THE CALL–SLA INTERFACE: INSIGHTS FROM A SECOND-ORDER SYNTHESIS

Luke Plonsky, University College London
Nicole Ziegler, University of Hawai‘i at Mānoa

The relationship between computer-assisted language learning (CALL) and second language acquisition (SLA) has been studied both extensively, covering numerous subdomains, and intensively, resulting in hundreds of primary studies. It is therefore no surprise that CALL researchers, as in other areas of applied linguistics, have turned in recent years to meta-analysis as a means to synthesize quantitative results across studies. To date, nearly 30 CALL–SLA syntheses and meta-analyses have been conducted, covering topics from hypertext glosses (Yun, 2011) to synchronous computer-mediated communication (Ziegler, 2015) to game-based learning (Chiu, Kao, & Reynolds, 2012). Despite these individual contributions, the overall effects of CALL on SLA across domains have yet to be addressed. In this ‘second-order’ review, we provide a thorough account of substantive findings and a critical description and evaluation of methodological practices of CALL–SLA meta-analyses. We begin by describing the generally substantial effects of CALL on L2 learning along with an examination of different types of technology such as CALL glosses and computer-mediated communication (CMC). Results of the methodological review reveal wide variability overall and in several practices associated with rigor, transparency, and utility of meta-analytic reviews. At the close of our article, we provide empirically identified recommendations for future primary and meta-analytic research.

Keywords: Meta-Analysis, Second Language Acquisition, Research Methods

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

Received: March 29, 2015; **Accepted:** October 28, 2015; **Published:** June 1, 2016

Copyright: © Luke Plonsky & Nicole Ziegler

INTRODUCTION

The last two decades have witnessed a coming of age for research in computer-assisted language learning (CALL). The field has turned from examining questions about whether CALL is effective for language learning to how the affordances of technology might best be exploited to provide learners with optimal language learning opportunities. As methodologies and investigative techniques have evolved, increasingly incorporating design features and constructs from mainstream second language acquisition (SLA) research (e.g., Chapelle, 2009; Smith, 2004, 2009), CALL research has shifted from the fringes to the forefront of applied linguistics.

Research in CALL, which can be defined as “any process in which a learner uses a computer, and as a result, improves his or her language” (Beatty, 2010, p. 7), has also been bolstered by improvements in research techniques, enabling more accurate examinations of computer-mediated materials and applications (Chapelle, 2001), leading to appreciable changes in the nature of computer-based L2 instruction. Evidence of this domain’s maturity can also be found in the existence of several well-established and respected academic journals and conferences as well as the increased emphasis placed on technology-related research by multiple graduate programs in applied linguistics (e.g., Iowa State, University of South Florida). Yet another sign that CALL research has come of age, we would argue, is the growing interest in taking stock of the current state of the field. Scholars are addressing the evolution and development of the field from a critical and reflective perspective, demonstrated by the growing number of syntheses and meta-analyses published in recent years.

The interest in synthetic research in the realm of CALL has coincided with a broader movement taking place in applied linguistics (e.g., Norris & Ortega, 2006; Oswald & Plonsky, 2010). Indeed, a number of research syntheses and meta-analyses have been conducted on a broad range of computer-assisted activities, materials, and technologies for second and foreign language learning, often at the interface of CALL and SLA. Felix's (2005) initial search for meta-analytic research on CALL, revealed only a single study (Zhao, 2003). Since then, the field has expanded to the point where enough primary literature exists to support over 20 syntheses and meta-analyses, including multiple meta-analytic replications, focusing on the use of technology for language learning.

Statement of the Problem and Research Questions

As described in our synthesis below, studies have found generally positive effects for a wide variety of language learning outcomes based on an equally wide range of tools, including mobile assisted language learning (Burston, 2015), computer-mediated communication (Lin, 2012, 2014), glossing (Abraham, 2008; Taylor, 2006, 2009, 2013; Yun, 2011), and gaming (Chiu et al., 2012). However, although much of the research suggests positive effects, studies vary along a variety of dimensions, including modalities, outcomes, and both the size and the nature of samples they include. Although the number of studies adopting a synthetic approach has grown exponentially during the last decade, there has not yet been a comprehensive *second-order synthesis* and analysis examining the overall effectiveness of CALL. As the name implies, a second-order synthesis presents a review of reviews within a given domain (for an example of this type of review in the broader context of educational technology, see Tamim, Bernard, Borokhovski, Abrimi, & Schmid, 2011). This article represents an effort to describe the findings and critically evaluate the methods of CALL–SLA syntheses and meta-analyses as a means to better understand what this body of research has to contribute to L2 theory, research, and practice. Because of the novelty of meta-analytic research in this domain, coupled with its visibility and potential, we also describe and evaluate methodological practices of CALL–SLA meta-analyses. Though much of our discussion here is focused on works already completed, we take a forward-looking perspective whenever possible, seeking to advance the field toward more informed, explanatory, rigorous, and transparent synthetic research, by asking the following research questions:

1. Compared to face-to-face (FTF) contexts, how effective is computer-assisted language learning in promoting L2 learning?
2. To what extent have methodological best practices been followed and reported thoroughly in computer-assisted language learning research?

PART 1: SYNTHESIS OF PUBLISHED SYNTHESES AND META-ANALYSES

Method

Data Collection and Inclusion and Exclusion Criteria

The same general systematic approach used in first-order (i.e., traditional) meta-analytic research was employed here. Multiple sources were used during the study identification and retrieval phase, including key word searches (e.g., *computer-assisted language learning*, *computer-mediated communication*, *language learning technology*, *meta-analysis*, *synthesis*, *review*, etc.) of targeted L2, applied linguistics, and CALL journals; online databases (Linguistics and Language Behavior Abstracts, Education Resources Information Clearinghouse, Education Full Text, ProQuest, and Dissertation Abstracts International); and personal communication with colleagues. Nearly 30 syntheses and meta-analyses examining the impact of technology on L2 learning were identified through the retrieval process and were reviewed in order to determine their potential relevance to the goal of the current review. In cases where multiple reports of the same meta-analysis were identified (e.g., Lin, 2012, 2014), only the most recently published version was included in the final sample. In addition, some studies included *meta-analysis* or *synthesis* in the title, but did not take an approach using effect sizes (e.g., Burston, 2015). Although these studies were not eliminated, they were included only in the narrative review and were not part of the statistical analysis. For the quantitative meta-analytic aspect of the present article, studies were reviewed using the following inclusion and exclusion criteria.

A report was included if the meta-analysis

- examined the effects of any form of computer technology compared to traditional, non-CALL instructional contexts on L2 learning outcomes;
- examined the effects of any form of computer technology on L2 learning outcomes; and
- reported an average effect size across multiple studies.

A report was excluded if the meta-analysis met any of the following criteria:

- It incorporated CALL in combination with traditional contexts, but the findings were not attributed to the computer-assisted aspects of the research.
- It did not examine L2 production, performance, or development. For instance, the focus may have been learners' perceptions of CALL or the paper may have focused on description or theoretical explication. Although valuable to the field as a whole, studies using only self-report measures or questionnaire data on learners' opinions or perceptions of the efficacy of CALL were excluded on the basis that they did not empirically examine the effects of technology on L2 production, performance, or development.
- It did not examine L2 learners as the target population, but rather focused on the use of CALL as a pedagogical training tool. In these studies, learners were secondary to the instructor-centered research questions and outcomes. Although this is an important area of synchronous computer-mediated communication (SCMC) research, educators' perceptions and experiences are not the focus of the current analysis.

Of the nearly 30 studies that were initially identified during the search and retrieval process, 14 quantitative meta-analyses met the inclusion criteria, each of which are marked with * in the References. These meta-analyses encompass 408 primary studies investigating the effect of various types of technology on L2 learning outcomes in a range of settings.

Statistical Procedures and Analysis

All analyses were conducted using the Biostat meta-analysis software program Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein, 2005). Quantitative meta-analyses are based on either the fixed-effects or random-effects model depending on the assumptions of the data[1]. Because the participants and treatments used in the sample studies varied in ways that may have influenced the outcomes (i.e., treatment length and intensity), it is logical to assume that there is not a single true, underlying effect size for the entire population of effects. In addition, the goal of the current analysis is to make generalizations regarding CALL across a range of populations and scenarios, rather than drawing conclusions for identical, narrowly defined participants. Thus, for this study, a random-effects model was selected over the fixed-effects model. In addition, the syntheses included in the current study have been sampled from a distribution of effect sizes and from the published literature, further supporting the selection of a random effects model (Borenstein, Hedges, Higgins, & Rothstein, 2009).

The final set of 14 meta-analyses yielded a total of 22 effect sizes. However, including an effect size for each outcome measure can lead to data dependencies and an underestimate of the standard error and a view of the summary effect that may be biased toward the findings of studies reporting a greater number of outcomes or effects (Borenstein et al., 2009). Effect sizes were therefore combined across similar outcomes measures. For example, in order to compute an accurate effect size for overall L2 learning outcomes, combined effect sizes were calculated from all of the outcome measures within each study, resulting in a "one study, one effect size" approach for the main analysis.

In order to obtain a deeper understanding of the overall effects of CALL on SLA, thereby examining broader research questions than those of the original, individual syntheses, the weighted average effect size was calculated for all of the quantitative research examining the impact of technology on L2 learning outcomes (see Plonsky & Brown, 2015). Next, studies were categorized according to type of technology, such as glossing, computer-mediated communication (CMC), game-based and mobile-assisted language learning. These categories represent broad strands of research in the field of SLA and, through the comparison and combination of effect sizes from the primary research, provide more conclusive answers regarding the efficacy and impact of these larger categories of computer-mediated interventions. Finally, these categories were qualitatively examined according to subgroups in order to provide a more fine-grained analysis of the role of specific forms of technology in foreign and second language learning. This multi-level analytical approach aims to provide a comprehensive global and local perspective regarding the efficacy of technology for SLA, informing researchers, educators, and policy makers about the current state of the field.

Results and Discussion

The Overall Impact of CALL on L2 Learning Outcomes

A total of 22 meta-analytic effect sizes were extracted from the 14 individual meta-analyses, covering 408 primary studies. In all cases, meta-analytic results were expressed using an effect size (*ES*) that expresses the mean difference between group scores (e.g., computer vs. non-computer treatment) in standard deviation units (typically *d* but sometimes *g*; see Plonsky & Oswald, 2014). Table 1 provides information on each individual study and its effect size(s).

Although only nine of the meta-analyses reported a total sample size, the overall number of learners included in this second-order analysis remains large, totaling over 14,000. In investigating the complete set of meta-analyses, it is clear that a wide range of technologies and computer-assisted approaches have been examined, with each study focusing on specific issues and characteristics of CALL. For example, some analyses have taken a more global perspective, examining the impact of CALL on general L2 learning outcomes (Grgurović, Chapelle, & Shelley, 2013, *ES* = .24), while others have elected to focus on the impact of CALL on more specific measures, such as Chiu's (2013) analysis of the efficacy of CALL for L2 vocabulary development (*ES* = .75). The meta-analysis conducted by Lin (2014) focused on the efficacy of CMC on overall L2 learning outcomes (*ES* = .44), while Ziegler (2015) concentrated on the effects of SCMC only on L2 learning outcomes (*ES* = .13). Taylor (2006) examined the effects of CALL L1 glosses on learners' reading comprehension (*ES* = 1.09), whereas Yun investigated the impact of text + visual hypertext glossing (*ES* = .46). The effect sizes of these studies vary in magnitude, and, taken together, provide a more global and accurate estimate of the overall impact of CALL as it has been examined in traditional and technology-enhanced educational contexts.

Table 1. *CALL–SLA Meta-analyses, Relevant Features, Effect Sizes*

Meta-analysis	Technology	Learning outcomes	*ES*
Abraham (2008)	CALL glossing	Reading comprehension	0.73
		Immediate vocabulary learning	1.40
		Delayed vocabulary learning	1.25
Chiu et al. (2012)	Game-based learning	General	0.53
Chiu (2013)	CALL	Vocabulary	0.75
Grgurović et al. (2013)	CALL	General	0.24
Lee et al. (2015)	CALL	Pronunciation	0.76*
Li (2010)	CALL	General	0.73*
Lin (2014)	CMC	General	0.44
Lin et al. (2013)	SCMC	General	0.33
Taylor (2006)	CALL glossing	Reading comprehension	1.09
Taylor (2009)	CALL glossing	Reading comprehension	0.92
Taylor (2013)	CALL glossing	Reading comprehension	1.44
Yun (2011)	Hypertext glossing	Vocabulary	0.46
Zhao (2003)	CALL	General	1.12
Ziegler (2015)	SCMC	General	0.13

*Note. *Extracted from a larger, non-CALL-specific meta-analysis*

According to the random-effects model, the overall weighted mean effect size of the 10 combined effect sizes for relative effects (*ES* = .512) indicates that there is a significant difference (*p* < .001) in learning outcomes when learners participate in CALL vs. traditional educational contexts. (We use *relative effect* to refer to the difference between types of treatments such as CALL-based and non-CALL-based instruction; for further details on such designs and effects, see Ziegler, 2015. We use *absolute effect* when comparing treatment and control group scores.) One view of this result would be that it represents a somewhat small difference between conditions (Plonsky & Oswald, 2014). This effect size also suggests that L2 learning in contexts using some form of technology provides a considerable advantage over traditional, non-technology based contexts in facilitating L2 learning outcomes. A relative effect size of .512 is associated with a U_3 value of 69.57%, where U_3 represents the degree of non-overlap between two groups (Cohen, 1988; Lipsey et al., 2012), suggesting that nearly 70% of groups participating in CALL contexts have significantly better learning outcomes than those learners in non-CALL contexts, providing support for the use of technology in the L2 classroom. Analyses also examined absolute effects, with findings indicating a much larger weighted mean effect size (*ES* = .84, U_3 = 79.95%) for the 8 combined effect sizes taken from the original meta-analyses, suggesting that CALL leads to significant improvements (*p* < .001) in L2 learner outcomes.

Positive effects of technology on language learning are also evident in the non-meta-analytic research syntheses that have appeared in the last decade or so. For example, the recent synthesis by Golonka, Bowles, Frank, Richardson, and Freynik (2014) examined over 350 studies encompassing research on a wide range of technologies, including classroom-based and independent learning tools aimed at improving a variety of L2 skills, such as pronunciation, vocabulary, and reading comprehension. After classifying the primary literature according to whether it focused on enhanced input and comprehension, output and interaction, feedback, affect and motivation, metacognition, or metalinguistic knowledge, results revealed the strongest support for the use of technology in automatic speech recognition (ASR) programs and SCMC. In other words, although Golonka et al. concluded that the evidence for the overall efficacy of technology may be limited, strong empirical evidence exists for technology-enhanced pronunciation practice and instruction over traditional instruction. However, results from Lee, Jang, and Plonsky (2015), which indicate a larger effect size for traditional methods of instruction over the use of technology (*ES* = .96 and *ES* = .76, respectively), seem to contradict these findings. Nevertheless, because technology continues to develop at a rapid rate, the potential for improved ASR and pronunciation software continues to grow, highlighting the prospective role of technology for independent and classroom based learning and emphasizing the need for continued research in this area. Furthermore, Golonka et al. (2014) concluded that chat programs led to increased quantity and quality in L2 production, a finding that has been supported by numerous other meta-analyses (e.g., Lin, Huang, & Liou, 2013; Ziegler, 2015) and syntheses (e.g., Lai & Li, 2011; Sauro, 2011). Results also found moderate support for the benefits of technology on speaking, reading comprehension, vocabulary, grammar, fluency, and promoting noticing and focusing on form.

Similarly, Zhao's (2003) narrative review demonstrated a consistent pattern of positive results across more than 150 studies from nearly all areas of language education, concluding that technology was not only useful for L2 development, but also specifically for the enhancement of input, the provision of feedback, and for supporting authentic communication. Zhao (2003) also points out, however, a number of methodological concerns, such as the lack of random sampling or the prevalence of small sample sizes, that may present threats to the internal validity of studies in this area. Similar methodological concerns were also raised by Felix (2005, 2008). Although positive effects were noted in her surveys on the effectiveness of CALL from 1981 to 2005, she also found a general lack of high methodological standards, raising concerns about the quality of the primary literature included in her analysis. The synthesis by Liu, Moore, Graham, and Lee (2002) of CALL research from 1990 to 2000 provided evidence for the efficacy of technology, although the researchers also highlighted the need for more rigorous research designs. Echoing early calls for more rigorous research designs (e.g., Chapelle, 2001), recent primary studies have also drawn attention to the need for improved empirical designs, particularly regarding internal reliability and validity (Cerezo, Baralt, Suh, & Leow, 2014) and better reporting practices (Ziegler, 2016a).

Type of Technology

As a means of organizing the remainder of this review of substantive findings, we have grouped studies according to type of technology, focusing on three specific areas that have inspired a substantial number of primary and secondary investigations in the field: (a) glossing, (b) CMC, and (c) game-based language learning and mobile-assisted language learning (MALL).

Glossing

Studies examining the efficacy of CALL glosses on reading comprehension and vocabulary development were one of the most well-represented research areas in the current sample, encompassing nearly half of the total number of meta-analyses (*k* = 5). Multimedia glosses, which can be defined as online tools that allow immediate or near-immediate access to the definitions of L2 words (Taylor, 2009), have been shown to largely improve learners' vocabulary. Results in this area generally demonstrate that both L1 and L2 CALL glossing can be beneficial to learners' vocabulary development (Taylor, 2006, 2009, 2013). For example, Taylor (2009) examined the absolute and comparative effects of CALL and non-CALL glosses on learners' reading comprehension in a study of 32 primary articles. The findings indicated a generally large effect for CALL glosses (*ES* = .92) and a somewhat smaller one for non-CALL glosses (*ES* = .43). Analyses also revealed a significant difference between CALL and traditional paper-based glosses: approximately 64% of learners using electronic glosses demonstrated greater performance than those without such glosses. Abraham (2008) provides additional support for the efficacy of online glossing, with findings indicating an average effect size of 0.73 in 11 studies, showing that learners with access to computer-mediated text glosses consistently and substantially outperform learners without such resources on measures of L2 reading comprehension. Measures of incidental vocabulary learning were also found to have similar results, with Abraham's findings indicating a large positive effect for glossing on immediate (*ES* = 1.40) and delayed (*ES* = 1.25) posttests of vocabulary knowledge. These results suggest not only that CALL glossing results in immediate positive benefits, but also that the effects are sustained over time.

Research has also demonstrated that CALL glossing in learners' L1 may be more effective than paper-based glossing, and as Taylor (2006) points out, L1 glosses may be more readily available and widely used than other types of comprehension aids, including L2 glosses, picture glosses, or audio glosses. Taylor's original (2006) meta-analysis examined 18 primary studies, producing a weighted average effect size of 1.09 for CALL glosses. Taylor (2013) provides an update to his earlier meta-analysis, expanding the number of included primary studies to 28 and nearly doubling the number of learners included in the sample size from 875 (2006) to 1458 (2013). A large effect (*ES* = 1.44) was found for L1 CALL glossing, providing support for Taylor's earlier findings. In addition, there was a significant difference between CALL and paper-based glosses (*ES* = .47), demonstrating the substantial and positive effects of CALL glosses on learners' L2 reading comprehension.

Building on many of the same findings, Yun (2011) sought to compare the benefits of hypertext glosses incorporating both text and visual input with text-only glosses on the development of L2 vocabulary knowledge. Because they can incorporate visual elements, hypertext glosses provide learners with multimodal learning opportunities that are more readily adaptable to learners' individual learning styles and strategies. Hypertext glosses may also be more flexible in terms of relevance to learners' needs and interests. Based on 10 primary studies, which yielded 35 effect sizes, results indicated a medium effect (*ES* = .46) for the use of both text and visual hypertext glosses for L2 vocabulary development. Proficiency was identified as a significant moderating variable, with beginning level learners seeming to benefit most from multi-modal hypertexts (*ES* = .70), while intermediate learners benefited the least (*ES* = .23). These results stand in contrast to those of Abraham (2008), whose findings indicated a large effect for intermediate learners on immediate (*ES* = 1.63) and delayed (*ES* = 1.43) measures of vocabulary. Although beginning level learners also experienced substantial effects on both immediate (*ES* = 1.00) and delayed (*ES* = .57) assessments, average effect sizes were smaller than those for higher proficiency learners.

Taken together, these studies provide strong support for the use of CALL glossing as a means to enhance and improve learners' vocabulary and reading comprehension. According to the random effects model, the overall weighted mean effect size of the four combined effect sizes (*ES* = .60) indicates that there is a significant and substantial difference ($p < .001$) on L2 learning outcomes between learners using CALL glosses compared to those using non-CALL glosses, with absolute effects demonstrating a generally large effect size (*ES* = .93) for CALL glossing. A somewhat smaller effect was found for CALL glosses (*ES* = .55, U_3 = 70.88%) for reading, while a larger effect was found for the use of CALL glosses for vocabulary development (*ES* = 1.33, U_3 = 90.82%) when compared to traditional, paper-based glosses. These results should be interpreted with caution, however, as the small sample size for both of these comparisons may have impacted the precision of the contrast. Furthermore, Yun points out that in the case of hypertext glossing, learning preference, rather than proficiency, may play an important role due to the multiple forms of input that learners may be exposed to. However, because Yun (2011) is the only meta-analysis to examine the relative effects of both visual and text versus text-only, additional studies examining potentially moderating variables, such as learning strategies, are needed to build a body of research large enough to use secondary and meta-analytic techniques to understand such relationships.

Computer-mediated Communication

CMC made up the next most common area of focus within the set of meta-analyses and syntheses. Interest in this area is likely due in part to the many similarities between CMC and FTF communication, in addition to a hypothesized ability to enhance the manner in which students produce, comprehend, process, and exchange information. Sauro's (2011) synthesis of 97 studies examined the relationship between SCMC and learners' grammatical, sociolinguistic, strategic, and discourse competences, finding that the majority of studies focused on the development of grammar (*k* = 48) and were theoretically grounded in the interaction approach to SLA (Gass & Mackey, 2015). Overall, Sauro concluded that SCMC was a productive context for investigating L2 outcomes, with many of the findings indicating a range of benefits for the use of CMC for SLA. Sauro also pointed out, however, the lack of research in K–12 contexts and the need for more research on the learning of a wider range of typographically different languages, points that were first raised by Liu et al. (2002) nearly ten years prior. This continued call for expanding the studied populations underscores the need for CALL researchers to give greater consideration to generalizability.

Lai and Li (2011) took a similarly qualitative approach, albeit with a more focused theoretical orientation. Their review examined the benefits and challenges to integrating technology into the task-based classroom, finding that although a number of empirical studies provide encouragement for the use of computer-mediated learning, there are substantial differences between task-based language teaching (TBLT) in FTF and technology-mediated environments (see also Plonsky & Kim, 2016; Ziegler, 2016b). Some findings have demonstrated greater performance in FTF interaction over computer-mediated environments (e.g., de la Fuente, 2003), suggesting that more research is needed to understand the relative efficacy of technology for L2 learning outcomes.

Turning to the quantitative meta-analyses of CALL, results generally support (with small to medium effects) the efficacy of CMC for L2 learning outcomes. For instance, Lin's (2014) meta-analysis of 59 primary studies revealed a medium relative effect (*ES* = .44) for CMC when compared with traditional FTF educational contexts, with a larger effect size (*ES* = .61) for asynchronous computer-mediated communication (ACMC) and a smaller effect size (*ES* = .31) for SCMC, suggesting that ACMC may be more effective for learning outcomes than contexts using only synchronous or ACMC+SCMC blended classrooms (*ES* = .46). Lin also identified proficiency and research setting as significant moderating variables, with findings demonstrating greater performance by beginning-level learners using CMC than advanced- and intermediate-level learners, as well as significant advantages for the use of CMC in foreign language (FL) settings over second language (SL) settings. In addition, interlocutor type was also found to moderate the effects of CMC, demonstrating an advantage for learners interacting with peers rather than with teachers (*ES* = .50) and a small to medium effect size for native speakers (*ES* = .49). Interestingly, no significant differences emerged across CMC modality, providing educators with encouraging information regarding the efficacy of both synchronous and asynchronous tools. In other words, the non-significant results seem to suggest that learners would be provided with similar learning opportunities regardless of the mode of communication, giving instructors a greater sense of flexibility when designing online or blended courses.

Taking a subset of this sample (*k*=10), Lin et al. (2013) focused on the impact of text-based SCMC on SLA, finding a small relative effect for SCMC (*ES* = .33) over FTF. Similar results to Lin (2014) were obtained, with the moderating variables of interlocutor and setting emerging as significant. Findings indicate a small effect size for learners interacting with other learners (*ES* = .28) and for learners participating in an SCMC context in a FL setting (*ES* = .28). These results provide additional support for the importance of the set of moderating variables identified in Lin (2014). Ziegler's (2015) meta-analysis of SCMC used a different set of inclusion and exclusion criteria, focusing only on those studies grounded within the interaction approach to SLA, and including text, video, and audio modalities. Results indicated a small advantage for SCMC both overall (*ES* = .13), as well as in FL (*ES* = .08) and SL (*ES* = .33) contexts. In contrast to the results of Lin et al. (2013), Ziegler's (2015) results suggest that mode of communication may have a greater impact on learners' performance in SL contexts. Findings also indicated a small relative effect size in favor of SCMC for both native (*ES* = .09) and non-native (*ES* = .11) speaking interlocutors, although neither comparison was statistically significant, suggesting that the observed benefits of SCMC for different interlocutors may not be stable. Given that these findings across all three meta-analyses are not conclusive, more studies are clearly needed in order to gain a better understanding of the role that type of interlocutor plays in the treatment effects of interaction in SCMC contexts.

Taken together, these meta-analyses encompass a total of 83 primary research reports with 4747 total students. Results from a random effects model indicate a small relative effect size (*ES* = .33) for CMC, suggesting that learning outcomes in CMC were .33 *SD* units larger than for traditional, FTF contexts. A relative effect size of .33 is associated with a U_3 value of 62.93%, meaning that 63% of learners in CMC contexts had learning outcome scores above the mean of those in FTF environments. In other words, these findings suggest that CMC contexts provide only a slight advantage—one that might be experienced by only a small number of students in a given L2 class, for example—over FTF contexts in facilitating L2 learning outcomes.

Game-based and Mobile-assisted Language Learning

Interest in the use of video games and mobile applications to teach a second language has grown rapidly (Aldrich, 2009; Burston, 2015; Thorne, Black, & Sykes, 2009). This now-popular modality provides interesting benefits, such as lowered affective filters, multiple practice opportunities in different contexts, and concentrated samples of targeted input (Reinders, 2012). Overall, results are promising and suggest that gaming provides a supportive environment for SLA. For instance, Suh, Kim, and Kim (2010) found that participation in synchronous game-based interaction led to higher scores in listening, reading, and writing, while the results of Rankin, Gold, and Gooch (2006) indicated improved vocabulary knowledge and target language output following synchronous game-play.

However, despite the growing body of research examining gaming and MALL, only two synthetic studies in this area were identified during the search and retrieval phase of the current study. Chiu et al. (2012) examined 14 primary studies encompassing 1116 students. The study found a medium positive effect (*ES* = .67) for digital game-based learning, which included both drill-based practice games and meaningful, action-based games. Meaningful and engaging games resulted in a larger effect size (*ES* = .84) than drill and practice games (*ES* = .41). This finding would suggest, as found outside of technology-based SLA research, that approaches grounded in interaction and meaningful communication are more effective for language learning (e.g., Mackey & Goo, 2007).

Burston's (2015) analysis of 19 primary reports addressed the implementation of MALL projects, focusing on the learning outcomes associated with mobile-based language applications. However, as Burston points out, the majority of the studies in the sample lacks rigorous research methodology, preventing a reliable and valid analysis using effect sizes. Taking a descriptive approach, Burston concluded that in approximately 80% of the studies examining the use of MALL for L2 learning, positive outcomes were reported. However, this finding should be interpreted cautiously, given not only the lack of robust research methodology in the primary research, but also the failure of these studies to examine the various moderating variables that may have impacted the results, providing information on not just *whether*, but also *how* the technology may have impacted learners' development.

Overall, despite the increase in studies examining gaming and mobile-assisted language learning applications, more primary research is needed to better understand the potential contributions of game-based interaction on L2 development. These tools, including multi-player games, virtual worlds, online collaborations, social media tools, and other user-driven technologies, provide new and exciting contexts for L2 instruction and research.

Summary of Part 1

In sum, statistical results indicate a small relative effect (*ES* = .512) for the use of technology in L2 learning, suggesting that learners participating in CALL contexts may have better learning outcomes than those in traditional educational contexts. Absolute effects also provide strong evidence for the efficacy of CALL (*ES* = .84). Results demonstrate positive benefits for CALL glossing (*ES* = .60) and CMC (*ES* = .33) relative to non-CALL contexts, although more research is needed to understand the full impact of game-based or mobile-assisted language learning. Syntheses also seem to support a general trend towards a developmental advantage for CALL (e.g., Lai & Li, 2011; Sauro, 2011; Zhao, 2003; although see Lee et al., 2015, for an exception).

PART 2: METHODOLOGICAL SYNTHESIS OF PUBLISHED META-ANALYSES

Complementary to the substantive focus in Part 1 of this paper, Part 2 presents a systematic review of meta-analytic practices employed in CALL research. This phase of our study is predicated on the notion that the quality of meta-analytic research in this area is unknown and, at the same time, critical given the visibility and high citation rate typically found for meta-analytic research (Cooper & Hedges, 2009). In parallel to Plonsky's (2013) work evaluating primary research, we define meta-analytic study quality as the combination of (a) contextually appropriate, methodological rigor and (b) transparency in the research process and the results they produce. The present study therefore seeks to describe and evaluate research and reporting practices in CALL meta-analyses (for a comparable review across various substantive domains, see Larson-Hall & Plonsky, 2015). Despite the retrospective nature of the study, we also seek to inform and improve the methodological practices in future applications of meta-analysis. In this sense, as with recent reviews of primary L2 research practices (e.g., Plonsky & Gass, 2011), our study is as much concerned with the future of meta-analysis in CALL as it is with its past. Based on the results of our review, we therefore provide a number of suggestions for improving the use of meta-analysis at the interface of CALL and SLA. With these motivations in mind, Part 2 of this study posed the following research question: To what extent have CALL–SLA meta-analyses been carried out according to standards of rigor and transparency?

Method

Sample

This part of the study began with the same set of research syntheses and meta-analyses identified in Part 1. Before data collection began, however, the sample was culled to include only meta-analyses, defined as secondary studies that aggregated effect sizes across primary studies. Also removed from the sample were reports that were later expanded upon in another form (e.g., Lin, 2012, 2014; Ziegler, 2013, 2015), with the most recent version remaining, as well as those conducted as part of a larger, non-tech-specific meta-analysis (Lee et al., 2015; Li, 2010). The final sample consisted of 10 meta-analyses (marked with ^ in the References).

Data Collection and Analysis

The data for Part 2 of this study were collected using a revised version of an instrument proposed by Plonsky (2012) for evaluating the rigor and transparency of meta-analyses in applied linguistics. The 17-item instrument (see Table 2), based in part on several previously developed instruments designed for similar purposes (Journal Article Reporting Standards Working Group, 2008; Moher, Liberati, Terzlaff, & Altman, 2009; Shea et al., 2007; see also Higgins et al., 2013) includes three sections, following the structure of meta-analytic reports: Introduction and Literature Review (items 1–3), Methods (items 4–11), and Results and Discussion (items 12–17). Each item addresses one aspect of meta-analytic rigor or transparency measured on a 3-point scale (*no* = 0, *somewhat* = 1, *yes* = 2). (For further information on the original design and development of the instrument, see Plonsky, 2012.)

We are sensitive to the potential for subjectivity in applying this instrument. Ideally, and in order to accurately and reliably determine methodological appropriateness, those implementing this coding scheme would be familiar with both the domain in question and with meta-analytic methods. With these concerns in mind, the coding process began with the first author and a research assistant, a PhD student with experience in synthetic data collection, both coding two studies. The results were then compared and item definitions were revisited and clarified as needed. The research assistant and the first author then independently coded the entire sample. Of the total of 170 items (17 items, 10 studies), the two coders disagreed on only 14 (92% agreement; κ = .87). In addition to these relatively high levels of consistency, agreement was also even across the instrument, with no individual item yielding more than two discrepancies.

Once each study was coded, we tallied scores for each individual study in the sample, as well as across items and sections. By doing so, we hoped to gain a fuller understanding of the strengths and areas needing improvement in CALL–SLA meta-analyses. (We did not conduct this study with the intention to draw attention to individual researchers or their meta-analyses. For this reason, we have replaced the names of the authors with numerical identifiers.) These scores were then explored with respect to guides for carrying out and reporting on meta-analyses from the broader synthetic literature as well as from within applied linguistics (e.g., Borenstein et al., 2009; Cooper, 2010; Plonsky & Oswald, 2015).

Table 2. *Instrument for Assessing CALL–SLA Meta-analyses*

Introduction and Literature review
1. Does the review address a focused and clearly defined question?
2. Are all (potential) moderator variables identified a priori and explained/motivated sufficiently?
3. Is the relevance of the study, theoretical and/or practical, presented?
Method
4. Was the search for relevant primary research reasonably exhaustive?
5. Are the inclusion/exclusion criteria sufficiently explicit and appropriate?
6. Was the potential of publication bias addressed adequately?
7. Was interrater reliability for coding measured and adequate?
8. Was the quality of primary studies assessed?
9. Were effect sizes from the same samples/studies dealt with appropriately?
10. Were effect sizes weighted appropriately?
11. Are all items on the coding sheet justified and available for inspection?
Results and Discussion
12. Are overall findings presented along with appropriate consideration of error (e.g., *CIs*)?
13. Does the review add new knowledge about the constructs of interest?
14. Are the results interpreted and contextualized appropriately?
15. Are the findings discussed in relation to the particular theory or model(s) being tested?
16. Are practical implications discussed?
17. Are the findings used to provide substantive and methodological suggestions for future research?

Results and Discussion

Looking across the sample, we can see that individual study scores vary substantially, from 13 to 33 (maximum possible = 34; see Figure 1). We should note, though, that two of the three lowest scoring meta-analyses were published in a "short-report" format, which likely limited the authors' ability to describe in full their studies' methods, results, and so forth. The ratings received by the studies in the sample also appear relatively evenly distributed across the scale, which can be taken as a sign that the instrument is sensitive to differences in meta-analytic quality. Unfortunately, wide variability in scores also indicates decidedly mixed quality in CALL–SLA meta-analyses conducted to date.

Looking more closely at the results, it is encouraging to see that the scores on Section 1 (Literature Review) are generally quite strong, with 7 out of 10 studies receiving 5 or 6 out of 6 possible points. However, wide variability is evident again in the two other sub-scales (Method and Results and Discussion, respectively). The scores in these sections range from 2 to 16 (maximum possible = 16) and from 6 to 11 (maximum possible = 12), respectively. These findings are somewhat concerning. Published meta-analyses appear to be lacking in rigor and transparency. We will now explore these results in a bit more depth by examining the meta-analytic practices of each section in turn.

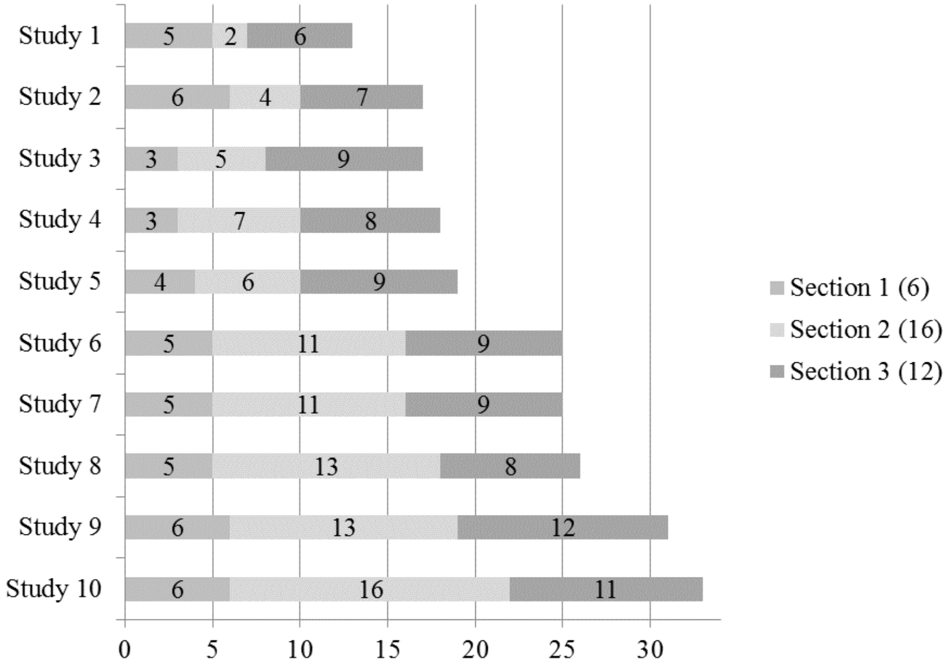

Figure 1. This graph shows the section scores for CALL–SLA meta-analyses. The maximum possible scores are found in parentheses in the legend.

The literature reviews in our sample appear to have generally met the standards tested by our instrument. Items 1 and 3 addressed whether the meta-analyses in the sample addressed a focused and clearly defined question and whether the theoretical or practical relevance of the study was made clear. Overall the sample scored high on these two items. One particular item stands out as problematic, though: the justification of moderator variables. Moderator variables help us to explain systematic variability in overall meta-analytic effects. Examples of such variables in this set of meta-analyses, as described in Part 1 of this study, might include learner proficiency, setting, length of treatment, and so forth. In meta-analysis, as in any type of empirical research, it is critical to motivate the inclusion of each moderator variable being examined. Further, that some studies fell short in this area highlights the importance of grounding the selection and investigation of moderator variables in the relevant theoretical foundations and empirical research, thereby ensuring that future synthetic research is able to provide answers to research questions that individual primary studies may be unable to answer on their own.

The results with respect to the Method section of meta-analytic reports vary considerably and fall far below what we might hope to find. Scores on this section of the instrument were actually the lowest of the three: The average score across the sample was only 8.8, or 55% of the total possible points. A meta-analysis' score on this section of the instrument is also an extremely strong predictor of its overall score. That is, the correlation between scores on Section 2 and overall scores in this study was r = .98. By comparison, the correlations for the Literature Review and Results and Discussion sections were r = .55 and r = .73, respectively. Part of the strength of this correlation can be attributed to the fact that the Method section is the longest sub-scale in the instrument (8 out of 17 total items). Even still, however, a correlation this strong would indicate that methodological rigor and transparency in different sections of meta-analytic reports are related, with the Method section being the best predictor of overall quality.

The results for this part of the study reveal two strengths in meta-analytic Method sections. The first is the provision of a sufficiently clear and explicit set of inclusion criteria used to determine which studies are eligible to be included in the meta-analysis. The search techniques employed to locate studies that fit the inclusion criteria represents the other bright spot in this section. These two steps are particularly important and can bear a significant influence on meta-analytic findings (Plonsky & Oswald, 2015). They are, however, just two of a series of decisions involved in conducting and reporting the meta-analytic process (Norris & Ortega, 2006; Oswald & Plonsky, 2010).

Findings related to other criteria that we coded for were much less favorable. For example, the studies in our sample paid very little attention to methodological quality in their respective domains. As both authors of this paper have argued and exemplified (e.g., Plonsky, 2013; Ziegler, 2013), it is critical that meta-analyses explore the substance (the what) as well as the methods (the how) in primary research. Failing to do so is a missed opportunity to gain a better understanding of the domain in question and, more importantly, to evaluate its research practices and provide empirically-grounded recommendations for improvements to design, analyses, reporting practices, and so forth. The need is perhaps particularly urgent in the realm of CALL research; indeed, a number of scholars have raised concerns regarding the methodological rigor in the domain. For example, following a review of methodological practices in CALL, Macaro, Handley, and Walter (2012) observed that researchers "often claimed that progress had been measured even though no pre-testing had apparently been carried out" (p. 26). This and other reviews also reported a lack of information regarding instrumentation such as piloting, power analysis, effect size reporting, and reliability (Ziegler, 2015, 2016a; see also Derrick, in press). Similar reviews have also expressed concerns over the narrow range of demographics sampled in CALL research, which may restrict the generalizability of results (e.g., Handley, 2014). As a way forward, several editorial and methodological papers have made a case for replication research in CALL (Chun, 2012; Plonsky, 2015; Porte, 2013; Smith & Schulze, 2013). Nevertheless, it will still be the responsibility of meta-analysts to describe and evaluate the research and reporting practices of the domains they review.

The majority of the remaining items corresponding to Section 2 of the instrument were only slightly better. The total number of points (out of a possible 20) for the item probing whether and how meta-analysts had addressed a potential publication bias in their studies was just 10. (See similar concerns expressed in the broader field of educational technology in Bernard, Borokhovski, Schmid, & Tamim, 2014). This is concerning given the inflating effects caused by the preference of authors and reviewers in favor of statistically significant findings (see Plonsky & Oswald, 2014). Other major and recurrent problems in this section include interrater reliability of meta-analytic coding (10/20), the handling of potential data dependencies (9/20), effect size weighting (10/20), and whether or not data collection instruments were justified and made available (10/20). These issues reflect a general lack of transparency in the methods of published CALL–SLA meta-analyses, highlighting the need for improved reporting practices in not just the primary literature (as pointed out by a number of the analyses included in the sample; see also Larson-Hall & Plonsky, 2015), but also in our sample of syntheses.

Scores on the Results and Discussion section of our instrument were higher than for the Method section, but still less than ideal. The average score across studies in this section was 8.8 or 73% of total possible points. In all or nearly all cases, the studies in our sample reported both (a) overall findings (a meta-analytic effect) along with an indication of the error around those findings and (b) findings that added new knowledge to the domain in question.

Somewhat lower results were obtained across the sample with respect to interpreting results. It is not sufficient for a meta-analysis to simply report quantitative findings and to label them as generically small, medium, or large. The size and importance of effect sizes, whether at the primary or meta-analytic level, must be interpreted in light of a number of considerations. One such consideration might be field-specific benchmarks such those proposed by Plonsky and Oswald (2014). But meta-analysts must also consider other factors such as the degree of experimental manipulation, the theoretical maturity of the domain, and the presence of publication bias. In the case of CALL–SLA meta-analyses and primary studies, the gains resulting from the use of any particular technological tool must be weighed against its cost, availability, and user-friendliness (Macaro et al., 2012). The inflating effects of publication bias must also be considered in fields like CALL where researchers may have predispositions regarding the effectiveness and utility of technological tools. Time is yet another variable worthy of consideration in interpreting meta-analytic effects. That is, there is good reason to expect that the effect sizes in a given domain will change over time due to methodological innovations, novel tools, and theoretical refinement. Such a pattern is very likely to be found in CALL, where technological tools are introduced and developed frequently and where individual teachers', researchers', and learners' familiarity with technology is constantly advancing. Unfortunately, none of the studies in our sample appear to have examined changes taking place over time in CALL effects.

Two other problematic trends in Results and Discussion sections of CALL meta-analyses can be seen in our findings. The first involves the use of meta-analytic findings to inform theory and practice. There is a need for CALL–SLA meta-analyses to go further here, utilizing their rich datasets and bird's eye views in coordination with researcher expertise to provide precision and nuance to the theoretical and practical issues facing the domains under investigation. Doing so involves returning to the theoretical motivations and issues usually raised in the Literature Review. The absence of such interpretations may be tied to a lack of theoretical support for primary studies, a critique often leveled in the context of CALL research (e.g., Chapelle, 2009; Handley, 2014). Viewing the results in this fashion also enables the researcher to provide explicit guidance for practitioners in the way of implementing CALL interventions. The results of moderator analyses can be used, for example, to point to those contexts or linguistic targets where different technology-based interventions can be most effective. Similarly, such results can also stem practices found by the meta-analysis to be less effective.

The remaining problematic trend in the Results and Discussion sections of CALL–SLA meta-analyses is a lack of recommendations provided for future research. It is incumbent upon meta-analysts to, again, take advantage of their perspective and data to guide future research toward more fruitful and under-explored effects, relationships, and so forth. Likewise, it is also the meta-analyst's duty to make recommendations for improving methodological practices in future empirical efforts. In order to do so, however, the researcher must code for substantive and methodological features of interest in primary studies, which was rarely the case in our sample.

FUTURE RESEARCH

It is clear from this review and from the sheer volume of research it covers that the CALL–SLA interface enjoys a level of maturity and productivity on par with or even surpassing many other areas within applied linguistics. Rather than looking only back at what has been done in this domain, however, we would like to look toward the future of this growing area of research. Part 1 of this article revealed a number of substantive features in need of further investigation at both the primary and meta-analytic levels. For example, because technology is a constantly evolving field, primary research is needed on how to best support and facilitate L2 learning with a variety of new and emerging tools, including multi-player games, virtual worlds, online collaborations, social media tools, and other user-driven applications. These technologies provide new and exciting contexts for L2 instruction. Of course, more research at the primary—and later, meta-analytic—level on their efficacy both in general and across learner demographics, target structures, and so forth is necessary. In addition, our review of synthetic research suggests that the majority of research at the CALL–SLA interface focuses on university age or adult learners, a finding identified by previous researchers as well (Grgurović et al. 2013; Liu et al., 2002; Zhao, 2003). The fact that this finding is consistent with previous research underscores the need for more research on primary and secondary learners in K–12 contexts, as the role of technology and its impact on learning may vary across these different populations. In addition, researchers need to expand the range of target languages under examination, as English and Spanish remain the most commonly investigated target languages (similar findings were reported by Felix, 2005 and Grgurović et al., 2013). The lack of expansion in target languages limits the generalizability of findings to a more diverse range of contexts, and should encourage future researchers to investigate the role of technology in less commonly taught languages, specifically those with different orthographic systems.

Turning to recommendations at the meta-analytic level, more careful consideration should be given to the selection and investigation of relevant moderating variables. For instance, given the conflicting results regarding proficiency, more research is needed to clarify how it might moderate the efficacy of technology for L2 learning outcomes. In addition, this review indicated contrasting findings for type of interlocutor and education setting, two variables that primary research has identified as influential on learners' production and development. Future meta-analysts should examine these important moderating variables in order to obtain a deeper understanding of their role in L2 learning. Finally, given its importance in curriculum and program development, more research examining the use of computer-adaptive testing, particularly at the meta-analytic level, would provide the field with useful information regarding the efficacy of technology not just as a learning tool, but as a method for assessment.

A number of recommendations were also revealed from our review of methodological practices observed in CALL–SLA meta-analyses. We would first encourage future meta-analysts to consider and address biases in their reviews, whether they are statistical in nature (e.g., in favor of $p < 05$), substantive (e.g., in favor of technology) in nature, or both. Doing so adequately will involve changes at multiple stages in the meta-analytic process including (a) more thorough inclusion criteria and search techniques, (b) analyses seeking to detect the presence of biases, and (c) interpretations that take into account the potential of biases in primary research to inflate or to otherwise alter meta-analytic results. Second, we recommend that future CALL–SLA meta-analyses take on the issue of methodological quality (i.e., rigor, transparency) in the primary literature. Doing so will enable meta-analysts to provide more informative and justifiable recommendations for future research, thereby hopefully improving such efforts. Third, we found the moderators in our sample to be generally well conducted and informative. As mentioned above, others could often have been included, and just as often, those moderators that were included were not well motivated in the literature reviews. Future meta-analyses would do well to explain and justify the inclusion of these critical analyses. Fourth, our sample fell short of what we might hope to find in terms of their transparency in the many decisions made when conducting a meta-analysis. In order for reviewers, editors, consumers, and practitioners to assess the validity of a meta-analysis, it is critical to provide, for example, information regarding the reliability of the data extraction. The items in the data collection instrument should also be justified and made available to readers. Our fifth and final recommendation to future meta-analyses is related to what researchers do with their results. Meta-analytic results can be complex and highly multivariate in nature. Therefore, in order to reach their maximum potential in their respective domains, the researcher must take great care in unpacking these results with respect to the theoretical, methodological, and practical issues at hand.

LIMITATIONS

Although the current second-order meta-analysis and synthesis provides critical information regarding both the efficacy of CALL for SLA and meta-analytic reports of such findings, there are limitations that must be noted. Due to the application of rigorous eligibility criteria, the sample of individual effect sizes calculated from the original meta-analyses is relatively small. Consequently, the identification of moderating variables across and within categories, such as proficiency and time of testing, was limited. Important constructs could not be analyzed because there were not enough meta-analyses examining these features. For example, Abraham (2008) was the only meta-analysis examining the impact of glossing on L2 development that included effect sizes for immediate and delayed posttests. Until more synthetic research is conducted examining potential moderating variables, firm conclusions regarding their overall role in L2 learning cannot be made.

As with any study, there are also limitations inherent to the design. Some researchers have criticized meta-analytic research for 'mixing apples and oranges' by combining different kinds of studies into one aggregate sample. These criticisms maintain that the average effect size will ignore potentially important differences across the collected studies (Borenstein et al., 2009). However, in a field as diverse as CALL and SLA, it is unavoidable that studies will differ across a number of characteristics, including participant population, research setting, educational context, and target language. In addition, treatments vary according to types of conditions and tasks, exposure time, and testing instruments. However, an important caveat to the lack of similarity across studies is that meta-analyses and syntheses are designed to ask broader research questions than the primary studies on which they are based. Furthermore, depending on the scope of the meta-analysis or synthesis, such study features can and should be investigated empirically as moderating variables.

We also recognize the challenges associated with the data collection instrument employed in Part 2 of the study. Different coders may arrive at different results. We would also like to point out, though, that no other instrument has been proposed for examining methodological quality at the meta-analytic level in applied linguistics. We encourage researchers in CALL and other areas of the field to explore the utility and validity of this instrument further in an attempt to validate it and improve the means by which we evaluate future meta-analyses.

CONCLUSION

This second-order review aimed to deepen our understanding of the CALL–SLA interface by examining the substantive and methodological features of meta-analytic research over the past two decades. Results indicated largely positive findings for the effects of CALL treatments on L2 learning outcomes, suggesting that learners in technology-mediated or technology-assisted contexts are likely to experience and perhaps surpass the positive developmental benefits associated with traditional FTF learning environments. Specifically, findings demonstrated a small relative effect (*ES* = .51) for overall L2 learning outcomes compared to traditional learning environments and a medium absolute effect (*ES* = .84) for CALL contexts. Turning to the effects of types of technology on overall L2 learning, a small to medium relative effect (*ES* = .60) was obtained for CALL glossing while a small relative effect (*ES* = .33) was found for CMC.

In addition, this study also identified areas of potential improvement in the methodology and possible future research agendas in CALL-based SLA meta-analyses and research syntheses. The issues regarding study design and the reporting of reliability and statistical measures identified here are in need of attention, as these concerns prohibit advancement of research findings. Furthermore, by not encouraging and requiring transparent and accurate reporting, researchers impede the precise interpretation of data and the identification of questions needing further empirical attention, potentially stifling growth within the field. We hope that these findings will encourage reflection among CALL researchers regarding current methodological practices, as the legitimacy of CALL research within mainstream SLA will be limited as long as research practices lag behind in innovation, execution, and reporting. It is our hope as well that the results of this study will inspire methodological improvements and provide the means through which future empirical investigations will help to move the field forward.

NOTE

1. Meta-analyses are based on either the fixed-effects or random-effects model depending on the assumptions of the data. Fixed-effects models assume that all studies included in the sample share a common effect size. In other words, the factors that might influence the effect size are constant across all the studies, making the true effect size the same. Random-effects models, on the other hand, assume that effects are normally distributed and vary due to heterogeneous factors. The decision regarding which model to use is based on a number of factors, including the researcher's understanding of whether the collected studies share a common effect size and the ultimate goals of the analysis.

ABOUT THE AUTHORS

Luke Plonsky (PhD, Michigan State University) is Senior Lecturer of Second Language Acquisition at University College London. His interests include SLA, L2 pedagogy, and quantitative research methods.

E-mail: l.plonsky@ucl.ac.uk

Nicole Ziegler (PhD, Georgetown) is Assistant Professor of Second Language Studies at University of Hawai‘i at Mānoa. Her research program focuses on adult and child instructed SLA, including mixed method and interdisciplinary research in L2 conversational interaction, TBLT, and CALL.

E-mail: nziegler@hawaii.edu

REFERENCES

* denotes studies included in the second-order statistical analysis in Part 1.
^ denotes studies included in the methodological analysis in Part 2.

*^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.

Aldrich, C. (2009). *Learning online with games, simulations, and virtual worlds: Strategies for online instruction*. San Francisco, CA: John Wiley.

Beatty, K. (2010). *Teaching & researching: Computer-assisted language learning*. New York, NY: Routledge.

Bernard, R. M., Borokhovski, E., Schmid, R. F., & Tamim, R. M. (2014). An exploration of bias in meta-analysis: the case of technology integration research in higher education. *Journal of Computing in Higher Education*, *26*(3), 183–209.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005). Comprehensive Meta-Analysis (Version 2) [Computer software]. Englewood, NJ: Biostat.

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

Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL*, *27*(1), 4–20.

Cerezo, L., Baralt, M., Suh, B. R., & Leow, R. P. (2014). Does the medium really matter in L2 development? The validity of CALL research designs. *Computer Assisted Language Learning*, *27*(4), 294–310.

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research.* Cambridge, UK: Cambridge University Press.

Chapelle, C. (2009). The relationship between second language acquisition theory and computer-assisted language learning. *Modern Language Journal*, *93*(S1), 741–753.

*^Chiu, Y.-H. (2013). Computer-assisted second language vocabulary instruction: A meta-analysis. *British Journal of Educational Technology*, *44*, E52–E56.

*^Chiu, Y.-H., Kao, C.-W., & 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*, E104–E107.

Chun, D. (2012). Review article: Replication studies in CALL research. *CALICO Journal*, *29*(4), 591–600.

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

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

Cooper, H., & Hedges, L. V. (2009). Introduction. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 3−16). New York, NY: Russell Sage Foundation.

de la Fuente, M. J. (2003). Is SLA interactionist theory relevant to CALL? A study on the effects of computer-mediated interaction in L2 vocabulary acquisition. *Computer Assisted Language Learning, 16*(1), 47–81.

Derrick, D. J. (in press) Instrument reporting practices in second language research. *TESOL Quarterly*.

Felix, U. (2005). What do meta-analyses tell us about CALL effectiveness? *ReCALL*, *17*, 269–288.

Felix, U. (2008). The unreasonable effectiveness of CALL: What have we learned in two decades of research? *ReCALL*, *20*, 141–161.

Gass, S. M., & Mackey, A. (2015). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed.), (pp. 180–206). New York, NY: Routledge.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer Assisted Language Learning, 27*(1), 70–105.

*^Grgurović, M., Chapelle, C., & Shelley, M. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*, 165–198.

Handley, C. (2014). Constructing an evidence-base for future CALL design with ‘engineering power’: The need for more basic research and instrumental replication. *The EUROCALL Review, 22*, 46–56.

Higgins, J. P. T., Lane, P. W., Anagnostelis, B., Anzures-Cabrera, J., Baker, N. F., Cappelleri, J. C., & Whitehead, A. (2013). A tool to assess the quality of a meta-analysis. *Research Synthesis Methods*, *4*, 351−366.

Journal Article Reporting Standards Working Group (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839−851.

Lai, C., & Li, G. (2011). Technology and task-based language teaching: A critical review. *CALICO Journal*, *28*, 1–24.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*(S1), 127–159.

*Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*, 345–366.

*Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*, 309–365.

Lin, H. (2012). The effectiveness of computer-mediated communication on SLA: A meta-analysis and research synthesis. In L. Bradley & S. Thouësny (Eds.), *EUROCALL proceedings* (pp. 177–181). Dublin, Ireland: Research-Publishing.net.

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

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

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms.* Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Liu, M., Moore, Z., Graham, L., & Lee, S. (2002). A look at the research on computer-based technology use in second language learning: A review of the literature from 1990-2000. *Journal of Research on Technology in Education*, *34*, 250–273.

Macaro, E., Handley, Z., & Walter, C. (2012). A systematic review of CALL in English as a second language: Focus on primary and secondary education. *Language Teaching*, *45*, 1–43.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–449). Oxford, UK: Oxford University Press.

Moher, D., Liberati, A., Terzlaff, J., & Altman, D. G. (2009). Preferring reporting items for systematic reviews and meta-analyses: The PRISMA statement. *British Medical Journal*, 339, 332−336.

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 3−50). Amsterdam, Netherlands: John Benjamins.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85−110.

Plonsky, L. (2012). Replication, meta-analysis, and generalizability. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 116–132). New York, NY: 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*, 655–687.

Plonsky, L. (2015). Quantitative considerations for improving replicability in CALL and applied linguistics. *CALICO Journal*, *32*, 232–244.

Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*, 267–278.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366.

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, *36*, 73–97.

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). New York, NY: Routledge.

Porte, G. (2013). Who needs replication research? *CALICO Journal*, *30*, 10–15.

Rankin, Y., Gold, R., & Gooch, B. (2006). Evaluating interactive gaming as a language learning tool. *Proceedings for ACM SIGGRAPH Conference*. Boston, MA.

Reinders, H. (Ed.). (2012). *Digital games in language learning and teaching.* New York, NY: Palgrave Macmillan.

Sauro, S. (2011). SCMC for SLA: A research synthesis. *CALICO Journal*, *28*, 369–391.

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., Porter, A. C., Tugwell, P., Moher, D., & Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*(10), 1471−2288.

Smith, B. (2004). Computer-mediated negotiated interaction and lexical acquisition. *Studies in Second Language Acquisition*, *26*, 365–398.

Smith, B. (2009). The relationship between scrolling, negotiation, and self-initiated self-repair in a SCMC environment. *CALICO Journal*, *26*, 231–245.

Smith, B., & Schulze, M. (2013). Thirty years of the CALICO Journal—Replicate, replicate, replicate. *CALICO Journal*, *30*, i–iv.

Suh, S., Kim, S. W., & Kim, N. J. (2010). Effectiveness of MMORPG-based instruction in elementary English education in Korea. *Journal of Computer Assisted Learning*, *26*, 370–378.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. (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*, 4–28.

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

*Taylor, A. M. (2009). CALL-based versus paper-based glosses: Is there a difference in reading comprehension? *CALICO Journal*, *23*, 147–160.

*^Taylor, A. M. (2013). CALL versus paper: In which context are L1 glosses more effective? *CALICO Journal*, *30*, 63-81.

Thorne, S. L., Black, R. W., & Sykes, J. M. (2009). Second language use, socialization, and learning in Internet interest communities and online gaming. *Modern Language Journal*, *93*, 802–821.

*^Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, *24*, 39–58.

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

Ziegler, N. (2013*). Synchronous computer-mediated communication and interaction: A meta-analysis* (Unpublished doctoral dissertation). Georgetown University, Georgetown, PA.

*^Ziegler, N. (2015). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, 1–34, doi:10.1017/S027226311500025X.

Ziegler, N. (2016a). Methodological practices in interaction in synchronous computer mediated communication: A synthetic approach. In A. Mackey & E. Marsden (Eds.), *Instruments for research into second languages: Empirical studies advancing methodology* (pp. 197–223). New York, NY: Routledge.

Ziegler, N. (2016b). Taking technology to task: Technology-mediated TBLT, performance, and production. *Annual Review of Applied Linguistics*, *36*, 136–163.