CrossMark

Lang. Teach. (2018), **51.3**, 417–429 © Cambridge University Press 2017
doi:10.1017/S0261444817000040

Replication Studies

Replication research in computer-assisted language learning: Replication of Neri et al. (2008) and Satar & Özdener (2008)

Zöe Handley Department of Education, University of York, UK
zoe.handley@york.ac.uk

This paper argues that key findings from computer-assisted language learning (CALL) research need to be replicated to permit the construction of a valid and reliable evidence-base which can inform the design of future CALL software and activities, together with language teachers' decisions about their adoption. Through the critical examination of two key studies, the limitations of much current CALL research with respect to the construction of such an evidence-base are highlighted and approaches to replication which may overcome these limitations are discussed. Like much CALL research, both papers are experimental studies. Key issues raised by the two studies are the need to: (1) conduct replications focused on attributes and affordances which transcend multiple technologies which might be employed in CALL, (2) collect process data to increase the explanatory power of studies, and (3) link CALL research to relevant second language acquisition (SLA) theory and research to increase the explanatory power of studies and integrate the findings from individual studies into the CALL and SLA evidence-base more broadly. Emphasis is also placed on the need to conduct replications with different populations in different settings in order to demonstrate the validity and reliability of current findings.

1. Introduction and background

Investment in the provision of technology in schools and universities has increased significantly in recent years (Macaro, Handley & Walter 2012). At the same time, the use of technology in language learning is now stipulated in a number of language standards (Macaro et al. 2012). In order to enable developers to make decisions about the design of future CALL software and activities as well as enable practitioners to make informed decisions about CALL adoption, the construction of a valid and reliable evidence-base with explanatory power is increasingly important. As discussed elsewhere, there are a number of different dimensions to the validity and reliability of an evidence-base (Cohen et al. 2011). One dimension which merits further consideration in the context of CALL research is external validity or generalizability. Generalizability in the context of CALL refers to not only the potential for the findings of research to translate to different educational contexts with

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

different learners, but also the potential for findings to apply to CALL software and activities beyond those specifically investigated in an individual study (Stockwell 2012). Explanatory power refers to the extent to which studies are able to provide reasons for any observed impact of the attributes and affordances of technology on language learning (Pederson 1987). Knowledge construction refers to the potential to integrate and synthesise the findings of different individual studies into a coherent evidence-base (Handley 2014).

This paper, however, argues that, in addition to lack of evidence of validity and reliability also observed in the wider field of SLA due to the lack of replication (Porte 2013), the current CALL evidence-base also lacks generalizability (to other CALL software) and explanatory power, and there is an absence of knowledge construction. It is argued that generalizability to other CALL software might be achieved by focusing in on attributes and affordances which transcend multiple technologies (Pederson 1987; Stockwell 2012), where attributes refer to 'features of the computer which have the potential to support and develop cognitive processing, such as symbol systems, multimedia and random access', and affordances refer to 'features of the computer which enable learners to engage in processes that support language learning' (Handley 2014: 49). Broad CALL versus non-CALL comparisons, however, dominate the literature (Felix 2008; Macaro et al. 2012), and consequently there are currently few insights into the impact of attributes and affordances which transcend multiple technologies. Explanatory power, it is argued, is best achieved when links are drawn with SLA theory and research, which might explain how different attributes and affordances of technology might either provide conditions or enable learners to engage in processes which are known to support language learning (Pederson 1987). Such links between SLA theory and research and the attributes and affordances of technology are, however, rarely made in the CALL literature and there is a lack of instrumental replication (Macaro et al. 2012), that is re-use of measures commonly used in SLA research. Knowledge construction, it is argued, is achieved by conducting more narrowly focused studies investigating the impact of one or more specific attributes or affordances of technology and engaging in instrumental replication, where instrumental replication refers to the re-use of measures employed in previous research. The former ensures the longevity of research and better enables researchers to build on one another's findings (Pederson 1987). The latter permits the integration of findings with those of other studies in CALL and the broader field of SLA. Further, with respect to the validity and reliability of the CALL evidence-base, and as has been observed in relation to SLA research more broadly (Porte 2013), there is a lack of replication in the field of CALL (Chun 2012).

In the remainder of this paper, these issues are discussed in more detail and exemplified through the critical examination of two key studies which make an important contribution to our understanding of the potential benefits of CALL. Both studies were selected on the basis of the 'weight of evidence' they had been attributed in Macaro et al.'s (2012) systematic review of research on the use of CALL in primary and secondary EFL teaching. In that study, weight of evidence was determined by the relevance of the study to the review, the appropriateness of its design, its trustworthiness or credibility, and its overall contribution to the review. Further factors which contributed to the selection of these studies were the significance of the research questions addressed by the studies and their findings as discussed below. The first study focuses on the use of Automatic Speech Recognition (ASR) to support

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

the development of pronunciation skills (Neri et al. 2008b), where ASR refers to software that 'listens' (Wachowicz & Scott 1999). Specifically, the study compares the effectiveness of ASR-based pronunciation training with standard non-CALL provision in an intervention involving young Italian learners of English. The second study focuses the use of synchronous computer mediated communication (CMC), i.e. chat, to support the development of speaking skills and reduce anxiety (Satar & Özdener 2008). It compares the impact of text chat on the speaking proficiency and anxiety levels of Turkish learners of English with the impact of audio chat on their speaking proficiency and anxiety levels.

Neri et al.'s (2008b) study merits consideration for two reasons. First, it is one of very few (classroom) evaluations of the impact of ASR on the quality of learners' pronunciation – a notable exception being Neri, Cucchiarini & Strik's (2008a) own evaluation of similar software with adult learners of Dutch. Classroom (or field) evaluations of the impact of ASR-based CALL software on language acquisition are important because ASR is based on probabilistic algorithms and the feedback provided is not entirely reliable. Whichever ASR system you employ, there is always the possibility of false positives, i.e. 'the system … determine[s] that the user pronounced a correct sound or word when in reality the sound was incorrect' (Eskenazi 2009: 834), and false negatives, i.e. 'the system … determine[s] that the user pronounced an incorrect sound or word when in reality it was correct' (Eskenazi 2009: 834). Second, Neri et al.'s (2008b) approach to assessing the quality of learners' pronunciations at pre- and post-test is rigorous in comparison with other studies in the field of CALL (see Macaro et al. 2012). Neri et al. (2008b) asked three native speakers of English to rate each learner production and to rate a small sub-sample a second time on a later occasion in order to assess the level of inter- and intra-rater reliability. Further, examples of both native speaker productions and 'very poor' learner productions were played to the native speakers at the start of each rating session to help them anchor their ratings.

Satar & Özdener's (2008) study is significant for three reasons. First, it focuses on a learner variable which has been found to be negatively associated with oral test performance, namely foreign language learning anxiety (Phillips 1992; Hewitt & Stephenson 2012). Second, it identifies a possible solution which may help break the 'vicious circle' of speaking anxiety and low speaking proficiency (Compton 2002: 25, cited in Satar & Özdener 2008). At post-test, Satar & Özdener (2008) found that learners who had completed tasks through the medium of text chat during their intervention had lower levels of anxiety than learners completing tasks through the medium of audio chat. They therefore suggest that text chat tasks might be used to reduce learners' anxiety before progressing to audio chat and face-to-face tasks. Third, it is one of only a few field evaluations of task-based language teaching (TBLT) – notable exceptions being Philp, Oliver & Mackey's (2006) investigation of the impact of planning time on task-based interactions and Li, Ellis & Zhu's (2016) comparison of task-based and task-supported language learning. To date, previous research on TBLT has predominantly focused on investigating the impact of different task designs and conditions on the types of interaction that learners engage in during speaking tasks (e.g. Slimani-Rolls 2005; Fujii & Mackey 2009; Park 2010) and the quality of output produced and the inter-relationship between complexity, accuracy and fluency (e.g. Ellis 2009; Ahmadian, Tavakoli & Dastjerdi 2012; Foster & Skehan 2013) in controlled laboratory studies, teacher and learner attitudes towards TBLT (e.g. Carless 2007; McDonough & Chaikit-Mongkol 2007), and teachers'

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

understanding (Zheng & Borg 2014) and adoption of TBLT (e.g. Andon & Eckerth 2009; van den Branden 2009). Field or classroom evaluations of computer-mediated TBLT are important because the findings of tightly controlled laboratory studies tend not to generalize to classroom studies; effect sizes for laboratory studies tend to be larger than those for classroom studies (Ziegler 2015). One reason for this might be that TBLT generally requires teachers to adopt a learner- rather than teacher-centred approach, which is a significant change in practice for teachers in many educational contexts.

In the following sections, the original studies mentioned above will be described in detail. An approach to replication will then be offered following a discussion of the original studies.

2. The original studies and suggested approach to replication

2.1 Neri et al. (2008b)

This study, published in *Computer-Assisted Language Learning*, is an evaluation of the ASR-based pronunciation training system PARLING (PARLa INGlese; Speak English). Designed for young Italian learners or English, each module focuses on a simplified version of a well-known story. When a module is loaded, PARLING displays the story on screen and plays a recording of a native speaker reading the story. Some of the words have hyperlinks which allow learners to access their meaning and a model pronunciation recorded by a native speaker. Learners can also record themselves pronouncing the words. PARLING provides the learner feedback on their pronunciation of the correct/incorrect type through the use of ASR, and asks learners to repeat any words it judges them to have pronounced incorrectly. When the learners have finished reading the story, the system presents them an adaptive ASR-based word game.

The study compared the use of the software with current classroom-based pronunciation teaching in a quasi-experiment with two intact classes of young (ten-year-old) Italian learners of English. The experiment had a between subjects design with a pre-test post-test format. During the pre-test learners were asked to read a list of 28 words selected from the training materials and designed to cover the most frequent phonemes of British English. During the intervention, the experimental group (N = 13) received four 30-minute sessions of training with PARLING, and the control group (N = 15) received four 60-minute sessions of training with the classroom teacher – sessions with the Computer-Assisted Pronunciation Training system were shorter because of the limited numbers of computers available to the class. The control group received similar training to the experimental group but delivered by their class teacher in a whole group setting. Specifically, the learners in the control group were given a printed copy of the story and the teacher read it aloud to them. The teacher also explained the meaning of some of the words in the story and provided learners with a pronunciation model which they were encouraged to imitate as a group. Having read the story, the learners completed a printed word game. During the post-test, learners were asked to read the list of 28 words again.

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

The pronunciation quality of each of the learners' productions was rated on a ten-point scale by three native speakers of British English, and a satisfactory level of inter-rater reliability as well as a high level of intra-rater reliability was achieved. Further, differences between the groups at pre-test were not statistically significant ($t = .32$, $p = .75$, $r = .06$).[1] Statistically significant improvements in pronunciation quality were observed for both groups from pre-test to post-test ($F(1.26) = 78.818$, $p < .05$, $r = .87$), but no effect of training group was observed ($F(1.26) = .610$, $p = .44$, $r = .15$), and there was no interaction between time and training group ($F(1.26) = .548$, $p = .45$, $r = .14$). The researchers argue that this finding suggests that the two forms of training are comparable, and highlight that this is a significant finding given that the learners in the experimental group received half as much training as those in the control group.

2.1.1 Approach to replication

One cannot deny the importance of conducting field evaluations of CALL software such as this. CALL versus non-CALL studies without doubt play a significant role in the CALL development life-cycle. Importantly, having optimized the design of a piece of CALL software or a CALL activity, they allow researchers to compare its effectiveness with current (non-CALL) provision (Grgurovic, Chapelle & Shelley 2013). As such, Neri et al.'s (2008b) study has high ecological validity.

However, if we put the authors' own research aims to one side, and consider the potential for the study to contribute to the construction of a valid and reliable CALL evidence-base, a number of replications of this study merit consideration.

First, in order to permit the generalization of findings to different CALL software a series of more narrowly focused conceptual replications in which key features of the design of ASR-based CALL software are isolated and manipulated is recommended (see Pederson 1987). The current study is an example of a broad comparison of CALL and current (non-CALL) provision. As has long been acknowledged in the field of CALL, the results of such broad comparisons are constrained to the specific software investigated (Pederson 1987; Garrett 1991; Handley 2014). This is because the CALL condition often differs in multiple ways from the non-CALL condition (Handley 2014). For example, in this study, in addition to whether or not ASR-based feedback on pronunciation quality is provided, the CALL condition differs from the non-CALL condition with respect to the voice in which the pronunciation models are presented, a variable which other research suggests has an impact on the effectiveness of pronunciation training (Probst, Ke & Eskenazi 2002). Further differences include amount of training (duration of the training and number of repetitions), nature of the training (individual vs group), and the nature of the word game (adaptive vs paper-based). In other words, there are a number of threats to the internal validity of this study, and it is not possible to definitively attribute the findings to the provision of ASR-based

[1] Following Plonsky (2015), effect sizes have been calculated to avoid a number of problems associated with Null Hypothesis Significant Testing (NHST) and the use of p-values. These problems include (1) the reduction of data to a significant/non-significant dichotomy, and (2) the sensitivity of p-values to sample size – given the same effect size, NHST is less likely to detect a difference between conditions with a smaller sample size than with a larger sample size.

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

feedback. In particular, a narrowly focused conceptual replication similar to Neri et al. 2008 is recommended. In this study, two versions of the CALL software were compared, one with ASR-based feedback and one without in which learners were required to judge for themselves whether their productions matched the models provided.[2]

Second, given the extent of individual variation in success among language learners (Dörnyei 2005) and the impact of different educational settings on the effectiveness of language learning interventions (Chapelle & Jamieson 1990), demonstrating the generalizability of the findings of CALL research through replications with other populations in other educational settings is also important (Polio & Gass 1997; Mackey 2012; Plonsky 2012; Polio 2012). This will become more important in the context of speech-enabled CALL as first language-independent (L1) approaches to automatic pronunciation scoring emerge (see for example Cincarek et al. 2009), and it is possible that the same software might be used with learners from a range of L1 backgrounds.

Replications with other populations are also important when working with intact classes as in Neri et al.'s (2008b) study. When one intact class is allocated to each experimental condition, the differences observed between the experimental groups may reflect a cluster or class effect, that is, differences between the classes in terms of pronunciation quality at pre-test and language learning experiences (Torgerson & Torgerson 2003) rather than the impact of ASR-based feedback on the quality of learners' pronunciation. Where the curriculum, timetabling and other constraints of the participating schools permit, in any future replications rather than allocating intact classes to conditions, a split class design in which an equal proportion of each intact class is assigned to each condition through random allocation is therefore recommended (Marsden 2007).

2.2 Satar & Özdener (2008)

This study, published in *The Modern Language Journal*, compared the impact of text and voice chat on learners' speaking proficiency and anxiety levels. The participants were ninety 16–17-year-old Turkish learners of English studying at a vocational high school. The experiment had a between-subjects design with a pre-test post-test format. At pre-test, learners were asked to complete a speaking test and a questionnaire focusing on foreign language learning anxiety. The speaking test consisted of five parts. In the first part of the test, the learners were asked to talk about themselves; in the second, they were asked to describe what the people in a series of pictures were doing; in the third, they were asked to discuss the similarities and differences between two pictures; in the fourth, they were asked to tell a story about something they had witnessed; and, in the fifth, they were asked to perform a role-play with a partner. The Turkish version (Aydin 2001) of Horwitz, Horwitz & Cope's (1991) questionnaire was used to measure foreign language learning anxiety. The learners were then allocated to the two experimental conditions according to their scores on the initial speaking test in order to

[2] A further narrowly focused evaluation with children is recommended because there is significant individual variation in success in language acquisition and individual differences has been found to play an important role in determining success (Polio 2012). Further, the nature of children's speech in comparison with that of adults, has meant that it has been necessary to develop dedicated algorithms to recognise and detect errors in their speech (Gerosa & Giuliani 2004).

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

Table 1 Results of pre-test post-test comparisons of learners' anxiety levels in Satar & Özdener (2008)

Condition	*Df*	*t*	*p*	*r*
Text chat	29	2.14	.04	.37
Voice chat	29	1.26	.22	.23
Control	29	− .29	.77	.02

ensure that they were matched. All learners received the same in-class instruction delivered by their class teacher, who was also one of the authors of the study. In addition to classroom instruction, the learners in the two experimental groups participated in one 40–45-minute out-of-class chat discussion per week for a period of four weeks. In both experimental groups, the chat discussions focused on a number tasks. Each session, the learners completed two tasks with a partner. The first two tasks were information gap tasks; the second two were problem-solving tasks; the next two were jigsaw tasks; and, the final two were decision-making tasks. The only difference between the two experimental groups was the mode of discussion: the first experimental group (N = 30) participated in text chat, while the second participated in voice chat (N = 30). The control group was not assigned any out-of-class activities, i.e. there was no provision for the control group. At post-test the speaking tests and the foreign language learning anxiety questionnaire were administered again.

Two language teachers, one of which was one of the researchers, rated the learners' productions using Hughes's (2003) general speaking proficiency scale and a language development scale created for the specific purposes of this study. Hughes's (2003) speaking proficiency scale comprises five sub-scales focusing on pronunciation, vocabulary, grammar, fluency, and understanding respectively. The language development scale comprised two sub-scales focusing on vocabulary and grammar development respectively. High levels of both inter-rater reliability and test-retest reliability were achieved for the speaking tasks, and the experimental and control conditions did not differ significantly in speaking proficiency at pre-test ($F(2,87) = 1.85$, $p = .16$, $w = .25$). At post-test, significant differences were observed across the experimental and control conditions ($F(2,87) = 23.76$, $p < .01$, $w = .33$),[3] with learners in both the text chat and voice chat conditions scoring significantly higher on the speaking test than learners in the control condition and no significant differences between learners in the voice chat and text chat conditions. The anxiety questionnaire was also found to have a high level of internal consistency. Learners' anxiety levels were found to decrease from pre-test to post-test in both experimental conditions, but this decrease in anxiety levels was only statistically significant for learners in the text chat condition (see Table 1).

On the basis of these results, the researchers concluded that text chat is as effective as voice chat in terms of impact on speaking proficiency, that transfer is possible from written to oral communication, and that engaging in text chat activities may help learners gain confidence in their language skills. Satar & Özdener (2008), however, note that it should be

[3] Following Plonsky (2015), the post-test effect size reported here has been adjusted for the effect size at pre-test.

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

taken into consideration that the learners in the text chat group had higher levels of anxiety than those in the voice chat group at pre-test. They also speculate that voice chat may have had a differential impact on the anxiety levels of more and less proficient learners, decreasing anxiety levels for more proficient learners and increasing them for less proficient ones.

2.1.1 Approach to replication

Again, the significance of this piece of research cannot be denied. As indicated above, it is one of few field evaluations of TBLT. Further, the results of the study suggest that text chat might help learners overcome the common problem of reluctance to speak in the target language.

However, if we put aside the authors' own specific aims in conducting this study, and focus on the aim of constructing an evidence-base which can inform the design of future CALL software and activities as well as decisions about their adoption, a number of replications of this study merit consideration.

First, a replication with a face-to-face control condition is recommended (see also Ziegler 2015).[4] As the authors argue, it is acknowledged that without CMC, learners have limited opportunities to engage in speaking practice outside the classroom. Therefore a no provision control condition is appropriate. However, if, as this study suggests, engaging in text chat mediated tasks does indeed reduce learners' levels of anxiety for speaking, and engaging in text chat mediated tasks has a comparable impact on speaking proficiency to engaging in voice chat mediated tasks, engaging learners in text chat-mediated tasks in the classroom may be beneficial and a comparison with standard (non-CALL) provision merits consideration. A replication in which text chat is compared with standard provision, i.e. face-to-face tasks, is also important to permit knowledge construction in the field. The inclusion of a face-to-face control condition would permit comparisons with the large body of TBLT research outside CALL (Ziegler 2015; and see Robinson 2011 for a review of TBLT research) and the integration of findings from this strand of research into meta-analyses of related research in that area (Mackey 2012).

Second, in order to validate the findings of this study, it is recommended that a replication is conducted in which learners' levels of anxiety are controlled for at pre-test. This might be achieved by matching learners on anxiety levels at pre-test or randomly allocating learners to conditions and including a measure of pre-test anxiety levels as a covariate in statistical analyses. Further, replications with more and less anxious learners are also recommended. As already stated, the finding that learners are less anxious is an important one and merits further investigation because it suggests that text chat might break the 'vicious circle' of speaking anxiety and low speaking proficiency (Compton 2002: 25). While a likely explanation for the finding is that text chat places less time pressure on learners than voice chat, as the authors themselves suggest, there are a number of other possible explanations for the findings. These include differences in the anxiety levels of the learners in the study at pre-test and consequently

[4] At face value this recommendation may appear to contradict earlier criticisms of CALL vs non-CALL studies. The original study, however, involves a narrow comparison in which only the mode of communication differs across experimental conditions. In contrast with most CALL vs non-CALL studies, the proposed replication would therefore have explanatory power.

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

the potential for their anxiety levels to decrease, as well as the possibility that text and voice chat have a differential impact on learners with different initial levels of anxiety:

> the pre-test anxiety level of the text chat group was higher than that of the voice chat group. This might have resulted in a larger drop in the anxiety levels of the text chat participants . . . than the voice chat participants . . . [A]nother possible reason for the smaller decrease in the anxiety levels of the voice chat participants could be that the voice chat affected learners with different proficiency levels differently: higher proficiency learners might have become more confident, but those with lower levels might have lost their confidence as they struggled to put their thoughts into words in the synchronous environment (Satar & Özdener 2008: 607).

The collection of observational or 'process' data, as suggested by Satar & Ozdner (2008) themselves, ought also to be considered in any replication of this study. The purpose of the collection of such data would be to permit the exploration of possible explanations for the differences observed at post-test (Marsden 2007; Révész 2014). Satar & Özdener (2008) speculate that a potential reason for learners' reduced anxiety in the text chat condition might be the reduced time pressure that learners experience. While it is acknowledged that there may be some interview evidence from learners in other studies (Beauvois 1997, cited in Satar & Özdener 2008) which supports the claim that there is less time pressure in text chat because there is greater tolerance of 'silence', a replication in which time pressure is quantified and associated with anxiety levels would strengthen these claims. This might be achieved by recording and transcribing voice chat conversations and comparing them with text chat logs obtained through the chat software or through screen-capture technology such as Camtasia.

Replications in which outcome measures typically used in TBLT studies are employed should also be considered to permit comparisons of results with the significant body of research on the impact of TBLT on the quality of the speech produced by language learners (see Robinson 2011 for a review of TBLT research) and the integration of CALL studies into meta-analyses focusing on TBLT (Mackey 2012). That is not to say that studies employing native speaker ratings, as in the original study, are not important. Indeed, such studies have an important place in the research cycle. Given their alignment with the way in which oral proficiency is assessed in high-stakes assessments, they play an important role in linking more theoretically oriented research with practice. Measures frequently used in TBLT, however, merit consideration to permit the comparison of results with this large body of research (Robinson 2011). These measures of proficiency typically include measures of complexity – 'the capacity to use more advanced language', accuracy – 'the ability to avoid error in performance', and fluency – 'the capacity to use language in real time' (Skehan & Foster 1999: 96). Further, each construct is considered to be multi-dimensional and has been operationalized in a number of different ways (Ellis 2009). The use of such measures, therefore, has the potential to provide more in-depth insights into the impact of different forms of chat on oral language proficiency than ratings. It is acknowledged, however, that such instrumental replication may be difficult because the tasks used to obtain outcome measures are often not provided in journal articles (Marsden & Mackey 2014) and information on coding procedures is often not sufficient to permit closer replication (Polio & Gass 1997). Notable exceptions include Foster, Tonkyn & Wigglesworth's (2000) article on coding for complexity using the

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

AS-Unit, and Foster & Wigglesworth's (2016) article on coding for accuracy. In order to permit more of this form of replication, which is essential for knowledge construction in the field, editors of journals in applied linguistics might be advised to encourage the inclusion of instruments in appendixes, or their submission to the Instruments for Research Into Second languages (IRIS) database (https://www.iris-database.org; Marsden & Mackey 2014), and the submission of methodological articles reporting and discussing data collection procedures, instrument design, coding and data analysis in depth.

Finally, like any study in CALL and SLA more broadly, replications with different populations in different settings ought to be considered in order to demonstrate the generalizability of findings, along with replications in which task variables and conditions that have been observed to impact on learning outcome in computer-mediated TBLT are manipulated. Examples of such variables include the focus target form and the type of interlocutor, native or non-native speaker (Ziegler 2015).

3. Conclusion

In conclusion, as the original studies reviewed here demonstrate, CALL has the potential to add significant value to language learning. Replication is, however, necessary to validate important findings in the field and demonstrate their reliability, including not only generalizability to different populations and educational settings but also generalizability to different CALL systems. A summary of recommended replications is as follows:

Neri et al. (2008b)

- Focused on the specific affordances of the medium to increase internal validity and explanatory power as well as permit knowledge construction
- With a split class design to increase internal validity
- With other populations in other educational settings to test external validity

Satar & Özdener (2008)

- With a face-to-face control condition to permit knowledge construction
- Using instruments and measures commonly used in related research in applied linguistics to permit knowledge construction
- With process data to increase explanatory power
- With other populations in other educational settings to test external validity
- With other task variables and task conditions to test external validity

Further, any future replications should be designed to contribute to the construction of an evidence-base upon which decisions about the design of future CALL software and activities can be made as well as decisions about investment in and use of CALL. That is, when planning replications, in fact any study, researchers need to carefully consider the potential for their study to build on existing knowledge in the field of CALL and applied linguistics more broadly, and for others to subsequently use and build on their findings.

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

Acknowledgements

The author is grateful to Dr Emma Marsden and Professor Leah Roberts for comments on earlier versions of this paper. She is also grateful to the authors of the studies recommended for replication in this article who kindly clarified the design of their studies and provided insightful comments on the proposed replications, as well as to the two anonymous reviewers who provided further constructive feedback on the paper.

References

Ahmadian, M. J., M. Tavakoli & H. V. Dastjerdi (2012). The combined effects of online planning and task structure on complexity, accuracy and fluency of L2 speech. *The Language Learning Journal* 43.1, 41–56.

Andon, N. & J. Eckerth (2009). Chacun à son gout? Task-based L2 pedagogy from the teacher's point of view. *International Journal of Applied Linguistics* 19.3, 286–310.

Aydin, B. (2001). *Konuşma ve yazma derslerinde yabani dil öğrenimindki kaygi nedenleri [Reasons for anxiety in foreign language learning in speaking and writing courses]*. Unpublished doctoral dissertation, Andalou University, Turkey, no. 1265.

Beauvois, M. H. (1997). Computer-mediated communication (CMC): Technology for improving speaking and writing. In M. D. Bush & R. M. Terry (eds.), *Technology-enhanced language learning*. Lincolnwood, IL: National Textbook Company, 165–184.

Carless, D. (2007). The suitability of task-based approaches for secondary schools: Perspectives from Hong Kong. *System* 35.4, 595–608.

Chapelle, C. & J. Jamieson (1990). Internal and external validity issues in research on CALL effectiveness. In P. Dunkel (ed.), *Computer assisted language learning and testing: Research issues and practice*. New York: Newbury House, 37–59.

Chun, D. (2012). Review article: Replication studies in CALL research. *CALICO Journal* 29.4, 591–600.

Cincarek, T., R. Gruhn, C. Hacker, E. Nöth & N. Satoshi (2009). Automatic pronunciation scoring of words and sentences independent from the non-native's first language. *Computer Speech and Language* 23.1, 65–88.

Cohen, L., L. Manion, K. Morrison, & R. Bell (2011). *Research methods in education*. London: Routledge.

Compton, L. (2002). *From chatting to confidence: A case study of the impact of online chatting on international teaching assistants' willingness to communicate, confidence level and fluency in oral communication*. Unpublished MA thesis, Department of English, Iowa State University, Ames.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. London: Routledge.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics* 30.4, 474–509.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication* 51, 832–844.

Felix, U. (2008). The unreasonable effectiveness of CALL: What have we learned in two decades of research? *ReCALL* 20.2, 141–161.

Foster, P. & P. Skehan (2013). Anticipating a post-task activity: The effects on accuracy, complexity, and fluency of second language performance. *Canadian Modern Language Review* 69.3, 249–273.

Foster, P. & G. Wigglesworth (2016). Measuring accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics* 36, 1–19.

Foster, P. Tonkyn, A. & G. Wigglesworth (2000). A unit for all reasons: The analysis of spoken interaction. *Applied Linguistics* 21.3, 354–374.

Fujii, A. & A. Mackey (2009). Interactional feedback in learner-learner interactions in a task-based EFL classroom. *International Review of Applied Linguistics in Language Teaching* 47.3–4, 267–301.

Garrett, N. (1991). Technology in the service of language learning: Trends and issues. *The Modern Language Journal* 75, 74–101.

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

Gerosa, M. & D. Giuliani (2004). Preliminary investigations in automatic recognition of English sentences uttered by Italian children. In Delmonte, R., P. Delcloque & S. Tonelli (eds.), *Proceedings of NLP and speech technologies in advanced language learning systems symposium*, Venice, Italy. Padova: Unipress, 9–12.
Grgurovic, M., C. A. Chapelle & M. C. Shelley (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL* 25.2, 165–198.
Handley, Z. (2014). Constructing an evidence-base for future CALL design with 'engineering power': The need for more basic research and instrumental replication. *EUROCALL Review* 22.2, 46–56.
Hewitt, E. & J. Stephenson (2012). Foreign language anxiety and oral exam performance: A replication of Phillip's *MLJ* study. *The Modern Language Journal* 96.2, 170–189.
Horwitz, E, M. Horwitz & J. Cope (1991). Foreign language classroom anxiety. In E. Horwitz & D. Young (eds.), *Language anxiety: From theory and research to classroom implications*. Englewood Cliffs, New Jersey: Prentice Hall, 27–36.
Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.
Li, S., R. Ellis & Y. Zhu (2016). Task-based versus task-supported language instruction: An experimental study. *Annual Review of Applied Linguistics* 36, 205–229.
Macaro, E., Z. Handley & C. Walter (2012). A systematic review of CALL in English as a second language: Focus on primary and secondary education. *Language Teaching* 45.1, 1–43.
Mackey, A. (2012). Why (or why not), when and how to replicate research. In G. K. Porte (ed.), *Replication research in applied linguistics*. Cambridge: Cambridge University Press, 34–69.
Marsden, E. J. (2007). Can educational experiments both test a theory and inform practice? *British Educational Research Journal* 33.4, 565–588.
Marsden, E. J. & A. Mackey (2014). IRIS: A new resource for second language research. *Linguistic Approaches to Bilingualism* 4.1, 125–130.
McDonough, K. & W. Chaikit-Mongkol (2007). Teachers' and learners' reactions to a task-based EFL course in Thailand. *TESOL Quarterly* 41.1, 107–132.
Neri, A., C. Cucchiarini & H. Strik (2008a). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL* 20.2, 225–243.
Neri, A., O. Mich, M. Gerosa & D. Giuliani (2008b). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning* 21.5, 393–408.
Park, S. (2010). The influence of pre-task instructions and pre-task planning on focus on form during Korean EFL task-based interaction. *Language Teaching Research* 14.1, 9–26.
Pederson, K. (1987). Research on CALL. In W. F. Smith (ed.), *Modern media in foreign language education: Theory and implementation*. Lincolnwood, Illinois: National Textbook Company, 99–131.
Phillips, E. M. (1992). The effects of language anxiety on students' oral test performance and attitudes. *The Modern Language Journal* 76.1, 14–26.
Philp, J., R. Oliver & A. Mackey (2006). The impact of planning time on children's task-based interactions. *System* 34.4, 547–565.
Plonsky, L. (2012). Replication, meta-analysis, and generalizability. In G. Porte (ed.), *Replication research in applied linguistics*. Cambridge: Cambridge University Press, 116–132.
Plonsky, L. (2015). Quantitative considerations for improving replicability in CALL and applied linguistics. *CALICO Journal* 32.2, 232–244.
Polio, C. (2012). Replication in published applied linguistics research: A historical perspective. In G. Porte (ed.), *Replication research in applied linguistics*. Cambridge: Cambridge University Press, 47–91.
Polio, C. & S. Gass (1997). Replication and reporting. *Studies in Second Language Acquisition* 19, 499–508.
Porte, G. (2013). Who needs replication? *CALICO Journal* 30.1, 10–15.
Probst, K., Y. Ke & M. Eskénazi (2002) Enhancing foreign language tutors – in search of the golden speaker, *Speech Communication* 37.3-4, 161–174.
Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics* 35, 87–92.
Robinson, P. (2011). Task-based language learning: A review of the issues. *Language Learning* 61.1, 1–36.
Satar, H. M. & N. Özdener (2008). The effects of synchronous CMC on speaking proficiency and anxiety: Text versus voice chat. *The Modern Language Journal* 92.4, 595–613.
Skehan, P. & P. Foster (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning* 49.1, 93–120.

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press

Slimani-Rolls, A. (2005). Rethinking task-based language learning: What we can learn from the learners. *Language Teaching Research* 9.2, 195–218.
Stockwell, G. (2012). Diversity in research and practice. In G. Stockwell (ed.), *Computer-assisted language learning: Diversity in research and practice*. Cambridge: Cambridge University Press, 147–163.
Torgerson, C. J. & D. J. Torgerson (2003). The design and conduct of randomised controlled trials in education: Lessons from health care. *Oxford Review of Education* 29, 67–80.
van den Branden, K. (2009). Mediating between predetermined order and chaos: The role of the teacher in task-based language education. *International Journal of Applied Linguistics* 19.3, 264–285.
Wachowicz, K. A. & B. Scott (1999). Software that listens: It's not a question of whether, it's a question of how. *CALICO Journal* 16.3, 253–276.
Zheng, X. & S. Borg (2014). Task-based learning and teaching in China: Secondary school teachers' beliefs and practices. *Language Teaching Research* 18.2, 205–221.
Ziegler, N. (2015). Synchronous computer-mediated communication and interaction. *Studies in Second Language Acquisition* 38.3, 1–34.

ZÖE HANDLEY is a lecturer in the Department of Education at the University of York and a member of the Centre for Research in Language Learning and Use. Her main research interests are in CALL and second language speech learning, including pronunciation training, fluency development and TBLT. She has published a systematic review on the use of technology in primary and secondary EFL classrooms and a number of articles on the use of text-to-speech synthesis in language learning.

https://doi.org/10.1017/S0261444817000040 Published online by Cambridge University Press