Cohen 1988) or the "percentage of scores in the lower-meaned group that are exceeded by the average score in the higher-meaned group" (Valentine and Cooper 2003, p. 3). For an effect size of $d = 0.50$, U_3 is approximately 69 % of area under the normal curve. This means that students at the average of the treatment outperformed students at the average of the control group (i.e., the 50th percentile) by 19 % (i.e., 69–50 % = 19 %). Care needs to be taken in interpreting these percentages because not all collections of effect sizes are normally distributed, as it is presumed in this approach to interpretation. Because of this it is generally more accurate for distributions of effect sizes rather than individual ones.

Extract effect sizes—application

Information for effect sizes was extracted by at least two independent coders. Included in this information were the sample size of each condition and the direction of the effect. The inter-rater reliability of this task was 91.90 % ($\kappa = 0.84$).

As a demonstration of how the basic statistics just presented appear in the software package *Comprehensive Meta-Analysis*™ (Borenstein et al. 2005), Fig. 1 shows the descriptive statistics associated with a subset of 21 effect sizes drawn from the complete distribution of 117 effect sizes. On the far left of the figure are the study names, in this case the author names and publication dates. In the center

Table 2 Study-level statistics used in meta-analysis and explanations

Equation number	Equation name	Equation	Explanation
Eq. 1	Cohen's *d* (standardized difference effect size)	$d = \frac{\overline{X}_E - \overline{X}_C}{SD_{Pooled}}$	Cohen's *d* is the basic unit of effect size in meta-analyses that compare an experimental condition with a control condition on a continuous-level dependent variable. The numerator is the ± difference between the means of the experimental condition and the control condition. The denominator is shown in Eq. 2
Eq. 2	Pooled *SD*	$SD_{Pooled} = \sqrt{\frac{(n_E-1)SD^2+(n_C-1)SD^2)}{(n_E+n_C)-1}}$	The pooled standard deviation of the experimental and control conditions' standard deviations is the denominator of *d*-type effect sizes
Eq. 3	Hedges' *g* ($df = N - 1$) Correction for small sample size	$g = d\left(1 - \frac{3}{4df-1}\right)$	Cohen's *d* is called a biased estimator because it does not correct for low sample size that tends to inflate their effect size. Hedges' g is the unbiased estimator
Eq. 4	Standard Error of *g*	$se_g = \sqrt{\frac{n_e+n_c}{n_e n_c} + \frac{d^2}{2(n_e+n_c)}} \cdot \left(1 - \frac{3}{4df-1}\right)$	The unbiased standard error of *g*, based largely on sample size, is the estimated "standard deviation" in the population
Eq. 5	Variance of *g* (Within-study variance)	$v_g = se^2$	The standard error (Eq. 4) is converted to a variance by squaring it
Eq. 6	*z* test (test statistic)	$z_g = \frac{g}{se_g}$	The test statistic *z* is constructed by dividing the effect size *g* by the standard error of *g* (se_g). It tests the null hypothesis (without degrees of freedom) that $g = 0$ (does not exceed chance expectations)
Eq. 7	Two-tailed test of z_g $^{+}(\alpha = .05)$	Null ($g = 0$): $z_g \geq \text{or} \leq \pm 1.96\,(\alpha = .05)$	The two-tailed *z* value null hypothesis $g = 0$ is tested using $z = 1.96$ ($p = .025$) as the critical value
Eq. 8	95th confidence interval	$Lower\,95th = g - (1.96 \cdot SE_g)$ $Upper\,95th = g + (1.96 \cdot SE_g)$	The upper and lower boundaries of the 95th confidence interval define the range within which the effect size is likely to reside. Intervals that cross zero (+ and − limits, or the reverse) are judged to be not significantly different from 0. This interpretation should match the *z* test

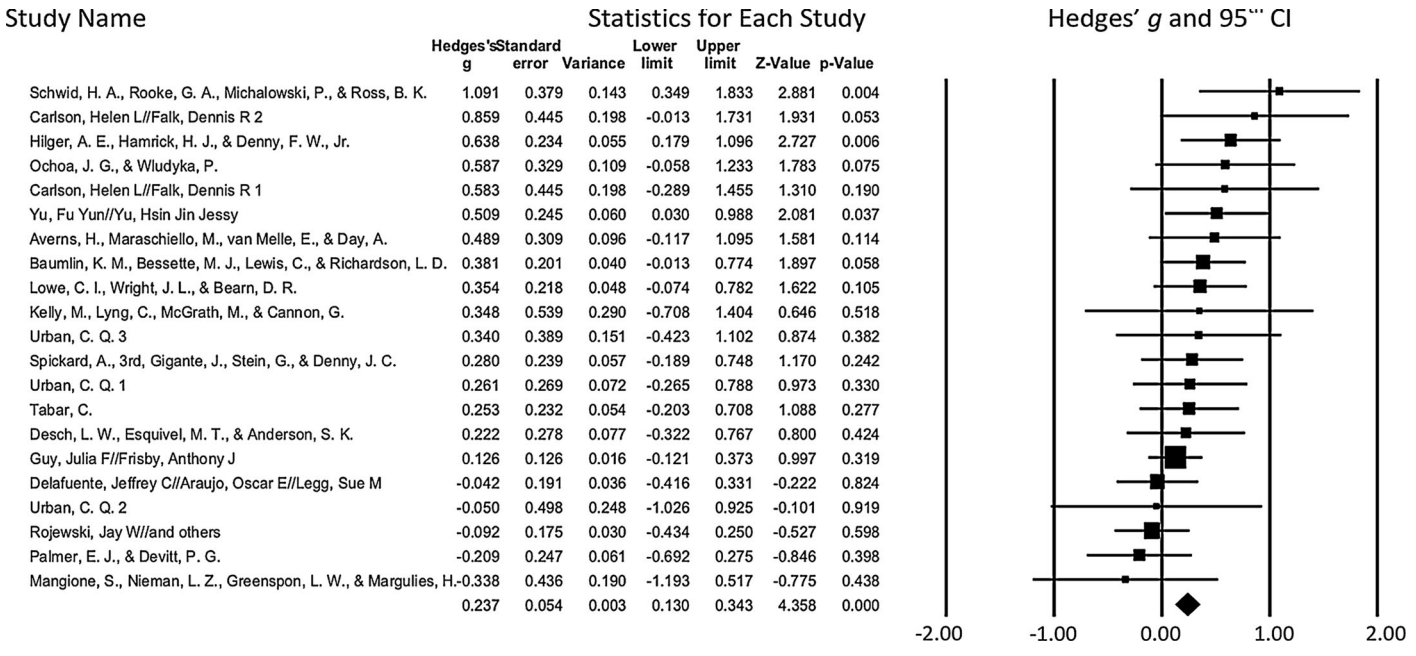

Fig. 1 Study-level statistics and Forest plot of 21 effect sizes from the distribution of 117 effect sizes

are the study-level statistics for these 21 effect sizes: Hedges g (Eq. 3), the standard error (*se*, Eq. 4), the variance (v, Eq. 5), the upper and lower boundaries of the 95th confidence interval, sometimes referred to as *CI*-95th (Eq. 8), the z value (Eq. 7) and its associated probability (p value). On the right side of the figure is a graphic representation called a Forest plot. The effect size for each study is depicted as a dot. The lines around it show the width of the 95th confidence interval for each study. Note that confidence intervals spanning 0.0 on the distribution are considered to be not significantly different from zero. The z test of these effect sizes also indicates that $p > .05$ for these studies. The dots that represent the effect size vary in size. Smaller dots are lower leverage effect sizes (i.e., smaller contributors to the weighted average effect size), while larger dots are higher leverage effects characterized by larger sample sizes.

Code study features and moderator variables—general

Study features can fall into four categories: (1) publication information (e.g., type of document, publication date); (2) methodological quality information (e.g., type of research design, measurement quality); (3) demographic information (e.g., grade level, subject matter); and (4) substantive moderators (i.e., instructional method, time-on-task). Moderators can be either categorical, ordinal or interval/ratio in terms of measurement level.

Moderator variable analysis of coded study features attempts to identify systematic sources of between-study variation, and so if Q_{Total} is not significant (i.e., the effect size distribution is homogeneous) there will be little chance that any moderators will be significant. One potential source of variation that is often referred to in the literature of meta-analysis derives from the presence of different research designs (e.g., Abrami and Bernard 2012). True experiments employing random assignment intended to neutralize selection bias are the most highly prized. Quasi-experimental designs that employ pretesting to establish group equivalence are considered to be reasonable alternatives to true experiments. Pre-experimental designs that contain no mechanism to ensure that selection bias is controlled are considered to be the weakest form of evidence (Campbell and Stanley 1963).

Code study features and moderator variables—application

Moderator analysis of coded study features was used to explore variability in effect sizes. These study features were derived from an ongoing analysis of the theoretical and empirical literature in the field and were based on several previous meta-analyses (Bernard et al. 2009; Schmid et al. 2014). Study features were of four major categories: methodological (e.g., research design); publication demographics (e.g., type of publication); course demographics (e.g., course level); and substantive (e.g., purpose of technology use). Among the latter, we were especially interested in the study features related to BL and interaction treatments. We considered the following dimensions on which experimental conditions consistently could be contrasted to control conditions: communication support; search and retrieval;

cognitive support; content/presentational support; and combinations of purposes. Inter-rater reliability for study features coding was 91.77 % (Cohen's $\kappa = 0.84$).

Step 4: evaluating the quality of studies (judge methodological quality, judge publication bias, perform sensitivity analysis)

Judge methodological quality—general

The methodological quality of included studies is part of the calculus that is used to judge the overall quality of a meta-analysis. If there is differential quality, especially if it favors one category of quality indicators, like research design, all of the findings can be jeopardized. Valentine and Cooper (2008) developed an instrument, called *The Study Design and Implementation Assessment Device* (Study DIAD), intended to improve the validity of judgments of the methodological quality of studies to be included in a systematic review. At the highest level, the device provides the possibility of assessment in four categories: (1) internal validity; (2) measurement and construct validity; (3) statistical validity; and (4) external validity. There are additional lower levels of assessment that when added together result in an overall score in each of the large categories.

Judge methodological quality—application

Our approach is based on the top-level structure of the Study DIAD. We assess each study in terms of six major qualities: (1) research design; (2) measurement quality; (3) effect size extraction precision; 4) treatment duration adequacy; (5) material equivalence; and (6) instructor equivalence. Each dimension is weighted according to its presumed importance to the validity of interpretation (i.e., how trustworthy each study is) and the weighted sum is tested in meta-regression against the effect sizes derived from each study. Table 3 shows the results of this analysis. The conclusion is that the study methodological quality index is not predictive of effect size (the slope is 0.0, $p > .05$). In other words, study quality does not differentially bias the findings of the meta-analysis. Based on this methodological quality analysis, there was no reason to apply any method of correction.

Judge publication bias—general

Analysis of publication bias seeks to determine if a sizable number of studies might have been missed or otherwise not included in a meta-analysis (Rothstein et al.

Table 3 Meta-regression analysis of methodological quality index by effect size (k = 117)

Regression model	Slope and standard error		Confidence interval		Significance test	
	Slope (*b*)	*SE*	Lower 95th	Upper 95th	*Z* value	*p* value
Slope	−0.02	0.02	−0.06	0.02	−0.83	.41
Intercept	0.62	0.35	−0.06	1.30	1.79	.07

2005) and that this number, if found and included, would nullify the average effect. There are various tools for assessing this bias, including the examination of a funnel plot (i.e., effect size by standard error) and statistical procedures like classic fail-safe analysis and Orwin's fail-safe procedure. The classic fail-safe procedure is used to determine how many null-effect studies it would take to bring the probability of the average effect to α. Orwin's procedure indicates the number of null studies needed to bring the average effect size to some standard of triviality (e.g., $g^{+} = 0.10$). Duval and Tweedie's (2004) procedure specifies the number of missing effect sizes necessary to achieve symmetry between effect sizes below and or above the mean. It then recalculates g^{+} considering the studies that were imputed (i.e., added mathematically to achieve symmetry). If no effect sizes are imputed, then symmetry is assumed.

Judge publication bias—application

The funnel plot for the current meta-analysis depicted in Fig. 2 is generally symmetrical around the mean of the distribution ($g^{+} = 0.334$). The following analytical statement about publication bias analysis appears in *Comprehensive Meta-Analysis*™:

> This meta-analysis incorporates data from 117 studies, which yield a z value of 14.97 and corresponding 2-tailed p value of 0.00000. The fail-safe N is 6,709. This means that we would need to locate and include 6,709 'null' studies in order for the combined 2-tailed p value to exceed .05. Put another way, 57.3 missing studies would be needed for every observed study for the effect to be nullified. The Orwin fail-safe for this study is $N = 254$. This means that we would need to locate 254 studies with a Hedges' g of 0.0 to bring the combined Hedges' g under 0.10. The trim and fill results suggest a similar pattern of inclusiveness. Under the fixed effect model the point estimate and 95 % confidence interval for the combined studies is 0.316 (Lower 95th = 0.28, Upper 95th = 0.36). Using trim and fill these values are unchanged. Under the random effects model the point estimate and 95 % confidence interval for the combined studies is 0.334 (Lower 95th = 0.26, Upper 95th = 0.41). Using trim and fill these values are unchanged.

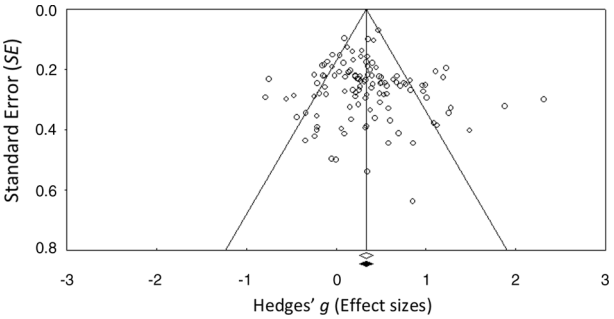

Fig. 2 Funnel plot with effect sizes (*horizontal axis*) and standard errors (*vertical axis*)

(*Note: the fixed effect model and the random effects model are discussed in detail in Step 5 below.*)

We judged that there was no publication bias present that might skew or compromise the results.

Perform sensitivity analysis—general

Outliers can play a significant role in distorting both the overall mean and variability of a collection of effect sizes. This is especially true for the fixed effect model where the inverse of within study variance is used to give priority to large studies with small standard errors and diminish the effect of smaller studies. The random effects model ameliorates this bias somewhat by incorporating average between-study variance (i.e., tau-squared) into the inverse variance weights. However, even with the random effects model, unrealistically large positive or negative effect sizes should be degraded in magnitude or removed from the collection. *Comprehensive Meta-Analysis*™ contains a calculation module called "One Study Removed." When activated, the routine recalculates the statistics with one study removed, in turn, for the entire distribution, so that the researcher can judge the influence or leveraging that individual studies have on the overall weighted average effect size. Large effect sizes, either positively or negatively signed, can affect the weighted average effect size, especially under the fixed effect model. When extreme effect sizes have large sample sizes, this influence is magnified, often skewing the weighted average effect size to a considerable degree. Since inverse variance weights are smaller under the random effects model, the exaggeration caused by "high leverage" effect sizes, is considerably reduced. It is important to examine distributions of effect sizes and to deal with aberrant effect sizes, either by removing them from the distribution (not the best solution), reducing their magnitude (e.g., reduce $g = 8.0$ to some smaller value within range of other large effect sizes) and/or reducing their influence (e.g., reduce sample size).

As just mentioned, occasionally, effect sizes can be so large (or small) as to stand out against the backdrop of more reasonably sized values. For instance, an effect size of ± 8.0 could be considered aberrant if all of the other large effect sizes are in the ± 2.0 to ± 2.5 range. The first course of action is to check for calculation errors, transfer errors or even reporting errors. Failing these explanations, a "one study removed analysis" may reveal that the leveraging effect is either too large or tolerable. If the outlying effects are judged to be too influential, they can be downgraded to some more reasonable value (e.g., ± 8.0 reduced to ± 2.5, as in the previous example) or removed from the analysis that follows. In all cases, a degree of balance should be maintained by considering the consequences of any adjustment.

Perform sensitivity analysis—application

The "One study removed" routine in Comprehensive Meta-Analysis was used to identify sources of potential anomalies in the dataset. When one study was removed

and the random effects means and standard errors were calculated, the lowest mean was $g^{+} = 0.318$, $k = 116$, $SE = 0.037$, and highest mean was $g^{+} = 0.344$, $k = 116$, $SE = 0.038$. Both of these newly calculated averages fall within the confidence interval of the total collection for $g^{+} = 0.334$, $k = 117$, $SE = 0.039$, Lower $95^{th} = 0.26$ and Upper $95^{th} = 0.41$. These data were judged to be extremely stable and reasonably unaffected by anomalous combinations of effect sizes and sample sizes.

Step 5: analyzing and integrating the outcomes of studies (synthesize studies, categorical and continuous moderator variables)

Synthesize studies—general

After all of the effect sizes have been extracted and the basic statistics calculated, the next step is to synthesize them. There are circumstances under which synthesis is not advised (e.g., very small samples of effect sizes, extreme variation in treatment definition or sample). Previously, we have discussed how effect sizes are derived, along with the statistics associated with them and so it is now time to examine how the effect sizes are combined.

Table 4 provides a nearly complete set of the equations for synthesizing a distribution of effect sizes under both the fixed effect and random effects models, along with explanations of their functions and importance. In the following sections, these equations will be to referred and discussed by number, just as they were in the previous section on study-level statistics.

At the study level (or the level of the individual effect size), there is only one source of variability (i.e., within-study variability) that is derived largely from the size of the sample and expressed as the *standard error* (*se*) and the *variance* (*v*). Large sample studies produce small standard errors and variances, while small sample-size produce the reverse. At the level of the meta-analysis (i.e., synthesis) there is another source of variability that can be identified. It is called the "between-study variance" (i.e., analogous to between-group variation in *ANOVA*) because it represents the differences among studies in the meta-analysis. It is how between-study variance is handled that defines the meta-analytic model that is applied and to a great extent, the results that are achieved. There are two analytical models—the fixed effect model and the random effects model—and it is important to understand which should be used in what circumstance. Both models are built around weighted effect sizes, but it how the study weights are constructed that constitutes the primary difference between them.

Fixed effect model The fixed effect model assumes that the distribution of effect sizes is so uniform that it can be described by a fixed weighted average (i.e., a single true point) around which there is only sampling error, and that all excess between-study variability, over and above sampling error, is distributional heterogeneity. This presumption about the nature of the meta-analysis affects how weighting of effect sizes is accomplished. Under the fixed model, the inverse of the within-study variance (Table 4, Eq. 9) is used to give larger studies greater weight and smaller studies less weight. Eq. 11 shows how the weights from all studies are summed.

Table 4 Meta-analysis level statistical equations and explanations

Equation number	Equation name	Equation	Explanation
Eq. 9	Fixed effect model inverse variance weight (where *Vi* is within-study variance)	$W_i = \frac{1}{v_i}$	Under the fixed model, studies are weighted by the inverse (reciprocal) of their within-study variance (se^2) to give them differential (proportional) weight in the synthesis
Eq. 10	Random effects model inverse variance weight (where τ^2 is average between-study variance)	$W_i = \frac{1}{v_i + \tau^2}$	Under the random model, studies are weighted by the inverse of the sum of their within-study variance plus the average between-study variance (τ^2), averaged over the entire distribution of effect sizes. The equations for this statistic are complex and can be found in Borenstein et al. (2009)
Eq. 11	Sum of the weights (fixed [Eq. 9] and random [Eq. 10])	$\sum_{i=1}^{k} W_i = W_1 + W_2 \ldots W_k$	The sum of the weights is the denominator of the weighted average effect size and is always positive
Eq. 12	Sum of the weights times g (fixed [Eq. 9] and random [Eq. 10])	$\sum_{i=1}^{k} W_i g_i = W_1 g_1 + W_2 g_2 \ldots W_k g_k$	The sum of the weighted *effect sizes* (i.e., the weight times g) is the numerator of the weighted average effect size. This sum can be either positive or negative and gives g^+ its positive or negative valence
Eq. 13	Weighted average of g (symbolized as g^+, $\bar{ES}$ and $\bar{T}$) (same for fixed and random models)	$g^+ = \frac{\sum_{i=1}^{k} W_i g_i}{\sum_{i=1}^{k} W_i}$	The weighted average effect size (g^+) is calculated by dividing Eq. 12 by Eq. 11. The numerator of the Eq. determines the $\pm$ sign of the weighted average effect size
Eq. 14	Variance of g^+ (same for fixed and random)	$V_{g^+} = \frac{1}{\sum_{i=1}^{k} W_i}$	The variance of the weighted average effect size is calculated as the inverse of the sum of the fixed or random weights (Eq. 11)
Eq. 15	Standard error of g^+ (same for fixed and random)	$SE_{g+} = \sqrt{V_{g+}}$	The standard error of the weighted average effect size is calculated as the square root of the variance
Eq. 16	Z value for g^+ (same for fixed and random)	$Z_{g+} = \frac{g^+}{SE_{g+}}$	The Z value associated with weighted average effect size is calculated by dividing the weighted average effect size by the standard error of g
Eq. 17	Two-tailed test of $Z_g{}^+$, $\alpha = .05$ (Same for fixed and random models.)	Null ($g = 0$): $Z_g \geq$ or $\leq \pm 1.96 (\alpha = .05)$	This two-tailed test of Z is performed in the same way as described in Eq. 7 (Table 2)
Eq. 18	Confidence interval of g^+ (Same for fixed and random models.)	$Lower\,95th = g - (1.96 \cdot SE_g)$ $Upper\,95th = g + (1.96 \cdot SE_g)$	The upper and lower boundaries are calculated in the same way as described in Eq. 8 (Table 2)

Table 4 continued

Equation number	Equation name	Equation	Explanation
Eq. 19	Heterogeneity of the distribution (g_i and V_i are for individual effect size) (Fixed effect model only.)	$Q_{Total} = \sum_{i=1}^{k} \frac{(g_i - g^+)^2}{V_i}$	Q_{Total} is a sum of squares calculated from the squared deviation around g^+ divided by the variance for each size (See Table 2, Eq. 5). It is the sum of within-study variability and is used to judge the heterogeneity of a distribution of effect sizes
Eq. 20	Test for significance of Q_{Total} (Fixed effect model only.)	$p = \chi^2(Q_{Total})(df = k - 1)$	Q_{Total} is tested for significance using the χ^2 distribution with $k - 1$ degrees of freedom. Note: Q_{Total} is prone to Type II error, especially when the distribution of effect sizes (k) is large
Eq. 21	I^2 (Percentage of heterogeneity) (Fixed effect model only.)	$I^2 = \frac{Q_{Total} - df}{Q_{Total}} \times 100.0\,\%$	I^2 is calculated from Q_{Total} and is interpreted as heterogeneous variation exceeding the expected degree of sampling error. Higgins et al. (2003) provides a qualitative scale that helps in interpretation. However, since it is based on Q_{Total} it has some of the same limitations

For a complete description of these statistical aspects of meta-analysis plus examples please see Borenstein et al. (2009)

Eq. 12, the term referred to as "weight times g" is also summed for each effect size. The weighted average effect size for the fixed model is then formed by dividing Eq. 12 *by* Eq. 11. In Table 4 the calculation of this weighted average is shown in Eq. 13.

The variance of g^+ (V) is formed from the reciprocal of the sum of the weights (Eq. 14) and the standard error (SE) is derived from the variance (Eq. 15). Using g^+ and the standard error, a Z test can be performed (Eq. 16) to test the two-tailed null hypothesis that $g^+ = 0$, and the lower and upper boundaries of the 95th confidence interval can be constructed (Eq. 17). If g^+ is significant (e.g., $g^+ > 0$), it is presumed to be located within the confidence interval range that does not include zero. If g^+ is not significant ($g^+ = 0$), no real assertions about the location of the mean can be made.

Under the fixed effect model, the second form of variation, between-study variation referred to above, is not used in the construction of the average effect size. Instead, between-study variability is summed in a statistic called Q_{Total} (i.e., sum of squares, Eq. 19) that is then tested for significance using the χ^2 distribution, with $k - 1$ (i.e., number of effect sizes minus one) degrees of freedom (Eq. 20). A significant Q value indicates that the distribution is heterogeneous beyond sampling error. One caveat to this is that Q-total is very sensitive to the number of effect sizes in the distribution (k), meaning that smaller distributions tend towards homogeneity while larger distributions are often heterogeneous. Higgins et al. (2003) developed a more descriptive derivative of Q-total referred to as I-square (Eq. 21). It is interpreted as the percentage of heterogeneity, out of 100 %, that exceeds sampling error.

Generally speaking, low heterogeneity comes from standardization of all of the features that contribute to differences among studies, like the nature of the samples, the similarity of the treatments, the precision and similarity of the instrumentation, etc. The more standardization and uniformity, the more viable the fixed effect model becomes. Studies in the health sciences (e.g., medical studies, drug trials) are notable for their intent towards standardization and so are more likely to qualify for the fixed effect model.

Random effects model By contrast, the random effects model does not assume that all effect sizes can be summarized with a common weighted average effect size, because there is no presumption of standardization or uniformity. Because they are so different, except for a general treatment definition, measurement instruments etc., each study is treated as if it was a random sample drawn from micro-populations of like (but unknown) studies, and as such should not be treated so differently in terms of weighting by sample size as any other study. Between-study variability is not summed, as it is in the fixed effect model (i.e., Q_{Total}), but instead distributed as average variability (called τ^2) and added to the within-study variability when forming inverse variance study weights (Eq. 10). These weights tend to be smaller than study weights applied under the fixed effect model, thus providing a more even weighting across the entire distribution. Once the random effects weights are constructed for each effect size, the other steps in constructing meta-analysis level statistics are exactly the same (Eqs. 11–18).

Synthesis strategy In the social sciences, including education, it is generally agreed (e.g., Hedges and Olkin 1985) that most collections of effect sizes around a common question should derive interpretation from the random effects model. This

is because most experimental literatures in education are comprised of primary studies that differ in a number of ways, so that establishing a fixed-point estimate of average effect sizes makes little sense. Examples of difference in primary studies are different research designs, different learners (e.g., age, background), different measures (i.e., some standardized and some not) and different reporting standards. However, this does not mean that the fixed effect model is not useful in interpreting a meta-analysis in education. The statistics Q_{Total} and I^2 tell us about how much variability is present in a collection of studies. These statistics also offer insight into whether it is possible that moderator variables (i.e., coded characteristics of the studies) might help explain how the overall average effect size might differ under different instructional, demographic or other conditions. Without sufficient variability, moderator variable analysis would be severely limited. In summary, both models can play an important role in the interpretation of a meta-analysis at the level of synthesis.

Synthesize studies—application

Data were analyzed using *Comprehensive Meta-Analysis*™ (Borenstein et al. 2005), a dedicated meta-analysis software. The summary statistics derived from $k = 117$ effect sizes are shown in Table 5. This table shows a fixed weighted average effect of $g^+ = 0.316$ (Eq. 13) and a random weighted average effect size of $g^+ = 0.334$ (also Eq. 13) (i.e., a low to moderately low average effect size by Cohen's qualitative standards), the standard error (Eq. 15), the lower and upper limits of the 95th confidence interval (Eq. 18) and the Z value along with two-tailed probability of g^+ (Eqs. 16 and 17). Both the fixed and the random weighted average effect sizes are significantly greater than zero. Heterogeneity statistics are shown for the fixed effect model, Q-total = 465.5, $p < .001$) (Eq. 19), and I-squared = 69.49 % (Eq. 21). (For more information, consult Hedges and Olkin 1985; Borenstein et al. 2009.) The Q-statistic tells us that the distribution is significantly heterogeneous and I-squared indicates that over 50 % of variability in the distribution is between-study variance (i.e., variability in effect sizes that exceeds sampling error). Higgins et al. (2003) call this moderately high between-study variability.

Table 5 Overall weighted average random effects and fixed effect sizes and homogeneity statistics

Analytical models	Effect size and standard error			Confidence interval	
	k	g^+	SE	Lower 95th	Upper 95th
Random effect model	117	0.334*	0.04	0.26	0.41
Fixed effect model	117	0.316**	0.02	0.28	0.36
Heterogeneity	Q-total = 372.91, $df = 116$, $p < .001$		I-squared = 68.89 %		$\tau^2 = 0.11$

* $z = 8.62$, $p < .001$; ** $z = 15.68$, $p < .001$

In this meta-analysis, random effects model statistics are most appropriate for interpretation because of the considerable differences among studies in terms of treatments, measures, participants, etc. (Borenstein et al. 2009). The significant variability indicated from the fixed effect analysis suggests that differences among samples exist. Some of the fluctuation might be accounted for by identifiable characteristics of studies, so that moderator variable analysis is warranted.

Analyze categorical and continuous moderator variables—general

Moderator variable analysis can take two forms: (1) analysis of nominal and/or ordinal-level coded variables using a meta-analysis analog to analysis of variance to determine if different levels or categories produce different results; and (2) weighted meta-regression to determine if continuous predictor variables or "dummy coded" categorical predictors are correlated, thereby accounting for between-study variability.

The most appropriate analytical model for categorical comparisons is the mixed model (Borenstein et al. 2009) which involves synthesizing effects sizes within categories using the random effects model (i.e., tau-squared is differential according to categories) and then comparing the categories using the fixed effect model. The result provides a test across categories (Q-between), in much the same way that *ANOVA* is used to compare levels of an independent variable. Q-between is evaluated using the χ^2 distribution with $p - 1$ degrees of freedom (i.e., number of categories of the moderator variable minus one).

Analyze categorical and continuous moderator variables—application

Six coded moderator variables were analyzed in an attempt to explain some of the previously detected variation in g. The mixed model was used for this purpose.

Categorical moderators The first two moderator variables, subject matter (i.e., STEM vs. Non-STEM) and course level (i.e., undergraduate vs. graduate courses) were not significant across levels. However, all levels of each moderator was significantly greater than zero and hover around $g^+ = 0.334$, the average for the entire collection. The only exception to this is for graduate level courses, with an average effect size of $g^+ = 0.15$, tended to underperform undergraduate courses. Likewise, there was a tendency for STEM subject matters to outperform Non-STEM subject matters.

In another set of analyses we explored the approximate amounts of time spent in BL in the treatment condition. The two categories were low-to-moderate (up to 30 % of course time) and maximum time (approaching 50 % of course time). Remember that studies with greater than 50 % of time online were beyond the scope of this meta-analysis. While there is a definite trend towards higher effect sizes for longer versus shorter time spent online, the Q-between was not significant for this variable (Q-Between $= 0.47$, $df = 1$, $p = .49$) indicating that within chance they are equal. However, this finding is suggestive enough, as well as being congruent with Means et al. (2013), that it should be followed up with further primary studies.

We also explored whether technology in the control condition made a difference compared to no technology. The answer is that it does not in this collection (Q-Between = 0.21, $df = 1$, $p = .65$). It is important to note that technology use in the control condition was never for the purpose of blending face-to-face and online study outside of class time. A similar finding was reported by Schmid et al. (2014), albeit with a much larger collection of effect sizes.

In Table 6a, the analysis of purpose of technology use in the treatment condition is shown. Four distinct levels were identified—communication support, search and retrieval support, cognitive support and content/presentational support. A fifth category, cognitive support + content/presentational support contained various combinations of the two. The Q-between for this variable was significant and a post hoc comparison of cognitive support versus content/presentational support revealed a significant difference between these two levels, $z = 2.28$, $p = .011$, with cognitive support outperforming content/presentational support. This finding replicates Schmid et al. (2014) for this variable. However, the combination category was also significantly lower than the cognitive support level and not significantly different from content/presentational level. The finding for this variable is difficult to interpret and any explanation would be speculative.

In Table 6b, four conditions of coded interaction treatments (IT) were compared. As a reminder, an IT is an arrangement of instructional/technology conditions that is designed to encourage student–student, student–teacher and/or student–content interaction (Bernard et al. 2009). Since in this study, the strength of ITs was not the defining characteristic of the treatment/control distinction, the control condition could contain ITs that the treatment condition did not contain. The first category in Table 6b reflects this and the average effect size of $g^+ = 0.11$ is not significantly greater than chance. The other results (lines 2–4 in Table 6b) strongly favor ITs in the treatment conditions, and in fact their average effect sizes increase incrementally with more ITs. Two ITs are significantly different from one IT ($z = 2.65$, $p < .001$). However, three ITs are not significantly different from two. All of the ITs in treatment conditions are significantly different from category 1. Unfortunately, the exact combinations of ITs could not be explored because of small cell sizes, with the exception of category 4 which contains all three kinds of ITS—student–student, student–teacher and student–content. These results are similar to those reported by Bernard et al. (2009) and in line with Anderson's (2003) hypotheses regarding the additive effects of ITs.

Continuous moderators In the final analysis, we wanted to know if the joint effects of blending with technology on achievement outcomes had changed during the period covered by this meta-analysis, 1990 up to 2010. To do this we ran weighted multiple meta-regression (i.e., method of moments random effects model) treating publication year, a continuous variable, as the predictor and Hedges' g (also continuous) as the outcome variable. This relationship is depicted in the scatterplot in Fig. 3. The results of this analysis revealed that effects associated with BL in higher education ($k = 117$) have not changed substantially over the years ($b_Y = 0.00$, $p = .41$, $Q_{\text{Regression}} = 1.00$, $Q_{\text{Residual}} = 142.00$). There is wide variability around the average effect size of 0.334, and the regression line is virtually flat. However, note the greater number of studies beginning just after the year 2000 and continuing to the year 2010.

Table 6 Mixed analysis of five coded instructional moderator variables

Categories of moderator variables	Effect size and standard error			Confidence interval	
	k	g^+	*SE*	Lower 95th	Upper 95th
(a) *Purpose of technology use*					
Communication support	9	0.31*	0.12	0.07	0.55
Search and retrieval support	3	0.54	0.43	−0.31	1.39
Cognitive support	14	0.59*	0.11	0.38	0.79
Content/presentational support (C/P)	11	0.24*	0.10	0.05	0.44
Cognitive support plus C/P support	63	0.22*	0.04	0.14	0.31
Q-between = 10.72, $df = 4$, $p = .03$	$k = 17$ effect sizes removed because of low cell counts or missing information				
(b) *Interaction treatments in the treatment and control conditions (ITs of student–student, teacher–student, student–content)*					
Treatment < IT in control	13	0.11	0.06	−0.02	0.23
One IT in treatment > control	60	0.26*	0.06	0.15	0.38
Two ITs in treatment > control	11	0.44*	0.07	0.31	0.57
Three ITs treatment > control	6	0.47*	0.16	0.16	0.78
Q-between = 14.40, $df = 3$, $p = .002$	$k = 27$ effect sizes removed when groups were equal in ITs or when different types of ITs favored different groups				

* $p < .05$

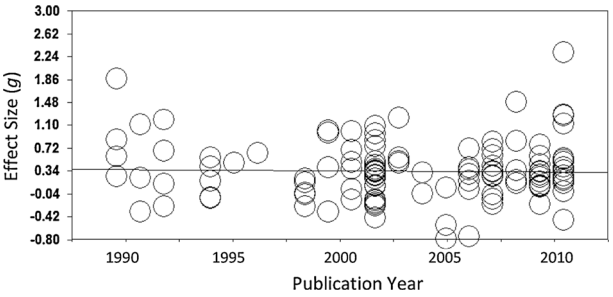

Fig. 3 Scatterplot of publication year by effect size ($k = 117$)

Step 6: interpreting the evidence (draw conclusions from the meta-analysis and discuss the results)

Draw conclusions from the meta-analysis and discuss the results—general

This step in a meta-analysis is fairly straightforward and not that different from the interpretation and discussion of results in any research project. Interpretation and discussion attempts to establish how collected and analyzed data inform the research question(s) that guided the review and explores the possible theoretical, conceptual and/or practical implications of the findings. The results are usually characterized in context with the literature and previous work that has been done, with an emphasis on the contributions that the current study makes to both theory and practice.

Draw conclusions from the meta-analysis and discuss the results—application

The most general conclusion that can be drawn from this meta-analysis is that improvement in achievement related to BL is low but significantly greater than zero. In terms of percentile difference (i.e., U_3 minus 50 %), students at the median of the control condition (i.e., no BL) would be expected to be 13.0 % higher in achievement had they experienced the BL conditions. The average effect size of $g^+ = 0.334$ ($k = 117$) is in the middle of the low category in terms of Cohen's (1988) qualitative strata of effect size magnitude (i.e., $g^+ > 0.20 < 0.50$). A caveat here is that a reasonable degree of between-study heterogeneity makes it more difficult to fix the exact location of the population mean aside from the probability that it resides between $g^+ = 0.26$ and 0.41 (i.e., lower and upper limits of the 95th confidence interval).

Interestingly, this average effect size is in the same ballpark as the findings for BL by Means et al. (2013) where a random effects average effect size of $g^+ = 0.35$ for $k = 23$ effect sizes was reported. This is despite the fact that our definition of BL was somewhat different, and more conservative, than theirs, where OL could take a larger proportion of total time (i.e., from 25 % to close to 100 %). And, it is very close to the overall average effect size from the study by Schmid et al. (2014) from which these BL results are derived ($g^+ = 0.33$, $k = 879$, $p < .001$). All of these

studies, including the present review, compared the presence of technology in the treatment condition to classroom conditions containing little or no technology.

We might surmise from this that the effects of technology integration in higher education, whether into full face-to-face classrooms or in distributed venues in the case of the blending of CI and online instruction, is effective to a modest but significant degree. It is a finding that is worth noting, although it is a difficult problem to explore experimentally, since students working outside of class may devote more or less time to studying online and use materials that may be different from or not available to students working solely in face-to-face classrooms. There is little doubt of confounding in the studies that form the bulk of this meta-analysis and generally speaking meta-analysis is ill-equipped to deal with confounds among substantive moderator variables, since there is no latitude for control as there is in true experiments. On the surface, these results appear to fly in the face of Richard E. Clark's (1983, 1994, 2009) hypothesis about the neutrality of technology use in education, but this cannot be determined for certain from this review, given the nature of the research literature that it draws upon.

However, these results do offer a reason for the continued investigation of BL as a viable, and possibly superior, option to straight face-to-face instruction and its alternatives, DE and OL. The researchable question, however, revolves around the balance between internal and external validity, as elaborated by Campbell and Stanley (1963) and Shadish et al. (2002). Should we control all potentially confounding factors in an effort to isolate the one active ingredient in learning success (a la Clark)—in this case the technology applied or the instructional approach used—risking results that are not replicable or even applicable to the real world of higher education (Abrami and Bernard 2006; Abrami and Bernard 2012)? Or should we consider technology and instructional method as an inseparable dyad that are used together to achieve the goals of education and use experimental replication as a means of determining which combinations work best (Ross and Morrison 1989)?

In many circumstances DE and CI are not true alternatives to one another, since they often serve distinctive clienteles—those who can attend face-to-face classes and those who cannot or prefer not to. BL is more like CI than DE, because it requires face-to-face attendance at particular times and at particular places, but if the online component is integrated successfully, it seems to add a dimension of independence from time and place that may turn out to be both more motivating and more facilitative of the goals of instruction than either CI or DE/OL. Therefore, new generations of primary research should address questions related to instructional design, particularly in regards to what mixes of CI and online conditions produce both *deep and meaningful learning* and *more satisfying educational experiences.* Understanding about student interaction and how to design *interaction treatments* that promote opportunities for students to interact among themselves, with the teacher and/or with course content is also central.

Abrami et al. (2011) argued that the next generation of designs for OL, and by extension BL, should help facilitate student motivation, self-regulated individual behavior and more purposeful student–student interaction. However, designing guided, focused and intentional student activity goes beyond just providing

opportunities for interaction (i.e., the definition of interaction treatments in Bernard et al. 2009). Researchers and designers (and instructors) must carefully consider why these forms of activity and/or mediated setups are desirable, and more importantly, how they can better facilitate learning based on theory, so that they are powerful and replicable. The literature is replete with examples of pedagogy and its attendant technology which fail, partly because the fundamentals are not grounded in theory and/or evidence-based principles of instructional design.

To realize evidence-based practice, we see theoretical work and research in three areas that are pertinent to future exploration and development in BL: (1) self-regulation design principles (e.g., Zimmerman 2000); (2) motivational design principles (e.g., Pintrich 2003); and (3) collaborative and cooperative learning design principles (e.g., Johnson and Johnson 2009), each discussed briefly below.

When working in the OL portion of BL, students need to learn skills, be provided with scaffolded experience and allowed to practice in real learning environments. This is important for students in BL settings because, by and large, these students are working outside of the orbit of direct teacher influence. Assignments that help students find value in goal setting, strategic planning, self-observation (i.e., self-reflection), etc., among the primary pillars of educating students in self-regulation, need to be promoted in the BL environment as well as in CI. Likewise, motivation is key to successful learning in BL environments. Structuring learning environments that encourage self-efficacy, stimulate interest and intrinsic motivation, and ensure task value are likely to be particularly effective under BL instructional conditions. Finally, cooperative and collaborative learning opportunities can be built into BL assignments (Bernard et al. 2000), can improve OL outcomes (Borokhovski et al. 2012) and can support OL/BL tool construction (Abrami 2010) to not only enhance learning, but also to strengthen elements of motivation and self-regulation. The literatures in these three areas of educational theory and practice intersect sufficiently that in combination they can provide the basis for a new and powerful theory of BL, thus laying the groundwork for future research agendas and even greater successes in practice.

Step 7: presenting the results

General

There are at least three audiences that may be interested in the results of a systematic review of the type described here. Practitioners, teachers in higher education contexts in this instance, may use the results to become knowledgeable of research findings and possibly modify or supplement their practices. Policy-makers, individuals who make purchase decisions and form policies that affect the large-scale adoption of innovations, may be informed by the results of a systematic review. It is recognized that research represents only one form of evidence for decision-making, but with access to the voices of researchers, broadly characterized, policy-makers are in a better position to make rational and informed choices. The third group that is likely to be affected by a systematic review is researchers who have contributed, or may potentially contribute studies to the growing corpus of

evidence that form the substance of reviews. Researchers need to know the directions of inquiry that are informing their field, in addition to the design, methodological and reporting conventions that make it possible for individual studies to be included in reviews.

For the first two groups, there may be an issue related to the form of reporting a systematic review or meta-analysis. As has been demonstrated here, a fair degree of knowledge and a sizable team of researchers is required to construct a review, and likewise there is some degree of knowledge related to its interpreting and applying its findings. Knowledge translation centers have been established to act as "go-betweens," of sorts, linking the researcher or meta-analyst with consumers of the results of systematic reviews or meta-analyses (e.g., What Works Clearinghouse).

Conclusion

In this article we have provided a brief description of meta-analysis as a methodology for quantitatively synthesizing the results of many comparative studies organized around a central question or set of questions. Seven steps were described in moderate detail and an example, an examination of BL teaching in higher education, was provided.[2] Using this methodology to address the question of whether technology has an impact on learning, we have found numerous examples of how meta-analysis both effectively synthesizes extant, empirical data, and perhaps even more importantly, serves as a heuristic to identify potent, causal factors that can inform practice and further research. We recognize that such procedures, as well as the meta-analyst who applies them, are "prisoners of the data": if the data are biased, so too will the results of a meta-analysis be biased. This illustrates the imperative that future primary researchers unequivocally isolate variables, create measures that are valid and reliable tools, and restrict the interpretation of outcomes to those that are core to the research question.

References

Note: A list of studies included in the meta-analysis is available upon request from the authors.

Abrami, P. C. (2010). On the nature of support in computer supported collaborative learning using gStudy. *Computers in Human Behavior, 26*(5), 835–839. doi:10.1016/j.chb.2009.04.007.

Abrami, P. C., & Bernard, R. M. (2006). Research on distance education: In defense of field experiments. *Distance Education, 27*(1), 5–26.

Abrami, P. C., & Bernard, R. M. (2012). Statistical control versus classification of study quality in meta-analysis. *Effective Education*, *4*(1), 43–72. doi:10.1080/19415532.2012.761889.

[2] More information about specific topics concerning systematic review and meta-analysis is available in the following publications: (1) history (Hunt, 1997; Glass, 1976); (2) general methodology (e.g., Cooper et al. 2009; Cooper 2010; Pettigrew and Roberts 2006; Lipsey and Wilson 2001; Cooper et al. 2009); and (3) statistics (e.g., Piggot 2012; Borenstein et al. 2009; Hunter and Schmidt 1990; Hedges and Olkin 1985).

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M., Tamim, R. M., et al. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage one meta-analysis. *Review of Educational Research, 78*(4), 1102–1134. doi:10.3102/0034654308326084.

Abrami, P. C., Bernard, R. M., Bures, E. M., Borokhovski, E., & Tamim, R. (2011). Interaction in distance education and online learning: Using evidence and theory to improve practice. *Journal of Computing in Higher Education, 23*(2/3), 82–103. doi:10.1007/s12528-011-9043-x.

Albrecht, B. (2006). Enriching student experience through blended learning. *ECAR Research Bulletin, 12*.

Allen, M., Bourhis, J., Burrell, N., & Mabry, E. (2002). Comparing student satisfaction with distance education to traditional classrooms in higher education: A meta-analysis. *American Journal of Distance Education, 16*(2), 83–97. doi:10.1207/S15389286AJDE1602_3.

Allen, M., Bourhis, J., Mabry, E., Burrell, N., Timmerman, E., & Titsworth, S. (2006). Comparing distance education to face-to-face methods of education. In B. Gayle, R. Preiss, N. Burrell, & M. Allen (Eds.), *Classroom and communication education research: Advances through meta-analysis* (pp. 229–241). Hillsdale, NJ: Erlbaum.

Allen, M., Mabry, E., Mattrey, M., Bourhis, J., Titsworth, S., & Burrell, N. (2004). Evaluating the effectiveness of distance learning: A comparison using meta-analysis. *Journal of Communication, 54*(3), 402–420. doi:10.1111/j.1460-2466.2004.tb02636.x.

Anderson, T. (2003). Getting the mix right again: An updated and theoretical rationale for interaction. *International Review of Research in Open and Distance Learning, 4*(2), 9–14. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/149.

Arabasz, P., & Baker, M. (2003). Respondent Summary: Evolving campus support models for e-learning courses. *EDUCAUSE Center for Applied Research.* http://www.educause.edu/ir/library/pdf/EKF/ekf0303.pdf.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*(2), 111–127. doi:10.2190/9LMD-3U28-3A0G-FTQT.

Bele, J. L., & Rugelj, J. (2007). Blended learning—an opportunity to take the best of both worlds. *International Journal of Emerging Technologies in Learning, 2(3).* doi:10.3991%2Fijet.v2i3.133.

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

Bernard, R. M., Abrami, P. C., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., et al. (2004). How does distance education compare to classroom instruction? A Meta-analysis of the empirical literature. *Review of Educational Research, 74*(3), 379–439. doi:10.3102/00346543074003379.

Bernard, R. M., Rojo de Rubalcava, B., & St-Pierre, D. (2000). Collaborative online distance education: Issues for future practice and research. *Distance Education, 21*(2), 260–277. doi:10.1080/0158791000210205.

Bernard, R. M., Zhang, D., Abrami, P. C., Sicoly, F., Borokhovski, E., & Surkes, M. (2008). Exploring the structure of the Watson-Glaser critical thinking appraisal: One scale or many subscales? *Thinking Skills and Creativity, 3*, 15–22. doi:10.1016/j.tsc.2007.11.001.

Bliuc, A. M., Goodyear, P., & Ellis, R. A. (2007). Research focus and methodological choices in studies into students' experiences of blended learning in higher education. *The Internet and Higher Education, 10*(4), 231–244. doi:10.1016/j.iheduc.2007.08.001.

Bonk, C. J., & Graham, C. R. (Eds.). (2006). *The handbook of blended learning: Global perspectives, local designs.* San Francisco, CA: Pfeiffer.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2005). *Comprehensive meta-analysis version 2.2.048*. Englewood, NJ: Biostat.

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

Borokhovski, E., Tamim, R. M., Bernard, R. M., Abrami, P. C., & Sokolovskaya, A. (2012). Are contextual and design student–student interaction treatments equally effective in distance education? A follow-up meta-analysis of comparative empirical studies. *Distance Education, 33*(3), 311–329. doi:10.1080/01587919.2012.723162.

Campbell, D., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Cavanaugh, C. S. (2001). The effectiveness of interactive distance education technologies in K-12 learning: A meta-analysis. *International Journal of Educational Telecommunications, 7*(1), 73–88. Norfolk, VA: AACE. Retrieved March 13, 2007 from http://www.editlib.org/p/8461.

Cavanaugh, C., Gillan, K. J., Kromey, J., Hess, M., & Blomeyer, R. (2004). *The effects of distance education on K-12 student outcomes: A meta-analysis*. Naperville, IL: Learning Point Associates.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445–459. doi:10.3102/00346543053004445.

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and Development, 42*(2), 21–29. doi:10.1007/BF02299088.

Clark, R. E., Yates, K., Early, S., & Moulton, K. (2009). An analysis of the failure of electronic media and discovery-based learning: Evidence for the performance benefits of guided training methods. In K. H. Silber & R. Foshay (Eds.), *Handbook of training and improving workplace performance, Volume I: Instructional design and training delivery*. Washington, DC: ISPI.

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

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.

Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J., & Montori, V. M. (2008). Internet-based learning in the health professions: A meta-analysis. *Journal of the American Medical Association, 300*(10), 1181–1196. doi:10.1001/jama.300.10.1181.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Los Angeles, CA: SAGE Publications.

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

Driscoll, M., & Carliner, S. (2005). *Advanced web-based training strategies. Blended learning as a curriculum design strategy* (pp. 87–116). New York, NY: ASTD Press.

Duval, S., & Tweedie, R. (2004). 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.

Garrison, D. R., & Vaughan, N. D. (2008). *Blended learning in higher education: Framework, principles, and guidelines*. San Francisco, CA: Jossey-Bass.

Glass, G. V. (1976). Primary, secondary and meta-analysis. *Educational Researcher, 5*(10), 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Graham, C. R. (2005). Blended learning systems. In C. J. Bonk, & C. R. Graham (Eds.), *The handbook of blended learning: Global perspectives, local designs*. Chichester: Wiley (originally Pfeiffer).

Hammerstrøm, K., Wade, C. A. & Jørggensen, A.-M. K. (2010). *Searching the literature: A guide to information retrieval for campbell systematic reviews 2010: Supplement 1*. Oslo, Norway: The Campbell Collaboration. doi:10.4073/csrs.2010.1.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., Shymansky, J. A. & Woodworth, G. (1989). *A practical guide to modern methods of meta-analysis*. (Stock Number PB-52). Washington, DC: National Science Teachers Association. (ERIC Document Reproduction Service No. ED309952).

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal, 327*, 557–560. doi:10.1136/bmj.327.7414.557.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York, NY: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: SAGE Publications.

Jahng, N., Krug, D. & Zhang, Z. (2007). Student achievement in online education compared to face-to-face education. *European Journal of Open, Distance and E-Learning*. Retrieved from http://www.eurodl.org/materials/contrib/2007/Jahng_Krug_Zhang.htm.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 368*(5), 365–379. doi:10.3102/0013189X09339057.

Keegan, D. (1996). *Foundations of distance education* (3rd ed.). London: Routledge.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. New York, NY: Sage Publications.

Lou, Y., Abrami, P. C., & D'Appollonia, S. (2001). Small group and individual learning with technology: A meta-analysis. *Review of Educational Research, 71*, 449–521. doi:10.3102/00346543071003449.

Lou, Y., Bernard, R. M., & Abrami, P. C. (2006). Media and pedagogy in undergraduate distance education: A theory-based meta-analysis of empirical literature. *Educational Technology Research and Development, 54*, 141–176. doi:10.1007/s11423-006-8252-x.

Machtmes, K., & Asher, J. W. (2000). A meta-analysis of the effectiveness of telecourses in distance education. *American Journal of Distance Education, 14*(1), 27–46. doi:10.1080/08923640009527043.

Marquis, C. (2004). WebCT survey discovers a blend of online learning and classroom-based teaching is the most effective form of learning today. *WebCT.com*. Retrieved from http://www.webct.com/service/ViewContent?contentID=19295938.

Means, B, Toyama, Y., Murphy, R. F., & Baki, M. (2013). The effectiveness of online and blended learning: A meta-analysis of the empirical literature, *Teachers College Record*, *115*(3), 1–47. Retrieved from http://www.tcrecord.org/library/content.asp?contentid=16882.

Means, B., Toyama, Y., Murphy, R., Bakia, M,. & Jones, K. (2010). *Evaluation of evidence-based practices in online Learning: A meta-analysis and review of online learning studies.* Technical Report. U. S. Department of Education, Washington, DC.

Moore, M. G. (1989). Editorial: Three types of interaction. *The American Journal of Distance Education, 3*(2), 1–6. doi:10.1080/08923648909526659.

Pettigrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford, UK: Blackwell Publishing.

Piggot, T. D. (2012). *Advances in meta-analysis*. New York, NY: Springer.

Pintrich, P. R. (2003). A Motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*(4), 667–686. doi:10.1037/0022-0663.95.4.667.

Rosen, Y., & Salomon, G. (2007). The differential learning achievements of constructivist technology-intensive learning environments as compared with traditional ones: A meta-analysis. *Journal of Educational Computing Research, 36*(1), 1–14. doi:10.2190/R8M4-7762-282U-554J.

Ross, S. M., & Morrison, G. R. (1989). In search of a happy medium in instructional technology research: Issues concerning external validity, media replications, and learner control. *Educational Technology Research and Development, 37*(1), 19–33. doi:10.1007/BF02299043.

Ross, S. M., Morrison, G. R., & Lowther, D. L. (2010). Educational technology research past and present: Balancing rigor and relevance to impact school learning. *Contemporary Educational Technology, 1*(1), 17–35. Retrieved from http://www.cedtech.net/articles/112.pdf.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, UK: 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.* Published online 13 September 2013. doi:10.3102/0034654313500826.

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.

Shachar, M., & Neumann, Y. (2003). Differences between traditional and distance education academic performances: A meta-analytical approach. *International Review of Research in Open and Distance Education, 4*(2). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/153/704.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Sitzmann, T., Kraiger, K., Stewart, D., & Wisher, R. (2006). The comparative effectiveness of web-based and classroom instruction: A meta-analysis. *Personnel Psychology, 59*(3), 623–664. doi:10.1111/j.1744-6570.2006.00049.x.

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.

Ungerleider, C. S., & Burns, T. C. (2003). Information and communication technologies in elementary and secondary education: A state of the art review. *International Journal of Educational Policy, Research & Practice, 3*(4), 27–54.

Valentine, J. C., & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinghouse.

Valentine, J. C., & Cooper, H. (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.

Watson, G., & Glaser, E. M. (1980). *Watson–Glaser critical thinking appraisal: Forms A and B*. San Antonio, TX: PsychCorp.

Williams, S. L. (2006). The effectiveness of distance education in allied health science programs: A meta-analysis of outcomes. *American Journal of Distance Education, 20*(3), 127–141. doi:10.1207/s15389286ajde2003_2.

Zhao, Y., Lei, J., Yan, B., & Tan, S. (2005). *What makes the difference? A practical analysis of research on the effectiveness of distance education*. Retrieved from http://ott.educ.msu.edu/literature/report.pdf.

Zhoa, Y., & Breslow, L. (2013). *Literature review on hybrid/blended learning*. Unpublished manuscript. Retrieved from http://tll.mit.edu/sites/default/files/library/Blended_Learning_Lit_Reveiw.pdf.

Zimmerman, B. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaertzs & P. R. Pintrich (Eds.), *Handbook of self-regulation* (pp. 13–39). New York, NY: Academic Press.

Robert M. Bernard Ph.D., is professor of education and Systematic Review Team Leader for the Centre for the Study of Learning and Performance at Concordia University. His research interests include distance and online learning and instructional technology. His methodological expertise is in the areas of research design, statistics and meta-analysis. This CSLP research team has published four meta-analyses in *Review of Educational Research* since 2004, six additional meta-analyses and systematic reviews in major research journals plus more than a dozen articles on various aspects of research and meta-analysis methodology. Members of the team have also conducted seminars, workshops and short courses for the Campbell Collaboration and in a number of research centers and universities in Canada, the United States and Europe and received awards for their outstanding contributions to research in educational technology.

Eugene Borokhovski Ph.D., is the Systematic Reviews 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, meta-analysis in particular.

Richard F. Schmid Ph.D., is professor of education, chair of the Department of Education, and Educational Technology Theme Leader for the Centre for the Study of Learning and Performance at Concordia University. His research interests include examining pedagogical strategies supported by technologies and the cognitive and affective factors they influence.

Rana M. Tamim Ph.D., is an assistant 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.

Philip C. Abrami Ph.D., is a research chair and the director of the Centre for the Study of Learning and Performance at Concordia University. His current work focuses on research integrations and primary investigations in support of applications of educational technology in distance and higher education, in early literacy, and in the development of higher order thinking skills.