ReCALL (2024), 36: 1, 58–71
doi:10.1017/S0958344023000162

RESEARCH ARTICLE

Sampling and randomisation in experimental and quasi-experimental CALL studies: Issues and recommendations for design, reporting, review, and interpretation

Oliver James Ballance
Massey University, Te Kunenga ki Pūrehuroa, New Zealand (O.J.Ballance@masey.ac.nz)

Abstract

The majority of research papers in computer-assisted language learning (CALL) report on primarily quantitative studies measuring the effectiveness of pedagogical interventions in relation to language learning outcomes. These studies are frequently referred to in the literature as *experiments*, although this designation is often incorrect because of the approach to sampling that has been used. This methodological discussion paper provides a broad overview of the current CALL literature, examining reported trends in the field that relate to experimental research and the recommendations made for improving practice. It finds that little attention is given to sampling, even in review articles. This indicates that sampling problems are widespread and that there may be limited awareness of the role of formal sampling procedures in experimental reasoning. The paper then reviews the roles of two key aspects of sampling in experiments: random selection of participants and random assignation of participants to control and experimental conditions. The corresponding differences between experimental and quasi-experimental studies are discussed, along with the implications for interpreting a study's results. Acknowledging that genuine experimental sampling procedures will not be possible for many CALL researchers, the final section of the paper presents practical recommendations for improved design, reporting, review, and interpretation of quasi-experimental studies in the field.

Keywords: research methodology; computer-assisted language learning; CALL; sampling; experimental design; quasi-experimental design

1. Introduction

Much like the use of randomised clinical/control trials in medical contexts, experimental studies are a high-status research genre within computer-assisted language learning (CALL) (Gillespie, 2020; Golonka, Bowles, Frank, Richardson & Freynik, 2014; Macaro, Handley & Walter, 2012). However, the distinction between experimental and quasi-experimental studies is not consistently recognised or understood within CALL literature, as many quasi-experimental studies are mislabelled as experimental studies (Section 2). The *APA Dictionary of Statistics and Research Methods* defines an *experiment* as

Cite this article: Ballance, O.J. (2024). Sampling and randomisation in experimental and quasi-experimental CALL studies: Issues and recommendations for design, reporting, review, and interpretation. *ReCALL* 36(1): 58–71. https://doi.org/10.1017/S0958344023000162

> a series of observations conducted under controlled conditions to study a relationship with the purpose of drawing causal inferences about that relationship. An experiment involves the manipulation of an independent variable, the measurement of a dependent variable, and the exposure of various participants to one or more of the conditions being studied. *Random selection of participants and their random assignment to conditions also are necessary in experiments* [emphasis added]. (Zedeck, 2014: 397)

In contrast, it describes a quasi-experimental design as an "experimental design in which assignment of participants to an experimental group or to a control group cannot be made at random" and notes that "this is usually the case in field research" (p. 872). Experiment-like designs that do not randomly select participants can also be considered quasi-experimental designs. This distinction between experimental and quasi-experimental designs is important when we reflect upon the logic of experimental reasoning. Randomised selection and assignment have major implications for probabilistic generalisation from a study's results. This is the case irrespective of any inferential statistics used, as the legitimacy of such analyses is contingent upon study design. As Gass, Loewen and Plonsky (2021) observe with reference to applied linguistics in general, accepting weak study designs and correspondingly dubious interpretations of findings is a threat to the credibility of the field. CALL, just like other fields of applied linguistics, needs to take such appraisals seriously. Consequently, it is essential that CALL researchers, reviewers and readers understand the difference between experimental and quasi-experimental studies.

This methodological discussion paper begins with an overview of common practice within the field based on recent review papers. It then outlines the role of randomisation in sampling, both assignation of participants to conditions and selection of the sample in the first place, clarifying the distinction between experimental and quasi-experimental designs. After considering the relationship between these sampling issues and experimental reasoning, the final section of the article attempts to reconcile the logic of experimental reasoning with the practical constraints that researchers work within, providing recommendations for the design, reporting, reviewing, and interpretation of quasi-experimental studies within the field.

2. Literature review: The state of the field

As Luke Plonsky's *Bibliography of Research Synthesis and Meta-Analysis in Applied Linguistics*[1] demonstrates, there are many recent review articles focused on CALL (e.g. Chen, Zou, Xie & Su, 2021; Gillespie, 2020; Golonka *et al.*, 2014; Hubbard, 2008; Macaro *et al.*, 2012; Plonsky & Ziegler, 2016) as well as its various subfields. For instance, recent papers have reviewed digital games (Peterson, White, Mirzaei & Wang, 2020), simulation games (Peterson 2021), computer-mediated communication (Avgousti, 2018; Çiftçi & Aslan, 2019; Lin, 2015; Mahdi, 2014; Wang & Devitt, 2022), mobile-assisted language learning (Booton, Hodgkiss & Murphy, 2023; Peng, Jager & Lowie, 2021; Shadiev, Hwang & Huang, 2017), flipped classrooms (Turan & Akdag-Cimen, 2020), intercultural education and CALL (Shadiev & Yu, 2022), automated writing evaluation (Fu, Zou, Xie & Cheng, 2022), corpora in language learning (Pérez-Paredes, 2022), and social media (Barrot, 2022; Manca & Ranieri, 2016). Together, these articles provide a powerful overview of several thousand CALL articles, revealing trends and patterns in the research culture as a whole.

Five of the reviews listed above claim that experimental studies are, or are becoming, the most common type of empirical study (Barrot, 2022; Fu *et al.*, 2022; Manca & Ranieri, 2016; Peterson *et al.*, 2020; Turan & Akdag-Cimen, 2020). This is not the case in all subfields, but three more papers that report relatively low numbers of experimental studies then call for more experimental studies (i.e. Shadiev *et al.*, 2017; Shadiev & Yu, 2022; Wang & Devitt, 2022). Most reviews reach similar conclusions about the common shortcomings of the studies that they identify as

[1] https://lukeplonsky.wordpress.com/bibliographies/meta-analysis/

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press

experimental: that they are conducted as relatively short-term interventions (Fu *et al.*, 2022; Gillespie, 2020; Golonka *et al.*, 2014; Lin, 2015; Peterson, 2021; Peterson *et al.*, 2020) and are conducted with small sample sizes (Fu *et al.*, 2022; Gillespie, 2020; Golonka *et al.*, 2014; Peng *et al.*, 2021; Peterson, 2021; Peterson *et al.*, 2020). It is commonly asserted that larger samples and more longitudinal studies are needed to provide more reliable and/or generalisable results (Fu *et al.*, 2022; Gillespie, 2020; Golonka *et al.*, 2014; Peterson, 2021; Peterson *et al.*, 2020). However, although study duration and sample size are both important issues, other sampling decisions have more direct relevance to the generalisability of a study's findings in relation to experimental reasoning.

Few review papers report explicitly on the sampling approaches adopted in the studies reviewed. Unusually, Wang and Devitt (2022) do directly discuss the sampling approaches taken in the subset of experiment-like studies that they reviewed. They note that only three papers randomly assigned participants to control and experiment conditions, correctly equating random assignation to conditions as central to experimentation. They report a further nine studies as quasi-experimental because their designs are similar to experiments in the use of pre- and post-tests, but they lack random assignation of test subjects to test conditions. A second review paper that reports on sampling approaches is Boulton and Cobb's (2017) meta-analysis of corpus use in language learning. Their analysis indicates that only 42% of the papers that reported on the constitution of their control and experimental groups used random assignment to conditions (14 out of 33 papers). But more worryingly, 42 of the 88 experimental groups that they identified gave no indication of how participants were assigned to groups at all. Together, these review papers suggest that between 25% and 16% of experiment-like CALL papers use random assignment. These figures also suggest that this problem is more common in CALL as compared with other areas of applied linguistics, as Plonsky and Gonulal (2015) reported that across subfields, 37% of applied linguistics studies used randomisation.

Several other CALL review papers allude to issues around sampling and distribution of test subjects to test conditions but do not report relevant data. For instance, both Lin (2015) and Manca and Ranieri (2016) mention experimental and quasi-experimental studies, although neither state the relative proportions observed. Macaro *et al.* (2012: 26) report that "in many studies . . . the basis of assignment to experimental and control conditions was not stated". In their overview of 350 studies, Golonka *et al.* (2014: 92) note "convenience samples of existing classes [and] lack of suitable control group" as significant challenges to the generalisability of the body of work reviewed, which suggests that sampling problems are widespread. Similarly, Gillespie (2020: 140) notes in his review of 777 articles that there is a "pattern of researchers being the teachers and using their own students as the subjects of investigation". At a minimum, this indicates that studies are commonly sampling from uninformatively narrow populations of students, but it seems more likely that Gillespie is actually identifying common use of convenience samples. Peterson *et al.* (2020) also seem aware of similar issues in the studies that they review, noting that many study results may be the product of research being conducted in particular contexts, presenting limits to their generalisability.

It is striking that few review papers even acknowledge the difference between experimental and quasi-experimental designs in the studies that they review, and that even when this distinction is acknowledged, only two papers then separated these two kinds of studies when reporting their findings (Boulton & Cobb, 2017; Wang & Devitt, 2022). Furthermore, none of these review papers reported on the sampling frames used, although the widespread use of convenience samples and intact classes was a common global evaluation of the literature under review, as indicated above. It is concerning that many review studies do not appear to recognise any distinction between experimental and quasi-experimental studies, referring to both types of design as experiments. Extrapolating from the recommendations for research practice provided in review papers, there seems to be a belief that more studies, in more settings, with more learners, will allow results to be aggregated and thereby strengthen the body of findings as a whole.

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press

It might be tempting to think that statistical meta-analyses will be able to resolve generalisability issues and make sense of a rapidly increasing, but frequently inconsistent, body of research literature. However, this is a misunderstanding of the intended purpose of meta-analysis and its vulnerabilities. Meta-analysis most clearly addresses issues of statistical power (Sheskin, 2020). That is, meta-analysis addresses insufficient sample size, not problems with sampling approach and valid interpretation. Ross and Mackey (2015: 220) make the following observations:

> A key advantage of meta-analysis is that it provides the basis for what researchers can expect to find if studies similar to those in the meta-analysis are replicated. When meta-analyses result in a nonzero effect size, further experimentation anticipating a zero-effect outcome, that is, one that establishes significance in relation to a null hypothesis, eventually becomes superfluous.

Meta-analysis assumes a high degree of homogeneity among the studies included in the analysis, and differences between the studies in a meta-analysis constitute a significant methodological challenge. In the *Handbook of Parametric and Nonparametric Statistical Procedures*, Sheskin (2020: 1725–1764) states that

> pooling the results of multiple studies that evaluate the same hypothesis is certainly not a simple and straightforward matter. More often than not there are differences between two or more studies which address the same general hypothesis. Rarely if ever are two studies identical with respect to the details of their methodology, the quality of their design, the soundness of execution, and the target populations that are evaluated. [This is problematic because . . .] studies should not only exhibit consistency with regard to the direction of their outcome, but more importantly should exhibit consistency with respect to the magnitude of the effect size present in the k studies . . . [but, realistically,] there are probably going to be differences in methodology and/or the types of subjects employed in some or all of the k studies, and it could be argued that the latter factors could be responsible for yielding different effect sizes.

Measures of variation within the pooled data of a meta-analysis, such as standard deviations and confidence intervals, are essential to interpreting the results of a meta-analysis. The more heterogenous the studies pooled, the wider such measures will be. Hence, it should not be entirely unexpected when Lin (2015: 276) reports of her meta-analytic review of computer-mediated communication that "there is a vast range in the magnitude of effects, clearly indicating a large standard deviation (SD = 0.65) – larger in fact than the value of the average effect size". Indeed, very large standard deviations and/or confidence intervals are typical of the results of meta-analyses reporting on the kind of heterogenous studies found in CALL (e.g. Boulton & Cobb, 2017; Cobb & Boulton, 2015; Lee, Warschauer & Lee, 2019; Zhao, 2004; Ziegler, 2016). Consequently, while we might be confident in the capacity of meta-analyses to indicate whether a certain type of CALL treatment can have a positive effect on learning overall, they are far less likely to identify the conditions under which a particular operationalisation of an intervention can be expected to be effective. Hence, meta-analyses in CALL will typically reveal that a type of treatment (e.g. gamification) can have a wide range of effects, sometimes very effective, sometimes marginally effective, sometimes less effective than the control. However, heterogeneity of studies notwithstanding, even modest meta-analytic claims about general effectiveness should be tempered by the well-recognised file-drawer issue (Sheskin, 2020). That is, studies that are available for inclusion in an analysis will tend to be positively biased because they are more likely to be published than studies with non-significant results. There are techniques for estimating the

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press

extent of such biases, but the file-drawer issue exacerbates interpretation issues when meta-analyses explore very heterogenous studies, reporting corresponding variability.

Meta-analytic reviews do have a valuable role to play, but they cannot be expected to solve the sampling issues confronting experimenters. As Plonsky and Ziegler (2016: 29) argue, one of the most important roles is to "describe and evaluate the research and reporting practices of the domains they review". But the informativeness of meta-analysis results themselves depends on the extent to which genuinely comparable studies are available. In practice, CALL meta-analyses rarely provide a clear indication of the effectiveness of any particular treatment due to major differences between the studies analysed. In the words of Sheskin (2020: 1746),

> one should view a combined effect size value with extreme caution when there is reason to believe that the k effect sizes employed in determining it are not homogeneous. Certainly, in such a case the computed value for the combined effect size is little more than an average of a group of heterogeneous scores, and not reflective of a consistent effect size across studies.

3. Experiments, sampling, and random assignation

Experimentation is concerned with observing the relationship between manipulated independent variables (typically pedagogical interventions in CALL) and dependent variables (for instance, some kind of learning outcome). The aim of experiment design is to create conditions in which sending a signal via manipulation of the independent variables allows the researcher to clearly observe the response in the dependent variables. Variation in the observations that is not caused by manipulation of the independent variable is considered background noise, obscuring observation of the signal–response relationship. For this reason, other sources of variation are also referred to as nuisance variables (Winer, Brown & Michels, 1991). We can think of an experiment design as trying to control, or neutralise, noise from nuisance variables so that a signal–response relationship between the independent and dependent variables can be observed.

Nuisance variables can make it impossible to tell whether changes in the dependent variable resulted from changes in the independent variable or were due to variation in the nuisance variables. For example, if I am interested in using a new technology for language learning, but the motivation levels of my students vary, motivation would be a nuisance variable, and my study would want to ensure that the effect of motivation does not obscure observation of any differences in learning outcomes that may result from the pedagogical intervention. If I use the new technology with a group of highly motivated language learners, and use an old technology with a group of poorly motivated language learners, it is not clear whether the independent variable or the nuisance variable accounts for differences between the observations. The nuisance variable, motivation, is obscuring observation of the variable that the experiment sought to explore. When nuisance variables may have obscured our observations of the signal–response relationship between the independent and dependent variable, they are said to have presented a confound, invalidating the results of the experiment.

Major nuisance variables can be controlled through various different aspects of experiment design, such as blocking (a topic beyond the scope of this paper). However, it is widely recognised that it is not practically possible to control all possible nuisance variables (Shadish, Cook & Campbell, 2001; Winer *et al.*, 1991). Fortunately, randomisation provides a means of neutralising the detrimental effects of large numbers of nuisance variables:

> The principle in this latter case is that all variables not controlled experimentally or statistically should be allowed to vary completely at random … The outcome is that over large numbers of subjects the unique characteristics of subjects which are not controlled are

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press

> distributed evenly over the treatment conditions, the primary purpose of this being to remove bias from the estimates of treatment effects. (Winer *et al.*, 1991: 8)

As well as explaining why sufficiently large sample sizes are needed, the main point here is that randomisation is the researcher's best hope that noise from nuisance variables is unbiased, and so does not present a confound. In other words, with a sufficiently large sample, "Random assignment creates two or more groups of units that are probabilistically similar to each other on the average" (Shadish *et al.*, 2001: 13). With sufficiently large samples, random assignation of participants to control and experimental treatments should result in nuisance variables being evenly distributed across experimental and control conditions.

In relation to using intact classes in studies (i.e. when an entire class is assigned to a condition), the experimental unit is the class, not the individuals in that class. This sampling approach can be referred to as cluster sampling. Here, each class assigned to a condition should be treated as one experimental unit (i.e. as one participant). Power analysis provides the most defensible indication of the number of experimental units a study should use (Faul, Erdfelder, Lang & Buchner, 2007). A study using cluster sampling will need as many intact classes as the power analysis indicates. So, for instance, if a power analysis indicated 30 participants per condition were sufficient to achieve desirable statistical power, the experiment could proceed with 30 intact classes per condition under the same logic.

To illustrate the problem with using intact classes but analysing individual participant results, imagine two intact classes of students: they might have two different teachers, be taught at different times of day, have better or worse classrooms, different timetables – there are numerous variables within this context that might have an effect on the results observed. If the individual students are not assigned to conditions randomly, the potential nuisance variables are systematically distributed across conditions, representing a confound in the design. Hence, intact classes must be treated as individual observations from a design perspective. Studies with only two intact classes have only two experimental unit observations, which is clearly insufficient for random assignment to deliver probabilistic equivalence. This is a major reason for distinguishing between experimental and quasi-experimental designs:

> Quasi-experimental control groups may differ from the treatment condition in many systematic (non-random) ways other than the presence of the treatment. Many of these ways could be alternative explanations for the observed effect, and so researchers have to worry about ruling them out in order to get a more valid estimate of the treatment effect. By contrast, with random assignment the researcher does not have to think *as much* about all these alternative explanations. If correctly done, random assignment makes most of the alternatives less likely as causes of the observed treatment effect at the start of the study. (Shadish *et al.*, 2001: 14)

Random assignation at the level of experimental units is a hallmark of experimentation because it provides assurance that noise from nuisance variables is unbiased and affects both treatments equally.

Shadish *et al.* (2001) are, unsurprisingly, sympathetic to quasi-experimental designs, but they caution that quasi-experimental designs are fundamentally falsificationist. That is, reasonable interpretation of quasi-experimental results requires a programme of experimentation to systematically examine and dismiss alternative explanations derived from plausible nuisance variables. Without such an examination, the independent variable is just one of many possible explanations for the results observed. Shadish *et al.* (2001: 14–15) note that

> in quasi-experiments, the researcher has to enumerate alternative explanations one by one, decide which are plausible, and then use logic, design, and measurement to assess whether

> each one is operating in a way that might explain any observed effect. The difficulties are that these alternative explanations are never completely enumerable in advance, that some of them are particular to the context being studied, and that the methods needed to eliminate them from contention will vary from alternative to alternative and from study to study . . . Obviously, as the number of plausible alternative explanations increases, the design of the quasi-experiment becomes more intellectually demanding and complex – especially because we are never certain we have identified all the alternative explanations. The efforts of the quasi-experimenter start to look like attempts to bandage a wound that would have been less severe if random assignment had been used initially.

Clearly, designing and performing an effective programme of quasi-experimentation is extremely challenging in the context of CALL because of the vast range of factors we know to affect pedagogical outcomes. The commonly reported practice of using intact classes in CALL studies greatly reduces the interpretability of the findings that such studies report. The quality and interpretability of CALL studies would be greatly improved by using genuinely experimental designs that randomly assign participants to treatments. As Ryan (2007: 6) states, “Randomization should be used whenever possible and practical so as to eliminate or at least reduce the possibility of confounding effects that could render an experiment practically useless”. However, before turning to ways of mitigating such issues in quasi-experimental designs (Section 5), it is important to consider the second requirement of experimental sampling: participant selection from sampling frames.

4. Experiments, sampling, and random selection

Thinking about the logic of experimental reasoning, interpreting experiment results as generalisable to a wider population beyond the test subjects depends on the relationship between the sample of test subjects and the population they were drawn from. Under the heading *Sampling and causal generalization*, Shadish *et al.* (2001: 23) note that random selection of participants provides the strongest rationale for ensuring logical connections between an experiment’s result and claims to generalisability. They distinguish *random assignment of participants to treatment conditions* from *random selection of participants* in the first place. However, both are important and derived from the same core rationale: the need to neutralise the effect of unknown sources of variance.

Random selection of participants addresses questions about whether different results would be obtained by conducting the experiment with different participants. If participants are chosen at random from the whole population that an experimenter wishes to generalise to, randomisation provides a means of evenly distributing noise derived from differences among individuals in the population. That is, while random assignation of participants to conditions addresses biases derived from which participants were observed in which treatment conditions, random selection of participants from a population addresses potential biases derived from which participants were included in the study.

To illustrate, imagine a study that involves two classes of postgraduate students from a top university, taught by the researcher, in a particular country where graduation is impossible without passing a specified language exam. Regardless of whether participants are randomly assigned to treatment conditions, it is reasonable to question whether a study would observe a different pattern of results if conducted with participants of a different age, at a different education level, taught by a different teacher, in a different country, at a different institution, with a different relationship to high-stakes language assessments, and so forth. It is random selection of participants from the whole population that the experimenter wishes to generalise to that provides the logic for generalisation by promising to distribute differences between included and excluded participants evenly. The practical issues with such an approach to experimentation are well recognised, but this does not change the facts: "mere membership in the sample is not sufficient for accurately representing a population" (Shadish *et al.*, 2001: 472). Rather, it is "a researcher who randomly samples experimental participants from a … population [who] may generalize (probabilistically) from the sample to all the other unstudied members of that same population" (Shadish *et al.*, 2001: 22). This is a major practical and methodological barrier to conducting experimental research in CALL.

In terms of a practical barrier, random selection of participants has profound implications for the logistics of actually conducting experimental research on any broadly defined population. For instance, a study wanting to generalise to EFL learners would need to randomly sample from all EFL learners around the world. It is not clear how such a study could be conducted by an ordinary CALL researcher; it would be a challenging undertaking even for a large, well-funded international research group. However, the practical challenges also point to the more abstract methodological barrier.

Logically, we can only randomly select participants if we have a sampling frame: a list or some other representation of the population that facilitates selecting a sample in "such a way that (1) all elements of the sample have an equal and constant chance of being drawn on all draws and (2) all possible samples have an equal (or fixed and determinable) chance of being drawn, [as then] the resulting sample is a random sample from the specified population" (Winer *et al.*, 1991: 13). However, for many populations of language learners, it is not obvious how a sampling frame could be created. For instance, education ministries might be able to provide lists of students studying particular languages in state schools, and such a list would be suitable for delineating those populations of learners, in that country, studying those languages, at state schools, and so it would be possible to randomly select participants from this population, but it does not seem remotely feasible to draw up a list of language learners in some more general sense. In effect, this methodological constraint would seem to preclude experimental research on language learners in a broad, general sense. As Jessen (1978: 160) famously cautioned, "Some very worthwhile investigations are not undertaken at all because of the lack of an apparent frame; others, because of faulty frames, have ended in a disaster or in cloud of doubt". It may be inconvenient, but the warrant for generalisation is provided by random selection of subjects from a credible sampling frame: a convenience sample provides no defensible basis for making formal, probabilistic generalisations.

Unsurprisingly, the typical study within CALL employs no reported sampling frame (e.g. Macaro *et al.*, 2012; see also Section 2), and so it is not possible to generalise the results of such studies to any identifiable population. Technically, the results reported in such studies only inform us about the participants observed, a kind of quantitative case study. The application of inferential statistics to such studies, although commonplace, is not hugely informative, because we have little ability to identify the population that inferences can be applied to.

The relationship between sample and population is also the major reason why pre-testing to establish group equivalence is not sufficient, even though such tests should be encouraged as a check on whether random assignation has successfully neutralised nuisance variables within the sample itself. Such tests provide a good indication that the experimental and control group are equivalent on a selected variable, but they cannot assure us that the sample is probabilistically typical of a wider population.

In summary, an understanding of experimental reasoning clearly shows that randomisation is integral to the design of generalisable experimental studies, both randomisation of participants to conditions and randomised selection of participants from a specified population. Without such randomisation, a study is quasi-experimental at best, and results should be interpreted with corresponding caution. That is, although quasi-experimental studies can demonstrate that a particular outcome is possible under certain conditions, it cannot readily delineate the conditions under which such results can be obtained: in effect, it provides a proof of concept, but the absence

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press

of randomisation prevents such studies from providing generalisable insights. Studies that use intact classes, snowball sampling or other forms of convenience sampling are quasi-experimental, should be identified as such, and interpreted this way.

5. Ways forward

The position outlined above, although well established in technical research methods literature, is likely to be unpalatable to many CALL researchers. In fact, one reviewer of this paper questioned whether it even makes sense to pursue experimental methods in relation to something as complex as CALL, pointing out that many CALL researchers have rejected a research model positing dependent and independent variables, opting instead to explore learning environments from an ecological perspective in which the effectiveness of factors within an environment are viewed as inseparable and interdependent (e.g. Marek & Wu, 2014). But, sympathy for this position notwithstanding, it does not change the fact the experimental research paradigm is still the predominant research paradigm within CALL, as shown in the literature review above, and yet few CALL researchers are in a position to undertake experimental research once we recognise the necessity of randomised assignment and selection in experimental research. There are many practical constraints on experiment design in an area as complex as language learning, including very limited funding and resources, codes of ethics and professionalism that would seem to conflict with the implementation of randomisation procedures, and immense pressure to produce a high rate of publication in prestigious journals (Colpaert, 2012). But the gap between experiment design theory and achievable practice does not justify inaccurate or misleading reporting. Rather, it necessitates careful assessment of how research can be best conducted and reported within practical constraints. Because quasi-experimental methods are likely to remain the predominant quantitative paradigm within CALL, this section of the paper presents recommendations for how quasi-experimental studies can be conducted and reported in a way that will enhance their quality and informativeness. Three main recommendations are made for improving the design, reporting and reviewing of quasi-experimental studies:

1. Transparency in reporting and interpreting quasi-experimental studies.
2. Understanding the kind of research questions that basic quasi-experimental designs can address.
3. Triangulating quasi-experimental data with other non-experimental data via
   (a) more measures of participants in relation to a greater range of potential nuisance variables;
   (b) providing a thicker description of the participants and research context;
   (c) combining quasi-experimental methods with qualitative research instruments suited to exploring the potential for nuisance variables to have affected the results obtained from a quasi-experimental study.

The first and most obvious step towards better quality research of this kind in CALL would be for more researchers to acknowledge the distinction between experimental and quasi-experimental research. As discussed above, a real experiment is premised upon specific sampling procedures that provide a warrant for generalising the findings (Shadish *et al.*, 2001; Winer *et al.*, 1991). Even when readers are unaware of the logic that underpins experimental research, there is a common understanding that experiments provide generalisable findings. Consequently, misrepresenting quasi-experimental research as experimental research is hugely problematic. It suggests a level of generalisability to findings that is simply not warranted. Articles claiming to present experimental research should provide a clear statement on three crucial aspects of the design: first, how participants were assigned to experimental conditions or treatments; second,

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press

how participants were selected from the population of interest; and third, a description of the sampling frame used. To be considered experimental research, the first two statements need to meet a definition of randomness, as discussed earlier in this article. The third condition needs to be met for clarity around the population that experimental results are to be generalised to. When studies are unable to provide satisfactory statements on these key aspects of experiment design, they should be described as quasi-experimental studies. Furthermore, alongside issues of nomenclature, researchers should also make sure that their interpretation of findings parallels the warrant for generalisability inherent in the design decisions made. When quasi-experimental designs are used, researchers should be correspondingly modest in regard to the generalisability claims that they make.

In a strict sense, quasi-experimental studies are not generalisable. They provide evidence of whether a treatment *can*, not whether it *does*. This is what Shadish *et al.* (2001) mean when they point out that quasi-experimental designs are fundamentally falsificationist. Without a warrant for generalisation to a population, analysis of a quasi-experimental study simply helps clarify the pattern of results observed in that sample. So, for instance, suppose we questioned whether learners could learn from an online resource as easily as they could learn from a paper-based resource. A quasi-experimental study could attempt to falsify this claim. But if a quasi-experimental study shows a difference in results in favour of the online treatment, we have demonstrated that at least some learners in some circumstances *can* learn more that way. However, for the reasons discussed in this paper, the quasi-experimental design does not entitle the researcher to make any formal claims about whether this relationship exists more generally in some wider population: it cannot inform us about whether this relationship *does* occur generally, irrespective of the many variables that were not controlled for or neutralised. It does not inform us about when we should expect this relationship to hold.

Constraints on formal generalisation from quasi-experimental studies lead to the second recommendation: Researchers should think carefully about which research questions a basic quasi-experimental design can answer, and which of these research questions are genuinely worth asking. Let us consider a typical study in which a treatment (for instance, learning language via a mobile app) is compared against a control or alternative treatment (e.g. paper-based study). Does anyone seriously doubt that it might be possible, for some people, in some circumstances, to learn more using a mobile app than using paper-based resources? If such a proposition is not contested, there seems little justification for attempting to falsify it. Basic quasi-experimental studies have the most value in cases where there is legitimate doubt as to whether a treatment could possibly lead to an effect, or where a treatment is posited to always lead to a particular effect. Even just one study with results counter to such *always* and *never* hypotheses can disprove them. Basic quasi-experimental designs can address these kinds of research questions effectively.

To illustrate, Gillespie (2020) remarks positively that studies comparing paper-based with computer-based interventions are becoming increasingly rare, and so it seems highly doubtful that there really is anyone who seriously questions the potential for computer-based methods to be as effective as paper-based methods, in at least some circumstances. We might contrast this with learning vocabulary by studying concordance lines. At some historical point, it probably did seem entirely reasonable to question whether this was even possible. However, once a quasi-experimental study has shown that a particular relationship is possible (e.g. Cobb, 1999), what value is there in repeating this ungeneralisable test? As meta-analyses of corpus-based approaches have consistently shown (Boulton & Cobb, 2017; Cobb & Boulton, 2015; Lee *et al.*, 2019), learners most assuredly can learn vocabulary more successfully when studying concordances than under various alternative or control conditions. But they also indicate that the amount of benefit obtained varies considerably, and in some cases it proves ineffective. Hence, what we are currently unsure of is the generalisability of any particular observation, and to answer this question, experimental designs are needed. So, to summarise, when researchers are not in a position to conduct experimental studies, but they are able to conduct a basic quasi-experimental study, they

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press

need to think carefully about the kinds of questions that quasi-experimental studies can answer and which of those questions are of interest to the CALL community.

A third way we might look to overcome the limitations of quasi-experimental designs would be to enhance basic quasi-experimental designs by supplementing the basic designs with additional sources of information that can facilitate transferability, or naturalistic generalisation – informal processes of speculation about how the results might relate to other contexts (Duff, 2006). When researchers report that their study was conducted using a convenience sample, such as a snowball sample or using their own classes, their academic integrity should be applauded but, in effect, this just alerts the reader to the fact that randomisation was not used, the study is quasi-experimental, and so the results are not formally generalisable. In essence, quasi-experimental studies are quantitative case studies. While the careful researcher should certainly acknowledge that the generalisability of the findings reported are in doubt – for instance, by alerting the reader that it is unclear whether these results would hold for other groups of learners – such acknowledgement should not entirely preclude a degree of reasonable speculation about how quasi-experimental findings might relate to broader populations. However, if researchers do want to try to promote naturalistic generalisation or transfer from their quasi-experimental findings, then they need to actively facilitate this process for their reader. That is, they need to provide as much information about their participant sample and research context as possible. For every possible nuisance variable, a description or assessment of the participants in relation to that variable strengthens the reader's confidence in transferring the study results to similar populations. There are three main ways that researchers can do this, and all of them concern better reporting in relation to potential nuisance variables.

One possible means of facilitating transfer would be to report more measures of participants in relation to a greater range of potential nuisance variables. We already see this in some studies that report proficiency measures for participants as a means of demonstrating parity between control and experimental groups. While demonstrating parity in proficiency levels is a welcome design facet because it can reassure us about assignment of participants to conditions in relation to that particular variable, it is probably most valuable in quasi-experimental designs by helping readers naturalistically generalise. So, for instance, if we know the learners in a study had a very advanced knowledge of the target language, we might see the results as potentially relevant to an advanced class that we teach, but also be less confident in seeing those findings as relevant to a class of beginners, for example. There are many instruments available in the broader applied linguistics literature that might help researchers describe their participants more informatively, such as measures of motivation, vocabulary size, strategy use, and so on. This does, however, bring to mind once more Shadish *et al.*'s (2001: 14–15) observation about the impossibility of systematically addressing all possible variables, making neutralisation via randomisation look appealing once more.

A second way that researchers could improve the interpretability of quasi-experimental studies is simply providing a thicker description of the participants and research context. While it might not be possible to measure every possible nuisance variable and report this formally, there is still value in simply describing the participants and context in a way that helps readers get a clear picture of the who, where, why, what, when and how of the project. It is reasonably common for researchers to report on the gender of their participants, age, proficiency level and/or language background of participants, and this does go some way towards facilitating naturalistic generalisation, but it is far less common to find a description of the educational environment and prevailing attitudes to language education in that context, factors that seem just as likely to have a profound effect on how any pedagogical intervention plays out (The Douglas Fir Group, 2016; Duff, 2006). For instance, if a study is undertaken at a low-status, rather authoritarian, vocational training college where students lack motivation for education in general and teachers complain that students are only interested in passing the exam, this would seem to be very pertinent information for interpreting the study results, as we can imagine that such contextual factors are

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press

likely to have a strong influence on the results. Thicker descriptions of both participants and research contexts would go a long way to making quasi-experimental studies more genuinely interpretable and informative for readers.

Finally, going beyond informal but informative description, there is great potential for quasi-experimental research to be combined with qualitative research instruments (Hashemi & Babaii, 2013). The fundamental issue with quasi-experimental designs is that the absence of randomisation fails to neutralise nuisance variables, making their influence on the observed results unknown. In contrast, rather than testing the effect of a pre-selected variable, qualitative methods are well suited to exploratory research, asking what factors are operative in an environment. As such, qualitative research instruments have the potential to examine the extent to which nuisance variables may have affected the results obtained from a quasi-experimental study. For instance, researchers could use pre- and post-intervention interviews, observation, or stimulated recall to explore the extent to which unanticipated variables may have influenced the results obtained in the quasi-experimental analysis. This kind of triangulation of research instruments could go a long way to overcoming the most serious shortcomings of quasi-experimental designs. Similar recommendations have been made in the broader context of education research wherein Cebula (2018), for instance, connects the prevalence of quasi-experimental designs in education research to the ongoing (un)replicability crisis in psychology and other fields.

6. Conclusion

CALL has developed rapidly as a field, and it has been quick to adopt sophisticated research techniques from more established fields with long traditions of experimentation, but, crucially, some key elements of experimental design and reasoning have been largely ignored. The logic of experiment design necessitates the use of randomisation to control for nuisance variables: random selection (i.e. selection by lot) from a population, and random assignation of participants to control and experimental treatments or conditions. When these conditions are not met, we are actually employing a quasi-experimental design. At a bare minimum, this distinction should be understood and acknowledged. Quasi-experimental designs are never conclusive, even at a theoretical level, and whether similar results would be obtained in different samples affected by different distributions of nuisance variables remains an open question. Given the difficulty of designing and conducting a genuinely experimental study in an area as complex as CALL, it seems very unlikely that true experimental designs will be widely adopted. Instead, it seems almost certain that quasi-experimental designs will continue to be the mainstay of research in our field. Consequently, it is not only essential that researchers acknowledge when the designs that they use are quasi-experimental and moderate their claims for generalisability appropriately, but also important for researchers to adopt research practices that will facilitate transfer and naturalistic generalisation. This is not simply a question of larger samples or more studies in more diverse contexts. Rather, steps should be taken to directly address the shortcomings of quasi-experimental designs. To such an end, three practices appear promising: first, reporting participant metrics that help to define the sample examined, such as proficiency scores or attitudinal data; second, providing rich, thick description of the participants and context; and third, triangulating quasi-experimental designs with qualitative research instruments designed to explore whether nuisance variables may have significantly affected the results reported in the quasi-experimental portion of the study.

Finally, as one reviewer pointed out, although this paper has focused on research within CALL, it is by no means clear that experiment design practices are superior in other areas of applied linguistics, or even other fields, such as education, psychology, or sociology. The observations

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press

presented in this paper regarding experimental and quasi-experimental research and their respective interpretability are relevant to all fields of research, both within applied linguistics and beyond.

Ethical statement and competing interests. The author declares no competing interests.

References

Avgousti, M. I. (2018) Intercultural communicative competence and online exchanges: A systematic review. *Computer Assisted Language Learning*, 31(8): 819–853. https://doi.org/10.1080/09588221.2018.1455713

Barrot, J. S. (2022) Social media as a language learning environment: A systematic review of the literature (2008–2019). *Computer Assisted Language Learning*, 35(9): 2534–2562. https://doi.org/10.1080/09588221.2021.1883673

Booton, S. A., Hodgkiss, A. & Murphy, V. A. (2023) The impact of mobile application features on children's language and literacy learning: A systematic review. *Computer Assisted Language Learning*, 36(3): 400–429. https://doi.org/10.1080/09588221.2021.1930057

Boulton, A. & Cobb, T. (2017) Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2): 348–393. https://doi.org/10.1111/lang.12224

Cebula, K. (2018) Experimental and quasi-experimental research design in education. In Hamilton, L. & Ravenscroft, J. (eds.), *Building research design in education: Theoretically informed advanced methods*. London: Bloomsbury Academic, 49–68. https://doi.org/10.5040/9781350019539.ch-004

Chen, X., Zou, D., Xie, H. R. & Su, F. (2021) Twenty-five years of computer-assisted language learning: A topic modeling analysis. *Language Learning & Technology*, 25(3): 151–185.

Çiftçi, H. & Aslan, E. (2019) Computer-mediated communication in the L2 writing process: A review of studies between 2000 and 2017. *International Journal of Computer-Assisted Language Learning and Teaching*, 9(2): 19–36. https://doi.org/10.4018/ijcallt.2019040102

Cobb, T. (1999) Breadth and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning*, 12(4): 345–360. https://doi.org/10.1076/call.12.4.345.5699

Cobb, T. & Boulton, A. (2015) Classroom applications of corpus analysis. In Biber, D. & Reppen, R. (eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press, 478–497. https://doi.org/10.1017/cbo9781139764377.027

Colpaert, J. (2012) The "publish and perish" syndrome. *Computer Assisted Language Learning*, 25(5): 383–391. https://doi.org/10.1080/09588221.2012.735101

The Douglas Fir Group (2016) A transdisciplinary framework for SLA in a multilingual world. *The Modern Language Journal*, 100(S1): 19–47. https://doi.org/10.1111/modl.12301

Duff, P. A. (2006) Beyond generalizability: Contextualization, complexity, and credibility in applied linguistics research. In Chalhoub-Deville, M., Chapelle, C. A. & Duff, P. A. (eds.), *Inference and generalizability in applied linguistics: Multiple perspectives*. Amsterdam: John Benjamins, 65–95. https://doi.org/10.1075/lllt.12.06duf

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2): 175–191. https://doi.org/10.3758/bf03193146

Fu, Q.-K., Zou, D., Xie, H. & Cheng, G. (2022) A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2022.2033787

Gass, S., Loewen, S. & Plonsky, L. (2021) Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, 54(2): 245–258. https://doi.org/10.1017/S0261444819000430

Gillespie, J. (2020) CALL research: Where are we now? *ReCALL*, 32(2): 127–144. https://doi.org/10.1017/S0958344020000051

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L. & Freynik, S. (2014) Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, 27(1): 70–105. https://doi.org/10.1080/09588221.2012.700315

Hashemi, M. R. & Babaii, E. (2013) Mixed methods research: Toward new research designs in applied linguistics. *The Modern Language Journal*, 97(4): 828–852. https://doi.org/10.1111/j.1540-4781.2013.12049.x

Hubbard, P. (2008) Twenty-five years of theory in the CALICO Journal. *CALICO Journal*, 25(3): 387–399. https://doi.org/10.1558/cj.v25i3.387-399

Jessen, R. J. (1978) *Statistical survey techniques*. New York: Wiley.

Lee, H., Warschauer, M. & Lee, J. H. (2019) The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, 40(5): 721–753. https://doi.org/10.1093/applin/amy012

Lin, H. (2015) Computer-mediated communication (CMC) in L2 oral proficiency development: A meta-analysis. *ReCALL*, 27(3): 261–287. https://doi.org/10.1017/S095834401400041X

Macaro, E., Handley, Z. & Walter, K. (2012) A systematic review of CALL in English as a second language: Focus on primary and secondary education. *Language Teaching*, 45(1): 1–43. https://doi.org/10.1017/s0261444811000395

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press

Mahdi, H. S. (2014) The impact of computer-mediated communication environments on foreign language learning: A review of the literature. *World Journal of English Language*, 4(1): 9–19. https://doi.org/10.5430/wjel.v4n1p9

Manca, S. & Ranieri, M. (2016) Is Facebook still a suitable technology-enhanced learning environment? An update critical review of the literature from 2012 to 2015. *Journal of Computer Assisted Learning*, 32(6): 503–528. https://doi.org/10.1111/jcal.12154

Marek, M. W. & Wu, W.-C. V. (2014) Environmental factors affecting computer assisted language learning success: A complex dynamic systems conceptual model. *Computer Assisted Language Learning*, 27(6): 560–578. https://doi.org/10.1080/09588221.2013.776969

Peng, H., Jager, S. & Lowie, W. (2021) Narrative review and meta-analysis of MALL research on L2 skills. *ReCALL*, 33(3): 278–295. https://doi.org/10.1017/S0958344020000221

Pérez-Paredes, P. (2022) A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*, 35(1–2): 36–61. https://doi.org/10.1080/09588221.2019.1667832

Peterson, M. (2021) Digital simulation games in CALL: A research review. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2021.1954954

Peterson, M., White, J., Mirzaei, M. S. & Wang, Q. (2020) A review of research on the application of digital games in foreign language education. In Kruk, M. & Peterson, M. (eds.), *New technological applications for foreign and second language learning and teaching*. Hershey: IBI Global, 69–92. https://doi.org/10.4018/978-1-7998-2591-3.ch004

Plonsky, L. & Gonulal, T. (2015) Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning*, 65(S1): 9–36. https://doi.org/10.1111/lang.12111

Plonsky, L. & Ziegler, N. (2016) The CALL–SLA interface: Insights from a second-order synthesis. *Language Learning & Technology*, 20(2): 17–37. https://doi.org/10125/44459

Ross, S. J. & Mackey, B. (2015) Bayesian approaches to imputation, hypothesis testing, and parameter estimation. *Language Learning*, 65(S1): 208–227. https://doi.org/10.1111/lang.12118

Ryan T. P. (2007) *Modern experimental design*. Hoboken: Wiley-Interscience. https://doi.org/10.1002/0470074353

Shadiev, R., Hwang, W.-Y. & Huang, Y.-M. (2017) Review of research on mobile language learning in authentic environments. *Computer Assisted Language Learning*, 30(3–4): 284–303. https://doi.org/10.1080/09588221.2017.1308383

Shadiev, R. & Yu, J. (2022) Review of research on computer-assisted language learning with a focus on intercultural education. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2022.2056616

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2001) *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sheskin, D. J. (2020) *Handbook of parametric and nonparametric statistical procedures* (5th ed.). Boca Raton: Chapman & Hall/CRC. https://doi.org/10.1201/9780429186196

Turan, Z. & Akdag-Cimen, B. (2020) Flipped classroom in English language teaching: A systematic review. *Computer Assisted Language Learning*, 33(5–6): 590–606. https://doi.org/10.1080/09588221.2019.1584117

Wang, M. & Devitt, A. (2022) A systematic review of computer-mediated communications in Chinese as a foreign language from 2008 to 2022: Research contexts, theoretical foundations and methodology, affordances and limitations. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688221132475

Winer, B. J., Brown, D. R. & Michels, K. M. (1991) *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Zedeck, S. (ed.) (2014) *APA dictionary of statistics and research methods*. Washington: American Psychological Association. https://doi.org/10.1037/14336-000

Zhao, Y. (2004) Recent developments in technology and language learning: A literature review and meta-analysis. *CALICO Journal*, 21(1): 7–27. https://doi.org/10.1558/cj.v21i1.7-27

Ziegler, N. (2016) Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, 38(3): 553–586. https://doi.org/10.1017/S027226311500025X

Oliver James Ballance is a lecturer in applied linguistics and English for academic purposes at Massey University. He teaches postgraduate courses on language teaching methodology and curriculum design and supervises postgraduate research projects. His main research interests concern the interface between corpus linguistics and language for specific purposes.

Author ORCiD. Oliver James Ballance, https://orcid.org/0000-0003-4695-6406

https://doi.org/10.1017/S0958344023000162 Published online by Cambridge University Press