ReCALL (2022), 34: 2, 147–168
doi:10.1017/S0958344021000240

RESEARCH ARTICLE

MALL language learning outcomes: A comprehensive meta-analysis 1994–2019

Jack Burston
Cyprus University of Technology, Cyprus (jack.burston@cut.ac.cy)

Konstantinos Giannakou
European University Cyprus, Cyprus (K.Giannakou@euc.ac.cy)

Abstract

The aim of this study is to comprehensively evaluate quantitative experimental mobile-assisted language learning (MALL) studies published between 1994 and 2019 that meet minimal conditions of research design and statistical analysis. Starting with a bibliographical database of 1,144 references to experimental MALL implementations, of which there were 700 objectively substantiated by quantitative experimental language learning outcomes, only 84 experimental MALL studies met the inclusion requirements. Their analysis addresses two critical sets of research questions. First, what are the general characteristics of the selected studies and, second, what are their language learning outcomes in terms of measured effect size. Nine general characteristics are considered: publication source, chronological distribution, country of origin, institutional environment, sample size, intervention duration, targeted language, language learner competence level, and learning focus. Effect size was calculated separately for between-group (independent, experimental) and within-group (quasi-experimental) treatment studies. In both cases, the overall results were quite large: 0.72 for the former and 1.16 for the latter. An analysis of four critical moderator variables (language learner competence level, language area focus, institutional environment, and intervention duration) revealed similarly large effect sizes. Notwithstanding, analysis of the data also confirmed obvious publication bias and a very high level of heterogeneity that frequently approached 100%. The relevance of positive language learning outcome conclusions thus needs to be tempered by these shortcomings.

Keywords: mobile-assisted language learning; MALL; language learning outcomes; effect size; research design

1. Introduction

Mobile-assisted language learning (MALL) has been the subject of over 3,800 studies since the first article appeared in 1994. Moreover, in the last dozen years, 58 overviews and meta-analyses of MALL implementation studies have been published, 35 in the past three years (see Appendix 1 in the supplementary material). Although these experimental studies have covered a considerable variety of topics, learning outcomes have quite naturally been a frequent focus of attention. In fact, they are considered in 78% (45/58) of the meta-analyses. Eleven meta-analyses are specifically devoted to the evaluation of learning outcomes.

By all accounts, MALL learning outcomes are overwhelmingly positive, which would make anyone wonder why MALL overviews continue to be undertaken with such frequency. More

Cite this article: Burston, J. & Giannakou, K. (2022). MALL language learning outcomes: A comprehensive meta-analysis 1994–2019. *ReCALL* 34(2): 147–168. https://doi.org/10.1017/S0958344021000240

© The Author(s), 2021. Published by Cambridge University Press on behalf of European Association for Computer Assisted Language Learning. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is included and the original work is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use.

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

pertinently, it begs the question of what justifies this meta-analysis in particular. Given the considerable attention that has been accorded to them, it could be assumed that positive MALL learning outcomes are beyond dispute. In fact, as this meta-analysis endeavors to demonstrate, all previous overviews have been based on very incomplete data. Some privilege MALL studies that have appeared in a small number of computer-assisted language learning (CALL) journals. Others add publications in educational technology journals. Still others rely on academic databases for their bibliographical resources. Almost none take into consideration MALL studies that have been published in mobile technology journals. And even fewer consider research undertaken in master's or doctoral dissertations. Likewise, MALL studies written in languages other than English are all but completely ignored. Needless to say, such omissions call into question the validity of previous evaluations of MALL learning outcomes.

When one of the most comprehensive MALL overviews (Burston, 2015) was reanalyzed based on a much more extensive research database, although many observations were confirmed, significant discrepancies were also revealed. For example, the proportion of studies with longer duration was shown to be considerably greater than previously reported. Despite the continued preponderance of mobile phone–based studies, a much smaller proportion of these involved SMS/MMS than previously claimed. Positive learning outcomes for vocabulary acquisition were even greater in the extended database, but this needed to be viewed against a much more varied learning focus than previously acknowledged. In view of the demonstrated shortcomings in the research coverage in Burston (2015), the accuracy of the learning outcomes reported in other previous MALL meta-analyses, however numerous, cannot be taken for granted. It is for this reason that the following comprehensive meta-analysis of experimental MALL learning outcomes has been undertaken.

To ensure maximum coverage, the intent of this meta-analysis is, first, to compile data from the most extensive range of bibliographic resources possible, most particularly through the systematic extraction of references from MALL studies themselves. Second, through a process of comprehensive keyword searches within these sources, this study identifies descriptions of experimental MALL implementations. The quantitative basis of this meta-analysis is then complemented by a qualitative assessment of the studies so identified. Through a close reading, these experimental interventions are subjected to a set of rigorous inclusion/exclusion criteria to ensure that they meet the requirements of sample size, treatment duration, research design, and statistical analysis. The language learning results of studies that meet all the selection requirements are then pooled to determine the effect size (ES) of outcomes. Comparisons with previous MALL meta-analyses are made where this is relevant to the research database underlying this study and the inclusion/exclusion criteria of studies selected for the analysis of language learning outcomes. However, no detailed analysis of previous meta-analyses is undertaken as each is based on a very different database set, both quantitatively and qualitatively, thus precluding any meaningful comparisons.

2. Methodology

2.1 Definition: Experimental MALL implementation

The undertaking of a comprehensive meta-analysis of MALL language learning outcomes requires, first, a clear definition of what constitutes an experimental MALL implementation. In this meta-analysis, an experimental MALL implementation is identified as one that involves the application of mobile-based or mobile-accessible apps and/or mobile device affordances (e.g. audio/video recording, picture/note-taking) for the teaching and learning of languages in a defined learning environment with specified participants and learning conditions. Specifically excluded from the present investigation are the many studies having to do simply with application designs, prototype evaluations, mobile app reviews, mobile device ownership, teacher training, institutional infrastructure, instructional technology needs, motivational effects, and general surveys of teacher and student perceptions of MALL unrelated to any specific MALL implementation.

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

2.2 Underlying MALL studies database

A meta-analysis of experimental MALL implementation outcomes is only as comprehensive as the research database from which it derives. The bibliography underlying this investigation consists of 3,503 MALL studies of all types, from all sources, written in any language and targeting any language, first language (L1) as well as second language (L2), which appeared from 1994 through to the end of 2019. The starting point is, thus, Callan (1994), the first publication about the use of handheld computers to support language acquisition. The end point of this study is determined by the inevitable time lag encountered in tracking down and analyzing MALL references, 2019 being the most recent complete year for which a comprehensive bibliography could be compiled.

Bibliographical references were sought through two processes: manual bibliography mining and extensive searches of academic databases. Bibliography mining essentially involves recursively extracting bibliographies from published MALL studies. It provides comprehensive, narrowly targeted MALL references from sources across a very broad spectrum. However, because of the inevitable time lag involved with published studies, such references are less current than academic databases, which are updated on an ongoing basis. Together, the two processes complement each other well and provide an effective combination of relevancy and currency.

The original source of the bibliography mining was the 575 references underlying Burston (2013). This was augmented by the references from the 58 MALL meta-analyses published between 2006 and 2020. The references in these studies were manually searched using all the obvious keywords: *mobile-assisted language learning*, *MALL*, *m-learning*, *mobile learning*, *language learning*, *mobile device*, *mobile phone*, *iPod*, *iPad*, *iPhone*, *smartphone*, *tablet*. In the process, other less obvious keywords also came to light: *ubiquitous*, *seamless*, *flipped*, *augmented reality*, *virtual reality*, *audience response system*, *student response system*, *clicker*, *digital pen*, *wearable*. The references extracted from published MALL studies were then completed by a similar keyword search of five academic databases: SCOPUS, SSCI, ERIC, ScienceDirect, and ProQuest Dissertations. In addition, mentions of papers in Researchgate.net and Academia.edu citing Burston's MALL publications provided a substantial ongoing stream of studies for further bibliographical searches. Of the total 3,503 MALL studies, 67 proved to be duplicates, typically PhD/MA theses or conference presentations/proceedings subsequently reappearing as a journal publication. The remaining 3,436 distinct studies form the basis of this meta-analysis. In all, it was possible to obtain and consult 85% (2,930/3,436) of the distinct MALL studies that appeared through 2019. As no restriction was placed on the languages in which these studies were written, those which the authors did not know were converted using Google Translate and analyzed in English.

As described in Burston (2021), the sources of MALL studies are many and varied. The most obvious are CALL, educational technology and mobile technology journals, and associated conference presentations. The distribution of MALL studies in the more prominent of these journals is summarized in Table 1.

As can be seen, between 1999 and 2019, 359 MALL studies appeared in these sources. Notwithstanding, this represents just 10% (359/3,436) of the distinct MALL studies that appeared through to the end of 2019. The remaining MALL studies are mostly to be found in other journals and conferences, the majority of which have nothing to do with language teaching or learning (Burston, 2021). A small number also appear in PhD dissertations, master's theses, and sundry reports.

2.3 Experimental MALL implementation database

2.3.1 Consulted studies

Of the 3,436 distinct studies that could be identified for this meta-analysis, only 1,144 meet the definition of experimental MALL implementation previously indicated. Based on the titles and abstracts of publications that were not obtainable, an additional 118 MALL studies also appear

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

Table 1. Prominent MALL studies sources

CALL	#Studies	Educational technology	#Studies	Mobile technology	#Studies
CALICO Journal 2005–2019	15	*British Journal of Educational Technology* 2009–2019	20	*International Journal of Interactive Mobile Technologies* 2007–2019	30
Computer Assisted Language Learning 2000–2019	62	*Computers & Education* 2004–2019	30	*International Journal of Mobile and Blended Learning* 2009–2019	20
International Journal of Computer-Assisted Language Learning and Teaching 2011–2019	21	*Educational Technology & Society* 2008–2019	34	*International Journal of Mobile Learning and Organisation* 2007–2019	27
JALT CALL Journal 2005–2019	28	*System* 2005–2019	10		
Language Learning & Technology 1999–2019	35				
ReCALL 2005–2019	27				
Total # studies	188		94		77

to conform to this definition. The consulted studies thus represent 91% (1,144/1,262) of all distinct implementation studies, so may be considered highly representative. As with all existing MALL meta-analyses, the starting number drops rapidly as soon as critical inclusion and exclusion criteria are applied.

2.3.2 Meta-analysis inclusion criteria

The most critical inclusion consideration is that the studies must report learning outcomes. When this is applied, only 814 implementation studies remain. A second critical inclusion condition is that reported learning outcomes must be substantiated by objective measurements specifically related to the experimental intervention. When 114 results determined uniquely by subjective instructor appraisals, self-reported student evaluations, and generic assessments such as mid-term/final tests and course grades are disregarded, 700 studies remain. Further culling of studies needs to be done based on two other criteria: research design shortcomings and inadequate statistical analysis.

As other MALL meta-analyses have attested, experimental MALL studies suffer from numerous design-related shortcomings (Burston, 2015; Chwo, Marek & Wu, 2018; Elgort, 2018; Lee, 2019; Shadiev, Liu & Hwang, 2020; Viberg & Grönlund, 2012). Most pervasive is small sample size and short intervention duration. Small sample sizes are problematic because they are likely to exaggerate the perceived effectiveness of outcomes (Cheung & Slavin, 2012; Liao, 1999). Although there is no fixed rule, some meta-analyses (Cheung & Slavin, 2013; Sung, Lee, Yang & Chang, 2019) recommend a sample size of at least 30, thus necessitating 60 participants for a typical experimental/control group treatment. However, MALL studies are very much determined by available class sizes, which are rarely so large. Even having access to two classes of 30+ students in many cases would not be possible. Creswell (2014, p. 164) recommends a more practical minimal sample size of 15 for a single treatment group and 30 when a control group is included. Applying this inclusion criterion reduces the number of valid MALL experimental studies to 572.

To offset the overly positive influence of the novelty effect, it is recommended that experimental treatments last at least eight weeks (Chwo *et al.*, 2018; Clark & Sugrue, 1991). This

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

recommendation is convincingly justified by Peng, Jager and Lowie (2020), who confirm that shorter treatment durations greatly exaggerate apparent language learning outcomes. According to their calculations, the ES of studies lasting less than four weeks was twice as great as those with a duration of four to eight weeks and four times greater than studies lasting longer than eight weeks. Unfortunately, there is no standard measurement unit for intervention duration in MALL studies. This can be indicated in classes, days, weeks, and months, as well as academic terms, semesters, and years. For purposes of calculating intervention duration, two months is taken to equal eight weeks and academic terms, semesters, and years all more than eight weeks. On this basis, a further 333 studies have been disregarded either because they did not meet the eight-week criterion or gave no information at all about treatment duration. Of the remaining 239 studies, a further 114 must be excluded owing to faulty research procedures. The most common of these is the failure to account for the extra time on language learning tasks accorded to experimental MALL groups, who work out of class in addition to whatever the control group did. This uncontrolled time-on-task variable is often compounded by the co-occurrence of additional instructor and/or peer mentoring and feedback, which the control group does not receive. Another frequent procedural shortcoming is the failure to specify the activities of control groups, which are simply described as receiving "traditional" instruction. In some cases, no information is given about what the experimental group did other than to say that it used a particular app. In other cases, when apps are accessible via multiple platforms, actual mobile device usage is merely assumed and not verified even by student self-reports.

Lastly, 41 studies must be excluded from consideration owing to the inadequate statistical analysis of their reported results. To meaningfully evaluate learning outcomes, it is essential that a study provide sufficient data to allow the calculation of ES (i.e. the relative magnitude of observed differences in post-intervention results). This requires knowing the exact size of intervention groups (i.e. sample size) and determining the mean and standard deviation (*SD*) of the mean for each intervention group. The *SD* measures how closely grouped the actual results are to the mathematical mean. The smaller the *SD*, the more closely the group average reflects actual individual performance. The larger the difference between the means and the smaller their *SD*, the greater the relative magnitude of the observed effects. In some cases, studies fail to meet the requirements for the calculation of ES owing to an indeterminate sample size; for example, the total number of participants is divided into subgroups of unspecified size (e.g. 154 participants split into four experimental and three control groups). Other studies give only raw scores or percentages for pre-/post-treatment results. Sometimes means are given without indicating their *SD*. In the end, of the original 700 experimental MALL implementations reporting objectively determined learning outcomes, only 84 (12%), can be considered to have a research design and provide a level of statistical analysis sufficiently robust to merit inclusion in the following meta-analysis (Appendix 2).

Though quite limited compared to the total number of experimental MALL studies, the 84 included in this meta-analysis are considerably more than what has appeared in the 11 existing MALL meta-analyses that specifically target language learning outcomes (Table 2).

In part, this is due to the longer time frame involved: 26 years here, compared to between three and 21 years for the others. In some cases, the focus in other overviews is restricted to English or specific domains (e.g. vocabulary, reading). Of greater importance, however, the present meta-analysis derives from a much larger database of 814 experimental MALL studies with reported language learning outcomes. The database of MALL publications that meet the initial selection criteria of the other 11 learning outcome meta-analyses identifies only between 64 and 367 studies. Kamasak, Özbilgin, Atay and Kar (2020) indicate only the number of MALL studies identified (982) prior to applying any selection criteria.

It also needs to be kept in mind that the exclusion criteria in the other previous meta-analyses are much less restrictive than that of the present study. Burston (2015) requires a minimal sample size of only 10 and an intervention duration of four weeks. The other 10 impose no such

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

Table 2. MALL meta-analyses with a learning outcome focus

Meta-analysis	Publication range	# Studies meeting initial selection criteria	# Studies analyzed
Burston & Giannakou, 2021	1994–2019	814	84
Lee *et al.*, 2014	1993–2013	288	44
Burston, 2015	1994–2012	291	35
Sung *et al.*, 2015	1993–2013	119	45
Cho *et al.*, 2018	2005–2017	367	20
Chen *et al.*, 2020	2008–2018	218	80
Eutsler *et al.*, 2020	2007–2019	102	61
Guanuche *et al.*, 2020	2010–2019	305	39
Kamasak *et al.*, 2020	2010–2020	–	13
Mahdi, 2017	2008–2017	219	16
Peng *et al.*, 2020	2008–2017	135	17
Klimova & Zamborova, 2020	2018–2020	64	9

Table 3. Inclusion rate from previous meta-analysis studies

Meta-analyses	# Included studies
Lee *et al.*, 2014	–
Burston, 2015	4/35 (11%)
Sung *et al.*, 2015	1/45 (2%)
Cho *et al.*, 2018	6/20 (30%)
Chen *et al.*, 2020	5/80 (6%)
Eutsler *et al.*, 2020	7/61 (11%)
Guanuche *et al.*, 2020	1/39 (3%)
Kamasak *et al.*, 2020	2/13 (15%)
Peng *et al.*, 2020	–
Mahdi, 2017	3/16 (18%)
Klimova & Zamborova, 2020	1/9 (11%)

conditions. In fact, Eutsler, Mitchell, Stamm and Kogut (2020) include a study with a single participant (McClanahan, Williams, Kennedy & Tate, 2012). Sung, Chang and Yang (2015) include a study that lasted only 30 minutes over two sessions (Huang, Liang, Su & Chen, 2012). In addition, only Burston (2015) considers research design shortcomings. The others simply accept reported results at face value. As a consequence of their less rigorous exclusion criteria, very few of the experimental MALL studies included in the 11 previous meta-analyses focusing on language learning outcomes appear in the present study (Table 3).

Neither Lee *et al.* (2014) nor Peng *et al.* (2020) identify the experimental MALL studies in their meta-analysis, but presumably they would have fared no better than the others. As can be seen, the greatest number of experimental MALL studies accepted in the present meta-analysis from any

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

one of these was seven, and even this represents only 11% (7/61) of that study's total database. In sum, the database underlying this meta-analysis is not only considerably more selective but also more extensive than any of its predecessors.

3. Research questions

This meta-analysis seeks to address two critical sets of research questions. First, what are the general characteristics of the selected studies and, second, what are their language learning outcomes in terms of measured ES. More specifically,

> Research Question 1: What are the general characteristics of the selected studies as these relate to their (a) publication source, (b) chronological distribution, (c) country of origin, (d) institutional environment, (e) sample size, (f) intervention duration, (g) targeted language, (h) language learner competence level, and (i) learning focus?

> Research Question 2: What is the overall ES of these studies and how does this relate to the following moderator variables: (a) language learner competence level, (b) language area focus, (c) institutional environment, and (d) intervention duration?

4. Results

4.1 General characteristics of selected MALL studies

4.1.1 Selection

In analyzing the 1,144 experimental MALL studies that underlie this meta-analysis, a database of nine critical features was compiled for comparison based on general background information (RQ#1a–d), relevance to research design (RQ#1e–f), and language acquisition parameters (RQ#1g–i).

4.1.2 Publication source

Aside from their very low number, experimental MALL studies meeting rigorous selection criteria for the reporting of language learning outcomes are notable with regard to their publication sources. Although it might be thought that the prominent publications identified in Table 1 would make up in quality for what they lack in overall quantity, in reality, MALL studies from these sources account for only 15 of the 84 (18%) (Table 4).

Nine of these appeared in three of the leading CALL journals: four each in *Computer Assisted Language Learning* and *Language Learning & Technology* and one in *ReCALL*. The remaining six included studies that were published in three educational technology journals: three in *Computers & Education*, two in *Educational Technology & Society*, and one in the *British Journal of Educational Technology*. At best, only 9% of the total MALL studies published in any given CALL journal and 11% in educational technology journals met the inclusion criteria for this meta-analysis. MALL studies in mobile technology publications are notable by their total absence among the included research. By far the great majority (82%; 69/84) of the included MALL experimental studies instead appeared in other journals, conference proceedings, and MA/PhD dissertations.

4.1.3 Chronological distribution

As concerns the distribution of selected MALL studies over time, as shown in Figure 1, none published before 2008 are included in this meta-analysis. The fact that nearly three quarters (61/84) of the included studies have appeared since 2015 may be taken as a sign of the increasing

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

Table 4. MALL studies meeting meta-analysis selection criteria

CALL	Selected studies	Educational technology	Selected studies
CALICO Journal 2005–2019	0/15 (0%)	*British Journal of Educational Technology* 2009–2019	1/20 (5%)
Computer Assisted Language Learning 2000–2019	4/62 (5%)	*Computers & Education* 2004–2019	3/30 (11%)
International Journal of Computer-Assisted Language Learning and Teaching 2011–2019	0/21 (0%)	*Educational Technology & Society* 2008–2019	2/34 (7%)
JALT CALL Journal 2005–2019	0/28 (0%)	*System* 2005–2019	0/10 (0%)
Language Learning & Technology 1999–2019	4/35 (9%)		
ReCALL 2005–2019	1/27 (4%)		
Total selected studies	9/188 (5%)		6/94 (6%)

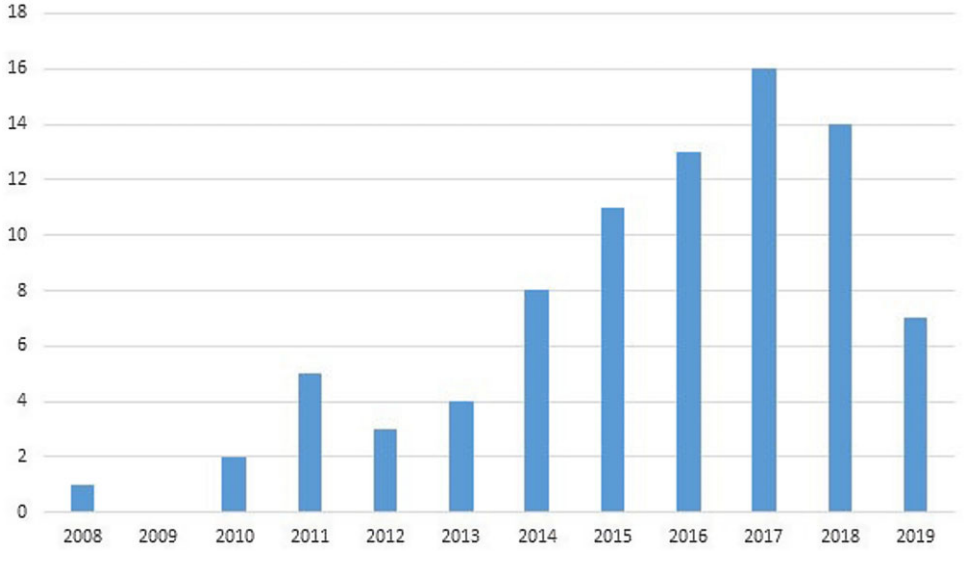


Figure 1. Chronological distribution of selected MALL studies

sophistication of MALL experimental studies in recent years. Given the inevitable time lag involved in tracking down MALL studies, the decrease in their number in 2018–2019 is very likely more apparent than real and could be expected to increase when bibliographical compiling eventually catches up with actual publications.

4.1.4 Country of origin

In this meta-analysis, as in all previous MALL overviews, Asian countries represent the greatest source of experimental MALL studies (Figure 2). Together, Taiwan, China, Japan, Korea, and Thailand account for 35% (29/84) of the included studies. In fact, however, on its own, Iran accounts for the greatest number of included experimental MALL studies, 24% (20/84). When

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

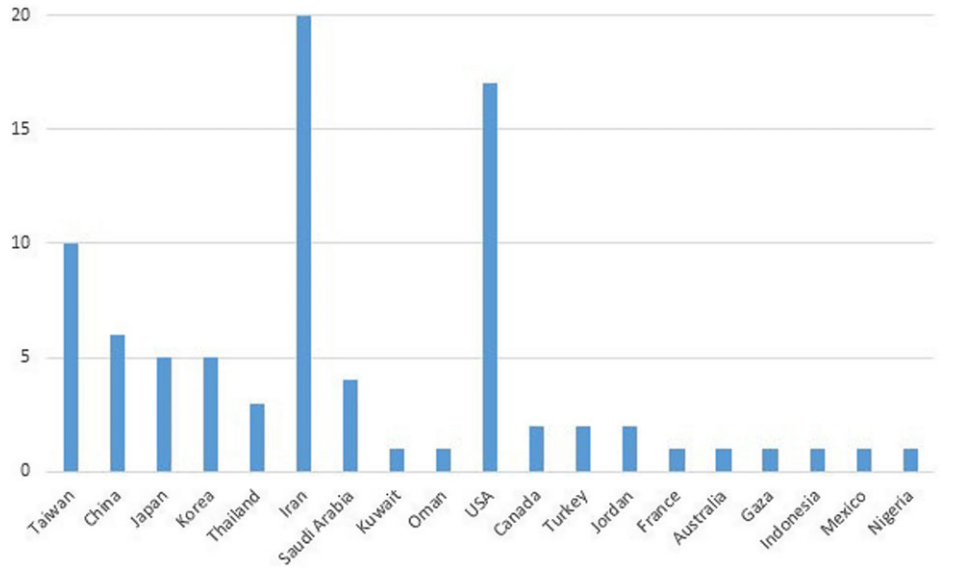


Figure 2. Countries of selected experimental MALL studies

combined with Oman, Kuwait, and Saudi Arabia, the total for Gulf state countries amounts to 31% (26/84). Together then, Asian and Gulf state countries account for nearly two thirds (55/84) of the experimental MALL studies evaluated in this meta-analysis. Thanks to seven master's/doctoral dissertations, the USA comes in at second place among single countries with 20% (17/84). Canada, Turkey, Jordan, France, Australia, Gaza, Indonesia, Mexico, and Nigeria together contribute the remaining 12 studies (14%): two each from the first three and one each from the others.

4.1.5 Institutional environment

As has been the case in all other overviews that track the institutional environment of MALL studies, at 47% (40/85) tertiary-level students represent by far the largest cohort of participants in the experimental MALL implementations included in this meta-analysis (Figure 3). It is to be noted that the number of institutional environments is one greater than the total number of studies because one study involved two institutional levels. Studies undertaken at preschool and primary school combined account for 25% (21/85) and secondary school another 17% (14/85). Language institutes and other adult education locations combined constitute 8% (7/85). The institutional environment of a further 4% (3/85) was not specified.

4.1.6 Sample size

In conformity with the inclusion criterion for experimental treatment sample size, the smallest in this meta-analysis is 15, the largest 279. Because some studies involved multiple intervention groups, the total number of sample sizes amounts to 93. The most frequent experimental group size is between 30 and 39 subjects, which represents 37% (34/93) of all the cases (Figure 4). At 24% (22/93), the second most frequent sample size was 20–29. When the 11% (10/93) of sample sizes between 15 and 19 are added, 71% of the studies in this meta-analysis involved less than 40 experimental participants. Intervention groups of between 40–59 and 60–79 accounted for 15% (14/93)

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

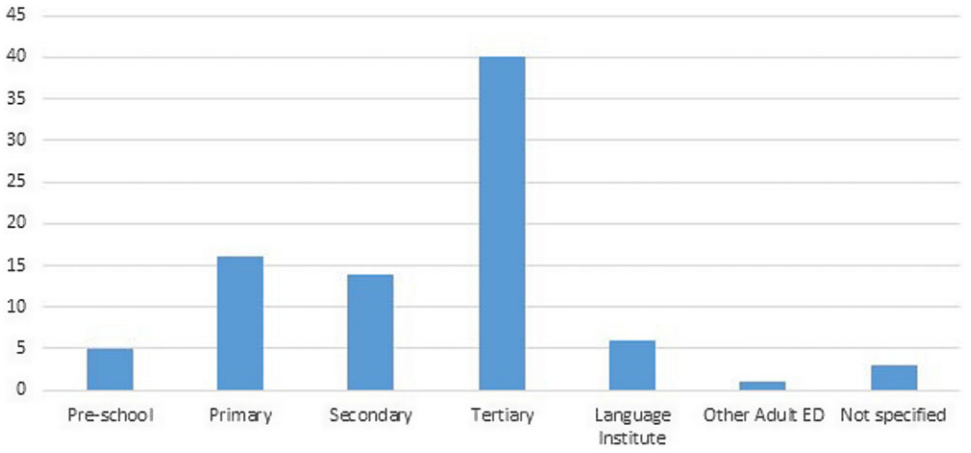


Figure 3. Institutional environment of selected experimental MALL studies

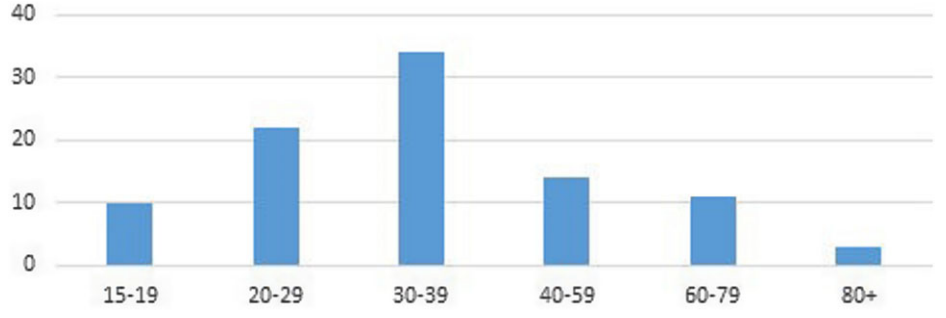


Figure 4. Sample size of selected experimental MALL studies

and 11% (10/93) respectively. At 3% (3/93), sample sizes of 80 or more were the least frequently encountered.

4.1.7 Intervention duration

The tabulation of intervention duration in this meta-analysis is complicated by the lack of consistency in the time units that are reported. As previously indicated, this can be in days, weeks, or months, as well as in terms, semesters, and years. To facilitate comparability, and not penalize MALL implementations that involved more frequent interventions over shorter periods of time, it was assumed that classes met three days per week, and days were thus converted to weeks on this basis (i.e. 24 days = 8 academic weeks). Accordingly, three studies of 6–7 weeks duration with 30–35 daily interventions were accepted for inclusion. Months were converted to units of weeks based on the ratio of months to a year (e.g. 3 months: 3/12 x 52 = 13 weeks). Terms were equated with a 10-week period, semesters to 14 weeks, and 8–13 months to an academic year (Figure 5). As can be observed, the three most frequent intervention durations are very similar, with the shortest 8–9 weeks at the top with 30% (25/84) and term (29%; 24/84) and semester length (26%; 22/84) closely behind. The three combined representing (85%), short treatment durations thus account for the great majority of all the studies. It is to be remembered, of course, that experimental treatments

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

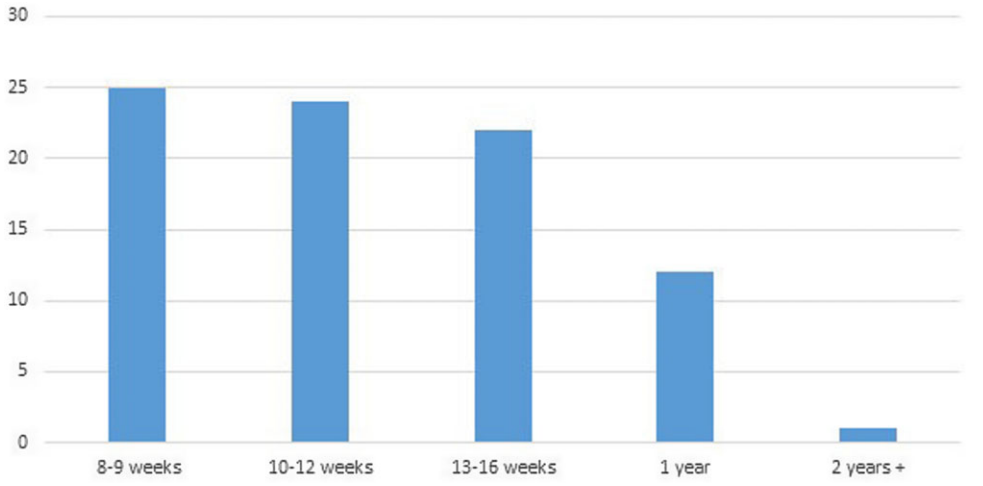


Figure 5. Intervention duration of selected experimental MALL studies

equivalent to less than 8 weeks were excluded from this meta-analysis. As previously mentioned, these represent another 333 very short-term studies, emphasizing all the more the prevailing short-term nature of most quantitative MALL studies. In fact, only 15% (13/84) of all experimental studies in this meta-analysis were of long duration. These ranged from 17 to 30 weeks, typically two terms or semesters to two academic years.

4.1.8 Target language

Although MALL studies presented in any language were considered for this meta-analysis, in fact all but two were written in English, one in Korean, and one in Turkish. Also while no restrictions were put on the targeted language of instruction, English accounts for 95% (81/85) of the cases, 15 of which were as an L1. Note that the number of languages is one greater than the number of studies because one study included both L1 and L2 English language learners. The remaining four studies are represented by one L1 and two L2 Spanish studies and one L2 German study.

4.1.9 Language learner competence level

In considering the language proficiency level of the participants in the MALL studies included in this meta-analysis, it is first necessary to distinguish between L1 and L2 implementations. Sixteen studies fall within the first category, involving 15 L1 English speakers and one L1 Spanish speaker. Of these, the language proficiency level of eight was not identified. In the remainder, six were classified as preliterate children and two weakly literate (i.e. disabled/struggling), one a child and the other an adolescent (Figure 6). Regarding the L2 language proficiency level in the remaining 69 studies, as with the L1 studies, its most striking feature is the inconsistency with which it is reported. In 39% (27/69) of the cases, it is either unmentioned or simply equated with school grade level or previous years of study. Moreover, even when L2 proficiency level is specified, this is done with reference to a variety of tests (i.e. TOEFL, TOEIC, IELTS, CET, etc.) and descriptors (e.g. Preliterate, Basic, Beginner, Intermediate, A1, B2, Advanced, Proficient, etc.). The best that can be done in summarizing this information is a simple classification in terms of preliterate, beginner, intermediate, and advanced. Although the L2 level is specified in 42 studies, in five cases mixed ability groups are described, resulting in a total of

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

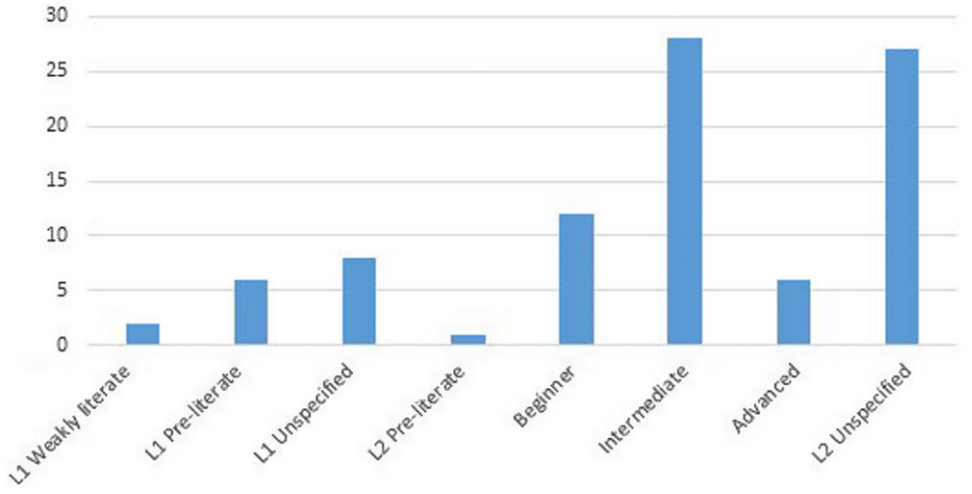


Figure 6. Target proficiency level of selected experimental MALL studies

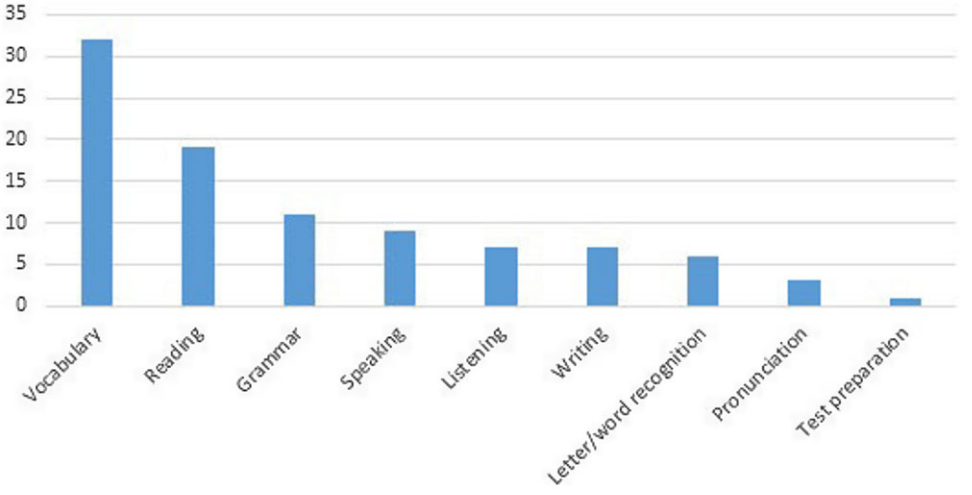


Figure 7. Focus of selected experimental MALL studies

47 specifications. Most notable is the small proportion of advanced-level learners. Even though nearly half the participants in experimental MALL studies were university students (Figure 3), advance-level learners represent only 13% (6/47) of the participants. Intermediate-level learners account for 60% (28/47) of the participants. Beginners represent 26% (12/47) and one preliterate child the remaining 2%.

4.1.10 Learning focus

As numerous MALL implementations have more than a single learning focus, a total of 95 are attested in the studies included in this meta-analysis (Figure 7). As is the case in all other

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

MALL overviews that track a full range of learning focuses, vocabulary is by far the most frequently targeted area, representing 34% (32/95) of the total. This is more than the second and third categories combined: reading 20% (19/95) and grammar 12% (11/95), respectively. Likewise, the six remaining learning focuses combined account for less than vocabulary on its own.

4.2 Effect size of learning outcomes

4.2.1 Calculation procedures

The calculation of ES presupposes, of course, a comparison of pre-/post-intervention results. As is usual practice, the MALL implementation studies included in this meta-analysis do this either based on a single group in which all participants receive the same treatment (within-group) or based on a comparison of subgroups that receive different treatments (between-groups). Most frequently, only two subgroups are involved: the experimental and control. Occasionally, the comparison may involve more than one experimental condition.

Of the 84 studies included in this meta-analysis, 21 involved only a single (within-group) treatment study and 63 two or more (between-groups). However many intervention groups there were, a pre-test related to the targeted language area was administered to establish a baseline for comparison with a post-test and, when more than one intervention group was involved, to verify the initial equivalence of the groups. As previously indicated, to be included in this meta-analysis, all studies had to report an exact sample size, mean scores, and a corresponding *SD*. To ensure the accuracy of calculations, these data were extracted from the selected MALL studies individually by each author. The two resulting databases were then compared and any discrepancies resolved by referral back to the MALL studies concerned.

In a meta-analysis, the calculation of ES is a two-stage operation. First, this must be done internally for each individual intervention result based on the data indicated previously. In the second stage of analysis, individual ES results are pooled to arrive at an overall estimation of the relative magnitude of intervention outcomes. This meta-analysis takes the pooling stage one step further by combining overall results relative to four moderator variables: language learner competence level, learning focus, institutional environment, and intervention duration.

4.2.2 Individual learning outcome selection

When extracting the learning outcome data, two decisions had to be made about how to treat studies with multiple learning focuses involving the same participants. According to Borenstein, Hedges, Higgins and Rothstein (2009), treating such outcomes as separate studies leads to inaccuracies when making ES estimates. However, pooling such results is problematic, too. First, in cases where major language skill areas are involved (e.g. writing and speaking), combining the results removes relevant data when ES is analyzed relative to the moderator variable of learning focus. Second, the procedure recommended for combining individual learning outcome results could negatively impact the validity of the results because of overestimated standard errors (Matt & Cook, 2009; Moeyaert et al., 2017). For these reasons, within this meta-analysis, different learning outcomes involving the same participants are not pooled but treated as separate study results.

The second issue regarding the treatment of multiple language focuses within the same study involved deciding which comparative results to retain for ES calculations. Such cases all presented two sets of data. On the one hand, pre-/post-test results for the experimental treatment are given for the participants in each of the targeted language areas. On the other hand, post-test results for the participants of each experimental intervention are also compared against those of a matched control group. Because the pre-/post-test outcomes allowed the results to be more clearly seen

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

within the experimental group, these are the comparisons that were retained for the ES calculations.

Based on these decisions, it was possible to calculate the ES of 140 language learning outcomes reported in the 84 studies analyzed in this meta-analysis. Despite the importance of determining ES, and the availability of the necessary data, only 19% (16/84) of the primary studies in this meta-analysis actually do so.

4.2.3 Individual treatment effect size calculations

The calculation of ES can be determined by several methods: Pearson's *r*, Cohen's *d*, Glass's Δ, Hedges's *g*, and so on. Two factors argued in favor of using Hedges's *g* for this meta-analysis: heterogeneity and sample size. Experimental MALL studies are most notable for their pervasive heterogeneity. In fact, almost all are different regarding their combination of number of intervention groups, sample size, intervention duration, targeted language skills, language competence level, institutional environment, pedagogical interventions, and assessment procedures. Also, as indicated previously, experimental MALL studies are characterized by small sample sizes. Given this disparity in research designs and small sample sizes, the calculation of ES was based on the Hedges's *g* method to avoid upward bias (Borenstein *et al.*, 2009; Ellis, 2010).

For maximum validity, when pooling individual ES estimates to calculate an overall effect value, it is important not to combine the results of quasi-experimental studies (i.e. within-group) and experimental (i.e. between-group) study trials. For this reason, standardized mean differences (Hedges's *g*, expressed as 95% confidence intervals [95% CI]) were calculated separately for the two study designs.

Due to diverse population demographics and research methods, the random-effects model for the meta-analyses was used. This assumes that, when estimating a mean effect, true effects vary between studies (Borenstein *et al.*, 2009; Field & Gillett, 2010). To estimate the ES from studies with a within-group repeated measures design, the pre-post correlation is required to impute the within-groups *SD* from the *SD* of the difference. Considering that studies did not routinely report this data, a conservative estimate of $r = 0.7$ was used, based on within-group test–retest correlations for the standardized measures utilized in this meta-analysis (r range: typically, > 0.7). If a study had numerous time periods (e.g. delayed post-test), only the pre-intervention to immediate post-intervention strength outcomes were extracted and entered for analysis.

Extracted data were analyzed using the software package Meta-Essentials: Workbooks for Meta-Analysis for differences between independent and dependent groups' continuous data (Suurmond, van Rhee & Hak, 2017). ES is measured in terms of standardized mean difference (SMD) values on a negative/positive scale that usually ranges between –3 to + 3, although lower and higher values are also possible. A value at or below 0.2 is considered trivial, above 0.2 to 0.5 is regarded as small, and above 0.5 to 0.8 moderate. An ES above 0.8 is considered large (Cochran, 1954; Cohen, 1988; Fisher, Frey & Hattie, 2016).

4.2.4 Overall MALL studies effect size calculation

Using the Meta-Essentials: Workbooks for Meta-Analysis software package, the individual ES calculations shown in Appendices 3–4 were pooled to produce a separate overall ES estimate for between-group (Table 5) and within-group (Table 5) experimental MALL studies.

A random-effects model was used to synthesize the results of the included studies. Even with the stricter selection criteria that applied to this meta-analysis, the overall ES is at the top end of moderate for between-group and large for within-group treatment studies. More specifically, for between-group treatment studies, the ES was calculated to be 0.72 (95% CI [0.34, 1.09]; $I^2 = 90\%$, P heterogeneity < 0.001). The overall ES for SMD calculated as Z was 3.75 ($P = 0.002$). Similarly,

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

Table 5. Overall effect size

Meta-analysis model	Random-effects model (between-group)	Random-effects model (within-group)
Confidence level	95%	95%
Combined effect size		
Hedges's *g*	0.72	1.16
Standard error	0.19	0.14
CI lower limit	0.34	0.89
CI upper limit	1.09	1.44
Z value	3.75	8.45
Two-tailed *p* value	0.000	0.000
Number of included subjects	6658	2031
Number of included studies	84	56
Heterogeneity		
Q	802.26	995.72
p_Q	0.000	0.000
I^2	89.7%	94.5%

for within-group studies, the ES was calculated to be 1.16 (95% CI [0.89, 1.44]; $I^2 = 94\%$, P heterogeneity < 0.001). The overall ES for SMD calculated as *Z* was 8.45 ($P < 0.000$).

It is to be noted that these calculations of overall ES are substantially larger than those reported in nearly all previous MALL meta-analyses that specifically target learning outcomes. Of the 11, five (Burston, 2015; Eutsler *et al.*, 2020; Guanuche, Eiriz & Espí, 2020; Kamasak *et al.*, 2020; Klimova & Zamborova, 2020) do not calculate ES at all. For the six that do, the overall effectiveness of experimental MALL implementations in all but one is calculated as moderate, varying between 0.51 and 0.722. Exceptionally, Peng *et al.* (2020) indicate a large overall ES of 0.94 for between-group studies.

4.3 Publication bias and heterogeneity

To properly understand the significance of pooled ES estimates, it is essential to view the results in relation to two critical factors: publication bias and heterogeneity. Given the very high proportion of positive experimental outcomes in the data, a well-attested indicator of the "file drawer" problem (i.e. a tendency not to publish negative results), a check was made for publication bias. This was done by visual inspection of funnel plots and evaluated formally with Egger's regression asymmetry test (Egger, Davey Smith, Schneider & Minder, 1997; Sterne, Egger & Davey Smith, 2001). The shape of the funnel plot for both between-group (Appendix 5) and within-group treatments (Appendix 6) indicated obvious asymmetry, and Egger's test also provided statistical evidence ($P < 0.000$) of publication bias in the current meta-analysis. All of which is to say that there is good reason to believe that the language learning outcomes in the studies of this meta-analysis are very likely to be less positive than the published research would indicate.

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

When pooling ES calculations in a meta-analysis, the validity of the result very much presupposes that the research parameters found in individual studies are similar enough to be confident that a combined estimate will be a meaningful description of the set of studies. This is a crucial factor when considering the overall estimate of ES in experimental MALL studies, which, as noted previously, are most notable for their pervasive disparity. To determine the heterogeneity underlying individual ES estimates, the Cochran Q test was used to measure the I^2 value for inconsistency. This ranges between 0% and 100% and is the ratio of between-study variance over the sum of within-study and between-study variances (Higgins & Thompson, 2002). Values exceeding 50% are usually considered to represent large and over 75% very large heterogeneity. As indicated in Table 5, with a heterogeneity value of nearly 90% for between-groups and over 94% for within-groups, the overall ES of the experimental MALL studies in this meta-analysis needs to be regarded with due caution.

4.4 Pooled results by moderator variables

Although the problem of heterogeneity cannot be eliminated entirely, its effects can be mitigated by limiting the pooling of ES results to more constrained subclasses. Of the many variables found in the selected studies in this meta-analysis, four in particular merit a closer analysis: language learner competence level, language area focus, institutional environment, and intervention duration. The first two relate directly to language acquisition parameters: the starting point and targeted skill area. The second two reflect the learning environment: where and for how long the intervention took place. Arguably, these constitute the most critical parameters for language learning outcomes in any experimental MALL intervention.

Regarding language competence, the ES results of six levels were individually pooled in this meta-analysis. For L1 studies, these are Preliterate and Weakly literate and for L2 studies they are Beginner, Intermediate, Advanced, and Mixed (i.e. intervention groups involving more than one proficiency level). So, too, the ES results of eight language area focuses were independently pooled: Vocabulary, Reading, Grammar, Listening, Speaking, Writing, Letter/Word Recognition, and Other (i.e. pronunciation and test preparation). As with the pooling of overall ES results, this was done separately for between-group and within-group treatment studies. The determination of whether the ES of any particular group was influenced by a moderator variable was based on a heterogeneity analysis (using the test statistic Q_B). The individual ES results of language learner competence level, language area focus, institutional environment, and intervention duration are summarized in Appendix 7 for between-group and in Appendix 8 for within-group studies.

4.4.1 Between-group results

For between-group treatments (Appendix 7), the ES of specified language proficiency levels is small to moderate (0.30 and 0.61) for L1 studies and varies between moderate to large (0.59–1.34) for non-mixed levels in L2 studies. It is to be noted that the ES for Beginner (1.24) and Intermediate (1.34) with non-mixed L2 groups is more than twice that of Advanced-level participants (0.59). So, too, except for Advanced-level groups (69%), the I^2 value for the heterogeneity of the L2 studies is between 95% and 99% compared to 13% at most for L1 studies.

With regard to language area focus, the ES covers the full range from small (Writing, 0.28) to large (Grammar, 3.49). The I^2 value improves to a moderate level with three language target areas: Letter/Word Recognition ($I^2 = 28\%$), Writing ($I^2 = 32\%$), and Listening ($I^2 = 41\%$). It remains very large for all the others.

Like language area focus, the ES relative to institutional environment covers the full range from small (Primary, 0.30) to large (Secondary, 1.39). Except for Adult Education ($I^2 = 0\%$) and Preschool settings ($I^2 = 17\%$), the I^2 value is large or very large for all institutional

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

settings: Language Institute ($I^2 = 55\%$), Primary ($I^2 = 60\%$), Tertiary ($I^2 = 90\%$), and Secondary ($I^2 = 92\%$).

The effect of intervention duration upon ES with between-group studies produces mixed results. As might be expected, the smallest ES is associated with the shortest treatment period, 8–9 weeks (0.50). However, with studies of 13–16 weeks' duration, the ES is only slightly greater (0.57). Studies of 10–12 weeks' duration exhibit the same ES as those that lasted two semesters or more. The I^2 value remains large or very large: 8–9 weeks ($I^2 = 76\%$), 13–16 weeks ($I^2 = 77\%$), 10–12 weeks ($I^2 = 93\%$), and two semesters or more ($I^2 = 95\%$).

4.4.2 Within-group results

As with between-group studies, in within-group interventions (Appendix 8) the ES of specified language proficiency levels is small to moderate (0.39 and 0.56) for L1 studies. On the other hand, in L2 studies all non-mixed levels evidence a quite large ES of between 1.44 and 1.82, notably higher than that of between-group studies. Although greater than that of between-group studies, the I^2 value for heterogeneity remains quite small for within-group L1 studies (19%–26%), but very large (84%–95%) for L2 studies.

Although at a moderate level and greater than with between-group studies, both Letter/Word Recognition (0.73) and Listening (0.61) continue to manifest the smallest ES. All the other targeted language areas demonstrate a large ES of between 1.07 and 1.98. Notably, ranging between 93% and 96%, the I^2 value for heterogeneity is very large for all language area focuses.

Regarding the institutional environment moderator, from Preschool through Tertiary settings, within-group studies manifest a direct correlation between academic level and ES. This ranges from small at Preschool (0.39) to moderate at Primary (0.69), large at Secondary (0.92), and even larger at Tertiary level (1.38). The largest ES in specified settings is found in Language Institute (2.58), although it should be noted that there are only two of these in the database. As with the between-group studies, Preschool environments manifest the smallest I^2 value for heterogeneity ($I^2 = 26\%$). All the other specified institutional settings are very large: Primary ($I^2 = 85\%$), Language Institute ($I^2 = 89\%$), Secondary ($I^2 = 92\%$), and Tertiary ($I^2 = 95\%$).

As is the case with between-group studies, the effect of within-group intervention duration upon ES produces mixed results. While all the ES are large, again, the smallest ES (0.84) is associated with the shortest intervention duration of 8–9 weeks. The largest ES (1.67) is found in studies of 10–12 weeks' duration, followed closely by 13–16 weeks (1.47). Interestingly, studies lasting two semesters or more manifested an ES (0.87) nearly the same as the shortest intervention period. Without exception, the I^2 value for heterogeneity of all within-group intervention durations is very large (85%–96%).

4.5 Statistical significance of moderator variables

As can be observed in Appendix 7, the ES of only one of the four moderator variables was found to be statistically significant for between-group interventions. Specifically, this was for institutional environment ($Q_B = 15.8$, $p < 0.001$). The remaining moderators exceed the limit for statistical significance: language learner competence level ($Q_B = 1.48$, $p > 0.05$), language area focus ($Q_B = 1.74$, $p > 0.05$), and intervention duration ($Q_B = 0.50$, $p > 0.05$). In contrast, as can be observed in Appendix 8, only intervention duration fails to meet the requirement for statistical significance ($Q_B = 3$, $p > 0.05$) for within-group interventions. All the other moderators show statistically significant results: language learner competence level ($Q_B = 1.48$, $p < 0.001$), language area focus ($Q_B = 10.8$, $p < 0.001$), and institutional environment ($Q_B = 165$, $p < 0.001$).

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

5. Discussion

5.1 Research question 1

In considering the general characteristics of the studies included in this MALL meta-analysis, it can be seen that their numbers have been steadily increasing since 2013, with apparent declines in 2018–2019 almost certainly due to the time lag in compiling bibliographical resources. Asian countries continue to be the most prolific sources of experimental MALL studies, although on their own Iran and the USA are the single greatest contributors. Further, the greatest number of experimental MALL studies have been undertaken in college and university settings. Notwithstanding, only a small percentage of MALL interventions involve advanced-level learners. The most striking feature of these studies is the great disparity of their research design. In fact, the only thing that really unifies these studies is their almost exclusive focus on English as a target language. There is also a marked focus on vocabulary acquisition, but to a much lesser degree than the language bias.

In most respects, these findings mirror what has already been reported in the 11 previously published MALL meta-analyses that focus on language learning outcomes. Critically, however, this meta-analysis differs from its predecessors in the comprehensiveness of the database from which its observations derive. As has been shown, few experimental MALL studies are to be found where previous meta-analyses have looked. Prominent CALL, educational technology, or mobile technology journals, and their associated conferences, account for only about 10% of published MALL studies. Likewise, a great many MALL studies escape discovery in the major academic databases. As a result of its much larger underlying database, this meta-analysis has brought to light a substantial body of hitherto overlooked experimental L1 MALL studies, which represent nearly a fifth of the total. Furthermore, the underlying resource database in this meta-analysis has been subject to far more rigorous inclusion and exclusion criteria. Only studies meeting strict conditions of sample size, treatment duration, research design, and statistical analysis have been considered for inclusion. As a result, though similar to previous observations, the general characteristics of the MALL studies described in this meta-analysis can be regarded as much more precisely defined and meriting considerably greater confidence.

While much has been done in MALL over the past two decades, several critical areas still await the attention of researchers and language teachers. First, the nearly exclusive focus on English as a target language needs to expand to other languages, both major and less commonly taught. The latter, in particular, could profit from the motivational boost that MALL has been demonstrated to engender. Moreover, the fixation on vocabulary acquisition needs to give way to other aspects of language learning and usage. Likewise, to date, MALL implementations have been resolutely tutorial in nature. Much more exploitation is needed of the sociocultural, communicative affordances offered by mobile technologies. This is especially so regarding the much-neglected domain of advanced-level language learners, who have the greatest potential to use MALL to establish and maintain, in the everyday and professional world, meaningful human-centric contact with their adopted culture and language.

5.2 Research question 2

As the foregoing analyses have shown, the calculation of overall ES is larger in this study than that in any previously reported meta-analyses specifically targeting language learning outcomes. This result is all the more notable considering the greater number of experimental MALL studies and more stringent inclusion criteria upon which this calculation is based in all 140 experimental MALL treatments involving 8,689 participants. Moreover, differentiating between the results of between-group and within-group treatments has demonstrated that much larger ESs are obtained from the latter (1.16) than the former (0.72). It needs to be borne in mind, however, that the higher level of positive outcomes may very well derive from the absence of control subjects in the within-group studies. Tempering all these very positive ES calculations, however, is an

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

obvious underlying publication bias and the extreme degree of heterogeneity evidenced in the primary studies of this meta-analysis.

A closer look at four critical moderator variables (language learner competence level, language area focus, institutional environment, and intervention duration) confirmed positive ES results whatever the variable and provided greater insight into these results. First, regarding language learner competence level, it showed that the ES is almost always smaller in L1 than in L2 studies, regardless of intervention design. In the case of between-group interventions, the L1 ES was small and moderate (0.30 and 0.61) compared to the moderate to large L2 ES (0.59–1.34). For within-group interventions, the difference in ES between L1 and L2 studies was even more marked, with the former only small to moderate (0.39 and 0.56) and the latter uniformly large (1.44–1.82). These more detailed results also attest to the very broad range of ES across the various language proficiency levels, extending from small for L1 Preliterates to quite large with Beginner- and Intermediate-level L2 learners. The range of ES was similarly broad for language area focus, particularly in between-group interventions where it extends from 0.28 for Writing to 3.49 for Grammar. Regardless of intervention type, institutional environment was shown to play an important role. The largest ES is to be found at institutional settings above the elementary level, most notably in within-group treatments. Intervention duration was also shown to affect ES outcomes, with the shortest durations producing the smallest ES regardless of intervention type. With the notable exception of L1 and Preschool studies, as with the overall ES findings, the detailed analysis confirmed a large to very large heterogeneity level across all the moderator variables. So, too, the very positive ES results of the detailed ES analysis need to be tempered by the lack of statistical significance in three of the between-group variables (language learner competence level, language area focus, and intervention duration) and one of the within-group moderators (intervention duration).

6. Conclusion

In the process of undertaking this meta-analysis of quantitative experimental MALL studies, aside from the information extracted relating to general characteristics and language learning outcomes, several other important facts have also been brought to light. Most notably, it has been shown the extent to which previous meta-analyses have been based on very incomplete data and how their inclusion criteria did not adequately take account of shortcomings in the research design and statistical analysis of included studies. As in all other preceding overviews, the number of included studies here is but a small fraction of those that meet the definition of a quantitative experimental MALL study. This raises the very serious question of why so many published studies have failed to be included in MALL meta-analyses. Their disparity, and the corresponding disparate nature of the sources in which they were published, undoubtedly are prime factors explaining why so many studies have gone unreported in MALL meta-analyses. That, however, is not the case here, where 700 objectively substantiated quantitative experimental MALL studies from all manner of publication sources were initially identified. The fact that 616 (88%) of these could not meet very basic inclusion criteria points to a fundamental problem of research design and statistical analysis in experimental MALL studies. This problem is not unique to MALL, and in fact the issue of research quality has been raised recently in the domain of mobile-assisted learning generally (Sung *et al.*, 2019). The causes and extent of research quality problems in quantitative experimental MALL studies would profit from a similar investigation.

6.1 Limitations of this study

Despite the comprehensiveness of the database underlying this study, its findings are nonetheless subject to several limitations. The most critical involves the extent to which observations have had to rely on subjective interpretations. In the absence of any commonly accepted evaluation metric,

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

the proficiency level of participants in experimental MALL studies can only be approximated: preliterate, beginner, intermediate, and advanced. Likewise, lacking any standardized measure of treatment duration, time periods can only be roughly estimated in terms of weeks. Furthermore, the lack of frequency information within treatment periods makes it difficult to determine just how much exposure to the MALL intervention participants have had. Knowing when an intervention has lasted the equivalent of eight weeks is very much subject to interpretation. It, thus, for example, has been assumed that treatments that occurred daily for six to seven weeks involved as much MALL exposure as those, of unknown frequency rates, that lasted at least eight weeks.

Above all, there are two important aspects of the experimental MALL studies in this meta-analysis to which no consideration has been given. First, nothing has been indicated about the theoretical framework underlying the MALL implementations in this study. In fact, MALL studies are profoundly atheoretical; less than a third in this meta-analysis identify any theoretical underpinnings at all. To be filled, these lacunae must await a much greater awareness of the relevance of theory to practice in MALL studies. Second, there is no indication of the extent to which the primary research studies analyzed here reflect the degree to which experimental MALL implementations have resulted in actual curricular innovation. The description of experimental MALL implementations is rarely followed by any mention of subsequent curricular integration. As the inquiry undertaken by Burston (2014) discovered, in 40% of the cases, MALL experiments did not lead to curricular modifications. On the other hand, there is no inherent reason why the adoption of MALL into the curriculum would necessarily engender published research. Logically, the integration of MALL into the curriculum could be much greater than what published research would indicate. That itself certainly is a question worthy of future MALL research.

Supplementary material. To view supplementary material referred to in this article, please visit https://doi.org/10.1017/S0958344021000240. Note that the authors have provided the following website address from which the appendices may be obtained: https://www.academia.edu/49301516/Burston_and_Giannakou_2021_Appendices

Ethical statement. There are no ethical issues or conflicts of interest involved in this study.

References

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

Burston, J. (2013) Mobile-assisted language learning: A selected annotated bibliography of implementation studies 1994–2012. *Language Learning & Technology*, 17(3): 157–225. https://www.lltjournal.org/item/2830

Burston, J. (2014) A survey of MALL curriculum integration: What the published research doesn't tell. *CALICO Journal*, 31(3): 303–322. https://doi.org/10.11139/cj.31.3.303-322

Burston, J. (2015) Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL*, 27(1): 4–20. https://doi.org/10.1017/S0958344014000159

Burston, J. (2021) Unreported MALL studies: What difference do they make to published experimental MALL research results? In Morgana, V. & Kukulska-Hulme, A. (eds.), *Mobile assisted language learning across educational contexts* (Chapter 2). Abingdon: Routledge. https://doi.org/10.4324/9781003087984-2

Burston, J. & Giannakou, K. (2021) MALL language learning outcomes: A comprehensive meta-analysis 1994–2019. *ReCALL Journal*, to appear.

Callan, S. (1994) *Can the use of hand-held personal computers assist transition students to produce written work of excellent quality?* Ontario: Wentworth County Board of Education.

Chen, Z., Chen, W., Jia, J. & An, H. (2020) The effects of using mobile devices on language learning: A meta-analysis. *Educational Technology Research and Development*, 68: 1769–1789. https://link.springer.com/article/10.1007/s11423-020-09801-5

Cheung, A. C. K. & Slavin, R. E. (2012) *Effectiveness of educational technology applications for enhancing reading achievement in K-12 classrooms: A meta-analysis*. Baltimore: Center for Research and Reform in Education, Johns Hopkins University.

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

Cheung, A. C. K. & Slavin, R. E. (2013) The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, 9: 88–113. https://doi.org/10.1016/j.edurev.2013.01.001

Cho, K., Lee, S., Joo, M.-H. & Becker, B. (2018) The effects of using mobile devices on student achievement in language learning: A meta-analysis. *Education Sciences*, 8(3): 105–121. https://www.mdpi.com/2227-7102/8/3/105/pdf

Chwo, G. S. M., Marek, M. W. & Wu, W.-C. V. (2018) Meta-analysis of MALL research and design. *System*, 74: 62–72. https://doi.org/10.1016/j.system.2018.02.009

Clark, R. & Sugrue, B. (1991) Research on instructional media. In Anglin, G. J. (ed.), *Instructional technology: Past, present, and future*. Englewood: Libraries Unlimited, 327–343.

Cochran, W. G. (1954) The combination of estimates from different experiments. *Biometrics*, 10(1): 101–129. https://doi.org/10.2307/3001666

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

Creswell, J. W. (2014) *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Upper Saddle River: Pearson Education.

Egger, M., Davey Smith, G., Schneider, M. & Minder, C. (1997) Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109): 629–634. https://doi.org/10.1136/bmj.315.7109.629

Elgort, I. (2018) Technology-mediated second language vocabulary development: A review of trends in research methodology. *CALICO Journal*, 35(1): 1–29. https://journals.equinoxpub.com/index.php/CALICO/article/view/34554/pdf

Ellis, P. D. (2010) *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511761676

Eutsler, L., Mitchell, C., Stamm, B. & Kogut, A. (2020) The influence of mobile technologies on preschool and elementary children's literacy achievement: A systematic review spanning 2007–2019. *Educational Technology Research and Development*, 68(4): 1739–1768. https://doi.org/10.1007/s11423-020-09786-1

Field, A. P. & Gillett, R. (2010) How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63(3): 665–694. https://doi.org/10.1348/000711010X502733

Fisher, D., Frey, N. & Hattie, J. (2016) *Visible learning for literacy, grades K-12: Implementing the practices that work best to accelerate student learning*. Thousand Oaks: SAGE.

Guanuche, A., Eiriz, O. & Espí, R. (2020) Corrective feedback through mobile apps for English learning: A review. In Narváez, F. R., Vallejo, D. F., Morillo, P. A. & Proaño, J. R. (eds.), *Smart technologies, systems and applications: First international conference, SmartTech-IC 2019: Quito, Ecuador, December 2–4, 2019: Proceedings*. Cham: Springer, 229–242. https://doi.org/10.1007/978-3-030-46785-2_19

Higgins, J. P. T. & Thompson, S. G. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11): 1539–1558. https://doi.org/10.1002/sim.1186

Huang, Y.-M., Liang, T.-H., Su, Y.-N. & Chen, N.-S. (2012) Empowering personalized learning with an interactive e-book learning system for elementary school students. *Educational Technology Research and Development*, 60(4): 703–722. https://doi.org/10.1007/s11423-012-9237-6

Kamasak, R., Özbilgin, M., Atay, D. & Kar, A. (2020) The effectiveness of mobile-assisted language learning (MALL): A review of the extant literature. In Moura, A. S., Reis, P. & Cordeiro, M. N. D. S. (eds.), *Handbook of research on determining the reliability of online assessment and distance learning*. Hershey: IGI Global. 194–212. https://doi.org/10.4018/978-1-7998-4769-4.ch008

Klimova, B. & Zamborova, K. (2020) Use of mobile applications in developing reading comprehension in second language acquisition – A review study. *Education Sciences*, 10: 391–411. https://doi.org/10.3390/educsci10120391

Lee, S.-M. (2019) A systematic review of context-aware technology use in foreign language learning. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2019.1688836

Lee, Y.-S., Sung, Y.-T., Chang, K.-E., Liu, T.-C. & Chen, W.-C. (2014) A meta-analysis of the effects of learning languages with mobile devices. *Lecture Notes in Computer Science*, 8699: 106–113. https://link.springer.com/chapter/10.1007%2F978-3-319-13296-9_12

Liao, Y.-K. (1999) Effects of hypermedia on students' achievement: A meta-analysis. *Journal of Educational Multimedia and Hypermedia*, 8(3): 255–277. https://pdfs.semanticscholar.org/2f27/97b15cbd249367f04897089df0e739888e3b.pdf

Mahdi, H. (2017) Effectiveness of mobile devices on vocabulary learning: A meta-analysis. *Journal of Educational Computing Research*, 55(3): 1–21. https://kku-sa.academia.edu/HassanMahdi

Matt, G. E. & Cook, T. D. (2009) Threats to the validity of generalized inferences. In Cooper, H., Hedges, L. V. & Valentine, J. C. (eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation, 537–560. https://www.scholars.northwestern.edu/en/publications/threats-to-the-validity-of-generalized-inferences

McClanahan, B., Williams, K., Kennedy, E. & Tate, S. (2012) A breakthrough for Josh: How use of an iPad facilitated reading improvement. *TechTrends*, 56(3): 20–28. https://doi.org/10.1007/s11528-012-0572-6

Moeyaert, M., Ugille, M., Beretvas, S. N., Ferron, J., Bunuan, R. & Van den Noortgate, W. (2017) Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press

multilevel meta-analysis. *International Journal of Social Research Methodology*, 20(6): 559–572. https://doi.org/10.1080/13645579.2016.1252189

Peng, H., Jager, S. & Lowie, W. (2020) Narrative review and meta-analysis of MALL research on L2 skills. *ReCALL*. Advance online publication. https://doi.org/10.1017/S0958344020000221

Shadiev, R., Liu, T. & Hwang, W.-Y. (2020) Review of research on mobile-assisted language learning in familiar, authentic environments. *British Journal of Educational Technology*, 51(3): 709–720. https://doi.org/10.1111/bjet.12839

Sterne, J. A. C., Egger, M. & Davey Smith, G. (2001) Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal*, 323(7304): 101–105. https://doi.org/10.1136/bmj.323.7304.101

Sung, Y.-T., Chang, K.-E. & Yang, J.-M. (2015) How effective are mobile devices for language learning? A meta-analysis. *Educational Research Review*, 16: 68–84. https://doi.org/10.1016/j.edurev.2015.09.001

Sung, Y.-T., Lee, H.-Y., Yang, J.-M. & Chang, K.-E. (2019) The quality of experimental designs in mobile learning research: A systemic review and self-improvement tool. *Educational Research Review*, 28: 1–21. https://doi.org/10.1016/j.edurev.2019.05.001

Suurmond, R., van Rhee, H. & Hak, T. (2017) Introduction, comparison, and validation of *Meta-Essentials*: A free and simple tool for meta-analysis. *Research Synthesis Methods*, 8(4): 537–553. https://doi.org/10.1002/jrsm.1260

Viberg, O. & Grönlund, Å. (2012) Mobile assisted language learning: A literature review. In Specht, M., Sharples, M. & Multisilta, J. (eds.), *mLearn: Mobile and contextual learning: Proceedings of the 11th International Conference on Mobile and Contextual Learning*. Aachen: CEUR-WS.org, 9–16. https://www.researchgate.net/publication/277832223_Mobile_Assisted_Language_Learning_A_Literature_Review

About the authors

Jack Burston holds the position of Honorary Research Fellow in the Language Centre of the Cyprus University of Technology. His current research is focused on mobile-assisted language learning and advanced-level foreign language instruction. Jack is a member of the Editorial Board of the *ReCALL* journal and *Language Learning & Technology*.

Konstantinos Giannakou holds the position of Lecturer in Public Health at the European University Cyprus. He is a specialist in the systematic assessment of meta-analyses using advanced meta-research methods. He holds a PhD in Environmental and Public Health from the Cyprus International Institute for Environmental and Public Health, Cyprus University of Technology.

Author ORCiD. Jack Burston, https://orcid.org/0000-0003-2905-5585
Author ORCiD. Konstantinos Giannakou, https://orcid.org/0000-0002-2185-561X

https://doi.org/10.1017/S0958344021000240 Published online by Cambridge University Press