ReCALL **27**(1): 4–20. 2015 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344014000159
First published online 24 March 2014

Twenty years of MALL project implementation: A meta-analysis of learning outcomes

JACK BURSTON

Cyprus University of Technology, Cyprus
(email: jack.burston@cut.ac.cy)

Abstract

Despite the hundreds of Mobile-Assisted Language Learning (MALL) publications over the past twenty years, statistically reliable measures of learning outcomes are few and far between. In part, this is due to the fact that well over half of all MALL-related studies report no objectively quantifiable learning outcomes, either because they did not involve MALL implementation projects, or if they did, learning gains were only based on subjective teacher assessments and/or student self-evaluations. Even more so, the paucity of statistically reliable learning outcome data stems from the short duration of projects and small numbers of students involved. Of the 291 distinct studies examined in this review only 35 meet minimal conditions of duration and sample size, i.e., ten experimental subjects over a period of at least a month. Sixteen of these suffer from serious design shortcomings, leaving only nineteen MALL studies that can reliably serve as a basis for determining the learning outcomes of mobile-based language applications. Of these studies, fifteen can be considered to report unequivocal positive results, with those focusing on reading, listening and speaking without exception evidencing a MALL application advantage. Four studies, all focusing on vocabulary, reported no significant differences.

Keywords: MALL, mobile-assisted learning, learning outcomes, grammar, vocabulary acquisition, reading, listening, speaking, writing

1 Introduction

With ever more sophisticated and affordable smartphones and tablet computers overcoming the technological and economic constraints that have hampered the widespread application of MALL, attention is increasingly being turned to the use of mobile devices as language teaching tools. However, as was the case with the emergence of Computer-Assisted Language Learning (CALL) in the 1980s, technological enthusiasm remains to be supported by objective evidence of the pedagogical effectiveness of MALL. Despite the appearance of over 600 MALL publications over the past twenty years, including three major overviews of the field (Chinnery, 2006; Kukulska-Hulme & Shield, 2008; Burston, 2014), no study has systematically evaluated the learning outcomes of MALL implementation projects. That is the focus of this paper.

2 Data selection

To date, most of what has been reported about the pedagogical effectiveness of MALL is to be found in the literature reviews of MALL implementation studies themselves. By their

very nature, such reviews are fragmentary, since they only relate to applications of relevance to the specific projects being undertaken. So, too, they tend to be brief, rarely more than a few paragraphs, and merely summarize what is reported in the original publications without critical evaluation.

In order to arrive at a comprehensive objective assessment of MALL learning outcomes, it is first necessary to extract from the published literature the papers which deal specifically with implementation projects, i.e., those which involve actual field testing of an application. In reality, some forty percent of MALL publications are unrelated to implementation projects and focus instead on such things as mobile device ownership, technological infrastructure and design issues, pedagogical methodology and teacher training, among others.

As detailed in Burston (2013), of some 575 MALL publications between 1994 and 2012, only 347 actually describe implementation projects. It is these MALL applications that form the basis of the following analysis. One particularly notable aspect of these works is the disparate nature of their publication origins, with only about ten percent to be found in established CALL journals. The remainder appear in publications relating to distance learning, mobile learning, educational technology, multimedia, telecommunications, and lexicography, to name just the most frequent. MALL implementation projects are also characterized by small group sizes, short durations, and a greater focus on research trialling and device experimentation than curriculum integration (Burston, 2014). However small the group or brief the duration of the study, claimed learning improvements are the norm. It is interesting to observe in this respect that the few studies that report no significant difference in MALL learning outcomes are all among the most recent (Derakhshan & Kaivanpanah, 2011; Osman & Chung, 2011; Brown *et al.*, 2012; Chiang, 2012).

While the number of MALL implementation studies totals 347, the learning outcome information of only 315 of these was accessible for this analysis. Nonetheless, this accounts for 91% of the database, so can be taken as highly representative of published MALL research. A close reading of the MALL implementation database reveals that, due primarily to multiple conference presentations and proceedings subsequently being republished as journal articles, the data of 24 studies appear more than once. So as not to over-represent the number of distinct projects, such duplicates are not included in this analysis, which is thus based on a total of 291 studies.

3 Data analysis

3.1 Elimination of compromised studies

3.1.1 Inadequate treatment duration/student numbers. For learning outcome results to be objectively meaningful, they must, of course, be based on statistically valid analyses, which requires that projects be of reasonable duration and involve an adequate number of subjects. In this respect, MALL implementation studies fare very poorly (Burston, 2014). Only about a quarter of all MALL applications have taken place over an entire academic quarter or more (see Figure 1). About 30% were trialled for only a week or less, with more than three-quarters of these lasting less than three hours and some no more than five to ten minutes. So, too, the number of learners involved in MALL implementations has been limited (see Figure 2). Only eight percent of the cohorts consisted of more than 100 participants.

https://doi.org/10.1017/S0958344014000159 Published online by Cambridge University Press

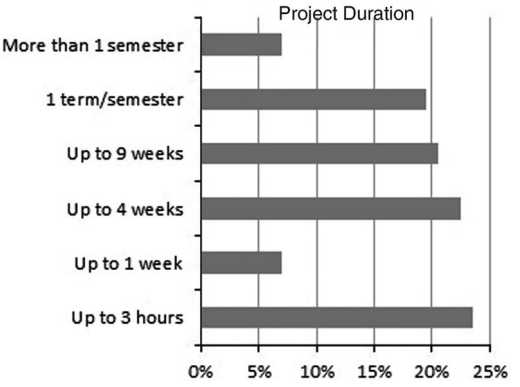

Fig. 1. MALL project duration

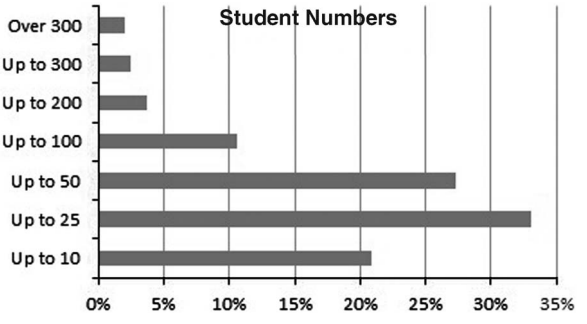

Fig. 2. MALL group sizes

Over half involved no more than 25, with well over a third of these groups consisting of no more than ten learners and some as few as four.

Making matters worse, of the 291 studies in the MALL implementation database, 61 specify neither the duration of the project nor group numbers. Another 27 specify only one or the other. Thus a total of 88 studies must be excluded from consideration due to their failure to specify the duration of treatment and/or sample size, leaving only 203 upon which a quantitative analysis of MALL learning outcomes can be based.

Although the minimal duration of an experimental treatment and number of subjects required for statistical results to be generalizable is debatable, for purposes of this analysis of MALL learning outcomes it was judged reasonable to exclude from consideration all implementation projects that lasted less than a month (i.e., twenty class days) or involved less than ten experimental subjects. Because of the preponderance of small numbers and short durations in the database, this resulted in the elimination of 94 studies. More stringent conditions for inclusion (e.g., an N of at least 20) would, of course, have resulted in even more exclusions.

Of the remaining 109 MALL implementation studies, 74 contain no quantitative learning outcome data. In many of these cases, the focus is on such topics as mobile device

https://doi.org/10.1017/S0958344014000159 Published online by Cambridge University Press

usage, student/teacher attitudes towards MALL, and learning strategies, so no learning outcomes at all are reported. In the others, learning outcome claims (invariably positive) are made, but are based only on subjective teacher impressions (including course grades) and/or student self-evaluations.

In sum then, of the 291 application studies in the MALL database under analysis, only 35 involve projects that report learning outcomes from implementations that lasted at least a month and involved ten or more experimental subjects, the minimal requirements set for the statistical generalizability of the results.

3.1.2 Design shortcomings. Among the studies that meet the minimal conditions of duration and sample size for inclusion, the results of sixteen have to be eliminated from consideration due to various design shortcomings.

3.1.2.1 Failure to track actual usage. In a study involving Japanese L2 English university students, Furuya, Kimura & Ohta (2004) evaluated the effects of practice exercises for the Test of English for International Communication (TOEIC) delivered to mobile phones via Simple Messaging Service (SMS). Although a pre-/post-test comparison confirmed a significant improvement, the actual usage of the program was not tracked nor was the possible influence of other external contributing factors. The failure to track actual usage also undermines the positive MALL results of Song (2008), who examined the effectiveness of a combination of a website program plus mobile phone SMS for L2 English vocabulary learning. Thematically related words were sent via SMS to volunteer Chinese adult learners. These were intended as concise reminders of the more extensive materials available on the website. Pre-/post-test results demonstrated a marginal improvement in performance. However, since the actual usage of neither the web-based program nor the SMS was tracked, it is not possible to attribute this gain to one factor or the other (or indeed either of the two). Similarly, Saran, Seferoglu & Cagiltay (2009) sought to measure improvement in L2 English pronunciation of Turkish prep-school students using a multimedia program which was delivered via Multimedia Messaging Service (MMS) on phones as well as online and in the form of a printed handout. Again, no actual usage data was collected, leaving attribution of the results in doubt. So, too, the fact that users of the audio-enhanced web-based program scored no better than those who only had access to a printed version of the program calls into question the extent to which students actually availed themselves of the alternative treatments.

3.1.2.2 Presence of uncontrolled variables. In Chen *et al.* (2009), Personal Digital Assistants (PDAs) enhanced with Radio Frequency ID (RFID) readers were used to provide L1 Chinese writing practice to primary school pupils in a collaborative, context aware, learning environment. The pre-/post-test results of the experimental group evidenced significant improvements in ten writing parameters. However, no learning outcome data are given for the control group which, in any event, was taught using a quite different pedagogical approach. Differences in pedagogical approach similarly undermine the findings of Oberg & Daniels (2013), who report the L2 English learning gains of Japanese university students who used an i-Pod Touch-based version of a textbook program in class. Whereas the experimental group was free to study the chapter contents at its own pace, in

https://doi.org/10.1017/S0958344014000159 Published online by Cambridge University Press

whatever sequence students chose, the control group was obliged to work with the textbook materials in lock-step fashion as directed by the instructor.

Thabit & Dehlawi (2012), working with Saudi Arabian university students, provided an experimental group of L2 English learners with MP4 players so they could watch self-study video-based learning materials. The control group, on the other hand, received no supplementary pedagogical materials. The experimental group significantly outperformed the control group on a final exam given four weeks later. These results, however, are very much called into question by the failure to take into consideration critical extraneous factors such as the MP4 program content relative to the tests and time on task.

Anaraki (2009) describes the design and development of a suite of twelve mobile phone Flash-based multimedia lessons for the learning of L2 English. The system was tested by Thai university students, who downloaded to their smartphones (or PDAs) three lessons a week for independent study. Pre-/post-testing was undertaken with the same twenty-question assessment, which confirmed a significant score increase for all students as well as an overall time on task reduction. However, these results are problematic in that the authors specify neither the language level of the students nor the language skills that were tested.

3.1.2.3 Inadequate control group descriptions. Başoğlu & Akdemir (2010) describe a pilot test that compared an L2 English flashcard application (ECTACO) used by Turkish university students to its printed counterpart used by a control group of the same size. Post-testing confirmed that using the flashcards on mobile phones was more effective in improving students' vocabulary learning than using flashcards on paper. Notwithstanding, these results are questionable in that the difference in gain made by the experimental group was only about five points on a scale of 100, with a very large standard deviation of 9.77 compared to 8.19 for the control group. Moreover, the only thing known about the control group is that it was taught using "traditional vocabulary acquisition techniques".

Baleghizadeh & Oladrostam (2010) investigated the effect of using mobile phones to record L2 English class discussions intended to elicit grammatical forms under review. Iranian university students made two-to-three-minute recordings of their speech on their mobile phones and as an out-of-class assignment analyzed their spoken mistakes and commented on them in a subsequent session. These students demonstrated significantly better grammatical accuracy on a twenty item multiple-choice grammar post-test compared to a control group of the same size that did not engage in these review activities. Once again, however, the validity of the reported results must be called into question because all that is known about the control group is that it received "the conventional way of grammar instruction".

Gabarre & Gabarre (2010) report the outcome of a mobile phone-based video recording project in a Malaysian university course for L2 French tourism and hospitality students. Using their phones, students worked together in groups of three or four to create a five-to-ten-minute narrated video promoting a Malaysian tourist attraction. According to instructor assessment, the videos submitted were of excellent quality with positive results reported in particular for pronunciation, vocabulary and grammar. The project, however, incorporated neither any pre-treatment assessment of student skills nor any control group in comparison to which learning gains could be established.

3.1.2.4 Presence of confounded variables. The claimed MALL learning gains in three studies are compromised by the effect of confounding variables. In Osman & Chung (2011), an experimental group of Malaysian university students was sent a variety of SMS intended to improve their L2 English communication skills. They were required to respond to these, then follow up on a class wiki. While records were kept of SMS responses and wiki usage, there is no way of attributing improvement in communicative competence to one as opposed to the other. Moreover, the learning conditions of the control group were not specified. In Al-Jarf (2012), an MP3-based self-study program was developed for Saudi Arabian university students to foster listening comprehension and speaking in L2 English. A pre-/post-test comparison confirmed a significant improvement. Since the program was accessible to students online via PCs as well as smartphones and MP3 players, and the devices used to access the self-study programs were not tracked, it is impossible to attribute the positive results to the use of any particular listening device.

Azabdaftari & Mozaheb (2012) report the results of a study that compared the L2 English vocabulary acquisition of Iranian university students. Half of these formed an experimental group that used a phone-based vocabulary program, the Spaced Repetition System (SRS), complemented by SMS exchanges with the instructor and Internet resources. The control group used printed flashcards containing English words with pronunciation on one side and corresponding L1/L2 equivalents on the other. Although the experimental group significantly outscored the control group on a twenty item multiple-choice post-test, it is not possible to attribute this to any one of the experimental conditions, all the more so since no information is provided about the contents or structure of the SRS, SMS or Internet resources. Moreover, no pre-test was undertaken to establish the L2 English competence of the experimental as opposed to the control group, it simply being presupposed that there were no significant differences between them at the outset.

3.1.2.5 Inadequate statistical analysis. Three studies focusing on English L2 vocabulary acquisition report positive findings, but fail to provide adequate statistical analysis of the data to substantiate the claim. Tan & Liu (2004) describe an experimental L2 English vocabulary learning system (MOBILE) for primary school children based on web-enabled student PDAs linked to a multimedia resource database on a teacher's notebook computer. Taiwanese students trialled the system, which allowed them to download learning materials, browse the web, take notes and do tutorial exercises. A series of six pre-/post-tests confirmed that use of the system resulted in significant vocabulary gains. However, the only data given are simple class averages with no statistical analysis of significant differences or standard deviations. Shimoyama & Kimura (2009) investigated the effectiveness of a flashcard program presenting English/Japanese word pairs as text with audio compared to the same with an additional graphics illustration or an example sentence. The study involved L2 English Japanese university students. For all but one group of five students, the scores of all subjects improved on a post-test, though there were no notable differences related to the different presentation formats. Once again, however, only raw test scores are given, with no analysis of statistical significance or standard deviations. Likewise, reporting positive results based only on group averages, Gutiérrez-Colon Plana, Gallardo Torrano & Grova (2012) describe a project that involved Spanish university students of L2 English. Via SMS the students were sent three exercises per week based on class content to which they were expected to respond immediately without consulting any outside resources. Students who

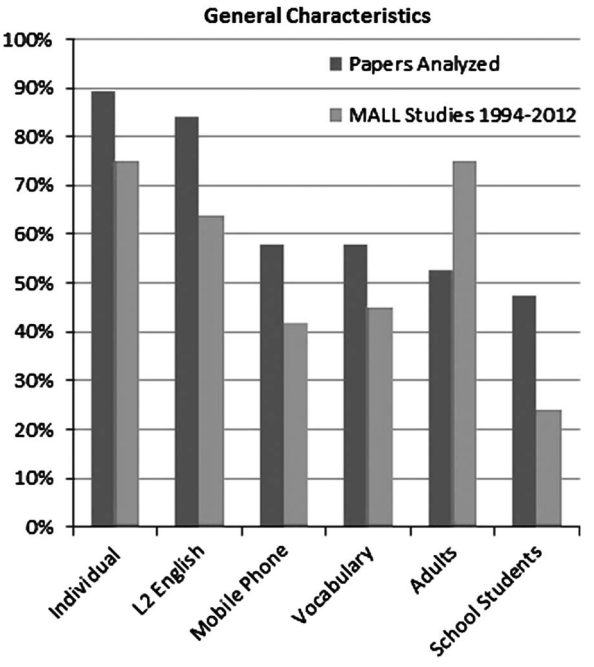

Fig. 3. General characteristics

took part in the project outscored a control group on a pre-/post-test comparison after the second semester. However, both groups scored considerably lower on the post-test than the pre-test, though the SMS group less so than the control. Besides the lack of statistical analysis, no indication is given about the activities of the control group except to say that it did not take part in the SMS project.

3.2 General characteristics of the studies analyzed

With the elimination of sixteen compromised studies, only nineteen MALL implementation projects remain that can reliably serve as a basis for determining the learning outcomes of mobile-based language applications. As can be seen in Figure 3, almost 90% (17/19) of the projects focus on individual students with nearly 85% (16/19) working on L2 English. Participants are nearly equally divided (53%/47%) between adult learners (10/19) and school students (9/19). At 58%, basic mobile phones (11/19) account for the majority of mobile devices used and vocabulary acquisition (11/19) is the single largest skill targeted. Compared to the full MALL implementation database of 347 projects (Burston, 2014), except for the proportion of adult versus non-adult subjects, the relative percentage of all these categories is greater in the studies under analysis.

Nearly three quarters of the projects (14/19) lasted between four and six weeks, with only one taking place over an entire term or semester (see Figure 4). Treatment durations, though acceptable for statistical analysis, thus remain short. At 37%, the most common sample size was between 25 and 49 subjects (7/19), with about a quarter (6/19) between 15 and 25 (see Figure 5). Group numbers are thus reasonable for statistical validity.

https://doi.org/10.1017/S0958344014000159 Published online by Cambridge University Press

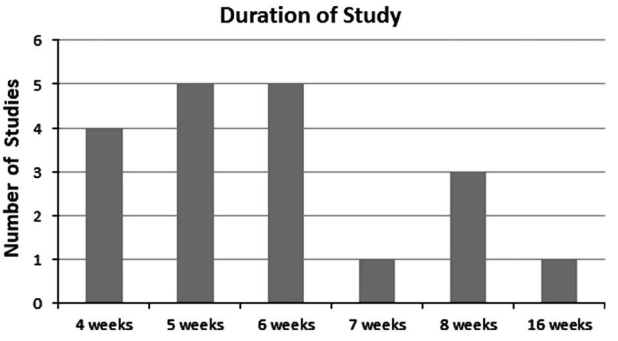

Fig. 4. Duration of study

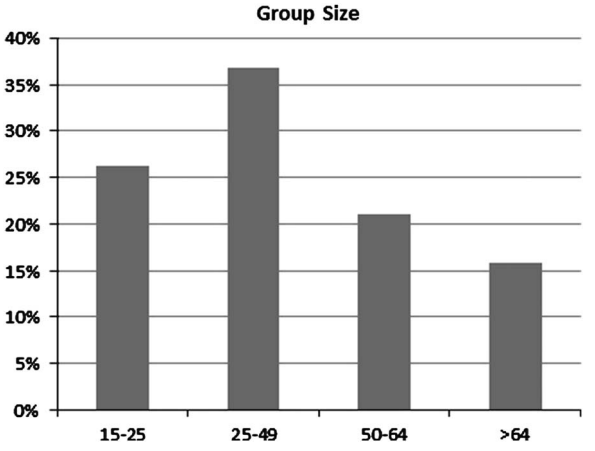

Fig. 5. Group sizes

With the exception of two studies from 2004, all of the projects were undertaken since 2008 with 63% (12/19) of the total appearing in 2011–2012 (see Figure 6). The database under analysis thus is much more representative of more recent studies.

3.3 Research findings

When doing a meta-analysis of research findings, it is normal procedure to calculate the effect size of results. An effect size statistically standardizes outcomes resulting from different sample sizes across studies so that they may be compared against a common yardstick. More specifically, it indicates the extent to which pre-/post-treatment groups differ in the mean values of a dependent variable at the end of a treatment phase. To be meaningful, effect sizes have to be based on properly designed research and, unfortunately, that is not the case for most of the following studies. For example, as can be seen in the appendix summary, nearly half (9/19) of the studies fail to specify the language level involved, leaving only ten of the studies to which an effect size analysis could potentially be applied with any confidence. Even these, however, are spread across five different

https://doi.org/10.1017/S0958344014000159 Published online by Cambridge University Press

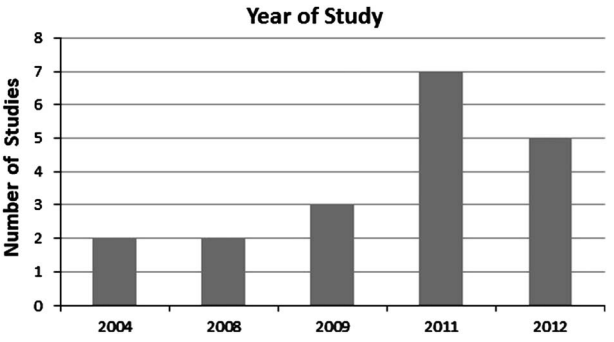

Fig. 6. Year of study

dependent variables (vocabulary, reading, speaking, listening, writing). So, too, the marked bias towards positive outcomes in the latter studies (9/10) calls into question the validity of any effect size calculation owing to the "file drawer effect" bias, i.e., studies with negative results not being published. For these reasons, effect sizes have not been calculated in the analysis which follows.

3.3.1 Vocabulary acquisition

3.3.1.1 SMS/MMS applications. As indicated in Figure 3, 58% (11/19) of the MALL studies under analysis are concerned with vocabulary acquisition. All of these projects targeted L2 English, nearly three-quarters of which involved the use of SMS or MMS on mobile phones (8/11) to teach simple L1/L2 word pairings, or L2 definitions, accompanied by example sentences.

3.3.1.2 Significant difference. With three exceptions, the results of L2 English vocabulary projects based on SMS/MMS programs report positive advantages. Three of these studies involved intermediate level Iranian L2 English learners. In one study (Motallebzadeh & Ganjali, 2011) for sixteen sessions, three times a week over a period of five weeks, 34 female university students were presented a total of 50 words with definitions and example sentences. Half of the group received these via SMS, the other half as a printed hand-out. Based on the results of a post-test, participants in the SMS group showed significantly better vocabulary retention than the ones in the printed paper group. In a second study (Motallebzadeh, Beh-Afarin & Daliry Rad 2011), for five weeks 40 female university students received twice weekly seven collocations with definitions and example sentences. Half of the group received these via SMS, the other half as a printed hand-out. Students took two quizzes in the same format as the presentation mode. Participants in the SMS group showed significantly better vocabulary retention than the ones in the printed paper group. The third Iranian study (Tabatabaei & Goojani, 2012) was a two-month project involving 60 male Iranian high school pupils who studied in class five or six English synonyms and antonyms per week, twice a week, for ten weeks. Out of class, each week students wrote one sentence for each word with half of the students sending these to their instructor and three fellow students via SMS, to which the instructor replied more or less

https://doi.org/10.1017/S0958344014000159 Published online by Cambridge University Press

immediately with corrections. The other half of the class, serving as a control group, did likewise by submitting their sentences to the instructor and exchanging written papers in class. While both groups showed higher performance on a post-test, the experimental group scored significantly better than the control group.

Two MMS-based projects also report positive MALL learning outcomes. The four-week long project of Saran, Seferoglu, & Cagiltay (2012) involved a total of 103 Turkish prep-school students, 53 of whom were beginners and 50 pre-intermediate level L2 English learners. In this study, the effectiveness of using MMS in learning vocabulary was compared to delivery through web pages and printed form. The MMS included the definitions of words, exemplary sentences, related visual representations, word formation information, and pronunciation. A pre-/post-test comparison confirmed that students in the MMS group at both levels learned more words than those who studied the web- and paper-based materials. Lin & Yu (2012) report on a MMS-based L2 English vocabulary learning program that was trialled by 32 Taiwanese junior high school pupils for four weeks. Nine words a week were delivered in one of four modes: text (syntactic category, Chinese translation, example sentence), text plus audio (word/sentence pronunciation), text plus image, and text plus audio plus image. Learning gains are claimed for all four conditions, but the effects of different presentation modes on vocabulary learning were not significantly different.

3.3.1.3 No significant difference. In two studies, the outcomes of delayed post-tests are at variance with the findings of immediate post-testing. In a Chinese study comparing the effectiveness of SMS versus printed presentation (Zhang, Song & Burston, 2011), one group of 32 L2 English university students studied a total of 130 words delivered via SMS five-at-a-time twice daily for 26 days. A control group of 30 received the same vocabulary on a printed word list, which participants studied at their own pace. Both groups showed significant improvement on an immediate post-test, with the SMS students outperforming the control group. However, a delayed test only four days later indicated no significant difference in vocabulary retention rates between the groups. In the Iranian project reported in Alemi, Sarab & Lari (2012), 28 intermediate level L2 English university students received ten words and example sentences twice a week via SMS for sixteen weeks. Their learning of 320 head words was compared to that of a control group of seventeen who studied the same words using only a dictionary. Both groups improved on a post-test, but with no significant difference between the two. It is claimed, however, that the SMS group showed significantly better vocabulary retention on a delayed post-test four weeks later. Inasmuch as delayed post-test scores seldom match, let alone exceed, those of immediate post-testing, these results are remarkably out of line with normal expectations.

Derakhshan & Kaivanpanah (2011) is the only study which shows no significant differences on any performance evaluation. The authors describe a seven-week program that used SMS for L2 English vocabulary acquisition with Iranian university students. An experimental group of 21 and a control group of 22 were both taught 15–20 words per session. Students wrote one sentence for each word for their instructor and three classmates. The experimental group sent these via SMS and the control students brought them to class on paper. A post-test and a delayed post-test administered two weeks later both showed no significant difference in word retention between the two groups.

3.3.1.4 PDA/Smartphone applications. In the remaining three vocabulary studies, which involved PDAs and smartphones, the outcomes gave mixed results. Chen & Chung (2008) developed an L2 English vocabulary learning system based upon Item Response Theory algorithms and a learning memory cycle. It operated via PDAs linked to a remote management server, client mobile learning system and three database agents: one that recommended vocabulary, one that generated tests, and one that assessed performance. The system was trialled by fifteen advanced level Taiwanese university students for five weeks and the results revealed significant, though modest (~5%), pre-/post-test enhancement of vocabulary abilities. A more recent PDA-based vocabulary study is that of Hwang & Chen (2013), published online in 2011. This Taiwanese project describes the learning of beginners' level L2 English in a situated learning environment by primary school children using a PDA-based multimedia program designed to teach vocabulary. This was done in the context of activities in which subjects listened to lessons and recorded their reading of basic words and the completion of simple sentences having to do with their lunch menu. A group of 30 pupils trialled the system during their lunch hour, four days per week, for two months. This group made significantly higher gains in their vocabulary acquisition as well as listening and speaking skills compared to a control group of equal size who studied without PDA support.

Vocabulary learning outcomes for the smartphone-based project (Fisher *et al.*, 2009) was linked to reading activities. It compared the effect upon L2 English vocabulary acquisition using paper books, e-books with dictionaries, and e-books with adaptive software (ELMO). The experiment was conducted over a period of six weeks with three groups of thirteen Japanese high school students, each of which used all three resources for two weeks. The end result was essentially no vocabulary gains for any of the participants. Most students read only three pages or less out of some 100 pages in each book and learned, on average, only one new word over each two-week period, regardless of the technology.

3.3.2 Reading competency. In addition to vocabulary acquisition, the database under analysis also includes four MALL projects involving reading competency, all of which were PDA-based and all of which showed a positive advantage from use of the MALL application. The two earliest, Zurita & Nussbaum (2004a, 2004b), describe the experimental use of wirelessly linked PDAs to foster the pre-reading word construction ability of Chilean L1 Spanish primary school children. Both trials ran for twenty days and took place entirely in class, during which the subjects worked collaboratively in triads to construct Spanish words from three syllables given on printed cards and via a PDA program. The first trial involved 21 children, all of whom worked in sessions that lasted 35–45 minutes, first with printed cards then with the MALL application. In the second trial, half of a group of 24 children worked with the PDA program in sessions that lasted fifteen minutes while the other half served as a control working only with printed cards in sessions that lasted 25 minutes. In the post-tests of both trials, significant learning gains were noted for both conditions compared to pre-tests, though these were substantially greater when using the PDA program. The second trial also demonstrated that the PDA group made greater learning gains with less time on task and less teacher support.

The remaining two MALL reading projects involved Taiwanese L2 English university students. Chen & Hsu (2008) experimented with a prototype web-enabled PDA-based reading/vocabulary system (PIMS) that was trialled by fifteen advanced level students for

five weeks. Using a fuzzy Item Response Theory algorithm that determined users' reading abilities, PIMS recommended English news articles to learners and automatically identified unfamiliar words for study. A pre-/post-test comparison confirmed a significant gain in reading comprehension ability. More recently, Wu *et al.* (2011) developed a server-based system accessible via PDAs (and smartphones) equipped with radio frequency id (RFID) tag readers and WiFi network connectivity to provide learners with location-appropriate L2 English texts to read. The system offered translations, pronunciation and explanations of words, sentences, paragraphs, and articles. Additionally, a reading guidance algorithm proposed texts based on a dynamically maintained learner portfolio. The system was trialled for eight weeks by three groups of students, one consisting of 38 subjects who read texts in printed form only with no environmental contextualization, one of 39 who used the location-aware reading program without the guidance function and one of 36 who used the reading program with the guidance function. Whereas the post-test reading scores of the paper-only group increased by only about one point on a scale of 100, that of the two experimental groups improved substantially, more than sixteen points without the guidance function and nearly twenty points with it.

3.3.3 Listening/speaking skills. Listening and speaking skills were the focus of three MALL projects which used mobile devices as audio recording tools. All of these studies report learning gains for the MALL application. The most technologically and pedagogically ambitious of these is Liu (2009), which describes the pilot testing of a server-based mobile learning system (HELLO) for L2 English. It consisted of three games, two of which involved location-aware task-based activities: one was played individually with a virtual learning tutor, the other collaboratively with other learners who used their mobile devices to record their interactions. The system was trialled in a Taiwanese school for eight weeks by 64 seventh graders equally divided into an experimental group which accessed the HELLO system via telephone-enabled PDAs and a control group which used only classroom resources (printed materials, CD/MP3 players, digital voice recorders). All test results of the HELLO group were significantly better than those of the control group.

In a study involving eleven Australian post-primary schools, Robertson *et al.* (2009) describe a six-week pilot test of the commercial mobile phone-based *Learnosity* language learning system that was trialled by a total of 95 L2 Indonesian students from grades 7 to 11. The project involved students viewing stimulus materials (photographs, a map, a menu, a travel brochure), listening via a mobile phone to questions in Indonesian about those materials, and recording their oral responses in Indonesian. Compared to a pre-test, students' listening/speaking scores on a post-test improved about eleven percent overall.

Papadima-Sophocleous *et al.* (2012) report the results of an experiment that sought to measure the impact of the iPod Touch upon L2 English oral reading skills. The six-week project involved fifteen intermediate level Cypriot university students who downloaded three texts with accompanying audio recordings that served as models of pronunciation. Participants used the iPods to listen to the models and record their own pronunciation. Based on a comparison of the first versus second readings of the three texts, the iPod-supported activity helped students significantly increase their automaticity in speed and the accuracy of the segmental and prosodic features of their oral reading.

3.3.4 Writing skills. Hwang, Chen & Chen (2011) describe a situated learning system that operated on an unspecified mobile device. The program included vocabulary, phrases, and sentence patterns designed to help Taiwanese elementary school children create beginning level written L2 English sentences. The six-week study compared the results of 28 pupils who used the system with 31 who did not. There was a significant difference in pre-/post-test writing performance between the two groups in favor of the experimental condition in terms of such factors as the number of sentences produced, reasoning, communication, and organization.

4 Conclusion

Despite the hundreds of MALL publications over the past twenty years, statistically reliable measures of learning outcomes are few and far between. In part, this is due to the fact that well over half of all MALL related studies report no objectively quantifiable learning outcomes, either because they did not involve MALL implementation projects, or if they did, learning gains were only based on subjective teacher assessments and/or student self-evaluations. Even more so, the paucity of statistically reliable learning outcome data stems from the short duration of projects and small numbers of students involved. Of the 291 distinct studies examined in this review only 35 meet minimal conditions of duration and sample size, i.e., ten experimental subjects over a period of at least a month. Sixteen of these suffer from serious design shortcomings, leaving only nineteen MALL studies that can reliably serve as a basis for determining the learning outcomes of mobile-based language applications. Of these studies, fifteen can be considered to report unequivocal positive results, with those focusing on reading, listening and speaking without exception evidencing a MALL application advantage. Four studies, all focusing on vocabulary, reported no significant differences.

In sum then, from what little is known with reasonable certainty, the learning outcomes of MALL implementations are unquestionably positive in nearly 80% of the cases. That being said, the difficulty involved in extracting such modest findings from MALL publications underscores two fundamental problems that permeate MALL evaluation studies: inadequate research design and technocentricity. As was the case during the first twenty years of CALL research, the majority of those undertaking MALL evaluation studies evidence a serious lack of training in experimental research methods. As Pederson (1988) concluded in reference to the first two decades of CALL research, the most pervasive failing of MALL evaluation studies is the nonreplicability of reported results. With the focus so much on the technology, scant attention is paid to a host of unacknowledged and uncontrolled variables that could influence learning outcomes as much as, if not more than, mobile device usage itself: novelty effects, actual content, the nature of feedback, the personal influence of the instructor, learner expectations, motivation, etc. Almost without exception, MALL implementation studies have fallen into the trap of attempting to attribute learning gains to the technology itself rather than to the way the technology was manipulated to affect achievement.

So, too, technocentricity is largely responsible for the lack of pedagogical innovation and failure of even the most recent MALL projects to exploit the communicative affordances of mobile devices. Nearly all presuppose a behavioristic paradigm involving rote learning and structuralistic tutorial exercises. While a few projects attempt to incorporate context

awareness and situated learning (Chen *et al.*, 2009; Liu, 2009; Wu *et al.*, 2011; Hwang, Chen & Chen, 2011; Hwang & Chen, 2013), only one in this review (Liu, 2009) involves out-of-class inter-student communication. What Godwin-Jones (2011) observed, in summarizing the findings of Kukulska-Hulme & Shield (2008), is equally true of the studies examined here:

> *… for the most part uses of mobile devices were pedestrian, uncreative, and repetitive and did not take advantage of the mobility, peer connectivity, or advanced communication features of mobile devices. Most activities were teacher-led and scheduled, not leveraging the anytime, anyplace mobile environment. Oral interactions and learner collaboration were infrequently used. The problem is less one of hardware/software shortcomings and more in developers' conceptualization of how language learning could be enhanced in new, innovative ways with the assistance of mobile devices.* (*op. cit.*: 7)

MALL thus has a long way to go to realize its pedagogical potential and justify the current interest in mobile-assisted learning. As more recent and innovative MALL implementations attest (Tai, 2012), it is possible to effectively exploit mobile devices in conformity with learner-centered, constructivist, collaborative methodologies. There is every reason to expect that MALL can make significant contributions to improving language learning, most particularly by increasing time spent on language acquisition out of class, by exploiting mobile multimedia facilities to complete task-based activities, and by using the communication affordances of mobile devices to promote collaborative interaction in the L2.

However, as the pedagogical approach and technological exploitation of MALL projects improve, so too must the reporting of their outcomes. MALL proponents have much work to do in learning how to undertake methodologically sound, statistically reliable studies that account for more than just technology usage.

References

Al-Jarf, R. (2012) Mobile technology and student autonomy in oral skill acquisition. In: Díaz-Vera, J. (ed.), *Left to my own devices: Learner autonomy and mobile-assisted language learning innovation and leadership in English language teaching*. Bingley, UK: Emerald Group Publishing Limited, 105–130.

Alemi, M., Sarab, M. and Lari, Z. (2012) Successful learning of academic word list via MALL: Mobile Assisted Language Learning. *International Education Studies*, **5**(6): 99–109.

Anaraki, F. (2009) A Flash-based mobile learning system for learning English as a second language. *Proceedings, International Conference on Computer Engineering and Technology*. Singapore, 400–404. http://www.journal.au.edu

Azabdaftari, B. and Mozaheb, M. (2012) Comparing vocabulary learning of EFL learners by using two different strategies: mobile learning vs. flashcards. *The EUROCALL Review*, **20**(2): 47–59. http://eurocall.webs.upv.es/index.php?m=menu_00&n=news_20_2#mozaheb

Baleghizadeh, S. and Oladrostam, E. (2010) The Effect of Mobile Assisted Language Learning (MALL) on Grammatical Accuracy of EFL Students. *MEXTESOL Journal*, **34**(2): 77–86.

Başoğlu, E. and Akdemir, O. (2010) A comparison of undergraduate students' English vocabulary learning: Using mobile phones and flash cards. *Turkish Online Journal of Educational Technology*, **9**(3): 1–7.

Brown, M., Castellano, J., Hughes, E. and Worth, A. (2012) Integration of iPads in a Japanese university's freshman curriculum. *Proceedings of the JALT CALL Conference 2012*. http://journal.jaltcall.org

Burston, J. (2013) Mobile-Assisted Language Learning: A selected annotated bibliography of implementation studies 1994–2012. *Language Learning & Technology, Special volume on Mobile-Assisted Language Learning*, **17**(3): 157–224.

Burston, J. (2014) The reality of MALL project implementations: Still on the fringes. *CALICO Journal*, **31**(1): 43–65. https://www.calico.org

Chen, C-M. and Chung, C-J. (2008) Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. *Computers & Education*, **51**(2): 624–645.

Chen, C-M. and Hsu, S-H. (2008) Personalized intelligent mobile learning system for supportive effective English learning. *Educational Technology and Society*, **11**(3): 153–180.

Chen, T-S., Chang, C-S., Lin, J-S. and Yu, H-L. (2009) Context-aware writing in ubiquitous learning environments. *Research and Practice in Technology Enhanced Learning*, **4**(1): 61–82.

Chiang, M-H. (2012) Effects of reading via Kindle. In: Colpaert, J., Aerts, A., Wu, W-C. V. and Chao, Y-C. J. (eds.), *The Medium Matters. Proceedings, 15th International CALL Conference*. Taichung, Taiwan: Providence University, 176–179.

Chinnery, G. (2006) Going to the MALL: Mobile Assisted Language Learning. *Language Learning & Technology*, **10**(1): pp. 9–16.

Derakhshan, A. and Kaivanpanah, S. (2011) The impact of text-messaging on EFL freshmen's vocabulary learning. *The EUROCALL Review*, **19**: 39–47.

Fisher, T., Pemberton, R., Sharples, M., Ogata, H., Uosaki, N., Edmonds, P., Hull, A. and Tschorn, P. (2009) Mobile learning of vocabulary from reading novels: A comparison of three modes. In: Metcalf, D., Hamilton, A. and Graffeo, C. (eds.), *Proceedings of 8th World Conference on Mobile and Contextual Learning*. Orlando, FL: University of Central Florida, 191–194.

Furuya, C., Kimura, M. and Ohta, T. (2004) Mobile language learning - A pilot project on language style and customization. In: Richards, G. (ed.), *E-Learn 2004, Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*. Chesapeake, VA: Association for the Advancement of Computers in Education, 1876–1880. http://www.editlib.org

Gabarre, C. and Gabarre, S. (2010) An innovative assessment method for real world learning: Learner created content with a cell phone, YouTube and an LMS. *Proceedings of Global Learn Asia-Pacific 2010 - Global Conference on Learning and Technology* Penang, Malaysia: AACE, 1202–1210. http://www.editlib.org

Godwin-Jones, R. (2011) Emerging technologies: Mobile apps for language learning. *Language Learning & Technology*, **15**(2): 2–11.

Gutiérrez-Colon Plana, M., Gallardo Torrano, P. and Grova, M. (2012) SMS as a learning tool: an experimental study. *The EUROCALL Review*, **20**(2): 33–47.

Hwang, W-Y. and Chen, H. (2013) Users' familiar situational contexts facilitate the practice of EFL in elementary schools with mobile devices. *Computer Assisted Language Learning*, **26**(2): 101–125.

Hwang, W-Y., Chen, C-Y. and Chen, H. (2011) Facilitating EFL writing of elementary school students in familiar situated contexts with mobile devices. *Proceedings, 10th World Conference on Mobile and Contextual Learning (mLearn)*. Beijing, China: Beijing Normal University, 15–23. http://mlearn.bnu.edu.cn

Kukulska-Hulme, A. and Shield, L. (2008) Overview of Mobile Assisted Language Learning: Can mobile devices support collaborative practice in speaking and listening? *ReCALL*, **20**(3): 271–289.

Lin, C-C. and Yu, Y-C. (2012) Learning English vocabulary on mobile phones. In: Colpaert, J., Aerts, A., Wu, W-C. V. and Chao, Y-C. J. (eds.), *The Medium Matters. Proceedings, 15th International CALL Conference*. Taichung, Taiwan: Providence University, 416–420.

Liu, T-Y. (2009) A context-aware ubiquitous learning environment for language listening and speaking. *Journal of Computer Assisted Learning*, **25**(6): 515–527.

Motallebzadeh, K., Beh-Afarin, R. and Daliry Rad, S. (2011) The effect of short message service on the retention of collocations among Iranian lower intermediate EFL learners. *Theory and Practice in Language Studies*, **1**(11): 1514–1520.

Motallebzadeh, K. and Ganjali, R. (2011) SMS: Tool for L2 vocabulary retention and reading comprehension ability. *Journal of Language Teaching and Research*, **2**(5): 1111–1115.

Oberg, A. and Daniels, P. (2013) Analysis of the effect a student-centred mobile learning instructional method has on language acquisition. *Computer Assisted Language Learning*, **26**(2): 177–196.

Osman, M. and Chung, P. (2011) Language learning using texting and wiki: A Malaysian context. *e-CASE & e-Tech International Conference*. Knowledge Association, 1888–1903.

Papadima-Sophocleous, S., Georgiadou, O. and Mallouris, Y. (2012) iPod impact on oral reading fluency of university ESAP students. *Proceedings, GLoCALL Conference*. Beijing, China (CD-ROM).

Pederson, K. (1988) Research on CALL. In: Smith, W. (ed.), *Modern Media in Foreign Language Education: Theory and implementation*. Lincolnwood, IL: National Textbook Co., 99–131.

Robertson, L. and The Le@rning Federation. (2009) *Mobile application for language learning: MALL Research Project Report*. Adelaide, South Australia: Curriculum Corporation, 1–48. http://www.ndlrn.edu.au/verve/_resources/MALL_Report_2009.pdf#search=Robertson

Saran, M., Seferoglu, G. and Cagiltay, K. (2009) Mobile assisted language learning: English pronunciation at learners' fingertips. *Eurasian Journal of Educational Research*, **34**: 97–114.

Saran, M., Seferoglu, G. and Cagiltay, K. (2012) Mobile language learning: Contribution of multimedia messages via mobile phones in consolidating vocabulary. *The Asia-Pacific Education Researcher*, **21**(1): 181–190.

Shimoyama, Y. and Kimura, M. (2009) Development of and effectiveness in vocabulary learning content for mobile phones in Japan. *World CALL 2008 Conference*. Kyushu-Okinawa, Japan: The Japan Association for Language Education and Technology, 138–141. http://www.j-let.org

Song, Y. (2008) SMS enhanced vocabulary learning for mobile audiences. *International Journal of Mobile Learning and Organisation*, **2**(1): 81–98. http://inderscience.metapress.com

Tabatabaei, O. and Goojani, A. (2012) The impact of text messaging on vocabulary learning of Iranian EFL learners. *Cross Cultural Communication*, **8**(2): 47–55.

Tai, Y. (2012) Contextualizing a MALL: Practice design and evaluation. *Educational Technology & Society*, **15**(2): 220–230.

Tan, T-H. and Liu, T-Y. (2004) The mobile-based interactive learning environment (MOBILE) and a case study for assisting elementary school English learning. *Proceedings of the 2004 IEEE International Conference on Advanced Learning Technologies*. Los Alamitos, CA: IEEE Computer Society, 530–534. http://ieeexplore.ieee.org

Thabit, K. and Dehlawi, F. (2012) Towards using MP4 players in teaching English language: An empirical study. *Journal of Engineering*, **2**(8): 25–28.

Wu, T., Sung, T., Huang, Y., Yang, C. and Yang, J-C. (2011) Ubiquitous English learning system with dynamic personalized guidance of learning portfolio. *Educational Technology & Society*, **14**(4): 164–180.

Zhang, H., Song, W. and Burston, J. (2011) Reexamining the effectiveness of vocabulary learning via mobile phones. *Turkish Online Journal on Educational Technology*, **10**(3): 203–214.

Zurita, G. and Nussbaum, M. (2004a) Computer supported collaborative learning using wirelessly interconnected handheld computers. *Computers & Education*, **42**(3): 289–314.

Zurita, G. and Nussbaum, M. (2004b) A constructivist mobile learning environment supported by a wireless handheld network. *Journal of Computer Assisted Learning*, **20**: 235–243.

Appendix

Summary of MALL Studies Analyzed

Author(s)	Language	Level	L1/L2	Language Focus	Sample Size	Project Duration	Learning Environment	Result
Motallebzadeh & Ganjali 2011	English	intermediate	L2	vocabulary	17 experimental 17 control	5 weeks	university	+
Motallebzadeh, Beh-Afarin & Daliry Rad 2011	English	intermediate	L2	vocabulary	20 experimental 20 control	5 weeks	university	+
Tabatabaei & Goojani 2012	English	intermediate	L2	vocabulary	30 experimental 30 control	2 months	high school	+
Saran, Seferoglu & Cagiltay 2012	English	beginner	L2	vocabulary	18, 17, 18	4 weeks	prep-school	+
		pre-intermediate			17, 17, 16			+
Lin & Yu 2012	English		L2	vocabulary	32	4 weeks	junior high school	+
Zhang, Song & Burston 2011	English		L2	vocabulary	32 experimental 30 control	26 days	university	NSD
Alemi, Sarab & Lari 2012	English	intermediate	L2	vocabulary	28 experimental 17 control	16 weeks	university	NSD
Derakhshan & Kaivanpanah 2011	English		L2	vocabulary	21experimental 22 control	7 weeks	university	NSD
Chen & Chung 2008	English	advanced	L2	vocabulary	15	5 weeks	university	+
Hwang & Chen 2013	English	beginner	L2	vocabulary	30 experimental 30 control	2 months	primary school	+
Fisher *et al.* 2009	English		L2	vocabulary	13 experimental 13 experimental 13 experimental	6 weeks	High school	NSD
Zurita & Nussbaum 2004a	Spanish		L1	reading	21 experimental 20 control	20 days	primary school	+
Zurita & Nussbaum 2004b	Spanish		L1	reading	12 experimental 12 control	20 days	primary school	+
Chen & Hsu 2008	English	advanced	L2	reading	15	5 weeks	university	+
Wu *et al.* 2011	English		L2	reading	38 experimental 39 experimental 36 experimental	8 weeks	university	+
Liu 2009	English		L2	listening/speaking	32 experimental 32 control	8 weeks	post-primary school	+
Robertson *et al.* 2009	Indonesian		L2	listening/speaking	95	6 weeks	post-primary school	+
Papadima-Sophocleous *et al.* 2012	English	intermediate	L2	listening/speaking	15	6 weeks	university	+
Hwang, Chen & Chen 2011	English	beginner	L2	writing	28 experimental 31 control	6 weeks	primary school	+