ReCALL **30**(3): 253–277. 2018 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344017000271
First published online 11 October 2017

Transparency of reporting in CALL meta-analyses between 2003 and 2015

HUIFEN LIN

National Tsing Hua University, Taiwan
(email: huifen@mx.nthu.edu.tw)

TSUIPING CHEN

Kun Shan University, Taiwan
(email: tsuiping0925@gmail.com)

HSIEN-CHIN LIOU

Feng Chia University, Taiwan
(email: liuhsienc@gmail.com)

Abstract

Since its introduction by Glass in the 1970s, meta-analysis has become a widely accepted and the most preferred approach to conducting research synthesis. Overcoming the weaknesses commonly associated with traditional narrative review and vote counting, meta-analysis is a statistical method of systematically aggregating and analyzing empirical studies by following well-established procedures. The findings of a meta-analysis, when appropriately conducted, are able to inform important policy decisions and provide practical pedagogical suggestions. With the growing number of publications employing meta-analysis across a wide variety of disciplines, it has received criticism due to its inconsistent findings derived from multiple meta-analyses in the same research domain. These inconsistencies have arisen partly due to the alternatives available to meta-analysts in each major meta-analytic procedure. Researchers have therefore recommended transparent reporting on the decision-making for every essential judgment call so that the results across multiple meta-analyses become replicable, consistent, and interpretable. This research explored the degree to which meta-analyses in the computer-assisted language learning (CALL) discipline transparently reported their decisions in every critical step. To achieve this aim, we retrieved 15 eligible meta-analyses in CALL published between 2003 and 2015. Features of these meta-analyses were extracted based on a codebook modified from Cooper (2003) and Aytug, Rothstein, Zhou and Kern (2012). A transparency score of reporting was then calculated to examine the degree to which these meta-analyses are compliant with the norms of reporting as recommended in the literature. We then discuss the strengths and weaknesses of the methodologies and provide suggestions for conducting quality meta-analyses in this domain.

Keywords: transparency, reporting, meta-analysis, CALL, systematic review, synthesis

1 Introduction

When research studies on certain topics in a field accumulate, there is a need for researchers to examine and compare the findings of these studies to either confirm/reject a hypothesis or to advance a theory. The history of using a systematic and quantitative method to review a large body of studies can be dated back to the 1930s (Liao & Hao, 2008). Since then, researchers and statistical experts have endeavored to develop systematic and professional statistical tools to combine and analyze results from empirical studies. These methods enable researchers to summarize primary studies in a replicable way, thus producing more supportable findings than narrative reviews and vote counting, both of which have been long used to synthesize cumulated studies (Norris & Ortega, 2000). Meta-analysis, as a statistical analysis of primary studies to integrate findings, was first proposed by Glass (1976). Since then, it has been used increasingly, and the findings of meta-analyses are widely cited. Nowadays though, Glass's model of research synthesis is no longer considered appropriate, as new methods for retrieving, integrating, and interpreting research findings have been developed (Cooper, 2007). Researchers have argued that the decision rules outlined by the originators of meta-analysis such as Glass, McGaw and Smith (1981) back in the 1980s should be modified and expanded as new meta-analytic methodologies are developed (Cooper, 2007). This need to revise the practice of meta-analysis has arisen also due to the many unanticipated results arrived at by meta-analyses on the same topic. Little consistency in the application of meta-analytic methods and the many variations a meta-analyst can adopt in the procedures and decision points mean that the results of meta-analyses are neither replicable nor comparable (Rothstein & McDaniel, 1989). Norris and Ortega (2000), for example, meta-analyzed 49 unique sample studies on the effectiveness of instruction in L2 learning. They categorized the studies in the sample into four groups based on the level of explicitness of instruction (i.e. explicit vs. implicit) and attention to form (focus on form vs. focus on formS). They compared these four types of instruction to baseline/comparison instructions (i.e. no instruction or non-focused exposure to the structures received by the experimental groups) and found that, overall, explicit types of instruction are more effective than implicit types. As an extension of Norris and Ortega's study, Goo, Granena, Yilmaz and Novella (2015) also meta-analyzed 34 unique sample studies, in which 11 were from Norris and Ortega's meta-analysis, and tried to scrutinize the relative effect of implicit and explicit L2 instruction. Both meta-analyses revealed somewhat similar results in that, overall, explicit instruction was more effective than implicit instruction. However, Goo *et al.* (2015) found both explicit and implicit instructions led to a large effect size on immediate post-test, whereas in Norris and Ortega (2000), the large effect size was associated only with explicit instruction. Goo *et al.* attributed the differences in findings to inherent differences in sampling the eligible studies. Norris and Ortega included all experimental and quasi-experimental studies in which either explicit or implicit instruction was compared with a control/comparison group, while Goo *et al.* only included studies where both explicit and implicit instruction were designed and compared. This example illustrates how different decisions in meta-analytic process can affect the outcomes.

To enhance the comparability, interpretability, and replicability of meta-analyses across disciplines, the analytical procedures used have to be clear, consistent, and, most important of all, transparent to readers and consumers of the meta-analyses. It is recommended that meta-analysts explicitly describe their procedures, offer justification or a rationale for decisions when alternatives are possible, and explain how different approaches might have affected the conclusions. In other words, the causes of inconsistency can be found and resolved as long as the authors make transparent their decisions at every judgment call.

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

2 Literature review

Meta-analysis gained popularity as a systematic form of secondary review in the 1980s. Since then, researchers have started to discuss and formulate procedures and examples for conducting such secondary reviews (Liao & Hao, 2008). However, it was not until the mid 1990s that the number of meta-analyses burgeoned (Littell, Corcoran & Pillai, 2008). This increasing interest in employing meta-analysis as a research synthesis method in second language learning/teaching revealed researchers' recognition of its validity in terms of being able to scientifically aggregate and analyze study findings. Meta-analysis is also able to identify gaps between available studies and to suggest future research directions or even formulate research agendas. As a systematic review, meta-analysis employs statistical methods to integrate and summarize primary studies on a particular topic by comprehensively locating research studies using "organized, transparent, and replicable procedures at each step in the process" (Cooper, 2007: 1). Meta-analysis, like most primary studies and any form of systematic review, follows similar steps: topic formulation, treatment design, sampling, data collection, data analysis, and reporting of results. In the topic formulation stage, research questions, hypotheses, and research purposes are proposed based on research interest and theoretical rationale. Involved in the overall study design are tasks such as developing a protocol, and specifying problems, conditions, sampling procedures, and outcomes of interest. Most important of all, study inclusion and exclusion criteria have to be proposed. A sampling plan has then to be developed in which the study is the sampling unit. Potentially all relevant studies will have to be searched for and obtained. In the data collection step, data are extracted from the primary studies and are integrated following a standardized format. Different approaches to analyzing data extracted from included primary studies in a meta-analysis are possible. However, basic and common steps include the provision of descriptive data on study features and intervention characteristics, examining heterogeneity from the obtained effect sizes, conducting moderator analysis[1] and sensitivity analysis,[2] and detecting publication bias.[3] In the final step of the meta-analysis, tables and graphs are employed to describe the results, interpretation and discussion of findings are presented, and the implications for policy, practice, and future research suggestions are proposed (Littell *et al.*, 2008).

2.1 Current state of the art

As mentioned earlier, inconsistency or conflicts in the conclusions of meta-analyses conducted on the same or similar topics can be attributed to several factors, such as more than

[1] A moderator is a variable that is hypothesized to affect the relationship between the independent and dependent variable. A moderator analysis is typically performed to examine if certain moderators such as treatment duration, intensity, and characteristics of the sample or study setting explain the effectiveness of the treatment (Shadish & Sweeney, 1991).

[2] When conducting a meta-analysis, the analysts may be confronted with a number of choices such as analysis model (i.e. random effects model or fixed effects model). Different choices may affect the results of the analysis. Thus a sensitivity analysis is usually necessary to detect how results may be different depending on the choices made (Elvik, 2005).

[3] Publication bias refers to a problem in meta-analysis, which is more likely to include in its review studies with significant results and thus biased in favor of studies with positive outcomes (Copas & Shi, 2000).

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

one alternative in major procedures. The call for complete and transparent reporting of decision-making in these critical steps has led to the development of standards or instruments to guide meta-analysts regarding what to report in each stage. Cooper (2007) developed a checklist of 20 questions to evaluate the validity of the research synthesis conclusions. Based on Cooper's checklist and other documents related to reporting standards, a working group on journal article reporting standards (JARS), commissioned by the American Psychological Association Publications and Communications Board, established the Meta-Analysis Reporting Standards (MARS) to recommend information to be included when reporting meta-analyses. These standards are much more comprehensive, covering what to describe/report in each section of a paper or topic. Other more concise and recent measurement tools such as A Measurement Tool to Assess Systematic Reviews (AMSTAR) have been proposed since MARS (e.g. Aytug *et al.*, 2012; Plonsky, 2012; Shea, Hamel, Wells, Bouter, Kristjansson, Grimshaw, Henry & Boers, 2009), all reacting to the impetus for more detailed and complete reporting of how meta-analyses are conducted and what they find.

We used MARS as a basis to develop a framework against which four other assessment tools/instruments were compared. The number of items in the surveyed instruments ranged between 17 and 54. All items can be classified into Introduction, Literature search, Method, and Discussion/Conclusion with some degree of variation, with the exception of AMSTAR, which created items to assess information related to data sources, analysis of individual studies, meta-analysis, reporting, and interpretation. It also asks for a summary judgment for each section. We examined the nature of the items and tallied the number that was deemed to be important by at least three of the five instruments that we surveyed. We found that in the Introduction section, meta-analysts need to specify the questions under investigation and related theory/policy or practical issues for such a synthesis. In the Method section, details such as inclusion and exclusion criteria, operational definition for both independent and dependent variables and moderator/mediator analysis need to be provided. In terms of searching for eligible literature, information on references, citation databases, and registries searched, as well as efforts to retrieve all available studies need to be supplied; the process of determining study eligibility needs to be described as well. In coding procedures, inter-coder reliability or agreement, and ways to assess study quality and handle missing data need to be explained; in the section that reports the statistical method, effect size metrics and averaging and/or weighting method, effect size confidence intervals or standard error need to be provided, and the meta-analysts also need to explain how to deal with studies with more than one effect size and what analysis model and assessment of heterogeneity were employed with appropriate justification. When reporting the results, a descriptive table (with effect size and sample size for each study) supplemented with tables or graphic summaries are recommended. When discussing the results, major findings, alternative explanations for observed results, study generalizability, limitations, implications, and interpretation for theory/policy or practice need to be addressed with guidelines for future research.

2.2 Second-order meta-analysis of meta-analysis in CALL

Second-order meta-analysis, also called "overview of reviews", "umbrella review", "meta-meta-analysis", and "meta-analysis of meta-analysis" (Schmidt & Oh, 2013: 204) is a research synthesis methodology that integrates evidence from multiple first-order meta-analyses with the aim of gauging the degree to which the variance in effect size calculated from the first-order meta-analyses was due to second-order sampling error, the estimate of which can better inform the precision of the effect sizes derived from the individual meta-analyses (Schmidt & Oh, 2013). An alternative focus of second-order meta-analysis could be on the way in which each meta-analysis was conducted. The authors were able to locate two such studies in the field of applied linguistics. Each study is briefly introduced as follows.

Plonsky and Ziegler (2016) used a revised version of Plonsky (2012) to evaluate the rigor and transparency of 10 meta-analyses in applied linguistics. The inter-rater reliability of the instrument was .87. Several observations of the meta-analyses reviewed in this second-order meta-analysis were presented: (1) the standards proposed in the instruments regarding the literature review were mostly met by the sample, except that most authors failed to provide justifications for inclusion of certain moderator variables; (2) the Method section is the area that needs much improvement; although the authors in the sample provided clear inclusion and exclusion criteria to screen eligible studies employing appropriate search techniques, not many studies employed a quality index to assess primary studies before or after they were integrated for further analysis; (3) there was a lack of discussion of how the findings were derived from the individual meta-analyses to inform theory and recommendations for future research.

Liou and Lin (2017) adopted an instrument developed by Aytug *et al.* (2012) to assess the transparency of reporting and the rigor of 13 meta-analyses on computer-assisted language learning (CALL). Their instrument consists of 18 items derived from a 54-item pool. These 18 items were endorsed by experts and are regarded to be “ethically imperative” (110); a meta-analytic report with no provision of information on these items would be considered as low quality and would be less likely to be replicated. This secondary meta-analysis found that the more recent meta-analytic reports were not more transparent or rigorous in their reporting and conduct than earlier ones, which is contrary to our hypothesis that the development and growth of meta-analytic research knowledge and techniques should enable the recent studies to be more finely tuned. The authors also found that the meta-analysts did not provide the keywords they used to search for relevant literature, nor did they provide justifications for analyzing certain moderator variables. Study features were normally not listed, and information on inter-rater reliability was either missing or such reliability was not checked.

2.3 Purpose statement

Although meta-analysis has become a widely accepted research synthesis method in the social science field, the inconsistent findings derived from multiple meta-analyses in the same research domain are regarded as a major weakness. Researchers have argued, though, that the inconsistencies in meta-analysis results are more easily resolved than those from narrative reviews, as long as meta-analysts “fully articulate” their decision rules (Rothstein & McDaniel, 1989: 766). Given the proliferation of meta-analyses conducted in disciplines in the social sciences, and the growing number of publications synthesizing the research in CALL, there is a need to formulate agreed-upon mechanisms and procedures for conducting such research syntheses. Accordingly, this research aims to seek answers to the following research questions:

1. How transparent and complete is the reporting of CALL meta-analyses with regard to the critical stages and procedures?
2. Are there correlations between transparency in reporting and number of citations, publication year, and word counts of the included reviews?

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

3 Method

In’nami and Koizumi (2009) provided guidelines for selecting databases for meta-analysis in applied linguistics. They first reviewed previous meta-analyses in this field, with the aim of understanding what databases were used. Initially, they located 24 journals that they believed targeted the applied linguistic audience and are more likely to publish meta-analyses in applied linguistics. The first stage of reading of the 24 journals identified 15 meta-analytic studies, of which 12 specified the databases that were used. They also compiled a list with journal coverage rates and periods of coverage in these databases. The authors finally recommended that Linguistics and Language Behavior Abstracts (LLBA), Educational Resources Information Center (ERIC), Modern Language Association (MLA), Linguistics Abstracts, and Scopus are ideal databases for retrieving meta-analyses in applied linguistics. As studies on CALL overlap considerably with applied linguistics in terms of the possible publication outlets, In’nami and Koizumi’s study provided a starting point from which we searched for possible eligible CALL meta-analyses. In the following, we detail the procedures we followed to retrieve the target studies.

3.1 Search for meta-analyses

The keywords used in previous meta-analyses were first examined, which revealed that *meta-analysis* was overwhelmingly the most frequently used keyword to identify a study as a meta-analysis. Other keywords were also observed, however, with a lower frequency, including research method, secondary research, research synthesis, quantitative research, research review, and effect size. To ensure comprehensive inclusion of meta-analyses conducted in the field of CALL, defined as “the search for and study of applications of the computer in language teaching and learning” (Levy, 1997: 1), the above keywords were used in combination with secondary-level identifiers such as technology, computer, computer-assisted instruction, computer-assisted language learning/teaching, language teaching/learning, L2, language acquisition, second/foreign languages, language skills (reading, writing, speaking, listening, pronunciation, etc.), with the aim of identifying an eligibly comprehensive sample.

We followed Aytug *et al.* (2012) and In’nami and Koizumi (2009) when selecting journals and databases to search for meta-analyses. We first reviewed the previous meta-analyses to identify the journals that published them. These journals were then searched issue by issue to retrieve more studies. The searches were conducted starting July 2014 and continued to June 2015. The searches did not exclude non-English research, but as the keywords we used were in English, it is possible that research conducted in languages other than English were filtered out. The journals we conducted manual searches on include those recommended by In’nami and Koizumi (2010) and our previous meta-analysis (Lin, 2015a, 2015b). The journals include the *Annual Review of Applied Linguistics* (ARAL), *Applied Language Learning* (ALL), *Applied Linguistics* (AL), *Applied Psycholinguistics* (AP), *Assessing Writing* (AW), *Canadian Modern Language Review* (CMLR), the *ELT Journal* (ELTJ), *Foreign Language Annals* (FLA), the *International Journal of Applied Linguistics* (IJAL), the *International Review of Applied Linguistics in Language Teaching* (IRAL), the *JALT Journal* (JALTJ), *Language Assessment Quarterly* (LAQ), *Language Learning* (LL), *Language Learning & Technology* (LLT), *Language Teaching* (LTea), *Language Teaching*

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Research (LTR), *Language Testing* (LTes), *The Modern Language Journal* (MLJ), *Reading Research Quarterly* (RRQ), the *RELC Journal* (RELCJ), *Second Language Research* (SLR), *Studies in Second Language Acquisition* (SSLA), *System*, *TESOL Quarterly* (TESOLQ), *Computers & Education* (C&E), *Educational Technology, Research & Development* (ETR&D), *Educational Technology & Society* (ETS), the *British Journal of Educational Technology* (BJET), and *Computer Assisted Language Learning* (CALL). We also conducted electronic searches on the databases recommended by In'nami and Koizumi (2010) to capture studies that might have been missed in the journal search. The databases we searched include Academic Search Premier, Comprehensive Dissertation Abstracts, ERIC, LLBA, MLA International Bibliography, Online Computer Library Center (OCLC) ProceedingsFirst, ProQuest Dissertations and Theses, PsycARTICLES, PsycINFO, ScienceDirect, and Social Sciences Citation Index (SSCI).

The keywords identified previously were also used in academic search engines such as Google Scholar to retrieve relevant studies. Furthermore, the bibliography on meta-analysis in applied linguistics compiled by Plonsky (2012) and provided on his personal website was also manually checked (http://oak.ucc.nau.edu/ldp3/bibliographies.html).

As the aim of this study was to examine the level of transparency and completeness in reporting meta-analytic procedures deemed to involve important decision-making and judgment calls in the CALL domain, studies had to meet the following criteria to be eligible for inclusion:

1. The meta-analysis had to synthesize studies on topics related to CALL.
2. The meta-analysis had to quantitatively synthesize the results of the included primary studies.
3. The meta-analysis was not reported across several sources; for a meta-analysis reported in more than one source, only one was included.

A meta-analysis was excluded if it was characterized with one of the following conditions:

1. The meta-analysis compared systematic reviews and meta-analyses (Littell *et al.*, 2008).
2. The meta-analysis aimed to describe the history and current status of the meta-analytic enterprise (Rosenthal & DiMatteo, 2001).
3. The meta-analysis proposed or recommended new procedures or stages of research synthesis (Cooper, 2003).

3.2 Codebook and transparency scale/score

Bearing in mind that the major purposes of this study were to understand the procedures and practices commonly used and followed by meta-analysts in CALL, and the degree of transparency in reporting important decision-making points in the report, we developed a codebook and a transparency measure/scale.

Previous research has revealed somewhat different stages and procedures in conducting a meta-analysis. Cooper (2003: 6), for example, proposed that a research synthesis should include the five stages of (1) problem formulation; (2) data collection, or the literature search; (3) data evaluation; (4) analysis and interpretation; and (5) presentation of results. The function that each stage serves is very similar to that of a primary study (Cooper, 1998).

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Rosenthal and DiMatteo (2001: 69–70), however, suggested five different stages of conducting a meta-analysis:

> Defin[ing] the independent and dependent variables of interest; collect[ing] the studies in a systematic way; examin[ing] the variability among the obtained effect sizes informally with graphs and charts; combin[ing] the effects using several measures of their central tendency; examin[ing] the significance level of the indices of central tendency; and using an examination of the binomial effect size display.

We reviewed previous studies that discussed meta-analytical procedures (Egger, Smith & Phillips, 1997; Wanous, Sullivan & Malinak, 1989), guidelines on how to conduct research syntheses (Plonsky, 2013), books and book chapters on meta-analysis (Lipsey & Wilson, 2001; Norris & Ortega, 2000, 2006), and meta-analytic practices and procedures from other fields (Aytug *et al.*, 2012) in designing our codebook and instruments. Specifically, we coded each meta-analysis based on features of the seven stages: Profile information, Literature search, Method, Results, Discussion, Conclusion, and Appendix, with each stage including three to 14 features to code. Table 1 presents the features and codes assigned at each stage. For each feature, we first determined whether the information was provided in the meta-analysis; we also noted down the page number, and each code was evaluated with the degree of certainty for each code, with 1 being "not so certain" and 3 "very certain".

The first author coded all of the meta-analyses included and the second and third authors served as second coders, each coding half of the studies. We first discussed the coding scheme and codebook; after reaching consistency in the meanings of the codes, we proceeded with the coding independently. The inter-coder reliability was calculated as the number of codes agreed upon by both coders divided by the number of all codes. For features that received different codes, a third coder (either the second or third author) was called upon, and discrepancies were resolved through discussion.

We modified the instrument that was developed by Aytug and his colleagues (2012) and constructed a transparency index consisting of 45 items that were each measured on a 3-point scale ("no" = 0, "partial" = 0.5, "complete" = 1). We coded whether the meta-analysts provided information on these items, irrespective of how they coded them. For example, 1 point was awarded if the meta-analyst reported the kind of statistical method that was used, irrespective of whether it was a fixed model, random-effects model, or mixed-effects model. If the meta-analysts reported the model that was employed, we assigned 1 to that item; on the contrary, 0 was awarded to the item if this information was not available, and 0.5 was awarded to items for which only partial information was provided. We then summed the scores of the 45 items and calculated a transparency score for each meta-analysis.

4 Results

In total, 15 individual meta-analyses were considered eligible for further analysis. These 15 meta-analyses were published between 2003 and 2015 and are marked with an * in the References. Of the 15 studies, eight were contributed by three authors: Lin (2014, 2015a, 2015b), Taylor (2006, 2010, 2013), and Chiu (2013) and her colleagues (Chiu, Kao & Reynolds, 2012). The topics of interest include computer-mediated communication on different aspects of learning (Lin, 2014, 2015a, 2015b; Lin, Huang & Liou, 2013); electronic/computer-mediated glosses on reading and vocabulary learning (Abraham, 2008;

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Table 1. *Forty-five items for the transparency analysis of the meta-analysis report*

	Profile information (7 items)	Code
1	Number of primary studies included in the review	Open-ended
2	Whether list of primary studies is available	Y/N
3	Total sample size of the meta-analysis	Open-ended
4	Effect size metric(s) used	Open-ended
5	Effect size averaging and weighting method(s)	Hunter–Schmidt, Hedges–Olkin, *p* values, other
6	Research synthesis method used	Random effects, fixed effect, fixed effect with subgroup analysis, other
7	Clear statement of the research question	Y/N
	Literature search (8 items)	
8	Reference and citation databases searched	
	Electronic database	Y/Searched, but specific databases are not listed/N
	Journal hand search	Y/Searched, but specific journals are not listed/N
	Reference list	Y/N
	Citation search	Y/N
	Conference programs	Y/Searched, but the list of specific conference programs is not provided/N
	Personal contacts	Y/N
	Websites/Internet	Y/Searched, but the list of websites is not provided/N
	Other	Open-ended
9	Types of studies included in the review	
	Journal articles	Y/N
	Book chapters	Y/N
	Books	Y/N
	Dissertations/theses	Y/N
	Conference abstracts	Y/N
	Government reports	Y/N
	Company reports	Y/N
	Unpublished – not further specified	Y/N
	Other	Open-ended
10	Time period covered by the search	Y/Only beginning or ending date is provided/N
11	Keywords used to enter databases and registries	Y/Some of them are provided/N
12	Date of the search	Y/N
13	Explicit list of inclusion criteria	Y/N
14	Explicit list of exclusion criteria	Y/N
15	Method of dealing with articles other than those in English	Y/N
	Method (13 items)	
16	Independent and dependent variables of interest	Y/N
17	Operational definitions of variables	Y/Some of them are provided/N
18	Number of coders used	Open-ended
19	Was the quality of the primary studies assessed?	Y/N
20	Reporting of inter-coder reliability (if more than 1 coder)	Y/N

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Table 1. *Continued*

	Method (13 items)	
21	Method of resolving disagreements (if more than 1 coder)	Y/N
22	Indicating any dependency in the data	Y/N
23	Description of how to handle data dependency	Y/N
24	Whether different study designs are combined	Y/N/Cannot tell
25	Did the study report what study features were coded?	Y/N
26	How to identify whether heterogeneity exists	Y/N
27	How to deal with heterogeneity	Y/N
28	Description of statistical formulas and/or software	Y/N
	Results (9 items)	
29	A descriptive table with the following information about the included studies	
	Study name	Y/N
	Sample size	Y/N
	Effect size(s) extracted from each study	Y/N
	Number of effect sizes contributed	Y/N
30	Tabular or graphic display of individual estimates	Y/N
31	Tabular or graphic display of overall estimate	Y/N
32	Reporting of amount of heterogeneity	Y/N
33	Rationale for the selection of moderators provided	Y/N
34	Reporting of publication bias analysis	Y/N
35	If so, types of publication bias analyses	Y/N
	Comparison of effect sizes by study source	Y/N
	Rosenthal's file-drawer fail-safe *N*	Y/N
	Trim and fill	Y/N
	Visual examination of funnel plot	Y/N
	Other	Y/N
36	Reporting of sensitivity analyses	Y/N
37	If so, types of sensitivity analyses	Open-ended
	Discussion/Conclusion (8 items)	Y/N
38	Statement of major findings	Y/N
39	General limitations	Y/N
40	Potential biases of the primary studies	Y/N
41	Consideration of alternative explanations for observed results	Y/N
42	Degree of heterogeneity was taken into account while discussing findings	Y/N
43	Generalizability of findings	Y/N
44	Implications and interpretation for theory, policy, or practice	Y/N
45	Future studies proposed	Y/N

Taylor, 2006, 2010, 2013; Yun, 2011); classroom applications of corpus analysis (Cobb & Boulton, 2015); effects of CALL on vocabulary learning (Chiu, 2013); digital game-based learning (Chiu *et al.*, 2012); strategy-oriented web-based English instruction (Chang & Lin, 2013); and general computer/technology-assisted language instruction (Grgurović, Chapelle & Shelley, 2013; Zhao, 2003). The journals that published these meta-analyses are *Language Learning & Technology* (three studies), *ReCALL* (two studies), *CALICO Journal* (four studies), the *British Journal of Educational Technology* (two studies), *Computer Assisted Language Learning* (two studies), and the *Australasian Journal of Educational Technology* (one study). Cobb and Boulton's (2015) study was published as a book chapter. The average number of primary studies per meta-analysis was 24.86, with an average sample size of 1,566 participants. Cohen's *d* was the effect size metric in 53% of the meta-analyses, whereas Hedges' *g* was the effect size of interest in 40% of the studies. Only one study used both Cohen's *d* and Hedges' *g*; 67% used Hedges and Olkin's methods. A few meta-analyses (2%) used both methods. One third (33%) of the meta-analyses in our sample used a random effects model, 6.7% used a fixed effect model, 13% used both models, and roughly 46% of the meta-analyses in our sample did not state the model used. Codings for the 15 included meta-analyses are provided as supplementary materials at https://doi.org/10.1017/S0958344017000271

In the following we report the answers to our two research questions.

4.1 How transparent and complete is the reporting of CALL meta-analyses with regard to the critical stages and procedures?

As shown in Table 2, out of a maximum score of 45, the average score of our sample is 22.27 with a standard deviation of 6.34, indicating a wide variability in the degree of transparent reporting. When closely examined, the lowest-scoring meta-analysis received a score of 13, the highest 35.5. We did not observe such a wide variability, though, in the individual sections. As shown, most of the meta-analyses provided sufficient information in the Profile section (86%) but not in the remaining sections. More precisely, except for Profile information, our sample reported more or less half of the information that was required to meet the standards. The Results (31%) and Method (37%) sections were the weakest, for which less than half of what is required to report was provided. The Literature search (55%) and Discussion/Conclusion (53%) sections were only slightly better, with most of the studies reporting slightly more than half of the information required.

Looking more closely, we found that in the Profile information section, all studies received at least 5 out of a possible 7 points, with one third of the studies receiving full points and one third missing only 1 point. This result is encouraging, as descriptive information provides the threshold information for readers to have a bird's-eye view of a meta-analysis. The scores in the Literature search section, however, warrant concern. Our sample revealed a lowest score of 2.5 and a highest of 6 out of a possible 8 points. About four studies received a score of 4 or less than 4 points, and only a third of the sample received 6 (the highest number of points in our sample). The same pattern was evident again in the Discussion/Conclusion section for which we see a lowest score of 2 and a highest of 7 out of a possible 8 points. However, in this section, about two thirds of the reports received at least 4 points.

In the following we discuss the finding of each section in more depth.

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Table 2. *Summary of transparency scores for all included studies by section*

Study/Section	Profile	Literature search	Method	Results	Discussion/ Conclusion	Total
Zhao (2003)	5	6	1	2.5	4	18.5
Taylor (2006)	7	3.5	2	2.5	4	19
Abraham (2008)	7	5.5	9	4.5	6	32
Taylor (2010)	7	2.5	3	2.5	5	20
Yun (2011)	7	5	4.5	4.5	5	26
Chiu *et al.* (2012)	5	3	5	2	2	17
Grgurović *et al.* (2013)	5	6	8	1	3	23
Chiu (2013)	5	4	1	1	2	13
Taylor (2013)	6	2.5	2	3.5	4	18
Chang & Lin (2013)	5	3	2	2.5	2	14.5
Lin *et al.* (2013)	6	5	8	4	4	27
Lin (2014)	7	6	7	2.5	4	26.5
Cobb & Boulton (2015)	6	4	2	2.5	6	20.5
Lin (2015a)	6	5	6	0.5	6	23.5
Lin (2015b)	6	5	12	5.5	7	35.5
Section average score	6.00	4.40	4.83	2.77	4.27	22.27
Maximum score	7	8	13	9	8	45
Percentage[a]	86%	55%	37%	31%	53%	49%

Note. [a]Section average score/maximum score for each section.

4.1.1 Profile information. This section consisted of seven items that asked mostly factual information of the meta-analyses, such as the number of primary studies included and a list of the studies, total sample size, the effect size metrics and average/weighting method, and the kind of synthesis method used. Generally, our sample scored high in this section, but two particular items stand out as problematic (see Table 3). Our item 6 asked about the research synthesis method used, for which nearly half of the studies ($n = 7$) did not provide an answer. The model selection is typically dependent on the results of a homogeneity test, which examines the variability of effect size distribution (e.g. whether the obtained effect size represents a common population effect, or the difference in effect size is due to sampling error only) (Li, Shintani & Ellis, 2012: 10). In meta-analysis, there are two models to analyze included studies, each with its own assumptions. A fixed-model is recommended if all included studies are identical and if the goal of the analysis is to compute a common effect size for the specified population, with no intention for the result to be generalized to other populations, as this model assumes that there is one true effect size for all included studies, and sampling error is the only reason that causes the effect size between studies to differ. On the contrary, a random effect model is recommended if we believe that within-study and between-study variability, in addition to sampling error, contribute to the variability in the effect size, and therefore the goal is not to estimate a true effect size but the mean of a distribution of effects (Berkeljon & Baldwin, 2009; Borenstein, Hedges, Higgins & Rothstein, 2009; Li *et al.*, 2012: 10). The model needs to be specified in the report because it reveals the goal of the meta-analysis and also entails totally different statistical procedures.

Another item that appears to be problematic is item 7, which assesses whether clear research questions are provided in the report. Four studies in our sample failed to meet

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Table 3. *Results for the Profile information section (7 items in percentages)*

Item	Item content	Yes	No
1	Number of primary studies included in the review	100	0
2	Whether a list of primary studies is available	86.7	13.3
3	Total sample size of the meta-analysis	86.7	13.3
4	Effect size metric(s) used	100	0
5	Effect size averaging and weighting method	100	0
6	Research synthesis method used	53.3	46.7
7	Clear statement of the research question	73.3	26.7

this requirement. When closely examined, researchers may believe it is sufficient to describe the overall goal of the study rather than provide narrow and specific research questions, as shown in Chiu (2013: E52): "This meta-analysis accounts for the overall effect of computer-mediated instruction in L2 vocabulary and specifically addresses the effects with regard to four factors: treatment duration, the educational level of participants, game-based learning and the role of teachers".

4.1.2 Literature search. This section asked for a detailed documentation of how potentially eligible studies were searched for and chosen for inclusion. This section does not judge the search strategies that were used but assesses if the listed procedures were reported. Unfortunately, as shown in Table 4, these 15 published meta-analyses did not provide satisfactory information regarding how they ended up with their final samples. Only a little more than half of the items were reported (4.4/8). Closely examined, we find that two items are missing from even the highest scoring studies in this section: date of search and method of dealing with articles other than those in English. Both items, to some degree, influenced the representativeness of the samples. Date of search, once reported, reveals information as to whether the identified studies and the total number of studies retrieved varied due to the date accessed. A systematic recording of the time of the search for eligible studies may help illuminate if there is instability in the sample. The method of dealing with non-English articles has always been an issue in meta-analysis, as excluding non-English articles may result in a biased sample not representative of meta-analyses conducted in a field. Although no researchers would explicitly state that non-English articles were excluded, it presents a challenge to search for and locate them. Once identified, the reading of the article surfaced as another obstacle to be overcome. Consensus regarding reporting still needs to be reached regarding whether non-English articles should be searched for, and if not, how this would potentially influence the representativeness of the sample and the results.

Keywords and explicit lists of exclusion criteria were another two aspects for which at least four studies in our sample did not provide information. Keywords serve as good signposts for retrieving studies that share certain characteristics; they are also useful for study replication. Without specifying the keywords used to retrieve eligible studies, readers may question the central constructs the meta-analysts have in mind when searching for eligible candidates of the study. All of the studies in our sample provided inclusion but not exclusion criteria. Researchers not specifying exclusion criteria might assert that studies that did not fit the inclusion criteria are automatically filtered out and that there is no need to

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Table 4. *Results for the Literature search section (8 items in percentages)*

Item	Item content	Yes	Partially	No
8	Reference and citation databases searched			
	Electronic database	66.6	6.6	26.6
	Manual journal search	40.0	0	60.0
	Reference list	40.0	0	60.0
	Citation search	20.0	0	80.0
	Conference programs	6.6	26.6	66.6
	Personal contacts	0	0	100
	Websites/Internet	33.3	6.7	60.0
9	Types of studies included in the review			
	Journal articles	100	0	0
	Book chapter	13.3	0	87.0
	Book	6.7	0	93.0
	Dissertations/theses	66.7	0	33.0
	Conference abstracts	26.7	0	73.0
	Government report	20.0	0	80.0
	Company report	0	0	100
	Unpublished – not further specified	60.0	0	40.0
10	Time period covered by the search	66.6	26.6	6.80
11	Keywords used to enter databases and registries	53.3	0	46.7
12	Date of the search	0	0	100
13	Explicit list of inclusion criteria	100.0	0	0
14	Explicit list of exclusion criteria	33.3	0	66.7
15	Method of dealing with articles other than those in English	0.0	0	100

Note. The percentages for each sub-item do not always add to 100%.

specify exclusion criteria; however, studies that meet the overall standard of inclusion may still need to be excluded due to technical details; for example, in Grgurović *et al*. (2013: 170), a study would still be excluded if it “did not report statistics or reported statistics that were insufficient to calculate the effect size” even though it might meet all of the inclusion criteria.

4.1.3 Method. Twelve items were assessed in the Method section (see Table 5), revealing a large gap in the scores of the studies ranging from 1 to 11. This section is also the second most poorly reported aspect of our sample in that the majority of the studies failed to report more than half of the items. The section assesses technical/statistical intent and the procedures employed by the meta-analysts; for example, it asked whether efforts were made to identify possible heterogeneity among studies, and if so, how. The same questions were asked about data dependency and how it was handled. The number of coders, the reporting of inter-coder reliability, and how coders resolved disagreement are also aspects that merit attention in this section. In meta-analysis, heterogeneity examines whether effect sizes calculated from individual primary studies are consistent. If a heterogeneity test result is significant, measures have to be taken to deal with it. In the same vein, data dependency, if not dealt with appropriately, would reduce estimates of variance, and inflate Type I errors (Borenstein *et al*., 2009; Scammacca, Roberts & Stuebing, 2014). In SLA/CALL research,

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Table 5. *Results for the Method section (13 items in percentages)*

Item	Item content	Yes	Partially	No
16	Independent and dependent variables of interest	80.0	0	20.0
17	Operational definitions of variables	33.3	6.6	60.1
18	Number of coders used	46.7	0	53.3
19	Was the quality of the primary studies assessed?	0	0	100
20	Reporting of inter-coder reliability (if more than 1 coder)	40.0	0	60.0
21	Method of resolving disagreements (if more than 1 coder)	40.0	0	60.0
22	Indicating any dependency in the data	13.3	0	86.7
23	Description of how to handle data dependency	13.3	0	86.7
24	Whether different study designs were combined	26.7	0	73.3
25	Did the study report what study features were coded?	53.3	0	46.7
26	How to identify whether heterogeneity exists	33.3	0	66.7
27	How to deal with heterogeneity	20.0	0	80.0
28	Description of statistical formula and/or software	80.0	0	20.0

however, data dependency is quite common and inevitable given the prevailing research design employed in this field.

Most empirical studies in SLA/CALL used more than one dependent variable and included more than just one treatment group to be compared with the control group. When the same participants are measured repeatedly or the same participants in the control group are compared in each comparison, the data become dependent (Scammacca *et al.*, 2014). Dependent data would seriously affect the validity of the meta-analysis results. Researchers have recommended several methods to deal with this issue (for a detailed comparison of available resolutions, refer to Scammacca *et al.*, 2014), and CALL meta-analysts should consider their overall purpose of the meta-analysis while taking into account their research questions and the nature of the data when deciding which measure to use to handle the data dependence. Three items that deal with coding also received little attention from the meta-analysts. Only a few studies reported inter-rater reliability and how disagreements between coders were resolved. Given the highly inferential and complex nature of data coding procedures involved in meta-analysis, as well as the many arbitrary decisions to be made along the way, it is advised that multiple coders be used, and inter-coder reliability in different sections and different analytical stages be reported.

4.1.4 Results. We assessed whether profile information such as sample size, extracted effect size(s), and the number of effect sizes contributed by each study was presented; we also assessed whether sensitive and publication bias analyses were conducted, and if so, how. Individual estimate and overall estimate of effect sizes calculated from individual studies and the entire sample were expected to be shown either in a table or graph. Furthermore, the amount of heterogeneity and rationales for the selection of moderators need to be reported as well. A sensitivity analysis is necessary in meta-analysis because of the alternatives available to meta-analysts. Most of the alternatives are not objective but arbitrary, which would result in inconsistencies in findings among meta-analyses on similar topics. The use of a sensitivity analysis is to detect whether there would be differences in results when the meta-analysis is repeated using alternative decisions or values instead of

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Table 6. *Results for the Results section (9 items in percentages)*

Item	Item content	Yes	No
29	A descriptive table with the following information about included studies		
	Study name	80	20
	Sample size	46.7	53.3
	Effect size(s) extracted from each study	66.7	33.3
	Number of effect sizes contributed	13.3	86.7
30	Tabular or graphic display of individual estimates	60	40
31	Tabular or graphic display of overall estimate	86.7	13.3
32	Reporting of amount of heterogeneity	13.3	86.7
33	Rationale for the selection of moderators provided	20	80
34	Reporting of publication bias analyses	33.3	66.7
35	If so, types of publication bias analyses		
	Comparison of effect sizes by study source	6.7	93.3
	Rosenthal's file-drawer fail-safe *N*	6.7	93.3
	Trim and fill	0	100
	Visual examination of funnel plot	6.7	93.3
	Other	0	100
36	Reporting of sensitivity analyses	0	100
37	If so, types of sensitivity analyses	0	100

the original ones. Heterogeneity results from the diversity in methodology in primary studies included in a meta-analysis, and can be observed if the obtained individual effect sizes are more different from each other than they should be due to chance (random error) alone (Higgins & Green, 2011). Publication bias analysis is to neutralize the effect represented by published studies when there is a consensus that the published studies are not representative of the entire population of studies done in an area (Rothstein, Sutton & Borenstein, 2006). Among the five sections, the Results section is the lowest scoring, with a highest score of 4.5 and a lowest score of 0.5 out of a maximum 9 points (see Table 6). No study in our sample conducted a sensitivity analysis, and hence no information for the type of sensitivity analysis was chosen. One third of our sample reported that they performed a publication bias analysis, but only three specified the type of publication bias analysis they employed. The amount of heterogeneity and the rationale for selecting moderators for subgroup analysis are also just as incomplete. The overall low scores in the Method and Results sections might suggest that meta-analysts in the CALL field were generally not equipped with sufficient skills or knowledge required to conduct complex meta-analysis, or were not well informed of the norm of reporting, especially for meta-analysis.

4.1.5 Discussion/Conclusion. This section asked whether the major findings and limitations of the meta-analysis were reported while taking into consideration the degree of heterogeneity and potential biases of the primary studies. We also examined whether classical components of a Conclusion section typically found in a primary study such as generalizability, implications for theory, policy, or practice, as well as recommendations for future studies were also evident in the meta-analyses. The result is not very encouraging, as shown in Table 7, with only a little over half of the items (53%) being reported. Specifically, all studies in our sample stated their major findings, and almost all provided implications for

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Table 7. *Results for the Discussion/Conclusion section (8 items in percentages)*

Item	Item content	Yes	No
38	Statement of major findings	100	0
39	General limitations	66.7	33.3
40	Potential biases of the primary studies	26.7	73.3
41	Consideration of alternative explanations for observed results	13.3	86.7
42	Degree of heterogeneity was taken into account while discussing findings	13.3	86.7
43	Generalizability of findings	26.7	73.3
44	Implications and interpretation for theory, policy, or practice	93.3	6.7
45	Future studies proposed	86.7	13.3

practice, policymaking, or theory. Recommendations for future studies were also proposed by most of the studies. Our sample is particularly weak, however, in the reporting of the potential biases of the primary studies, advancing alternative explanations for observed results, and asserting the generalizability of their findings. Meta-analysis is a more scientific way to manage large quantities of data objectively and effectively than narrative reviews, yet it still cannot refute the possibility that potential bias in primary studies can seriously affect the results. Characteristics of the study, funding sources, selective outcome reporting, and publication processes could all introduce bias into a primary study (Turner, Boutron, Hróbjartsson, Altman & Moher, 2013). Although such bias is not easy to detect, and there is no assessment regarding how it can be reduced or measured, meta-analysts need to inform readers of potential biases inherent in the primary studies, and how the results might have been influenced. Furthermore, a meta-analysis, combining studies with subtle differences in participants, study characteristics and research design, and other major aspects, tends to have greater generalizability than a single large-sample randomized primary study. The results synthesized from the combined studies across different populations and settings are generalizable to a broader range of participants, provided that no significant heterogeneity among studies is present (Heyland, n.d.). The generalizability of the results from a meta-analysis lies partly in how clearly the inclusion criteria are described, and how consistently they are followed in the study selection. Operational definitions of major factors/constructs examined in the synthesis also delimit the generalizability of the results. The meta-analysts need to discuss this for consumers of their findings, especially policymakers, who look for summarized evidence on a particular topic to guide their decisions (Garg, Hackam & Tonelli, 2008).

4.2 Are there correlations between transparency in reporting and number of citations, publication year, and word counts of the included reviews?

The number of citations for each of the 15 reviews was retrieved from Google Scholar (with a cut-off time of February 3, 2017), and word counts were derived by converting each review from a pdf file into a Word document and then running the word-count analysis. Table 8 shows the number of citations, word counts, and transparency scores for each of the included reviews. Pearson correlation analyses indicated that the number of citations was not significantly related to the transparency scores of reporting in all sections: Total,

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Table 8. *Citations, word counts, and transparency scores for each section*

Study	Citations	Word count	Total	Profile	Literature search	Method	Results	Discussion/ Conclusion
Lin, 2014	17	13,089	33.5	7	11.5	7	4	4
Grgurović *et al.*, 2013	70	14,183	28	5	11	8	1	3
Taylor, 2010	30	6,748	25.5	7	5.5	3	5	5
Taylor, 2013	7	10,432	22.5	6	4.5	2	6	4
Cobb & Boulton, 2015	19	10,371	25	6	6	2	5	6
Chiu, 2013	39	2,659	16	5	7	1	1	2
Abraham, 2008	157	13,012	42	7	13	9	7	6
Lin, 2015a	8	12,574	28	6	8	6	2	6
Zhao, 2003	271	8,204	21	5	7	1	4	4
Chang & Lin, 2013	7	5,801	18	5	5	2	4	2
Yun, 2011	50	8,336	31.5	7	8	4.5	7	5
Taylor, 2006	74	3,941	24.5	7	6.5	2	5	4
Lin *et al.*, 2013	29	10,062	34	6	9	8	7	4
Chiu *et al.*, 2012	32	1,691	20	5	6	5	2	2
Lin, 2015b	9	14,211	41	6	10	12	6	7

$r(13) = .020$, $p = .944$; Profile information section, $r(13) = –.108$, $p = .702$; Literature search, $r(13) = .223$, $p = .425$; Method, $r(13) = –.142$, $p = .614$; Results, $r(13) = .061$, $p = .828$; Discussion/Conclusion, $r(13) = .019$, $p = .945$. However, word count was found to be highly correlated to the level of transparency in reporting. In particular, word counts were correlated with total transparency score, $r(13) = .838$, $p = .002$; Literature search, $r(13) = .666$, $p = .007$; Method, $r(13) = .678$, $p = .005$; Discussion, $r(13) = .645$, $p = .009$.

Several correlation analyses between publication year and overall quality of reporting (as demonstrated in the total transparency score) and between publication year and different sections (as demonstrated in the section total score) were conducted to explore whether recent publications revealed more transparent and complete reporting than older ones. The results show that there was no significant correlation between publication year and overall transparency, $r(13) = .130$, $p = .322$, or various sections, Profile information, $r(13) = –.146$, $p = .302$; Literature search, $r(13) = .080$, $p = .388$; Method, $r(13) = .371$, $p = .086$; Results, $r(13) = –.194$, $p = .25$; Discussion and Conclusion, $r(13) = .054$, $p = .425$. However, we did find positive correlations in reporting between specific sections: Method and Literature search, $r(13) = .794$, $p = .000$. Method and Discussion/Conclusion, $r(13) = .457$, $p = .043$; Profile information and Results, $r(13) = .645$, $p = .005$; Results and Discussion/Conclusion, $r(13) = .548$, $p = .017$; Profile information and Discussion, $r(13) = .589$, $p = .010$.

5 Discussion and conclusion

Systematic reviews, taking various forms, have become increasingly important as stakeholders, practitioners, and researchers seek evidence to make important decisions regarding the kinds of investments to make in transforming a classroom into one in which technological tools play a major and significant role in enhancing both the teaching and learning of

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

second or foreign languages. With the increasing number of meta-analyses published in the literature, there has been a recognized need for a consistent framework of reporting for such work (Willis & Quigley, 2011). Our study explored the reporting quality of 15 meta-analyses in the field of CALL as assessed by adopting a transparency index constructed in previous studies.

The results generally endorsed those found in previous second-order analyses in different disciplines (e.g. Ahn, Ames & Myers, 2012; Aytug *et al.*, 2012; Plonsky & Ziegler, 2016). Of all five sections, the Method and Results are two areas that warrant much improvement. Specifically, in the Method section, operational definitions of variables, the quality of the primary studies, data dependency handling, and heterogeneity identification and analysis in the included studies need to be considered and described in appropriate depth. When reporting results, we hope to see information provided for the number of effect sizes contributed by each study, the meta-analyst's rationale for the selection of moderators, and the results of publication bias analyses and sensitivity analyses, if they are conducted. Procedures for dealing with heterogeneity among studies need to be addressed as well. The reporting of these items requires a meta-analyst's professional knowledge of both statistics and procedures. Strengthening the knowledge base in these areas might be possible via consulting published books or guides on how to conduct meta-analyses, or seeking assistance from statisticians.

As an extension of Plonsky and Ziegler (2016), the findings of the present study coincide with most of theirs; for instance, both syntheses found: (1) a wide variety in quality of reporting (as measured by transparency/rigor index); (2) Method and Results sections are areas in which practice of reporting needs the greatest improvement; (3) the quality of the primary studies was generally not assessed; and (4) justification of moderator selection was not provided. Employing a more complete transparency scale with 45 items to assess the reporting quality of 15 meta-analytic studies, out of which 10 were included in Plonsky and Ziegler, the unique contributions of our paper can be discussed from three aspects. First, our instrument, including about 2.5 times the number of items used in Plonsky and Ziegler, allowed us to conduct a more sensitive and complete assessment of current reporting practice of CALL meta-analyses. For example, in addition to assessing whether sensitivity/publication analyses were conducted, we also asked which type of sensitivity/publication analysis was used. This follow-up question is important because it echoes the basic premise of the article that the many choices made by meta-analysts can greatly affect the outcomes, and we need them to report exactly what their choices are in terms of the type of analysis decided. Second, the Profile information and Literature search sections, created as two separate independent sections for which most of the items were not included in Plonsky and Ziegler, asked for mostly factual information and a detailed documentation of the meta-analytic procedures. These two sections, although not intended to judge the way each meta-analysis was conducted, corresponded again to our above-mentioned premise that reporting of the meta-analysis needs to be as transparent as possible for cross-study comparison purposes. A third contribution of our paper lies in some of its conflicting findings with Plonsky and Ziegler. To the authors' best knowledge, Plonsky and Ziegler's paper, published in the 20th anniversary special issue of *Language Learning & Technology*, was the first second-order synthesis of meta-analyses in the CALL discipline. Most of the studies included in the sample of their study were also included in the present study as well. In the present research, we found a high percentage of studies reporting implications and

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

interpretation of their findings on theory, policy, and practice, as well as providing future study recommendations. This is not the case, though, in Plonsky and Ziegler's paper. Although the authors of both papers recognized the potential of subjectivity in applying respective instruments by employing multiple coders and establishing inter-coder reliability, the results are conflicting. The nevertheless contradictory findings have shed some light on the problem and alert us to ponder possible causes that might be attributed to using different sets of instruments and therefore diverse operationalization of the constructs behind them.

Improvement of reporting is possible for some items but not for others. For instance, although no agreed-upon position existed with regard to whether quality of the primary study should be listed as one inclusion criteria when sampling for a meta-analysis, some second-order syntheses do call for the assessment of quality in order to avoid the long-held criticism of "garbage in, garbage out". Even if this recognition is endorsed, assessments of study quality are mostly not available, and there is little or no consensus regarding the criteria for determining research quality. Two common variables related to study quality have been reported in the literature: whether subjects are randomly assigned to treatments and whether the instruments are reliable (Durlak, Weissberg & Pachan, 2010; Valentine, Cooper, Patall, Tyson & Robinson, 2010). These two pieces of information, despite their importance, were consistently missing from the primary studies, preventing meta-analysts from excluding possibly low quality studies. This in turn has flawed the meta-analytical procedures, resulting in conclusions that are not valid or are untrustworthy. Such information should be deemed as mandatory before a primary study is accepted for journal publication. Furthermore, our exploratory analysis revealed that word count is significantly related to the level of transparency and completeness of reporting. Although this finding is highly expected, journal editors might reconsider whether the word count restrictions commonly imposed on primary studies should be more flexible for systematic reviews, which require considerably more space if essential details are to be included. It was unexpected that the more recently published studies were no better reported than the earlier ones; nor were the highly cited studies more complete in their reporting. Our anticipation of a co-relationship between both variables and transparency in reporting lies in the increase in the publication of guidelines over the last decade, and the recent developments in the statistical methodology used in meta-analysis (Willis & Quigley, 2011). Such guidelines were initially developed for disciplines other than CALL, and therefore have not drawn sufficient attention from CALL researchers. This might explain why recent meta-analyses were no better than earlier ones. Furthermore, systematic reviews or meta-analyses are still in their infancy in CALL, and researchers might not be aware of the existence of such reporting practices.

5.1 Limitations and recommendations for future meta-analyses of CALL research

Using a 45-item checklist, our study set out to examine the reporting quality of meta-analyses in CALL. Reporting quality is defined as the extent of transparency and completeness of the reporting as measured by the degree of compliance with the checklist. As we indicated earlier, there have been published guidelines over the last decade; the one that we chose was originally developed for appraising meta-analyses in organizational science. There were a number of other well-developed guidelines or checklists published prior to or after the one we adopted.

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

By using different checklists, the results of our study may have been different. With this limitation in mind, we suggest that there should be a standardized checklist, as complete and comprehensive as possible, for specific disciplines so that results can be compared.

Our second limitation lies in the small sample size of the meta-analyses under review. We included only 15, which is a far smaller number than that in other fields. This small number may not represent all published meta-analyses in CALL, although we did try our best to identify all eligible studies. Furthermore, our studies included multiple meta-analyses conducted by the same authors. The reporting practice of these authors might have been over-representative of that practiced by other meta-analysts. We suggest the assessment of reporting quality of meta-analyses at regular intervals as more and more are conducted in this field.

Drawing on the findings of reporting practice examined in the present study, several recommendations for future meta-analyses of CALL research are in order. First, we recommend that meta-analysts consider provision of information with regard to the research synthesis method (i.e. fixed effect, random effect, or mixed effect model) they adopted in aggregating studies; date of search and strategy utilized to deal with studies other than in English, and whether and how they handle heterogeneity and data dependency among studies, as well as inter-coder disagreement. We also recommend that publication bias and sensitivity analysis be conducted, and that any moderator analysis for subgroup analysis be justified. Furthermore, if potential biases of the primary studies are detected, alternative explanations and the generalizability of their findings need to be reported. Second, we recommend meta-analytical studies of topics that go beyond those that were explored in the sample of the present paper. Additionally, second-order syntheses of qualitative meta-analyses or narrative reviews are another option to synthesize research in the CALL field. Our third recommendation stems from the challenge and difficulties when we applied our instruments to assess the reporting quality of the 15 studies. Recently, we have witnessed a trend of establishing reporting standards to regulate the ways manuscripts on meta-analysis should be prepared. Such an endeavor of generating agreed-upon reporting standards encourages researchers to carefully consider their design at every stage along the way when conducting meta-analyses. Many existing reporting standards, however, are designed for use in other disciplines, which might lose their sensitivity to appropriately capture the essence of research design followed by CALL researchers. We therefore recommend the development of agreed-upon and validated instruments or a set of assessment tools to improve the conduct and reporting of meta-analyses explicitly for CALL meta-analysts.

Acknowledgements

This study reported part of the data collected from a two-year grant research project supported by the National Science Council of Taiwan (grant number MOST103-2410-H-007-028-MY2). The remaining data will be presented as a book chapter: Liou and Lin (2017). We also express our gratitude to the anonymous reviewers and especially Editor Alex Boulton for their comments on an earlier draft of this paper.

Supplementary material

For supplementary materials referred to in this article, including the coding of the 15 meta-analyses, please visit https://doi.org/10.1017/S0958344017000271

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

References

*Abraham, L. B. (2008) Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, **21**(3): 199–226. https://doi.org/10.1080/09588220802090246

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

Aytug, Z. G., Rothstein, H. R., Zhou, W. and Kern, M. C. (2012) Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods*, **15**(1): 103–133. https://doi.org/10.1177/1094428111403495

Berkeljon, A. and Baldwin, S. A. (2009) An introduction to meta-analysis for psychotherapy outcome research. *Psychotherapy Research*, **19**(4–5): 511–518. https://doi.org/10.1080/10503300802621172

Borenstein, M., Hedges, L. V., Higgins, J. P. T. and Rothstein, H. R. (2009) *Introduction to meta-analysis*. Chichester, UK: Wiley. https://doi.org/10.1002/9780470743386

*Chang, M.-M. and Lin, M.-C. (2013) Strategy-oriented web-based English instruction: A meta-analysis. *Australasian Journal of Educational Technology*, **29**(2): 203–216. https://doi.org/10.14742/ajet.67

*Chiu, Y.-H. (2013) Computer-assisted second language vocabulary instruction: A meta-analysis. *British Journal of Educational Technology*, **44**(2): E52–E56. https://doi.org/10.1111/j.1467-8535.2012.01342.x

*Chiu, Y.-H., Kao, C.-W. and Reynolds, B. L. (2012) The relative effectiveness of digital game-based learning types in English as a foreign language setting: A meta-analysis. *British Journal of Educational Technology*, **43**(4): E104–E107. https://doi.org/10.1111/j.1467-8535.2012.01295.x

*Cobb, T. and Boulton, A. (2015) Classroom applications of corpus analysis. In In Biber, D. and Reppen, R. (eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press, 478–497.

Cooper, H. (1998) *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H. (2003) Editorial. *Psychological Bulletin*, **129**(1): 3–9. https://doi.org/10.1037/0033-2909.129.1.3

Cooper, H. (2007) *Evaluating and interpreting research syntheses in adult learning and literacy (NCSALL occasional paper)*. Boston, MA: National Center for the Study of Adult Learning and Literacy. http://www.worlded.org/WEIInternet/inc/common/_download_pub.cfm?id=16699&lid=3

Copas, J. and Shi, J. Q. (2000) Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*, **1**(3): 247–262. https://doi.org/10.1093/biostatistics/1.3.247

Durlak, J. A., Weissberg, R. P. and Pachan, M. (2010) A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. *American Journal of Community Psychology*, **45**(3–4): 294–309. https://doi.org/10.1007/s10464-010-9300-6

Egger, M., Smith, G. D. and Phillips, A. N. (1997) Meta-analysis: Principles and procedures. *British Medical Journal*, **315**: 1533–1537. https://doi.org/10.1136/bmj.315.7121.1533

Elvik, R. (2005) Can we trust the results of meta-analyses? A systematic approach to sensitivity analysis in meta-analyses. *Transportation Research Record: Journal of the Transportation Research Board*, **1908**: 221–229. https://doi.org/10.3141/1908-27

Garg, A. X., Hackam, D. and Tonelli, M. (2008) Systematic review and meta-analysis: When one study is just not enough. *Clinical Journal of the American Society of Nephrology*, **3**(1): 253–260. https://doi.org/10.2215/CJN.01430307

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

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

Goo, J., Granena, G., Yilmaz, Y. and Novella, M. (2015) Implicit and explicit instruction in L2 learning. In Rebuschat, P. (ed.), *Implicit and explicit learning of languages* (Vol. 48). Amsterdam: John Benjamins, 443–482. https://doi.org/10.1075/sibil.48.18goo

*Grgurović, M., Chapelle, C. A. and Shelley, M. C. (2013) A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, **25**(2): 165–198. https://doi.org/10.1017/S0958344013000013

Heyland, D. K. (n.d.) *Merits and limitations of meta-analyses*. http://scholar.googleusercontent.com/scholar?q=cache:pbTgrH5h_RUJ:scholar.google.com/+Merits+and+Limitations+of+Meta-analyses&hl=zh-TW&as_sdt=0,5

Higgins, J. P. T. and Green, S. (2011) *Cochrane handbook for systematic reviews of interventions* (Vol. 4). Chichester, UK: Wiley.

In’nami, Y. and Koizumi, R. (2009) A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, **26**(2): 219–244. https://doi.org/10.1177/0265532208101006

In’nami, Y. and Koizumi, R. (2010) Database selection guidelines for meta-analysis in applied linguistics. *TESOL Quarterly*, **44**(1): 169–184. https://doi.org/10.5054/tq.2010.215253

Levy, M. (1997) *Computer-assisted language learning: Context and conceptualization*. Oxford: Oxford University Press.

Li, S., Shintani, R. and Ellis, R. (2012) Doing meta-analysis in SLA: Practice, choices, and standards. *Contemporary Foreign Language Studies*, **384**(12): 1–17.

Liao, Y.-K. C. and Hao, Y. (2008) Large-scale studies and quantitative methods. In Voogt, J. and Knezek, G. (eds.), *International handbook of information technology in primary and secondary education* (Vol. 20). New York: Springer Science & Business Media, 1019–1035. https://doi.org/10.1007/978-0-387-73315-9_64

*Lin, H. (2014) Establishing an empirical link between computer-mediated communication (CMC) and SLA: A meta-analysis of the research. *Language Learning & Technology*, **18**(3): 120–147. http://llt.msu.edu/issues/october2014/lin.pdf

*Lin, H. (2015a) Computer-mediated communication (CMC) in L2 oral proficiency development: A meta-analysis. *ReCALL*, **27**(3): 261–287. https://doi.org/10.1017/S095834401400041X

*Lin, H. (2015b) A meta-synthesis of empirical research on the effectiveness of computer-mediated communication (CMC) on SLA. *Language Learning & Technology*, **19**(2): 85–117. http://llt.msu.edu/issues/june2015/lin.pdf

*Lin, W.-C., Huang, H.-T. and Liou, H.-C. (2013) The effects of text-based SCMC on SLA: A meta analysis. *Language Learning & Technology*, **17**(2): 123–142. http://llt.msu.edu/issues/june2013/linetal.pdf

Liou, H.-C. and Lin, H.-F. (2017) CALL meta-analyses and transparency analysis. In Chapelle, C. A. and Sauro, S. (eds.), *The handbook of technology and second language teaching and learning*. Hoboken, NJ: Wiley, 409–427. https://doi.org/10.1002/9781118914069.ch27

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

Littell, J. H., Corcoran, J. and Pillai, V. (2008) *Systematic reviews and meta-analysis*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195326543.001.0001

Norris, J. M. and Ortega, L. (2000) Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, **50**(3): 417–528. https://doi.org/10.1111/0023-8333.00136

Norris, J. M. and Ortega, L. (2006) The value and practice of research synthesis for language learning and teaching. In Norris, J. M. and Ortega, L. (eds.), *Synthesizing research on language learning and teaching*. Amsterdam: John Benjamins, 3–50. https://doi.org/10.1075/lllt.13.04nor.

Plonsky, L. (2012) Replication, meta-analysis, and generalizability. In Porte, G. (ed.), *Replication research in applied linguistics*. Cambridge: Cambridge University Press, 116–132.

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press

Plonsky, L. (2013) Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, **35**(4): 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L. and Ziegler, N. (2016) The CALL–SLA interface: Insights from a second-order synthesis. *Language Learning & Technology*, **20**(2): 17–37. http://llt.msu.edu/issues/june2016/plonskyziegler.pdf

Rosenthal, R. and DiMatteo, M. R. (2001) Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, **52**: 59–82. https://doi.org/10.1146/annurev.psych.52.1.59

Rothstein, H. R. and McDaniel, M. A. (1989) Guidelines for conducting and reporting meta-analyses. *Psychological Reports*, **65**(3): 759–770. https://doi.org/10.2466/pr0.1989.65.3.759

Rothstein, H. R., Sutton, A. J. and Borenstein, M. (eds.) (2006) *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: Wiley.

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

Schmidt, F. L. and Oh, I.-S. (2013) Methods for second order meta-analysis and illustrative applications. *Organizational Behavior and Human Decision Processes*, **121**(2): 204–218. https://doi.org/10.1016/j.obhdp.2013.03.002

Shadish, W. R. and Sweeney, R. B. (1991) Mediators and moderators in meta-analysis: There's a reason we don't let dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology*, **59**(6): 883–893. https://doi.org/10.1037/0022-006X.59.6.883

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., Henry, D. A. and Boers, M. (2009) AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology*, **62**(10): 1013–1020. https://doi.org/10.1016/j.jclinepi.2008.10.009

*Taylor, A. (2006) The effects of CALL versus traditional L1 glosses on L2 reading comprehension. *CALICO Journal*, **23**(2): 309–318.

*Taylor, A. M. (2010) CALL-based versus paper-based glosses: Is there a difference in reading comprehension? *CALICO Journal*, **27**(1): 147–160. https://doi.org/10.11139/cj.27.1.147-160

*Taylor, A. M. (2013) CALL versus paper: In which context are L1 glosses more effective? *CALICO Journal*, **30**(1): 63–81. https://doi.org/10.11139/cj.30.1.63-81

Turner, L., Boutron, I., Hróbjartsson, A., Altman, D. G. and Moher, D. (2013) The evolution of assessing bias in Cochrane systematic reviews of interventions: Celebrating methodological contributions of the Cochrane Collaboration. *Systematic Reviews*, **2**: 2–8. https://doi.org/10.1186/2046-4053-2-79

Valentine, J. C., Cooper, H., Patall, E. A., Tyson, D. and Robinson, J. C. (2010) A method for evaluating research syntheses: The quality, conclusions, and consensus of 12 syntheses of the effects of after-school programs. *Research Synthesis Methods*, **1**(1): 20–38. https://doi.org/10.1002/jrsm.3

Wanous, J. P., Sullivan, S. E. and Malinak, J. (1989) The role of judgment calls in meta-analysis. *Journal of Applied Psychology*, **74**(2): 259–264. https://doi.org/10.1037/0021-9010.74.2.259

Willis, B. H. and Quigley, M. (2011) The assessment of the quality of reporting of meta-analyses in diagnostic research: A systematic review. *BMC Medical Research Methodology*, **11**: 1–11. https://doi.org/10.1186/1471-2288-11-163

*Yun, J. (2011) The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, **24**(1): 39–58. https://doi.org/10.1080/09588221.2010.523285

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

About the authors

Huifen Lin is a Professor in the Foreign Languages and Literature Department at the National Tsing Hua University, Taiwan. She has several publications on CALL meta-analysis. Her research interests include technology-assisted language learning/teaching and quantitative research methods.

Tsuiping Chen is Associate Professor in the Applied English Department of Kun Shan University, Tainan, Taiwan. Her research focuses on using qualitative and quantitative meta-analysis methods to investigate peer feedback research conducted in the ESL/EFL writing classrooms.

Hsien-Chin Liou is a Professor in the Department of Foreign Languages and Literature at Feng Chia University, Taichung, Taiwan, and specializes in CALL and related topics such as corpus use, academic writing, and vocabulary learning. She has published numerous articles in various CALL or language learning journals, as well as book chapters.

https://doi.org/10.1017/S0958344017000271 Published online by Cambridge University Press