Quantifying CALL: significance, effect size and variation

Alex Boulton

Quantifying CALL: significance, effect size and variation

Alex Boulton[1]

Abstract. Good practice in primary research has evolved over many decades of research in applied linguistics to counter human fallibility and biases. Surprisingly, perhaps, synthesising such research in an entire field has only recently started to develop its own methodologies and recommendations. This paper outlines some of the issues involved, especially in terms of quantitative research and meta-analysis. A second-order synthesis of meta-analyses in Computer-Assisted Language Learning (CALL) provides only medium effect sizes, but the figures are interpreted in terms of realistic expectations. The inevitable variation in effect sizes can be attributed in principle either to the research methodologies (both primary and secondary) or – more interestingly – to real-world phenomena.

Keywords: meta-analysis, research synthesis, effect size, variation, CALL synthesis.

1. Primary studies and research synthesis

CALL research is often quantitative in nature, in line with the bulk of research in applied linguistics which sees measurement as a way to counter subjectivity and strive towards more 'scientific' rigour. This of course neglects the important insights that can only realistically be gleaned from qualitative studies, with their more frequent focus on emic, ecological, holistic considerations, and ability to account for complex, narrative, continuous data such as interviews which do not lend themselves easily to quantitative analysis. This is not to say that quantitative research should be abandoned, only that – like all research – it needs using and interpreting with caution.

A particular problem with primary quantitative research is that many studies adopt Null Hypothesis Significance Testing (NHST) as the standard model. NHST is

1. CNRS & University of Lorraine, Nancy, France; alex.boulton@univ-lorraine.fr

How to cite this article: Boulton, A. (2016). Quantifying CALL: significance, effect size and variation. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 55-60). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.538

entirely subject to sample size (any difference will be significant if the sample is large enough), it doesn't tell us anything about what we're really interested in (i.e. the effect of a particular variable), and it encourages dichotomous thinking on an arbitrary basis (with *p*-values typically set at 95% for no good reason). NHST has been heavily criticised on all these counts, with Plonsky (2015) claiming it has done "far more harm than good" (p. 242).

More useful are effect sizes such as Cohen's *d*: such measures do address the real issues, and have the substantial advantage that they can be pooled across studies. For this to be effective, we first however need to begin with rigorous trawls of research in a clearly-defined field to ensure that we do in fact cover what we set out to synthesise. Traditional surveys as found in the ubiquitous 'literature review' in primary studies are notoriously inadequate if one relies on personal interpretation of serendipitous collections. This is a major issue in the complex field such as education, where "everything seems to work in the improvement of student achievement… Teachers can thus find some support to justify almost all their actions – even though the variability about what works is enormous" (Hattie, 2009, p. 6). Research synthesis and specifically meta-analysis are attempts to make the procedures more scientific and transparent (Norris & Ortega, 2000).

2. Meta-analysis in CALL

Table 1 is an attempt to summarise meta-analyses in CALL, derived largely from Plonsky and Ziegler (2016) and Oswald and Plonsky (2010). The 12 meta-analyses show the size of the effect of CALL use in experimental groups compared to control or comparison groups. The first three give large effect sizes according Plonsky and Oswald's (2014) empirically-derived, field-specific benchmarks based on meta-analyses in Second Language Acquisition (SLA) ($d \geq .9$); the next two are medium ($d \geq .6$), followed by three small ($d \geq .4$) and four negligible effects ($d < .4$). The mean is .64, the pessimistic conclusion being that CALL work as a whole has barely a medium effect on learning. Worse, the few large effect sizes are derived from very small samples as shown in the column for *k*. Indeed, there is a large negative correlation (r=-.51) between the number of studies featuring in these meta-analyses and the effect size calculated.

This rather pessimistic discussion is based on the premise that we need to find large effect sizes for CALL. Minimally, however, what we would hope to find is that CALL is *at least as good as* traditional teaching, with an effect size of d=0 (or, according to Hattie, 2009, d=.4), which is the case here. Since most primary studies

are relatively focused and short-term, they will not show other benefits which we may want to impute to CALL. These might include cost or time efficiency, motivation or enjoyment, long-term retention or appropriation, learning-to-learn or becoming 'better language learners', increased autonomy or transferable skills, etc. Such conclusions are speculative at best since these things are notoriously difficult to research, and would have to be the subject of further studies.

Table 1. Meta-analyses in CALL (cf. Oswald & Plonsky, 2010; Plonsky & Ziegler, 2016)

study	year	source	focus	k	d
Abraham	2008	CALL	Glossing vocabulary	6	1.40
Zhao	2003	CALICO J	CALL general	9	1.12
Taylor	2006	CALICO J	Glossing reading comprehension	4	1.09
Chiu	2013	BJET	Vocabulary	16	0.75
Abraham	2008	CALL	Glossing reading comprehension	11	0.73
Chiu et al.	2012	BJET	Game-based learning	14	0.53
Taylor	2009	CALICO J	Glossing reading comprehension	32	0.49
Lin, H.	2014	LL&T	CMC	59	0.44
Yun	2011	CALL	Glossing vocabulary	10	0.37
Lin, W. et al.	2013	LL&T	SCMC	19	0.33
Grgurović et al.	2013	ReCALL	CALL general	65	0.24
Ziegler	2015	SSLA	SCMC	14	0.13

3. Variation

Meta-analyses need interpreting with caution: in particular, it is tempting to seize on a single figure as the ultimate answer to the question: Does it work? "Professionals in CALL often find this comparison question frustrating… but in a political sense, it would be useful if CALL specialists could answer it" (Grgurović, Chapelle, & Shelley, 2013, p. 2). More realistically, we need to look at *variation* in what works: different primary studies of ostensibly of the same phenomenon will provide different effect sizes – as will different meta-analyses in a given field (cf. Table 1).

Variation in meta-analyses may derive from the studies themselves. Clearly not all primary research is of similar quality, and a synthesist has to decide how to deal with this – typically by devising *a priori* inclusion criteria (e.g. whether to include conference proceedings), or by treating quality as a variable and subsequently examining its impact on the effect sizes (e.g. to compare papers in conference

proceedings against those in prestigious journal articles). This underlines the many choices that are to be made: quality is a consideration in secondary as much as in primary research. The increasing numbers of meta-analyses and the prominence given to them in such diverse places as the American Psychological Association manual or *Language Learning* editorials, along with handbooks (e.g. Cumming, 2012) and websites with methodological recommendations for good practice and increasingly sophisticated tools (cf. https://lukeplonsky.wordpress.com), may suggest that the practices are straightforward.

While SLA synthesis may have become something of a tradition in its own right, researchers have a tremendous range of options to choose from at all stages. How exactly do they define the field? What tools do they use to arrive at a (near-) exhaustive collection of studies in that field? What study types do they reject? Simply when deciding what studies to include, many meta-analyses are deliberately limited to control/experimental designs published in English in high-ranking journals in particular time periods for example, and so inevitably miss much of the field, the rationale being to increase study quality. How do they extract the data, and how do they calculate and interpret effect sizes? Primary research can be extraordinarily complex, with several experimental groups doing different things using different tools and procedures for different main objectives, and the resulting data presented in the form of raw data, descriptive statistics, *F*/*t*-scores, etc. The synthesist then has to decide the specific formula to use, and decide whether to weigh the results (e.g. for sample size) and how to deal with the extreme values for outliers. Checking is essential, usually from inter-rater reliability measures, but also in the form of funnel plots for publication bias in the studies sampled, and if possible, comparing effect sizes from different designs (control/experimental vs. pre/post-test and if possible delayed test designs).

Variation can also be examined to determine which moderator variables contribute more or less to overall effect sizes. It might be, for example, that large effect sizes only come from long-term studies, or from certain learner populations (proficiency, age, sex, cultural background, field of study, etc.), for certain linguistic objectives or tools or procedures, and so on. Promising categories therefore feature in a coding sheet, which is itself immensely complex and difficult to draw up rigorously. It is no surprise perhaps that among the main conclusions of syntheses are recommendations for better reporting practices in primary research, as it is only at this stage that it becomes really apparent how much information is missing, vague or unsubstantiated. Burston and Arispe (2016), for example, found that 50% of research articles in four major CALL journals targeting ‘advanced’ levels of proficiency had learner populations of B1 level only.

4. Conclusions

Research is extremely complex, not just in terms of choices and procedures but also in terms of the field itself – language, language use and language learning. Bias is inherent in primary research, and measures need to be taken to ensure that this is not exacerbated in an overall view of the field. Quantitative studies provide essential insights but do not capture the whole picture, while narrative syntheses are "inevitably idiosyncratic" (Han, 2015, p. 411) – both are essential to provide as full an understanding as possible. Synthesists need to be rigorously transparent in designing their studies, writing up their results, and providing supplementary materials for others to check, replicate, or modify as more research becomes available.

The observations here are in large part inspired from a meta-analysis of data-driven learning, i.e. the use of corpora in language learning (Boulton & Cobb, forthcoming). In addition to the above, the main conclusions are that the many choices are often glossed over; that single-figure main effects can be misleading and need careful interpretation; and that, despite the relatively low overall effect sizes reported, there are reasons for optimism in the field. Inevitably, more research is needed in all areas.

References

Boulton, A., & Cobb, T. (forthcoming). Corpus use in language learning: a meta-analysis. *Language Learning*.

Burston, J., & Arispe, K. (2016). The contribution of CALL to advanced-level foreign/second language instruction. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 61-68). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.539

Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology supported language learning. *ReCALL, 25*(2), 165-198. https://doi.org/10.1017/S0958344013000013

Han, Z. (2015). Striving for complementarity between narrative and meta-analytic reviews. *Applied Linguistics, 36*(3), 409-415. https://doi.org/10.1093/applin/amv026

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning, 50*(3), 417-528. https://doi.org/10.1111/0023-8333.00136

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: choices and challenges. *Annual Review of Applied Linguistics, 30*, 85-110. https://doi.org/10.1017/S0267190510000115

Plonsky, L. (2015). Quantitative considerations for improving replicability in CALL and applied linguistics. *CALICO Journal, 32*(2), 232-244.

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912. https://doi.org/10.1111/lang.12079

Plonsky, L., & Ziegler, N. (2016). The CALL–SLA interface: insights from a second-order synthesis. *Language Learning & Technology, 20*(2), 17-37.

Research-publishing.net

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; Voillans, France, info@research-publishing.net

CALL communities and culture – short papers from EUROCALL 2016
Edited by Salomi Papadima-Sophocleous, Linda Bradley, and Sylvie Thouësny

Typeset by Research-publishing.net
Cover design by © Easy Conferences, info@easyconferences.eu, www.easyconferences.eu
Cover layout by © Raphaël Savina (raphael@savina.net)
Photo "bridge" on cover by © Andriy Markov/Shutterstock
Photo "frog" on cover by © Fany Savina (fany.savina@gmail.com)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-43-8 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-44-5 (Ebook, PDF, colour)
ISBN13: 978-1-908416-45-2 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2016.