ReCALL (2024), 36: 1, 22–39
doi:10.1017/S0958344023000149

RESEARCH ARTICLE

Quantitative experimental L2 acquisition MALL studies: A critical evaluation of research quality

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

Androulla Athanasiou
Cyprus University of Technology (androulla.athanasiou@cut.ac.cy)

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

Abstract

With more than 1,200 publications over the past two decades, experimental mobile-assisted language learning (MALL) studies targeting second/foreign language (L2) acquisition outcomes are certainly not lacking in quantity. Their research quality, on the other hand, has often been brought into question, most notably with regard to the adequacy of their assessment instruments and statistical analyses. Yet limiting the determination of research quality to the evaluation of testing procedures, and the statistical analysis of the results they produce, ignores the critical relevance of the underlying research parameters that generate the results in the first place. A comprehensive evaluation of quantitative experimental L2 acquisition MALL research quality, encompassing design as well as assessment instruments and statistical analysis, thus remains to be undertaken. The present investigation endeavors to do so based on an extensive compilation of 737 MALL studies published between 2000 and 2021. The research quality of these publications is evaluated according to four main parameters: language acquisition moderators, treatment intervention conditions, assessment instruments, and statistical analysis. These are applied according to a modified version of the Checklist for the Rigor of Education-Experiment Designs (CREED), which classifies research design quality into five levels: low, medium-low, medium, medium-high, high. With over three quarters of all studies falling within the low category, the result leaves much to be desired. Since the modified CREED algorithm developed here can equally be applied to studies from their inception, it offers a way forward to improve the research quality of future experimental MALL studies.

Keywords: Experimental MALL; Language Acquisition Outcomes; Research Quality; CREED Evaluation Algorithm; Mobile assisted language learning (mall); Research methodology

1. Introduction

Although a relatively young field of research, over the past two decades, quantitative experimental mobile-assisted language learning (MALL) studies targeting second/foreign language (L2) acquisition outcomes have nonetheless been the topic of more than 1,200 publications. While certainly not lacking in quantity, the research quality of these studies has often been brought into question (Burston, 2015; Burston & Giannakou, 2022; Chwo, Marek & Wu, 2018; Elgort, 2018; Lee, 2019; Shadiev, Liu & Hwang, 2020; Viberg & Grönlund, 2012). However, to date, this failing

Cite this article: Burston, J., Athanasiou, A. & Giannakou, K. (2024). Quantitative experimental L2 acquisition MALL studies: A critical evaluation of research quality. *ReCALL* 36(1): 22–39. https://doi.org/10.1017/S0958344023000149

© The Author(s), 2023. Published by Cambridge University Press on behalf of EUROCALL, the European Association for Computer-Assisted Language Learning. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

has only been systematically considered in one recent evaluation (Hou & Aryadoust, 2021). Though an important contribution to the evaluation of MALL research quality, the analysis of the latter is limited in a number of critical respects. First, its bibliographical database is restricted to 174 studies initially located within a single academic database (Scopus). In reality, during the time period covered in this meta-analysis (2008–2020), some 870 quantitative experimental MALL studies appeared in a variety of journals, conferences and MA/PhD theses (Burston, 2021b). The database upon which this meta-analysis rests is further constrained to just L2 English studies. Moreover, the authors focus primarily on measures of test instrument reliability and validity. A comprehensive evaluation of experimental L2 MALL research quality, encompassing design as well as assessment instruments and statistical analysis, thus remains to be undertaken. This is the goal of the present study.

In order to adequately evaluate the quality of quantitative experimental L2 MALL research, three conditions must be met. First, the extent to which an evaluation accurately reflects the true quality of MALL research depends critically upon the comprehensiveness of the bibliographical database upon which it is based. To meet the requirement of comprehensiveness, publication references must be sought from the broadest range of sources possible. Second, a clear definition of what constitutes adequate research design and statistical analysis must be provided and justified. Third, a consistent and practical algorithm is needed that can serve to both evaluate the experimental rigor of existing MALL implementations and guide future studies. These three factors are the focus of the following sections of this paper.

2. MALL studies database compilation

To evaluate the research design and statistical analysis of published experimental quantitative L2 acquisition MALL studies, an initial database of 1,237 experimental MALL investigations for the period 1994–2021 was compiled from all sources: journals, conference presentations, conference proceedings, books, book chapters, doctoral dissertations, master's theses, project reports, blogs, newspaper articles, etc. This figure excludes 70 duplicate studies, typically conference presentations or proceedings later published as journal articles or book chapters. No restrictions were placed on the target language or the language in which the studies were written. These references were extracted from the *General MALL Bibliography 1994-2020* provided in Burston (2021b), updated through 2021.[1] The latter was compiled through a process of bibliography mining, as described in Burston (2021a). Essentially, this involves recursively extracting bibliographies from MALL studies. This was augmented by the references from the 82 MALL meta-analyses published between 2006 and 2022 (Burston, 2021b, augmented to include 2021–2022).

The references in these meta-analyses 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*. In addition, mentions of papers in ResearchGate and Academia.edu citing the first author's MALL publications provided a substantial ongoing stream of studies for further bibliographical searches using the same keywords. The process of bibliography extraction was recursively applied until no new publications were discovered.

For this investigation, it was possible to consult 1,185 of the 1,237 experimental L2 MALL studies that appeared between 1994 and 2021. Since this represents 96% of the database, it may confidently be regarded as highly representative. Of the total consulted, 749 published between 2000 and 2021 reported quantitative, objectively determined language learning outcomes – that is,

[1]Running to nearly 300 pages before the update, space limits preclude the inclusion of this bibliography here. It can, however, be accessed via the link in the reference to Burston (2021b).

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

Table 1. Objective quantitative experimental L2 MALL publications

523 Journal articles
137 Conference proceedings
15 Book chapters
18 Master's theses
19 Doctoral dissertations
712

based on formal assessments specifically related to the targeted language learning areas (e.g. vocabulary, grammar, reading, writing, etc.). The remainder involved subjective self-assessments of learning gains, evaluations of student perceptions of MALL treatments or language usage attitudinal changes. Among the objective studies, 737 (appearing in 712 publications) demonstrably came from peer-reviewed sources (journals, conference proceedings, book chapters) or supervised postgraduate research (master's theses/PhD dissertations). Since peer review and publication oversight are essential gatekeepers of research quality, only these are considered in this evaluation. Their distribution is summarized in Table 1.

3. Research design and statistical analysis criteria

3.1 Statistical analysis

The concern about research quality expressed regarding the primary studies in L2 MALL meta-analyses in fact reflects a problem affecting L2 experimental acquisition studies in general, independently of technology. Much of the research of Luke Plonsky and his associates has focused specifically on this issue (Larson-Hall & Plonsky, 2015; Plonsky, 2011, 2013, 2014; Plonsky & Gass, 2011). As with Hou and Aryadoust (2021) in MALL, however, nearly all the attention has been directed towards assessment prerequisites and statistical analyses. In particular, this body of research demonstrates (again and again) the lack of pre-tests and measures of test instrument reliability and validity. So, too, the failure to report means, standard deviations, *p* significance values and effect sizes is frequently attested.

Needless to say, the importance of these assessment prerequisites and statistical analyses is no less critical to any evaluation of MALL research quality. However, the research quality of quantitative experimental L2 MALL studies, and indeed L2 experimental acquisition studies in general, is not uniquely determined by assessments and their statistical analysis. As Light, Singer and Willet (1990: viii) not so delicately put it, "You can't fix by analysis what you bungled by design." In fact, two other major factors contribute crucially to the quality of reported outcomes: language acquisition moderators and treatment intervention conditions.

3.2 Language acquisition moderators

A number of elements relating to language acquisition can affect the learning outcomes of quantitative L2 MALL research. Four in particular need to be specified to be able to properly interpret assessment and statistical results. The influence of age and sex upon language learning has been recognized in second language acquisition (SLA) research for decades (Al Ghabra, 2015; Block, 2002; Collier, 1987; DeKeyser, Alfi-Shabtay & Ravid, 2010; Palea & Boştină-Bratu, 2015; Rahman, Pandian, Karim & Shahed, 2017; Ramsey & Wright, 1974; Zoghi, Kazemi & Kalani, 2013). Closely related to age is the learning environment, and related pedagogical methodologies, in which experimental treatments are undertaken. Interventions that work well with children in a primary school setting may prove ineffective with adolescents in a high school, just as those

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

applied to the latter may prove unsuitable for university postgraduates or migrants in adult education centers. Likewise, the L2 linguistic competence level of participants must be made known. Treatments intended for beginner-level language learners could be as inappropriate for intermediate-level learners as those targeting advanced-level learners. Moreover, the linguistic competence level of participants needs to be objectively substantiated, preferably by reference to a recognized standardized assessment metric (e.g. CEFR/ACTFL-aligned tests, TOEFL/TOEIC scores, etc.). Unless the age, sex, academic environment and substantiated L2 linguistic competence level of participants are specified, it is not possible to meaningfully evaluate the effectiveness of any MALL experimental treatment.

3.3 Treatment intervention conditions

Despite the considerable attention paid to research design in SLA literature, very little guidance is to be found relating to details of treatment intervention conditions. These omissions are so evident that it could be thought that they are simply regarded as too obvious to mention. For example, while the necessity of formulating explicit research questions is a sine qua non (e.g. Hudson & Llosa, 2015; Phakiti, De Costa, Plonsky & Starfield, 2018; Plonsky, 2015), not so the fact that the specific language focus targeted (e.g. reading, writing, speaking, listening, vocabulary, grammar, etc.) must also be identified. Likewise, mention is lacking of the need to provide a detailed description of the pedagogical materials upon which treatments are based (i.e. books, videos, CDs, internet sites, computer-based programs, mobile apps, etc.). Just as critical, and just as consistently passed over in silence, is the need to explain treatment procedures – that is, exactly how participants actually used the treatment materials.

While treatment duration is commonly acknowledged to affect experimental language learning outcomes and the validity of analytical results, aside from general advice that the longer the better, minimum requirements are not specified. In large part, this is a reflection of the practical institutional constraints on the conduct of SLA research. Equally omitted, with no such justification, is the need to indicate the frequency of the treatment intervention during an experiment. This is critical information, for without it there is no way to determine how much exposure participants had to the treatment. The same treatment applied an hour per day five days a week for four weeks involves far more exposure than one that was only applied a half hour per day once a week for 10 weeks.

As with treatment duration, sample size is known to affect the validity of experimental findings, notably because small sample sizes tend to exaggerate the perceived effectiveness of outcomes (Cheung & Slavin, 2012; Liao, 1999). Yet, because SLA studies usually take place in institutional educational environments, sample sizes are almost always determined by the practical limits of available class sizes. Typically, in any academic environment, this amounts to only a few dozen participants. As a consequence, again aside from the general advice that bigger is better, minimum sample size requirements are not specified. That being said, whatever the size, the exact number of participants in treatment groups must be specified in order for critical statistical analyses like means and standard deviations to be calculated.

The one area outside of assessment instruments and statistical analysis where SLA literature notably dwells on research design is the identification and control of independent treatment variables. These are the factors hypothesized to determine experimental outcomes – that is, the dependent variables. One frequent problem in this regard in experimental MALL studies is the failure to account for the effect of time on task. If, for example, over a 10-week period an experimental group spends five hours per week more on targeted language learning tasks than a control group, any differences in learning outcomes could simply be the result of the extra time spent learning rather than the experimental treatment. Another frequent failure is to describe control groups as following a program of "traditional" instruction without describing what this was. Lacking this information, it is simply not possible to know exactly how this variable affected

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

the dependent variable(s), if at all. Confounding variables pose a similar problem, for example, when an experimental group receives both peer and instructor feedback whereas a control receives only instructor feedback. The two types of feedback together, rather than the treatment itself, could be responsible for the observed outcomes.

In contrast with, and arguably because of, the substantial body of SLA literature on the subject, MALL literature on research design is almost non-existent. Aside from the description of specific implementations, MALL studies (and CALL studies, for that matter) have virtually nothing to say about it. Of the thousands of MALL publications, the recent Hou and Aryadoust (2021) is the only MALL study to systematically focus upon research design. The specific issue of test reliability and validity is raised in Lin and Lin (2019), who note that over half of the studies in their MALL meta-analysis failed to report them. Treatment duration and sample sizes are considered in the inclusion criteria of three other MALL meta-analyses. For inclusion, Burston (2015) required a minimum sample size of 10 per treatment group and a treatment duration of four weeks. Following Clark and Sugrue (1991), Chwo *et al.* (2018) imposed a minimum treatment duration of eight weeks in order to offset the novelty effect. Most recently, Burston and Giannakou (2022) also adopted the eight-week treatment duration requirement and, following Creswell (2015), imposed a minimal sample size of 15 for within-group studies and 30 for between-group studies.

In sum, although it may appear obvious to have to say so, in order to interpret the effect of the treatment in quantitative experimental L2 MALL, the pedagogical materials used and the manner in which they are used must be fully described. So, too, treatment duration as well as intervention frequency and sample size(s) must be fully specified. More specifically, educational researchers recommend a treatment duration of at least eight weeks and a minimal sample size of 15 per participant group. Lastly, experimental treatments must be free of unrecognized and uncontrolled independent variables.

4. Experimental L2 MALL research quality evaluation metric

4.1 Research quality evaluation criteria

Given the dearth of publications relating to research design in MALL generally, and quantitative experimental L2 MALL studies in particular, the establishment and justification of research quality evaluation parameters must of necessity be synthesized from criteria across a range of related academic disciplines (applied linguistics, SLA, instructional technology, education). In so doing, it is possible to judge quantitative experimental L2 acquisition MALL research quality relative to the adequacy of four major evaluation parameters: language acquisition moderators, treatment intervention conditions, assessment instruments and statistical analysis. Although other factors could have been taken into consideration, such as research preliminaries like the traditional literature review and explicit statement of research questions, the proposed evaluation metric demonstrably provides a firm, objective and practical basis for determining the research quality of experimental L2 MALL studies. In all, it operationalizes 20 evaluation criteria, as summarized in Table 2.

4.2 Checklist for the rigor of education-experiment designs (CREED)

While the features summarized in Table 2 provide broad-based evaluation criteria, an important question remains as to how they should be applied. In particular, the relative importance of the four main evaluation parameters must be determined. As Light *et al.* (1990) make explicit, no matter how sophisticated the statistical analysis of experimental outcomes, the research quality of a study is determined well before any number crunching. Sung, Lee, Yang and Chang (2019) take this observation much further through the elaboration of a hierarchical evaluation metric specifically designed to assess the research design quality of quantitative mobile-assisted learning (MAL) studies in general (Figure 1). Based on an extensive analysis of 342 mobile-learning studies

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

Table 2. Quantitative experimental L2 MALL evaluation criteria

Language acquisition moderators	Treatment intervention conditions	Assessment instruments	Statistical analysis
Participant age	Targeted language focus	Pre-test	Means
Participant sex	Materials description	Post-test	Standard deviation
Participant proficiency level	Usage procedures	Reliability	p value (< 0.05)
Institutional environment	Minimum duration (8 weeks)	Validity	Effect size
	Frequency of intervention		
	Minimal sample size (15/30)		
	Group size detail		
	Independent variable control		

written in English between 2006 and 2016, they elaborated a Checklist for the Rigor of Education-Experiment Designs (CREED). This classifies research design quality into five levels: low, medium-low, medium, medium-high, high. The framework underlying CREED was itself derived from three research design evaluation metrics: CONSORT (Moher, Schulz, Altman & Consort Group, 2001; Stone, 2003); Study Design and Implementation Assessment Device (Valentine & Cooper, 2008); and What Works Clearinghouse (2017).

As can be observed in Figure 1 (Sung *et al.*, 2019: 6), CREED operates on an eliminatory principle. A study must meet minimal criteria to be considered for evaluation at the next highest level. For CREED, the most basic evaluation parameter relates to treatment conditions, specifically the distinction between experimental (between-group) and quasi-experimental (within-group) studies. Without a control group, a study is assigned to the lowest category of research quality. Moreover, unless the control group was determined by random selection, a study can at best be considered of medium quality, and this only if two conditions are met. First, the baseline equivalence of experimental and control groups must be established through pre-testing. In the absence of baseline equivalence, a study is relegated to the lowest research quality level. Second, assessment instruments must meet minimal requirements of reliability and validity. If not, the study is classified as medium-low. To be considered of high quality, a study requires a minimal sample size of 30 for both treatment and control groups. Measurements of assessment instrument reliability and validity ultimately determine the assignment of a study to the high research quality level if these conditions are met or medium-high if they are not. Although basic statistical assumptions (e.g. p value, effect size) are also taken into consideration when evaluating the rigor of experimental designs, these are not incorporated into the CREED.

The contribution of Sung *et al.* (2019) is particularly relevant to the present evaluation of MALL studies for two important reasons. First, the CREED framework approach offers a sound basis upon which to develop an algorithm for the guidance and systematic assessment of the research quality of quantitative experimental L2 MALL studies. Second, the results of their analysis allow the quality of MALL research to be placed within the broader context of MAL.

Although CREED is an important step in the right direction, it fails to account for critical evaluation criteria. Treatment conditions do not include any consideration of intervention duration/frequency or the description of pedagogical materials or how they were used. Moreover, being intended as a generic guideline and evaluation metric for mobile learning, it does not consider any particular domain requirements, and thus does not account for language acquisition moderators. It also determines research quality uniquely on the basis of research design independently of the adequacy of statistical analysis (means, standard deviation, p value, effect size).

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

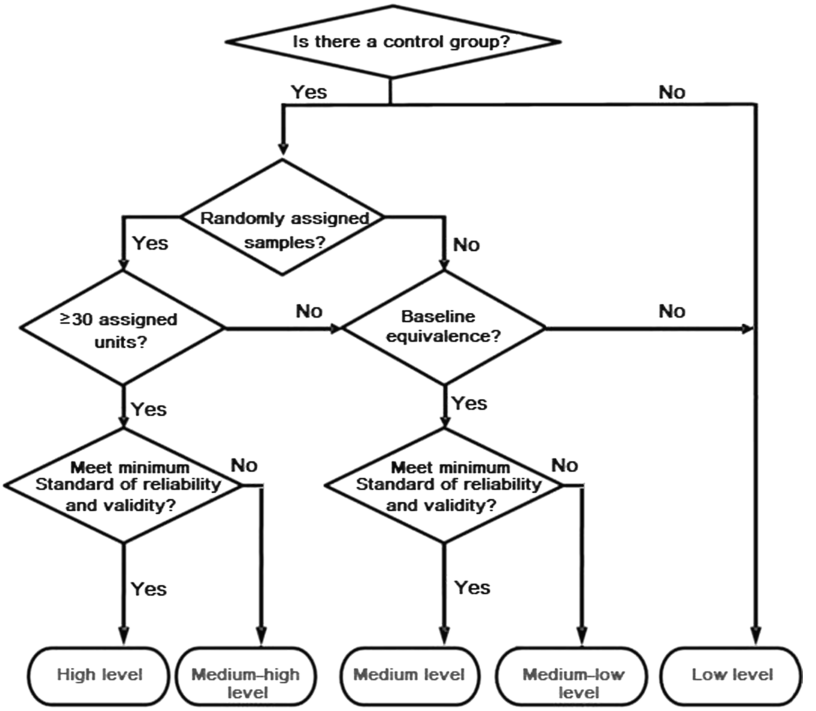


Figure 1. The CREED checklist for determining the levels of rigor of experimental designs (Sung *et al*., 2019)

4.3 Modified CREED framework

In sum, although the hierarchical, eliminatory evaluation procedure underlying the CREED approach has much to recommend it, the application of this framework to quantitative experimental L2 MALL studies requires considerable adaptation. The result is summarized in Figure 2.

In adapting the CREED framework to the evaluation of quantitative experimental L2 MALL studies, the most significant modification involves the incorporation of the missing language acquisition moderators and statistical analysis parameters. In so doing, these become the reference points for the minimal and maximal eliminatory criteria. The description of language acquisition moderators is very straightforward and requires no specialized research training. It is very much a matter of due diligence (not to say common sense), whence its justification as a minimal requirement. Statistical analysis, on the other hand, presupposes a firm understanding of statistical procedures and the ability to apply them properly. It also represents the ultimate stage of learning outcome assessment, whence its designation as a maximal requirement. This is necessarily, and logically, preceded by the assessment instruments parameter. Without reliable and valid pre-/post-tests, the statistical analysis of learning outcomes is meaningless. Treatment intervention conditions have been expanded to include treatment duration/frequency and the description of pedagogical materials, usage procedures and control of independent variables. As with the original CREED framework, experimental quantitative L2 MALL studies are classified into one of five levels: high, medium-high, medium, medium-low, low.

It is to be noted that the priority given to true experimental studies has been removed from the modified CREED framework. While it is always desirable for a quantitative MALL study to be based upon control and experimental groups, applying a true experiment condition as a primary requirement in itself would have prevented over a third (264/737, 36%) of the studies in the MALL

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

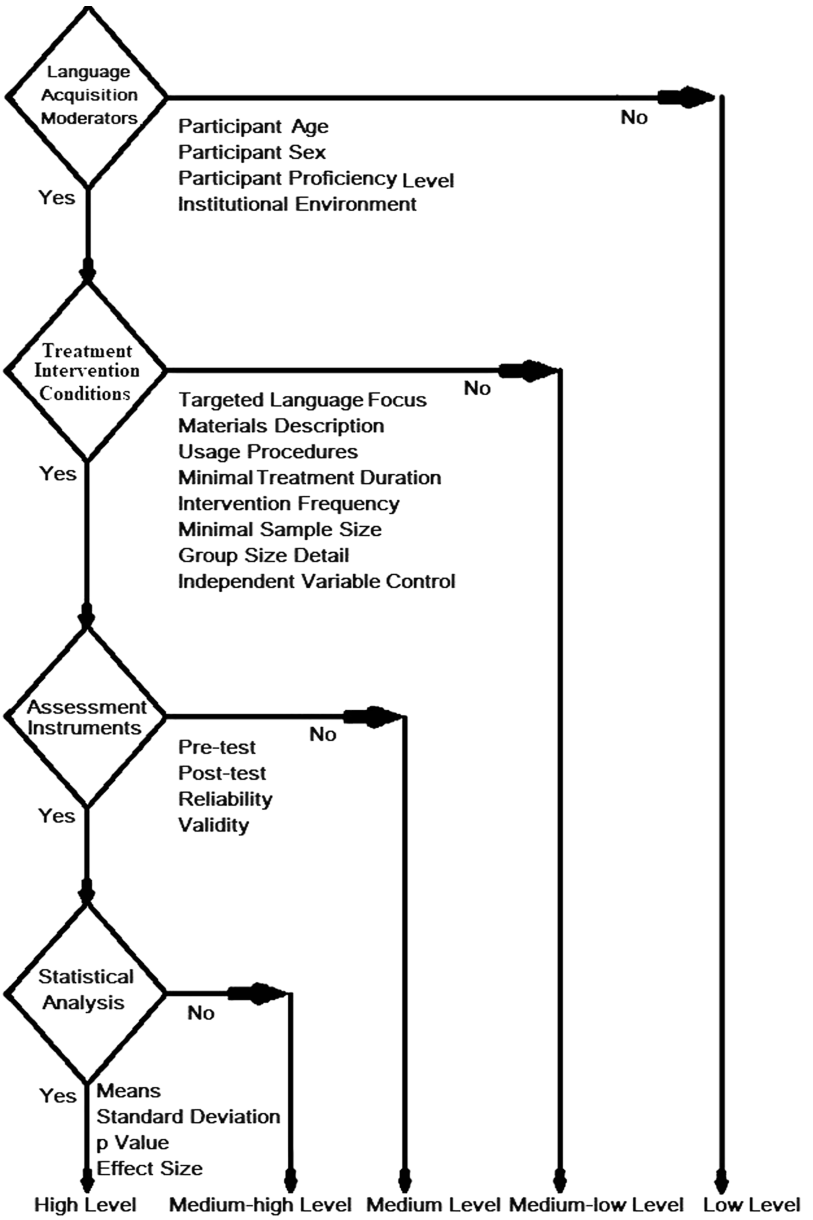


Figure 2. Modified CREED framework for quantitative experimental L2 MALL studies

implementation database assessed in this analysis from rising above the lowest ranking. The practical constraints imposed by the classroom-based environment in which experimental L2 MALL studies almost always take place argue against the imposition of this requirement. Available student numbers are frequently just not sufficient to support between-group studies, all the more so with a minimal sample size of 30 for each treatment group. For this reason, the minimal sample size of 30 has been reduced to a more attainable 15.

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

5. Experimental MALL study analysis

5.1 Application of the modified CREED algorithm

5.1.1 Language acquisition moderators

As indicated earlier, determining the status of language acquisition moderators is a relatively straightforward process, although locating this information often requires looking beyond the Participants section of the studies. Sometimes it is only given in the Abstract or needs to be extracted from data tables. The only problematic parameter here involves the substantiation of participant language competence level when it is not directly related to a recognized standardized evaluation metric (e.g. TESOL, IELTS, DALF, Cervantes, Goethe-Zertifikat, etc.). Where local proficiency tests are administered, unless a detailed description is provided, it can be difficult or impossible to determine whether the linguistic competency level condition has been met.

5.1.2 Treatment intervention conditions

Determining whether or not the description of materials and procedures has been adequately provided frequently requires a subjective assessment of the information provided. Many times, this has to be extracted from data tables or appendices. Another challenge is determining whether or not a study meets the eight-week duration requirement. Unfortunately, there is no standard measurement unit for intervention duration in MALL studies. This can be indicated in sessions, classes, days, weeks and months as well as academic terms, semesters and years. For the purposes of calculating intervention duration, where given in sessions or days (unless otherwise indicated), it is assumed that classes would normally meet three times a week, with 24 sessions/days thus equating to eight weeks. So, too, two months is taken to equal eight weeks and academic terms, semesters, and years all in excess of eight weeks.

Determining whether or not the effect of independent variables upon MALL outcomes has been comprehensively established requires careful attention to what has not been acknowledged in a study. When multiplatform apps are involved, has the actual usage of mobile devices been verified? When there are separate experimental and control groups, aside from the MALL treatment itself, are any advantages provided to the experimental group participants that those in the control group do not receive? Are the learning materials demonstrably equivalent? Do experimental groups receive extra mentoring or instructor feedback? Is time on task equivalent for both groups?

5.1.3 Assessment instruments

When determining whether or not a pre-test has been administered, it is important that any preliminary language proficiency test not be confused with a pre-test. These two kinds of tests serve very different purposes. That of a general proficiency test is to establish the equivalence of the language competency level of experimental and control groups before the treatment begins. A pre-test, on the other hand, must target the same language focus as the treatment to establish a baseline against which progress, or lack thereof, can be demonstrated by the results of a post-treatment assessment of the same targeted language focus. A delayed post-test may also be administered, but this is not a requirement. As with a language proficiency test, pre-/post-tests need to be properly identified where standardized assessments are used or adequately described when local tests are administered. Determining the adequacy of a test description is greatly facilitated when the actual tests (or at least a sample thereof) are attached as an appendix. In the absence of such data, determining the adequacy of a pre-/post-test description can frequently involve a subjective assessment.

Both pre-tests and post-tests need to be formally evaluated for reliability and validity. Reliability is a measure of the consistency of the research results and replicability of the research. It is typically determined by a test/re-test procedure to statistically calculate what is known as Cronbach's alpha. When inherently subjective assessments are being made, for example, of

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

pronunciation, written compositions, communicative competence, interrater reliability needs to be demonstrated. This shows how closely the results of different test assessors agree. Validity is a measure of the extent to which a test actually assesses the treatment focus. In language tests, this is an intrinsically subjective judgement that can only be made by one, or preferably more, experienced outside evaluators (i.e. other language teachers or test specialists).

5.1.4 Scoring procedures

Assessing the research design quality of over 700 quantitative experimental MALL studies relative to 20 evaluation parameters is a challenging undertaking for which special care needs to be taken to ensure accuracy. As indicated earlier, determining whether or not a criterial condition has been met not infrequently requires subjective judgements. So, too, the dispersal of essential information in many studies makes it all too easy to miss critical data. To reduce the effects of these potentially compromising factors to a minimum, the entire database was initially evaluated by the first author, then independently reassessed by the second author. Any discrepancies were resolved by mutual agreement. The third author then resolved any remaining issues relating to the evaluation of statistical analyses.

As originally formulated, the evaluation criteria in CREED were stated in absolute terms: either they were fully met or they were not. However, the modified version is noteworthy not only for the greater number of main parameters but also for the greater number of criteria within each. In order to avoid excessively penalizing partial omissions, application of the algorithm was changed in two ways. First, half credit was allocated in two cases: when the language proficiency level of participants was indicated, but not substantiated, and when pre-/post-tests were given but not adequately identified (in the case of standardized tests) or described (in the case of author-created tests). Second, the conditions of the four main evaluation parameters were deemed to have been met if 87.5% of the data for that parameter were essentially compliant. This percentage corresponds to the highest mathematically possible compliance rate below 100% in the first three main evaluation categories. There being only four subcategories in the statistical analysis category, the highest compliance rate below 100% would be 75% (3/4), which was deemed insufficient to qualify as complying with the evaluation parameter requirements. Relaxation of the all-or-nothing compliance requirement mostly affected the language acquisition moderators, with 13% (83/650) of the studies moving up one or more levels. Specifically, 6% (41/650) advanced to medium-low, 6% (36/650) to medium, and the remaining 1% to medium-high (5/650) and high (1/650).

5.2 Results of the modified CREED research quality evaluation

5.2.1 Overall results

Notwithstanding the more tolerant application of the modified CREED algorithm, quantitative experimental L2 acquisition MALL studies are characterized by the low quality of their research design and statistical analysis, with nearly 77% (567/737) of all studies falling within the low category. When studies evaluated as medium-low are added, the total approaches 88% (646/737). Only 10% (76/737) of MALL studies could be classified as being of medium research quality, and together medium-high and high accounted for just 2% (15/737) of all studies, with MALL studies of fully high quality representing a barely perceptible 0.4% (3/737). Needless to say, such results unequivocally justify concerns raised in the literature regarding MALL research quality.

It is interesting and informative to compare these overall MALL results with those obtained by Sung *et al.* (2019) for the research quality of experimental MAL studies generally. Though MAL studies fare better, low (30%) and medium-low ratings (42%) largely predominate. Medium (16%) and medium-high (10%) ratings together account for little more than a quarter of the MAL studies and fully high-rated studies (1%) remain virtually non-existent. Obviously, the problem of low research quality is thus not restricted to MALL but rather symptomatic of experimental MAL

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

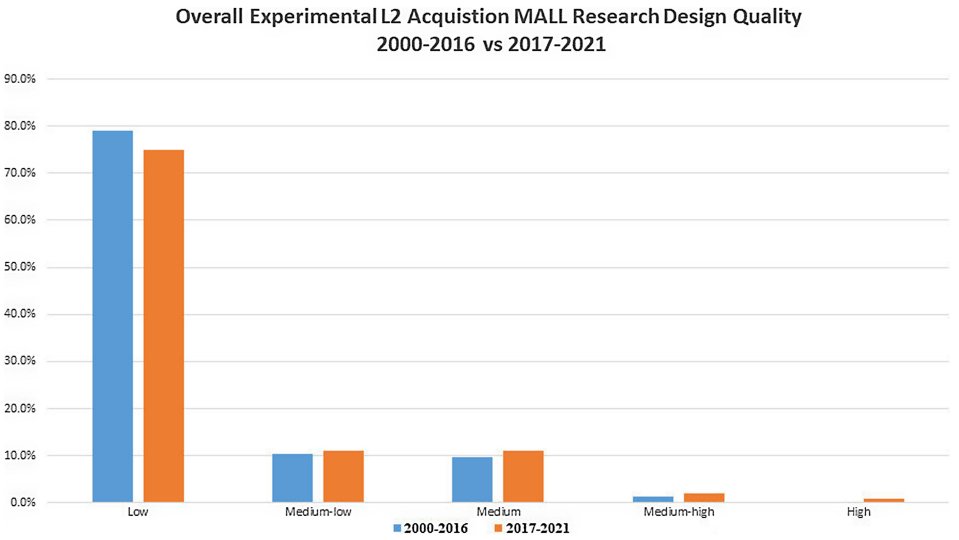


Figure 3. Comparative CREED results for mobile-assisted learning studies for the periods of 2000–2016 and 2017–2021

studies in general. That being said, the particularly low ratings of MALL compared to MAL studies need to be considered in context, for the MAL results are based on a much more limited data set. As previously mentioned, it covers a much shorter time period and is only a fraction the size of the database underlying this MALL evaluation. So, too, its evaluation parameters are much less extensive.

5.2.2 Recent MALL studies

Given that the modified CREED analysis of quantitative experimental L2 acquisition MALL studies here includes nearly 300 investigations that antedate the database underlying Sung *et al.* (2019), it is legitimate to question whether more recent MALL studies might fare better than the overall results of which they are a part. For this reason, the MALL CREED analysis was recalculated based on two time periods: 2000–2016 and 2017–2021 (Figure 3). As can be seen, the results are very much the same, with the most recent period scoring a slightly lower proportion of low ratings (75% vs. 79%) and correspondingly slightly greater proportion across the other ratings. Although all the high ratings are in the most recent period, medium-high and high rates there remain below 2% and 1%, respectively.

5.2.3 Overall results by publication source type

In order to get a clearer picture of the research quality of experimental quantitative L2 MALL studies, it is necessary to determine the effect of publication source upon the results. As is the case with Sung *et al.* (2019), meta-analytic studies frequently restrict their database to journal articles on the assumption that these assure the highest quality research. However, this assumption is not supported by the facts (Figure 4). As might be expected because they are commonly considered to be works in progress, the near totality (133/137, 97%) of conference proceedings fails to reach a medium level of research quality. Book chapters fare better, but still attest 80% (12/15) low ratings. Journal articles, the supposed gold standard for research quality, are only marginally superior with a low rating in nearly three quarters (387/523, 74%) of the studies. To their credit, journal articles do rise above the medium level, but in only 3% (14/523) of the publications. Though very limited in number, journal articles are notably the only MALL studies that receive a high ranking. The smallest proportion of low ratings and

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

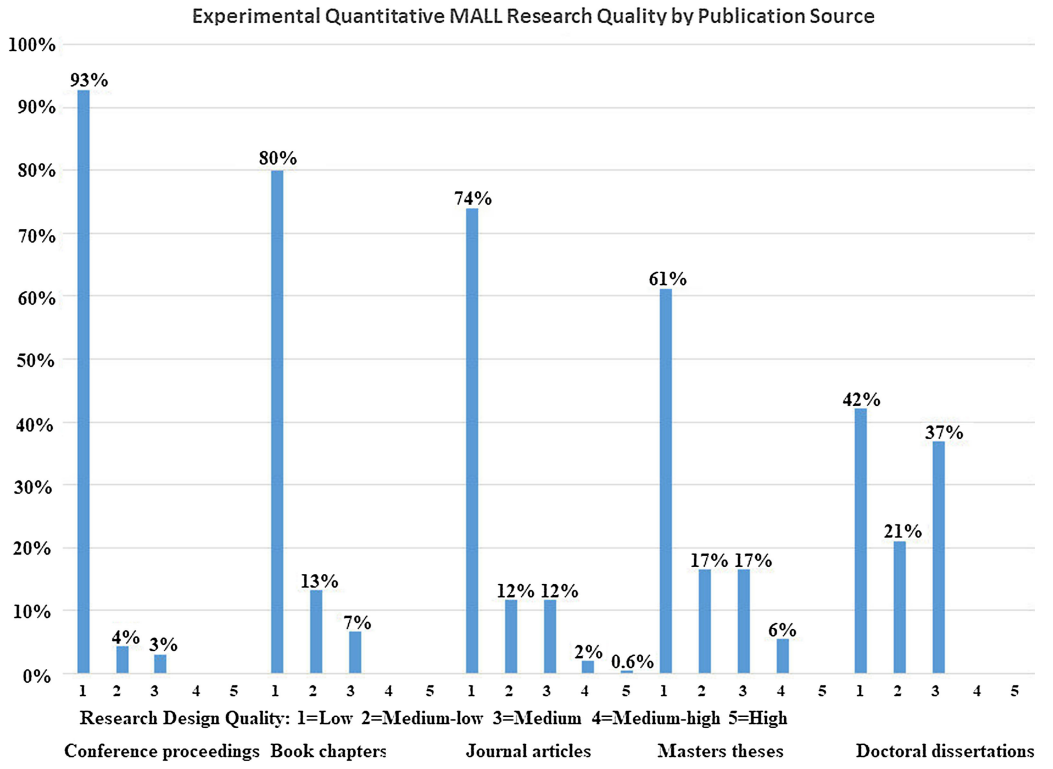


Figure 4. MALL research quality by publication source

greatest of medium research quality are to be found in graduate studies. MA theses attest 61% (11/18) low and 17% (3/18) medium and doctoral dissertations 42% (8/19) low and 37% (7/19) medium ratings. Moreover, the proportion of MA theses at the medium-high level (1/18, 6%) is nearly three times greater than that of journal articles (11/523, 2%).

5.2.4 MALL studies in prominent CALL journals

It might be argued that the poor showing of journal studies results from treating all journals equally, whereas quantitative experimental L2 acquisition MALL studies published in prominent CALL journals would be expected to demonstrate a higher level of research quality. Although the exact membership of CALL journals that should be considered prominent may be open to question, the following six would certainly need to be prime candidates for such recognition: *CALICO Journal*, *Computer Assisted Language Learning*, *EUROCALL Review*, *JALTCALL*, *Language Learning & Technology*, and *ReCALL*. The tabulation of the results for these CALL journals is informative in a number of respects (Table 3). First, it is to be observed that they account for only 13% (66/523) of all the journal articles included in the database underlying this analysis. The remaining 87% are mostly to be found in educational technology publications. A surprising number also appear in journals that have nothing in particular to do with either language, teaching or mobile technology (e.g. *Journal of Engineering*, *Journal of US-China Public Administration*, *Journal of Clinical and Counselling Psychology*, *Studies in Systems, Decision and Control*).

Also to be noted is the disproportional distribution of the MALL studies among the CALL journals. One publication, *Computer Assisted Language Learning*, on its own accounts for nearly

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

Table 3. MALL study research quality in prominent CALL journals

Journal	Low	Medium-low	Medium	Medium-high	High
CALICO Journal	3	1			
Computer Assisted Language Learning	19	4	6		3
EUROCALL Review	5				
JALTCALL	4				
Language Learning & Technology	7	1	4		
ReCALL	4	3	1	1	
Total	42	9	11	1	3
	64%	14%	17%	2%	5%

as many studies (32) as the other five combined (34). It is also the only CALL journal in which quantitative experimental MALL studies of a high research design level are attested. In fact, whatever the publication sources, only three such studies occur in the entire MALL database underlying this investigation and they are all found in *Computer Assisted Language Learning*. Lastly, when the research design rankings are recalculated by extracting the results of MALL studies in prominent CALL journals from the overall journal averages, the research quality differences between the two are even more marked (Figure 5). Specifically, a substantially smaller percentage of low ratings are attested in the prominent CALL publications (42/66, 64%) compared to the other journals (345/457, 77%). Correspondingly, the overall proportion of the three medium ratings is significantly greater in prominent CALL journal MALL studies (21/66, 32%) than in the other journals (112/457, 24%). As previously mentioned, high ratings for MALL studies are only found in prominent CALL journals. Notwithstanding the improvements observed in the research quality rankings of quantitative experimental L2 MALL studies that appear in prominent CALL publications compared to those that occur in other journals, with nearly two thirds of these studies rated low, it must be concluded that journal publication, even in prominent CALL publications, is no guarantee of high research quality.

6. Sources of MALL research quality shortcomings and possible solutions

Given the eliminatory nature of the CREED evaluation parameters, the fact that so many experimental quantitative L2 MALL studies fail to rise above the low level is directly attributable to their failure to adequately respond to the language acquisition moderators criteria. This is worrisome because it reflects a systematic failure to recognize the essential factors that ultimately determine the learning outcomes of a MALL implementation. On the positive side, the educational environment is nearly always specified (711/737, 97%). However, in nearly half the MALL studies in this analysis, no mention is made of the age (356/737, 48%) of the participants. An indication of their sex is omitted even more frequently (386/737, 52%). In nearly as many cases, participant language competency level is unmentioned (343/737, 47%), or simply equated with previous years of language study or enrolment in a language course at level X (191/737, 26%). Fortunately, these problems are potentially quite easy to fix. Noting the age and sex of participants requires only a minimal effort. Specifying and substantiating their L2 proficiency level is only slightly more demanding. Meeting the language competence criterion requires either identifying the assessment instrument (if it is a standardized test) or adequately describing its content and testing procedures (if it is a locally created test).

In principle, once researchers are made aware of the language acquisition moderators requirements, there is no reason for a MALL study to fall below a moderate-low ranking.

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

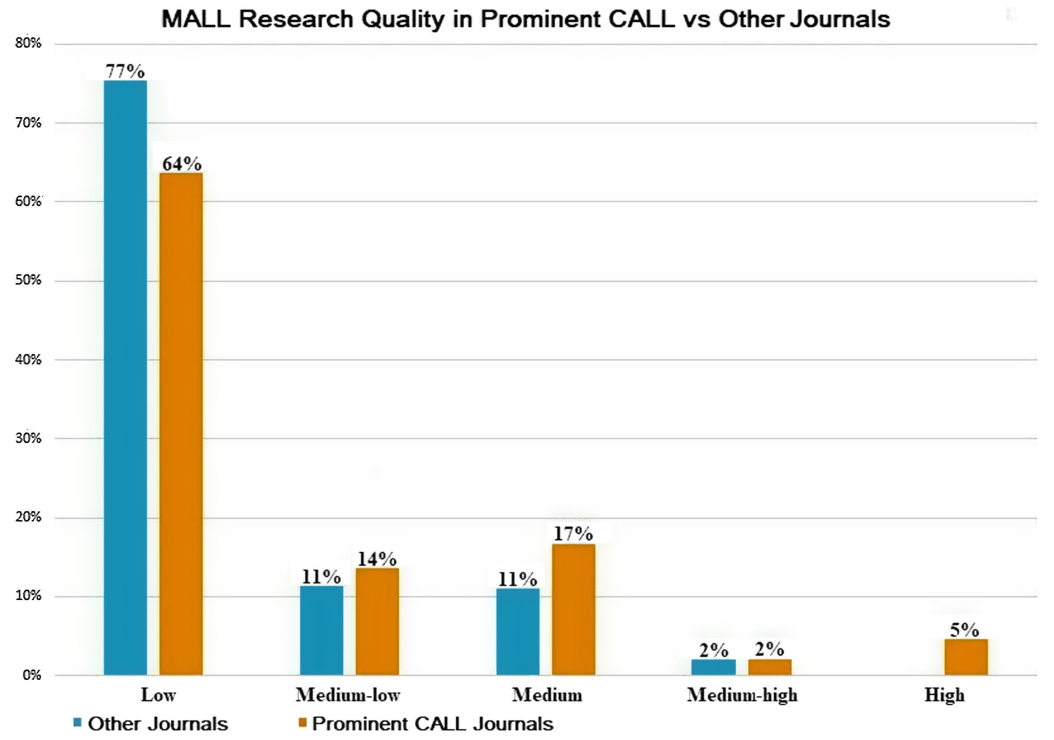


Figure 5. Prominent CALL versus other journal MALL study rankings

Reaching the moderate level, which at present 88% (646/737) of the MALL studies fail to do, is also, for the most part, not a difficult undertaking. Within the treatment intervention conditions evaluation parameter, which functions as the gatekeeper of this level, the language target focus is rarely left unspecified (14/737, 2%). In contrast, the greatest shortcoming encountered is that of short treatment duration, which is encountered 59% (432/737) of the time accompanied by the common failure to specify the frequency of treatment intervention (250/737, 34%). On the other hand, small sample size occurs as a problem in only 13% (92/737) of the studies and the failure to specify the precise number of participants within each group is rare (31/737, 4%). Shortcomings relating to the description of pedagogical materials and procedures are encountered 40% (291/737) and 27% (200/737) of the time, respectively. Problems with uncontrolled variables, especially time on task, occur in over a third of the studies (267/737, 36%). Most academic schedules are organized in quarters, terms or semesters, which typically last at least 10 weeks. So meeting the minimal eight-week requirement is mostly a matter of advance planning, for which it is critical to specify the frequency of treatment exposure. While it is certainly possible to undertake experimental MALL implementations with short durations (and small sample sizes), these should be considered as preliminary trials more appropriate for conference presentations and proceedings. Adequately describing pedagogical materials and procedures is simply something that has to be done and only requires more attention to detail. However, being aware of, and avoiding, possible uncontrolled variables requires considerable forethought, and this is where more experienced researchers need to be consulted before a MALL experiment is implemented. Failing to do so beforehand cannot be remedied afterwards.

Complying with treatment intervention conditions requirements, is thus something that any would-be MALL researcher could and should be able to do to reach a level of moderate research quality. Attaining a higher level, which 96% (63/66) of experimental MALL studies even in the

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

most prestigious CALL journals presently fail to do, is equally feasible, though it requires more specialist expertise in testing and statistical analysis. With regard to the assessment instrument parameter, despite the absolute necessity to pre-test in order to establish a baseline for post-treatment comparison, 18% (134/737) of the studies analyzed here fail to do so, automatically preventing them from advancing beyond a medium ranking. Another 5% (37/737) fail to post-test, relying instead upon progressive observations or the results of generic assessments, such as course exams or final grades. In about a third of the studies, although pre-tests and post-tests were administered, the assessment instruments are inadequately identified or described, post-tests (277/737, 38%) even more so than pre-tests (231/737, 31%). The greatest shortcoming regarding assessment instruments is the pervasive failure to report reliability (517/737, 70%) and validity (611/737, 83%). Although it is easy enough to identify (in the case of standardized assessments) or describe (with locally created assessments) pre-tests and post-tests, the calculation of reliability and determination of validity requires specialist training. In particular, these cannot be established merely by alluding to "the well-known" reliability of test X or the "X years of practical experience" of the researchers, which a number of MALL studies do.

Attaining a fully high level of MALL research quality requires complying with basic statistical prerequisites applied to test results. The easiest of these is the calculation of means, as reflected by the fact that this is reported in 87% (640/737) of the studies. The calculation of standard deviations (553/737, 75%) and p values (591/737, 80%) does not fare as well. However, the greatest shortcoming in statistical analysis is the failure to calculate effect sizes, which goes unreported in 82% (607/737) of the MALL studies in the present database. Although statistical packages are readily available to meet all the conditions of the statistical analysis evaluation parameter, MALL researchers may lack the expertise to avail themselves of these tools. In which case, the only alternative is to collaborate with colleagues who can or hire the outside statistical expertise required.

7. Discussion and conclusion

This investigation set for itself the goal of evaluating the research quality of quantitative experimental L2 acquisition MALL studies that have appeared over the past two decades. Of the 1,237 experimental MALL studies identified through a comprehensive bibliographical search, 737 reported objectively determined, peer-reviewed/academically supervised results and thus were retained for analysis as highly representative of publications in the field. Based on a synthesis of criteria proposed by research experts from a wide variety of disciplines of relevance to MALL, four main parameters, encompassing 20 criteria, were selected as the basis for evaluating the research quality of these MALL studies: language acquisition moderators, treatment intervention conditions, assessment instruments, and statistical analysis. An algorithm originally developed for the evaluation of MAL, CREED, was then adapted to provide a mechanism for the objective ranking of MALL studies into five levels of research quality (low, medium-low, medium, medium-high, high) on the basis of the four MALL evaluation parameters. As with the original CREED algorithm, these were applied in an eliminatory fashion. In order to advance from one level to the next highest, a study had to comply with at least 87.5% of the criteria in the evaluation parameter. As a result, it was shown that concern about the quality of experimental quantitative L2 MALL research was amply justified. Overall, 88% (646/737) of the investigations were ranked as being below medium quality. Moreover, this result showed only slight signs of improving in more recent studies, with low/medium-low ratings of 89% (335/376) before 2017 and 86% (311/361) after. The only real bright spot in the more recent MALL publications was the occurrence of the only high-ranking studies, of which there were but three out of the entire 737 study database, all published in the same journal. When publication sources were evaluated independently, it was shown that conference proceedings fared far worse than the other sources, with nearly 93% (127/137) ranked

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

as low. Journal articles, which account for about three quarters of the publications (523/712, 74%), in effect contributed the most studies from any source. By extracting the experimental MALL studies published in prominent CALL journals from this tabulation, it was confirmed that they were in fact of a higher quality and, notably, included the three high-ranked studies. However, with a 64% (42/66) low rating, this is not enough to justify the assumption that publication even in prominent CALL journals necessarily guarantees high research quality. The percentage of book chapters rated as low (12/15, 80%) was slightly above the overall result (567/737, 77%). While frequently neglected from MALL meta-analyses in preference to journal publications, graduate studies in fact made the best showing. Notwithstanding, 61% (11/18) of MA theses and 42% (8/19) of PhD dissertations were ranked as low. Only one MA thesis (1/18, 6%) and no PhD dissertations rated a medium-high level of research quality.

Upon closer inspection, it was found that the underlying cause of the high proportion of low ratings was directly attributable to easily remediated omissions in the reporting of the age, sex and L2 language competency of participants. Within the treatment intervention conditions parameter, the greatest shortcoming was that of the short duration of experimental interventions and the failure to specify treatment frequency. The failure to adequately account for uncontrolled independent variables was also problematic. While requiring more effort to fix, all of these shortcomings can be remedied by better planning and attention to detail.

Attaining a research quality rating above the medium level should likewise be within reach of most MALL practitioners. Although the results show that the administration of targeted pre-tests and post-test was generally well followed, the reporting of the reliability and validity of these assessment instruments was far too often neglected, and these omissions represent the primary reason why experimental MALL studies failed to reach a medium-high level of research quality. Overcoming these assessment instrument shortcomings is the first really serious hurdle MALL researchers face, for doing so requires technical expertise that may be lacking. The same is true with regard to the statistical analysis criteria that have to be met to attain the high level of research quality. While calculating means, standard deviations and p values appears not to be problematic for the majority of MALL researchers, the same is definitely not true with regard to the reporting of effect size. As indicated previously, the solution here is to either acquire the technical expertise required or collaborate with colleagues or hired professionals who can provide it.

8. The future of experimental quantitative L2 MALL research

As, hopefully, the above evaluation of existing quantitative experimental L2 acquisition MALL studies has demonstrated, their research quality leaves much to be desired. Moreover, the modified CREED algorithm developed to evaluate these studies allows the sources of research problems to be objectively and accurately pinpointed. As a consequence, the algorithm has the potential to be used not only for ex post facto evaluation but also to improve experimental MALL studies from the design stage as well as during and after implementation. Ultimately, this is the key to improving the research quality of published experimental MALL studies. Provided that publishers, conference organizers, reviewers and graduate studies supervisors accept and apply the CREED approach, or develop an evaluation metric variant more suitable to their needs, improvements in the quality of published MALL research are achievable. However, for this to happen, those responsible for maintaining the quality of published MALL research must first of all recognize the extent of the problem, concur on the sources of it, resolve to make this known to future authors, propose remedial action, and insist on compliance with evaluation criteria as a precondition for publication submission. Moreover, to the extent that the shortcomings identified in quantitative MALL research reflect underlying weaknesses in MAL, CALL, and SLA research more generally, the application of this algorithm has the potential to identify and rectify the same defects in these domains as well. Whatever its own shortcomings might be, the modified CREED

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

evaluation algorithm used in this study demonstrably offers a practical starting point for such an undertaking. Additional evaluation parameters, such as research preliminaries like the literature review and explicit statement of research questions, could and should be added. So, too, the experience gained by applying the algorithm in other fields will allow it to be more fine-tuned to meet specific domain requirements.

Ethical statement and competing interests. The authors declare that the research was conducted in the absence of any ethical considerations or competing commercial or financial relationships that could be construed as a potential conflict of interest.

References

Al Ghabra, H. (2015) *The influence of gender and age in SLA*. University of Debrecen Institute of English & American Studies. Department of English Linguistics. https://www.academia.edu/12657019/Gender_Age_and_Second_Language_Acquisition?email_work_card=view-paper

Block, D. (2002) Language & gender and SLA. *Quaderns de Filologia: Estudis Lingüístics*, 7: 49–73.

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. (2021b) *MALL research bibliographies*. https://cut.academia.edu/JBurston/MALL-Bibliographies#mallbibliographies

Burston, J. (2021a) 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*. New York: Routledge, 10–35. https://doi.org/10.4324/9781003087984-2

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

Cheung, A. C. K. & Slavin, R. E. (2012) *The effectiveness of education technology for enhancing reading achievement: A meta-analysis*. Baltimore: Center for Research and Reform in Education, Johns Hopkins University. https://files.eric.ed.gov/fulltext/ED527572.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. https://archive.org/details/instructionaltec0000unse_z4d7

Collier, V. P. (1987) The effect of age on acquisition of a second language for school. *New Focus: The National Clearinghouse for Bilingual Education Occasional Papers in Bilingual Education*, 2: 2–8. https://files.eric.ed.gov/fulltext/ED296580.pdf

Creswell, J. W. (2015) *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Upper Saddle River: Pearson Education. https://www.amazon.com/Educational-Research-Conducting-Quantitative-Qualitative/dp/9332549478

DeKeyser, R., Alfi-Shabtay, I. & Ravid, D. (2010) Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, 31(3): 413–438. https://doi.org/10.1017/s0142716410000056

Elgort, I. (2018) Technology-Mediated second language vocabulary development: A review of trends in research methodology. *CALICO Journal*, 35(1): 1–29. https://doi.org/10.1558/cj.34554

Hou, Z. & Aryadoust, V. (2021) A review of the methodological quality of quantitative mobile-assisted language learning research. *System*, 100: 1–38. https://doi.org/10.1016/j.system.2021.102568

Hudson, T. & Llosa, L. (2015) Design issues and inference in experimental L2 research. *Language Learning*, 65(S1): 76–96. https://doi.org/10.1111/lang.12113

Larson-Hall, J. & Plonsky, L. (2015) Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, 65(S1): 127–159. https://doi.org/10.1111/lang.12115

Lee, S.-M. (2019) A systematic review of context-aware technology use in foreign language learning. *Computer Assisted Language Learning*, 1–25. https://doi.org/10.1080/09588221.2019.1688836

Liao, Y.-K. C. (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

Light, R. J., Singer, J. D. & Willett, J. B. (1990) *By design: Planning research in higher education*. Cambridge, MA: Harvard University Press. https://www.amazon.com/Design-Planning-Research-Higher-Education/dp/0674089316

Lin, J.-J. & Lin, H. (2019) Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*, 32(8): 878–919. https://doi.org/10.1080/09588221.2018.1541359

Moher, D., Schulz, K. F., Altman, D. G. & Consort Group (2001) The CONSORT statement: Revised recommendations for improving the quality of reports of parallel group randomised trials. *The Lancet*, 357(9263): 1191–1194. https://doi.org/10.1016/s0140-6736(00)04337-3

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press

Palea, L.-L. & Boştină-Bratu, S. (2015) Age and its influence on second language acquisition. *Revista Academiei Forţelor Terestre*, 4(80): 428–432. https://www.armyacademy.ro/reviste/rev4_2015/Palea.pdf

Phakiti, A., De Costa, P., Plonsky, L. & Starfield, S. (eds.) (2018) *The Palgrave handbook of applied linguistics research methodology*. London: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-59900-1

Plonsky, L. (2011) *Study quality in SLA: A cumulative and developmental assessment of designs, analyses, reporting practices, and outcomes in quantitative L2 research*. Michigan State University, PhD dissertation. https://eric.ed.gov/?id=ED533659

Plonsky, L. (2013) Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35(4): 655–687. https://doi.org/10.1017/s0272263113000399

Plonsky, L. (2014) Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *The Modern Language Journal*, 98(1): 450–470. https://doi.org/10.1111/j.1540-4781.2014.12058.x

Plonsky, L. (ed.) (2015) *Advancing quantitative methods in second language research*. New York: Routledge. https://doi.org/10.4324/9781315870908

Plonsky, L. & Gass, S. (2011) Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, 61(2): 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Rahman, M. M., Pandian, A., Karim, A. & Shahed, F. H. (2017) Effect of age in second language acquisition: A critical review from the perspective of critical period hypothesis and ultimate attainment. *International Journal of English Linguistics*, 7(5): 1–7. https://doi.org/10.5539/ijel.v7n5p1

Ramsey, C. A. & Wright, E. N. (1974) Age and second language learning. *The Journal of Social Psychology*, 94(1): 115–121. https://doi.org/10.1080/00224545.1974.9923189

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

Stone, A. A. (2003) Editorial: Modification to "instructions to authors." *Health Psychology*, 22(4): 331. https://doi.org/10.1037/0278-6133.22.4.331

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

Valentine, J. C. & Cooper, H. (2008) A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods*, 13(2): 130–149. https://doi.org/10.1037/1082-989x.13.2.130

Viberg, O. & Grönlund, Å. (2012) Mobile Assisted Language Learning: A literature review. Proceedings 11th World Conference on Mobile and Contextual Learning, mLearn 2012 (pp. 1–8). https://ceur-ws.org/Vol-955/papers/paper_8.pdf

What Works Clearinghouse (2017) *What Works Clearinghouse standards handbook Version 4.0*. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Zoghi, M., Kazemi, S. A. & Kalani A. (2013) The effect of gender on language learning. *Journal of Novel Applied Sciences*, 2(S4): 1124–1128. http://www.jnasci.org/wp-content/uploads/2013/12/1124-1128.pdf

About the authors

Jack Burston holds the position of Honorary Research Fellow in the Language Centre of the Cyprus University of Technology. He is a language-teaching specialist with a formal background in theoretical and applied linguistics, second language acquisition and testing. His current research is focused on mobile-assisted language learning (MALL) and advanced-level foreign language instruction. Jack is a current member of the Editorial Board of the *ReCALL* journal, *Language Learning & Technology* journal and *The Journal of Teaching English with Technology*. He also served for many years on the Editorial Board of the *CALICO Journal* and was the Software Review Editor of the *CALICO Journal* for 13 years.

Androulla Athanasiou is an English Language Instructor at the Language Centre of the Cyprus University of Technology. She holds a PhD in English Language Teaching (Warwick University). Her research interests lie in material design, computer-assisted language learning, learner autonomy, collaborative learning, English for specific purposes in higher education and the use of the Common European Framework of References for Languages (CEFR).

Konstantinos Giannakou is an Assistant Professor in Epidemiology and Biostatistics at the European University Cyprus. He earned a BSc in Nursing from the Cyprus University of Technology in 2012, an MSc in Public Health from Oxford Brookes University in 2013, an MSc in Epidemiology from the London School of Hygiene & Tropical Medicine (LSHTM) in 2017, and a PhD in Environmental and Public Health from the Cyprus International Institute for Environmental and Public Health in 2019. Among his research interests is the use of advanced meta-research methods to synthesize evidence from published systematic reviews and meta-analyses.

Author ORCiD. Jack Burston, https://orcid.org/0000-0003-2905-5585
Author ORCiD. Androulla Athanasiou, https://orcid.org/0000-0002-0125-8033
Author ORCiD. Konstantinos Giannakou, https://orcid.org/0000-0002-2185-561X

https://doi.org/10.1017/S0958344023000149 Published online by Cambridge University Press