White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/127564/

Version: Accepted Version

Article:
Marsden, Emma Josephine, Morgan-Short, Kara, Thompson, Sophie et al. (2018) Replication in second language research: Narrative and systematic reviews, and recommendations for the field. Language Learning. pp. 321-391. ISSN: 0023-8333

https://doi.org/10.1111/lang.12286

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Replication in second language research:

Narrative and systematic reviews, and recommendations for the field.

EMMA MARSDEN,[1] KARA MORGAN-SHORT[2], SOPHIE THOMPSON[1], DAVID ABUGABER[2]

[1]*University of York, Centre for Research into Language Learning and Use, Department of Education, York, YO10 5DD, United Kingdom Email:* emma.marsden@york.ac.uk

[2]*University of Illinois at Chicago, Department of Hispanic and Italian Studies and Department of Psychology, United States of America*

Running head: Replication in second language research

Key words: Replication; Methodology; Systematic review; Research design; Publishing

Acknowledgements

The current systematic review was presented at two colloquia on replication convened by the first two authors, at the *American Association of Applied Linguistics* (Portland, Oregon, U.S.A., 2017) and the *European Second Language Association* (Reading, UK, 2017). We thank our co-presenters and the audiences at those colloquia for insightful discussion and feedback. We are grateful to Luke Plonsky for helpful advice during the initial stages of the current systematic review. An earlier version of a small subsection of this synthesis, partially funded by the UK *Economic and Social Research Council* (RES-062-23-2946), was presented at *The IRIS Conference*, University of York, 2-3 September 2013 and at *The International Symposium on Bilingualism*, Nanyang Technological University, Singapore, 10-13 June 2013. *Contact information*: Emma Marsden, emma.marsden@york.ac.uk and Kara Morgan-Short karams@uic.edu.

Replication in second language research:

Narrative and systematic reviews, and recommendations for the field.

Abstract

Despite its critical role for the development of the field, little is known about replication in second language (L2) research. To better understand replication practice, we first provide a narrative review of challenges related to replication, drawing on recent developments in psychology. This discussion frames and motivates a systematic review, building on syntheses of replication in psychology (Makel, Plucker, & Hegarty, 2012), education (Makel & Plucker, 2014), and L2 research (Polio, 2012b). 67 self-labelled L2 replication studies found across 26 journals were coded for 136 characteristics. We estimated a mean rate of 1 published replication study for every 400 articles, with a mean 6.64 years between initial and replication studies, and a mean 117 citations of the initial study before a replication was published. Replication studies had an annual mean 7.3 citations, much higher than averages in linguistics and education. Overlap in authorship between initial and replication studies and the availability of the initial materials both increased the likelihood of the replication supporting the initial findings. Our sample contained no direct (exact) replication attempts, and changes made to initial studies were numerous and wide-ranging, thus obscuring, if not undermining, the interpretability of replication studies. We end by proposing 16 recommendations, relating to rationales, nomenclature, design, infrastructure, and incentivization for collaboration and publication, to improve the amount and quality of L2 replication research.

Replication in second language research:

Narrative and systematic reviews, and recommendations for the field.

Replication studies are considered by many to have a fundamental role in any scientific endeavour. When using the same materials and procedures as a previous study, replication studies serve to test the reliability of the previous study's findings; when altering specific methodological or participant characteristics of a previous study, they serve to test generalisability of the earlier findings under different conditions. One indication of the importance of replication is found in the 50 or more calls for replication research in the field of second language (L2) research alone, from Santos (1989), through Polio and Gass (1997), to very recent proposals for specific replication studies, such as Vandergrift and Cross (2017), and even book length treatments (Porte, 2012) (see references for 50 calls and commentaries marked ° in 'Supplementary Materials 1: Included studies plus commentaries and exclusion criteria'). Beyond these calls, efforts to actively promote and facilitate replication studies have also emerged. For example, the IRIS (Instruments for Research Into Second Languages) repository (www.iris-database.org) was established in 2011 and holds, at the time of writing, over 3600 materials that can be used for replication, among other purposes, in L2 research (Marsden & Mackey, 2014; Marsden, Mackey, & Plonsky, 2016). The Open Science Framework (https://osf.io/), also established in 2011, provides a web infrastructure to facilitate collaboration and has been used for large replication efforts in psychology (e.g., Open Science Collaboration, 2015), which continue to make waves in academia (Laws, 2016; Lindsay, 2015; Martin & Clarke, 2017) and the general media (Baker, 2015; Devlin, 2016). In some fields, a flourishing meta-science (the scientific study of science; see Munafò et al., 2017) has included syntheses assessing the quantity and nature of

replication efforts, e.g., in education (Makel & Plucker, 2014) and in psychology (Makel, Plucker, & Hegarty, 2012; see also a special issue of *Psychological Bulletin*, 2018).

The driving force behind this battery of calls, commentaries, infrastructure, and meta-science is a perceived crisis in the state of replication research. The severe concerns underpinning the alleged crisis have several dimensions relating to: the (small) *amount* of published replication research; the (poor) *quality* of replication research; and the (lack of) '*reproducibility*' (i.e., the extent to which findings can(not) be reproduced in replication attempts that have been undertaken). These concerns speak to the very core of science, raising fundamental questions about the validity and reliability of our work. Indeed, some commentators have called replication the "Supreme Court" or "gold standard" of research evidence (Collins, 1985; Jasny, Chin, Chong, & Vignieri, 2011) and a "linchpin of the scientific process" (Nature, Editorial, July 2006).

In the field of L2 research, given the importance of replication and the 50 calls for replication in L2 research that we identified, we might expect a substantial number of published replication studies by now. However, a perceived lack of prestige, excitement and originality of replication plagues L2 research (Porte, 2012), as in other disciplines (Branco, Cohen, Vossen, Ide, & Calzolari, 2017; Chambers, 2017; Schmidt, 2009), and these perceptions are thought to have caused, at least in part, directly or indirectly, alleged low rates and a poor quality of published replication studies. However, a systematic meta-science on replication research is not yet established in the field of L2 research, leaving a poor understanding of the actual amount and nature of replication studies that have been published.

The current study begins to address this gap through narrative and systematic reviews. The narrative review (part one) considers challenges in replication research and is largely informed by commentaries and meta-science from psychology, given that the cognitive and social subdomains of psychology are highly influential in L2 research, and also from

education, another key sister discipline. The narrative review is organized around four broad themes: the quantity of replication research; the nature of replication research; the relationship between initial and replication studies; and the interpretation and extent of reproducibility of initial studies' findings. To gain insight into these issues in the context of L2 research, the systematic review (part two) provides a synthesis of L2 studies in journal articles that self-labelled as replications with research questions and methods that were largely determined by the narrative review but that also emerged through the design and pilot of the coding instrument. Finally (in part three), we offer further discussion and 16 recommendations for future replication work, which draw on our narrative and systematic reviews and on our experience of carrying out multi-site (Morgan-Short, Marsden, Heil, et al., 2018) and single site (Faretta-Stutenberg & Morgan-Short, 2011; McManus & Marsden, 2018; Marsden, Williams, & Liu, 2013; Morgan-Short, Heil, Botero-Moriaty, & Ebert, 2012) replications. We start from the widely agreed premise that testing the replicability of findings should have an essential role in the testing and refinement of theory, at least for hypothesis-testing epistemologies that seek to ascertain generalizability, and other epistemologies in which constructs are deemed to be definable and observable. Thus, our overall aim is to provide conceptual clarification and an empirical base for future discussion and production of replication studies, with a view to improving the amount and quality of L2 replication research.

Part One: Narrative Review of Concerns and Challenges Related to Replication.

The primary aim of this narrative review is to consider key issues related to replication research. In accomplishing this, we also prepare the way for our systematic review in part 2. Thus, in the course of part 1, we indicate how aspects of the narrative review inform the aims, scope, structure, and methods of our systematic review in part 2.

First, we clarify our use of the terms 'replicable/replicability' and 'reproducible/reproducibility', given some debate surrounding these terms (NSF, 2015; The National Academies, 2016). The term replicable/replicability commonly serves two functions and we have tried to ensure at each use whether we refer to either (a) the extent to which it is possible to carry out a study again (e.g., whether sufficient information and materials are available to allow replication of the *study* itself, also known as repeatability) or (b) the extent to which the results of a replication study are similar to those of the initial study (i.e., replication of *findings*). The term 'reproducible/reproducibility' is used in a more marked way to refer only to (b), the extent to which the results of a replication study are similar to those of an initial study (i.e., reproduction of *finding*s), in line with the recent developments in the field of psychology (e.g., Open Science Collaboration, 2015).

The Quantity of Replication Research

To understand the state of replication research in our field or other fields, we must first determine the quantity of replication research that has been undertaken. In order to do this, we must identify which studies should be counted as replication research. This is not a trivial matter. Given a broad definition (e.g., 'studies investigating related questions using similar designs and materials'), a very large number of studies could be called replications (see Plonsky's 2012 discussion of the extent to which studies included in a meta-analysis could be considered to be replications; and see VanPatten, 2002a&b for narrower conceptualisations). On the other hand, even studies that fall into a narrower definition of replication (e.g., investigating the same research questions with a design and materials that are as similar as possible to an earlier study) may not label themselves as replications. To illustrate, of the four studies that were part of the replicative sequence extended by Morgan-Short, Marsden, Heil, et al. (2018), only one (Wong, 2001) turned up in our systematic

review as a self-labelled replication. Given this subjectivity and inconsistency and, more importantly, given that we wanted in our systematic review (part 2) to ascertain the extent to which the term 'replication' has been used to label studies reported in journals, we used instead the self-identification of authors, i.e., studies that self-labelled in the title or abstract as a replication study. This is similar to the approach of Makel and Plucker (2014) and Makel et al. (2012), who examined the state of replication in psychology and education, respectively, and avoided the need to create bespoke definitions of replication studies. However, we acknowledge that this approach does not encompass all research that could be viewed as a replication, an issue considered more fully in Part 3. Note also that throughout we use the term 'replication' to refer to a 'replication study' i.e., one that attempted, to some degree, to replicate a previous study's aims and methods. Our use of the term 'replication' alone makes no allusion to whether the study succeeded (or indeed aimed) to replicate the methods exactly, nor to the extent to which earlier findings were reproduced.

In addition to identifying replications, one must consider the nomenclature of sub-types of replication. In the field of psychology, an early proposal of three subtypes was made by Lykken (1968): 'literal replication', in which additional participants were recruited to the same study; 'operational', which used the same methods and conditions; and 'constructive', where the claimed relation between constructs was tested using any methods the replicator wished. Others have converged on two (Makel et al., 2012; Schmidt, 2009): (a) 'direct', where there are no intentional or significant alterations of the initial study – considered "the means of establishing reproducibility of a finding with new data" (Open Science Collaboration, 2015, p. 1) and (b) 'conceptual', where there is intentional adaptation of the initial study to investigate generalizability to new conditions, contexts, or study characteristics. Using this distinction, Makel et al. (2012) found that 81.9% of replications in psychology were conceptual, 4.1% were categorised as both conceptual and direct, and 14%

were direct. The latter figure is most likely considerably higher now given the recent surge of direct replications (see below).

One problem with such dichotomous labelling is that for 'conceptual' replications the *amount* and *type* of changes to the initial study can vary and/or be vague, making it difficult to assess whether a study can test the effects of new constructs or of 'boundary conditions' (i.e., study features that help determine the limits of generalizability to, for example, different participant characteristics). Indeed, Earp and Trafimow (2015) provide a framework for conceptualizing different types of replication falling along a multi-dimensional spectrum with each type serving a different purpose.

In L2 research, issues of nomenclature for different types of replications have also been a source of confusion (Polio, 2012b). Porte (2012) provided a taxonomy of three broad types of replication: (a) exact or literal; (b) partial, approximate or systematic; (c) conceptual or constructive. However, the extent to which this recommendation has been adopted by the field in a systematic manner remains unclear. Thus, our synthesis aimed to examine the nomenclature used for self-labelled replication research and the extent to which different labels have reflected the amounts and types of change between initial and replication studies. With this insight, we go on to propose a clear and principled nomenclature for the field.

On a final note about nomenclature, in the current reviews we have used the term 'initial study' rather than 'original' when referring to studies that were replicated. This is because studies are rarely if ever truly 'original' in the sense of being a completely novel idea. Also, 'original' carries negative connotations for its 'replication,' as it could imply that anything that is not 'original' cannot share other characteristics broadly associated with originality, such as innovative, fundamental, or agenda-setting.

After having identified replications and classified them by type, issues of quantity can then be examined. In the field of education, Makel and Pluker (2014) found a replication

study publication rate of 0.13% (221 out of 164,589 articles) in the 100 highest impact journals between 1938 and 2014. In the field of psychology, Makel et al. (2012) estimated that among the top 100 journals between 1900-2010, the replication study publication rate was 1.07%, though this is now likely to be higher given recent multiple, direct replication projects: the Many Labs project (Klein et al., 2014); the Pipeline Project (Schweinsberg et al., 2016); Registered Reports (Nosek & Lakens, 2014); and the Reproducibility Project (Open Science Collaboration, 2015). In business, marketing, and communication journals, replication rates have ranged from 1% to 3% (Evanschitzky, Baumgarth, Hubbard, & Armstrong, 2007; Hubbard & Armstrong, 1994; Kelly, Chase, & Tucker, 1979). In the field of L2 research, the rate of replication studies is perceived as being low, but without systematic data on this, concerns to date have necessarily been speculative.

Attempts to improve rates of replication have met many challenges (Porte, 2012), including some imposed by publishing venues themselves. The quantity of replication is, perhaps, influenced by the extent to which journals encourage or discourage replication. To investigate how psychology journals approach this issue, Martin & Clarke (2017) reviewed the Scope sections of author guidelines of 1151 journals and found that 63% did not state they accepted replications but did not discourage them; 33% implicitly discouraged them by emphasizing originality, novelty or innovation of submissions; 3% of journals stated they accepted them; and 1% actively discouraged replications by stating they did not publish them. The fact that only 3% of journals stated they accepted replications may partly be due to the perceived impact, and hence prestige, of replication. However, this perception may not reflect reality. To illustrate this with an example from the field of education research, Makel and Plucker (2014) found the median citation count of replications was five (range = 0 to 135), compared to 31 for the initial studies (range = 1 to 7,644). However, this difference is not surprising, as initial studies have more time to be cited and high citation counts are often the

reason for replicating them in the first place. Furthermore, as Makel and Plucker note, five citations for replications is relatively high, given that only one of the top 100 education journals had a five-year impact factor (IF) higher than five. For the field of psychology, Makel et al. (2012) found that the median number of citations of replications was 17 (range = 0 to 409), compared to the initial studies' 64.5 (range = 1 to 2,099), and was also observed as being relatively high given that only three of the 100 analysed journals had a five-year IF greater than 17. Thus, contrary to expectations, replications had a higher impact than the average article in their field as represented by journal IFs.

Motivated and informed by previous work that quantifies replication in psychology and education, our systematic review addresses two key questions: To shed light on the *quantity* of replication in L2 research, our synthesis calculates the rate of replication, examines which journals have published replications, and documents whether journals discourage or encourage them. To gauge the *impact* of published replications, we also investigated the number of citations of replications and the IF of the journals that publish them.

The Characteristics of Research Studies that Warrant and Lead to Replication

Beyond questions about the quantity of replication, we consider what kind of research the field appears to support as meriting replication. The extent to which reproducible findings are deemed desirable can vary according to different ontological, epistemological, and methodological perspectives (Markee, 2017; Polio, 2012a; Porte, 2012; Porte & Richards, 2012). There is a high degree of consensus that replication, particularly when narrowly defined as direct or close replication, is not appropriate or useful for all types and stages of research (such as ideological or interpretative approaches; exploratory or grounded research; or case studies). There is also clear consensus that replication *is* of value for a large portion of research, usually that which involves some hypothesis testing and/or data that is quantitative

(either at collection or coding stages). This may be because for this type of research materials, measurements and analyses are designed to be reproducible so as to ascertain generalisability, in line with the epistemologies of such research. Putting those relatively well-rehearsed issues aside, and focusing mainly on the large body of research in which the desirability of replication is rarely controversial, a variety of suggestions have been made about characteristics of research that warrant replication endeavors, such as the significance and design of the initial study.

Regarding the significance of an initial study, Nosek and Lakens (2013) suggested that "important" research to replicate is "…often cited, a topic of intense scholarly or public interest, a challenge to established theories, but should also have uncertain truth value (e.g., few confirmations, imprecise estimates of effect sizes)." (p 59). Thus, one may turn to the number of citations as a warrant for replication. For example, Makel et al. (2012) suggest it would be surprising if replications had not been triggered after an (admittedly arbitrary) 100 citations of a study.

However, citation counts alone are unlikely to offer reliable or sufficient motivation for replication. Importance also stems from the research community's views on what research needs to be replicated in order to inform theory, method or practice. We briefly mention four possible approaches for establishing what is important. The journal *Language Teaching* includes an article type in which authors justify and describe specific replications that should be done, and indeed, twelve such articles have been published, at the time of writing (see Supplementary Materials 1). However, the extent to which this unique initiative leads to replication is unknown. Somewhat surprisingly, in our study sample (see Part 2 below), we found no published replications that followed the suggestions made, nor did we observe a general increase in the number of replications published after these articles types began (in 2014, with Basturkmen, 2014). Another approach is to crowdsource proposals for replication

(see 'PsychFileDrawer' http://www.psychfiledrawer.org/top-20/), whereby a social media platform allows people to propose and vote on the studies they would like to see undergo a replication study. Since it began in 2012, this archive of reports of replications currently holds 71 replication reports, but the extent to which such an initiative, which is outside the standard publication venues, will have a lasting impact on the amount of quality of replications is unclear. Another possibility is that journal editors invite replications of particular studies, as occasionally done by editors of the Registered Replication Reports in *Perspectives in Psychological Science*. This approach exerts strong editorial influence over the types of studies that are replicated and how they are replicated, and demands a heavy editorial role (Dan Simons, personal communication). A final possibility is that researchers themselves, 'bottom-up,' provide theoretical and methodological justifications in the rationales sections of their replication studies, and these arguments are evaluated via current peer-review mechanisms. All these approaches may help to establish what research merits replication, although data is needed to ascertain the extent to which they are effective mechanisms for improving the amount, quality, or perceived prestige of replications.

Another factor potentially indicating importance, and thus a need for replication, are 'surprising' findings (see Makel et al. 2012, p. 540; Porte, 2012, p. 7). 'Surprising' could be, for example, large effect sizes when a meta-analysis would predict them to be smaller (or vice versa). Laws (2016) described all of the 13 studies replicated in Nosek and Lakens' (2014) special issue as 'curios' (p. 2), with odd findings. Interestingly, 10 out of those 13 initial findings were not reproduced. Thus, one (arguably undesirable) downside to 'surprising' findings serving as a rationale for replication is that if the rate of reproducing findings from such research is unusually low, the overall rate of reproducibility for a field may appear to be lower than it actually is (Hartshorne & Schachner, 2012; Laws, 2016). Also, using the 'surprising findings' rationale, alone, as a warrant for replication could

introduce a type of 'reverse' publication bias, whereby finding 'no effect' in a replication (where 'an effect' or statistical significance was found in the initial study) is considered the more publishable and citable outcome (Ioannidis, 2005; Luijendijk & Koolman, 2012).

Finally, the statistical significance of a study's results may have (undue) influence on its perceived importance for replication. Publication bias (whereby journals tend to publish and/or researchers tend to submit statistically significant findings) is a widely acknowledged problem, and null findings are confined to the "file drawer" (a term coined by Rosenthal, 1979; the phenomenon documented by, for example, Bakker, van Dijk, & Wicherts, 2012; Schmidt & Oh, 2016; Sterling, Rosenbaum, & Weinkam, 1995; Sutton, 2009). Though the extent of field-wide publication bias in L2 research has not yet been systematically studied, its existence is likely (Fanelli, 2012; Plonsky, 2013) and several meta-analysts have found evidence of it in specific domains (Lee & Huang, 2008; Lee, Jang, & Plonsky, 2015; Plonsky, 2011). This means that even unintentionally, anyone choosing a study to replicate is likely, due to chance alone, to choose one with statistically significant findings. To give one example of this phenomenon, Laws (2016) noted that the four multi-site replications he reviewed almost entirely neglected null findings. Similarly, in the Open Science Collaboration (2015) project, only three of its 100 initial findings were null. Yet it is of course useful to carry out replications of studies with null or borderline findings. For instance, for the three null studies replicated by the Open Science Collaboration, the replications confirmed two as 'null' but produced statistically significant findings for the other (see also Morgan-Short et al., 2018). The need to replicate studies with null findings is particularly important in L2 research, where sample sizes are often too underpowered to reject the null hypothesis, with an average post-hoc power of 0.57 (Plonsky, 2013), the statistical equivalent of "tossing a coin in the air and hoping for heads" (Plonsky, 2015, p. 29). In sum, the absence of statistical significance in an initial study may (a) not validly

indicate the absence of an effect but rather be an artefact of other issues, such as small *n* or chance findings, (b) be a theoretically or practically useful finding that does merit corroboration via replication, and (c) lead to dichotomous rather than nuanced interpretations. Thus, statistical significance alone serves as a dubious warrant for replication.

Beyond the significance of an initial study, a warrant for replication must also consider research design. Indeed, suggestions have been made to select studies based on a set of problematic characteristics of a study and its findings. For example, Lindsay (2015) proposes being on the "lookout for this troubling trio: (a) low statistical power, (b) a surprising result, and (c) a *p* value only slightly less than .05" (p. 1827-8). Another proposal—a quantitative 'doping test for science' proposed by Schimmack (2016)—is known as the replicability index (R-index) and is used to evaluate the statistical replicability of a set of studies. It calculates the difference between median estimated power and likely rate of reproducing findings, which results in the so-called inflation rate. Results of studies with these concerning characteristics may be due to Questionable Research Practices (QRPs), such as not reporting all outcome measures or conditions, only reporting statistical tests that found statistical significance, data-peeking before deciding when to stop testing participants or whether to exclude (particular definitions of) outliers, and HARKing ('hypothesizing after the results are known') (Chambers, 2017; Kerr, 1998; Lindsay 2015). Thus, replication could help ascertain the likelihood of whether findings were actually valid or merely an artefact of such issues.

Even if a replication is warranted, other design characteristics of initial studies may affect the feasibility of replication. Practicalities of time and resources may impede certain studies from being replicated, meaning that 'cheap and easy' (Laws, 2016) studies are replicated while replication in some sub-domains is "likely to remain castles in the air" (p. 3). One likely manifestation of these practical constraints was the Many Labs Replication Project

(Klein et al., 2014), which delivered a single 15-minute questionnaire (combining 13 earlier experiments) to 6,344 participants across 12 countries via 36 research groups. In L2 research, designs which are usually more costly involve longitudinal designs (such as experiments with pre-, post- and delayed post-tests as opposed to one-shot or cross-sectional designs), one-to-one measures such as oral production tests (versus group-delivered pen-and-paper or computer-based tests), equipment that is expensive to purchase or utilise (eye-tracking or neuroimaging hardware), and participant populations that are difficult to reach (e.g., rarer L1-L2 combinations, schools, heritage speakers, or participants linked to a specific history or culture). Replications with such designs may be underrepresented compared to more easily administered designs.

Another characteristic that affects whether, and how well, a replication study can be carried out is the transparency of the initial research, as availability of materials and data, as well as thorough reporting, are needed for replication, and are particularly important for independent or direct and partial replications. For example, the availability of data helps replicability and the evaluation of reproducibility as researchers can, for example: increase the sample size of previous research; combine their data with previous data in new analyses; reanalyse data to assess the reliability of the initial analyses (in fact specifically termed 'reproducibility' by NSF, 2015); and evaluate the parity of samples, which is particularly critical in second language learning research as participant demographics (e.g., proficiency, age, and L1) are known to affect learning. However, an academic culture in which there is little chance of replication happening or being published reduces the perceived need to make research replicable through materials and data availability and transparent reporting, as researchers might very reasonably ask themselves: "is anyone *really* going to attempt to replicate this?" This no doubt partially accounts for a history of inadequate reporting practices (as noted by e.g., Derrick, 2016; Han, 2016; Larson-Hall & Plonsky, 2015; Plonsky

& Derrick, 2016), poor transparency of materials (Marsden & Mackey, 2012; Marsden et al., 2016; Marsden, Thompson, & Plonsky, accepted), and very scarce availability of data (Larson-Hall & Plonsky, 2015; Larson-Hall, 2017; Plonsky, Egbert, & LaFlair, 2015) (see Marsden (in press) for an overview). For discussions of similar challenges in psychology, see Fecher, Friesike, & Hebing (2015), Lindsay (2017), Wicherts, Borsboom, & Molenaar (2006). Indeed, aiming to address this situation, the Transparency and Openness Promotion Guidelines (TOP) (Nosek et al., 2015) encourage journals to incentivise/require their authors to make their materials and data transparent. These guidelines also set explicit benchmarks about the levels to which journals promote replication (referred to again in part 3 below), thus drawing clear links between replication and the transparency of materials and data.

In sum, issues such as the initial reporting of methods, results and analysis; the availability of the initial materials and data; and the resources needed may all reduce the likelihood, quality, or usefulness of replication (even when a replication is clearly warranted). Motivated by these issues, in our synthesis we probe the question of what warrants and leads to replication by examining the characteristics of studies that have been replicated in L2 research, such as: their citation counts; their broad findings ('statistically significant' or 'null'); their designs, measures, and sample sizes (to investigate the extent to which replication has been concentrated on 'cheap and easy' designs); the transparency of their reporting; and the availability of their materials and data.

Extent of Change between Initial and Replication Studies

The rationale for replicating a study can also be determined by the nature of the specific changes made to the designs of the initial studies. Many researchers include caveats about their studies, suggesting that future research should replicate the study to test boundary conditions, i.e., the extent of generalizability to, for example, a different outcome measure, experimental design, first-language background, modality, target language, age or proficiency

of participants. However, making many or unacknowledged/unspecified changes to a study lies in tension with being able to account for whether differences in findings compared to the initial study are ascribable to the heterogeneity that was introduced (intentionally or otherwise) or to some other factor. This issue was tackled by Klein et al. (2015) in their direct replications, wherein heterogeneity between initial studies and replications was kept to a minimum except for two key variables (participant nationality and lab vs. online delivery). They estimated the proportion of variation in effect sizes attributable to heterogeneity of implementation rather than to chance, and showed that the effects of heterogeneity were non-existent or very small in most cases. A related issue is that even when maximum effort is made to maintain *homo*geneity of implementation between initial and replication studies, there may be auxiliary assumptions embedded in the hypotheses or design of the initial studies. Regardless of whether these assumptions are well understood or not, if the replication study violates them inadvertently, this can affect the outcomes and could result in findings that do not align with those of the initial study (as discussed by Trafimow & Earp, 2016). As a preliminary investigation into the extent and nature of heterogeneity in L2 replication research, in the current synthesis we sought to collect data on what types of changes have been made in replications and the extent to which heterogeneity between initial and replication studies was intentional (for partial or conceptual replications), explicitly acknowledged (for all types of replication), or not acknowledged by the authors.

Another common caveat in the concluding sections of articles is that replication is required due to the small sample size of the study. It might therefore be expected that self-labelled replications have larger *n* than initial studies. However, Tversky and Kahneman's (1971) survey found that most social scientists believed that if a finding had been observed with a certain *n*, the same outcome should be observed with a smaller *n*. Given a scenario in which an initial study ($N = 40$) produced statistically significant findings and a replication (*N*

= 30) did not, most respondents gave an explanation for this difference related to theory, measurement constructs, or participant characteristics, rather than an explanation related to, more simply, the higher power of the initial study. In order to eliminate low power as a potential explanation of non-reproduced results, the *n* of a replication study should be at least the same as the *n* of an initial study. Furthermore, it may be desirable for a replication to have a *larger* sample size: Earp, Everett, Madva, & Hamlin (2014) argue that publication bias and the concomitant issue of increased likelihood of results being statistically significant and/or effects being on the high end of the distribution can mean that the *same* sample size might fail to reproduce the earlier findings or detect an effect at all. Even with a larger sample size, a replication study may not have sufficient power to find an effect similar to that of the initial study (or any meaningful effect) if that effect is spurious or overinflated. A priori power analyses, at a minimum, can help to address the issue of whether a change in *n* is needed for a replication (see Simonsohn, 2016 for related discussion).

To investigate the heterogeneity and sample size issues discussed in regard to replication, the current study documented the nature and amount of changes between the initial and replication studies. We also explored whether these changes were associated with the nomenclature of replications (e.g., 'direct' versus 'conceptual') and the extent to which their findings were supportive of the initial studies, as discussed next.

Extent of Reproducibility

The extent to which replications demonstrate 'reproducibility' of earlier findings partly depends on how 'reproduced' is defined. When reproducibility has been quantified in syntheses and meta-analyses of replication in other fields, there has been a range of outcomes. For direct replications in psychology, the Many Labs project found that 10/13 replications reproduced the initial findings, whereas the Registered Reports project (Nosek & Lakens, 2014) found that 10/13 did not; meanwhile, Rohrer, Pashler and Harris (2015)'s four

high-powered replications found no support for earlier studies. The Open Science Collaboration (2015) used different measures of reproducibility for their direct replications and found that based on Null Hypothesis Significance Testing (NHST), only 36% of replications yielded significant results compared to 97% of the initial studies. However, NHST can only provide a dichotomous perspective ('significant' or 'not significant') (e.g., Norris & Ortega, 2000; Norris, Plonsky, Ross, & Schoonen, 2015), and does not allow for a more fine-grained measurement of the extent of reproducibility. Broader categories for assessing reproducibility are needed to provide a more tolerant, less rigid measure that reflects some of the variability inherent in many studies, particularly likely in research with human participants and/or multiple complex variables (for discussion, see Earp, 2016 and Trafimow & Earp, 2017). For example, another approach is to use subjective ratings of reproducibility. Interestingly, however, the Open Science Collaboration (2015)'s subjective ratings approach led to assessments of reproducibility that were very similar to their NHST approach: based on subjective ratings on a seven-point scale from 'virtually identical findings' to 'not at all similar,' 39% of replications were deemed to have reproduced the initial result (versus 36% according to NHST). Perhaps the similarity in the findings was obtained because subjective ratings may have largely relied on the NHST reported in the studies. Note, however, that different outcomes were found when using effect sizes to assess reproducibility. Reproducibility increased to 47% when based on whether the effect size fell within the 95% confidence interval of the initial effect size. Finally, using yet another measure of reproducibility, Patil, Peng and Leek (2016)'s re-analysis of the OSF (2015) data found that 77% of the effect sizes were within a 95% prediction interval of the initial effect size. (See Francis, 2012; Lindsay, 2015; Maxwell, Lau, & Howard, 2015; and Stroebe & Strack, 2014 for further discussion of ascertaining reproducibility; and Marsman et al., 2017 for Bayesian approaches to assessing reproducibility).

Other, broader syntheses of the replication effort within whole discipline domains have made estimates of the extent to which findings have been reproduced, as reported by the authors themselves, using subjective rating measures. As in our systematic review in Part 2, this is a suitable estimate mechanism given that the replications included in these syntheses were not direct and so a precise, quantitative assessment of reproducibility was not a key aim. In the field of education, Makel and Plucker (2014) used a subjective three-level scale to rate 'reported replication success' in existing replications, of which only 14% were direct. They found that 67.4% of replications reported 'successfully' replicating the initial findings, 19.5% replicated some but not all findings, and 13.1% failed to replicate the initial findings. Using a similar rating scale, Makel et al. (2012) found that 78.9% successfully reproduced the initial findings, 9.6% did not, and 11.4% reported mixed support. Overall, the reproducibility rate in these fields has been calculated to range from around 36% to 79%, but depended on how this was assessed, among several other factors.

One such factor is that reproducibility is likely to vary according to subdomain. For example, in the Reproducibility Project, 25% of effects in social psychology were replicated (according to the criterion $p < 0.05$), compared to 50% of effects in cognitive psychology (as noted above, however, there are problems with using the dichotomous cut-offs of NHST). A second factor may be the type of replication. For direct replications, where minor differences in implementation are not theorized to influence the findings, expectations for reproducibility are high. Although it cannot be expected that all direct replications would find the same magnitude of effects or patterns of statistical significance as the initial study (Francis, 2012; Laws 2016; Lindsay, 2015; Open Science Collaboration, 2015), one might predict effect sizes within the 95% confidence intervals of the initial effect size, and, at the very least, the same *direction* of differences or associations. On the other hand, for partial and conceptual replications, which intentionally introduce change to the initial study design, researchers may

make theoretical predictions about why the change may (or may not) make a difference to findings. That is, partial and conceptual replications introduce more than just incidental 'operational heterogeneity', sometimes with the expectation of not reproducing the initial findings. An example of this from our study sample is Ellis and Sagarra (2011), who intentionally introduced more verb inflectional diversity into their materials, and found the difference in findings compared to the initial study that they were expecting. However, the intuitive expectation of less supportive findings emerging from partial or conceptual replication studies, compared to direct replications, does not seem to be observed consistently. For example, Makel et al. (2012) found that, in fact, descriptively more conceptual replications supported initial findings at a *higher* rate than direct replications (82.8% versus 72.9%, respectively), whereas Makel and Plucker (2014) found the reverse (66% versus 71.4%). However, neither pattern was statistically significant. In light of these issues, in the current synthesis we avoid describing replications as 'failed' or 'unsuccessful'; given that our sample did not yield any direct replications, *not* reproducing findings (however that is measured) does not necessarily indicate 'flaws' in either the initial or replication studies, as it could in fact have been expected. That is, we did not set out to evaluate the overall level of reproducibility in the field as being 'good' or 'bad'.

A third factor in reproducibility may lie in the independence of the replication researchers in relation to the initial researchers. In education, Makel and Plucker (2014) found that nearly half (48.2%) of the replications were conducted by the same research team that published the initial research. When at least one author was on both the initial and replication articles, 88.7% of replications were supportive of the initial findings, although the rate dropped to 70.6% if the replication was published in a different journal. With no author overlap, the rate dropped further, with 54% of replications being supportive. In psychology, Makel et al. (2012) found 91.7% were supportive with author overlap, versus 64.6% with no

overlap. Given the high rate of reproducibility with author overlap, Koole and Lakens (2012) focussed only on independent replications in their set of recommendations for replication, arguing that "the most compelling direct replications are conducted independently by different researchers than the original study" (p. 609). This was a key motivator for the pre-registered multi-site replications published by *Perspectives in Psychological Science* (soon to move to *Advances in Methods and Practices in Psychological Science*), in which research teams all have access to the same materials but conduct the study independently (and in some cases, do not look at the data until it has been passed to the replication convener or coordinating editor).

Note, however, that independence of researchers does not necessarily reduce bias, as bias can also affect an independent replicator who may predict findings against others' work (Bakan, 1967). Indeed, author overlap may bring perceived advantages. In a climate where there is little sharing of materials and data, author overlap may increase the chances of better fidelity to the initial study's materials and protocols. Indeed, Makel et al. (2012) found that most *direct* replications were conducted by authors of the original study. Similar to the availability of data being associated with better reporting and stronger evidence (Wicherts, Bakker, & Molenaar, 2011), the availability of *instruments* may affect the nature of results too, by, for example, increasing the likelihood of demonstrating support for the initial study's findings. In our own synthesis, we explored this possibility, partly driven by a concern that although more supportive findings may be a perceived benefit of author overlap, this may not necessarily be beneficial for the speed and objectivity of a broader scientific, community-based endeavour, as allowing others to access materials may facilitate faster and, perhaps, less partisan replication efforts.

The current synthesis did not aim to evaluate the reproducibility of L2 research, partly determined by the fact that we found no direct replications. (An anonymous reviewer

wondered whether we might undertake this, but as measures and other variables were very often changed between the initial and replication studies, and other changes were unacknowledged, ascertaining a general level of reproducibility in existing L2 self-labelled replication would not have been informative. Note, the recent endeavours in the field of psychology undertook new, direct replications with the explicit goal of measuring reproducibility in mind). However, we do provide a preliminary examination of whether the extent to which replications were supportive of the initial findings, as claimed by the replicating authors, may have been associated with certain factors such as the sub-type of replication, the independence of researchers, and the availability of materials. This examination is based on subjective ratings that coded the extent to which the replications' findings were reported as supporting the initial findings, as used by Makel et al. (2012) and Makel and Plucker (2014). This coding therefore relied on how the replicating researchers presented and discussed their data and analysis in relation to the earlier study.

Part Two: A Systematic Review of Self-Labelled Replication

Aims

The previous narrative review of commentaries, meta-analyses, and meta-science on replication closely informed the research questions and methods for our systematic review of replication in L2 research. For example, Makel et al. (2012) and Makel & Plucker (2014)'s syntheses of replication in the fields of psychology and education closely informed our investigations into: the quantity and nomenclature of replications, their publishing outlets, and citation counts; relations between the authorship of replications and their initial studies; the extent of 'independent' replication (with/without authorship overlap, in same/different journals); and whether findings were interpreted by authors as being supportive or otherwise of the initial studies. In these respects, our systematic review is, in broad terms, a conceptual

replication of Makel et al.'s (2012) and Makel & Plucker's (2014) systematic reviews, sharing common aims though with numerous differences in context and methods.

Other issues identified in our narrative review informed our systematic review, but had not, to our knowledge, been systematically examined in previous synthetic work on replication. For example, our review of infrastructure and projects that have helped methodological transparency and collaboration in psychology led us to document the transparency of our L2 initial studies, such as their reporting and availability of their materials, data, and analyses. This allowed us to examine the impact that methodological transparency and authorship overlap may have had on replication research, such as: whether and how replicators accessed materials and data; the existence of 'inter-connected' series of initial and replication studies; and associations between materials transparency and the extent to which replication findings supported the initial findings. Also, we wanted to estimate the time between a study and its replication when replications were published in separate articles to their initial studies (rather than within the same multi-experiment article, of which we found very few, in any case). Addressing these issues gave us insight into the infrastructural and cultural change that might be necessary to enhance the amount and quality replication research.

Other aspects of our systematic review were also indirectly informed by the narrative review above but were sharpened a great deal during the process of doing the systematic review itself. For example, when our search did not yield any self-labelled direct replications, then documenting heterogeneity (the amount and nature of changes that had been introduced into the replication studies compared to the initial studies) became a major undertaking in coding the articles. This led us to examine whether the amount of these changes was related to self-labelling nomenclature and to the extent to which a replication supported the initial study's findings. Additionally, a small number of issues were incorporated into our review

during the development of the coding scheme to document the kinds of studies that have been replicated in L2 research. These issues related to characteristics specific to L2 research, such as study design, measures, and participant characteristics.

In these ways, our systematic review converged on the following questions:

RQ1 How much self-labelled L2 replication has been published, and in which journals?

a) Which replication labels have been used?
b) Which journals have published replications?
c) What are the citation counts of replications, of their initial studies, and of the journals in which the replication and initial studies have been published?
d) To what extent have closely inter-connected series of initial and replication studies been conducted?

RQ2 What kinds of L2 studies have been replicated?

a) Have the findings from initial studies tended to be 'statistically significant' or 'null'?
b) What were the designs and contexts of the initial studies?
c) What were the participant characteristics in the initial studies?
d) To what extent were the initial study's materials accessible?

RQ3 To what extent and how did researchers change the initial L2 studies?

a) To what extent did the amount of change between initial and replication studies relate to nomenclature of replications?

RQ4 To what extent did L2 replications support the initial studies' findings?

a) How did authors compare their findings with the initial findings?
b) Which factors might have been associated with the extent to which replications were supportive of initial findings:
    i. author overlap?

ii. amount of change from the initial study?

iii. transparency of the initial study's materials?

Methods for the Systematic Review

Searching.

We focussed our search on academic, peer-reviewed journals, as we wanted to examine the extent of self-labelled replication in this medium, which has been identified as the primary channel for disseminating L2 research (Smith & Lafford, 2009; VanPatten & Williams, 2002). We therefore excluded replications in books, dissertations, conference proceedings, etc., following procedure in previous syntheses in the field (e.g., Plonsky & Gass, 2011; Plonsky, 2013). Admittedly, this leaves our sample susceptible to the effects of potential publication bias among journals. However, this would be particularly concerning for studies carrying out quantitative meta-analyses of substantive findings (as effect sizes are likely skewed upwards due to publication bias), but less of a concern here as we did not undertake such a meta-analysis. Nevertheless, we note that the 'file drawer problem' *is* likely to affect replications as much as, if not more than, other studies, due to concerns about manuscript rejection when findings do not align with those of the initial researchers (who might be chosen to peer-review the manuscript).

First, our review of commentaries about L2 replication yielded six empirical replications for potential inclusion. We then performed a keyword search for articles in LLBA and PsycINFO databases that contained in their title or abstract the word *replicat** AND either *second language* or *foreign language*, with no date restrictions. After combining results and removing duplicates, this yielded 891 hits (as of October 9, 2016). A Google Scholar search using these same keywords yielded a prohibitively high number of results (>18,000), and as we felt that our previous 891 hits provided a sufficiently representative picture for our purposes, these Google Scholar results were not used.

We then selected only articles that were in SSCI journals and written in English. To be included in our review, articles had to present empirical research with data from L2 learners, educators, or materials. We excluded many false hits of *replicat**, partly because researchers used the term to point to the need for replication of their own study or to claim their findings aligned with ('replicated') some earlier findings, though the study itself was not a replication attempt (see also Makel et al., 2012 who found that only 68% of articles using *replic** were actual replications). More details about exclusions, with examples, can be found at the end of 'Supplementary Material 1: Included studies and exclusion criteria.' After implementing these exclusion criteria, we ultimately identified 67 replication articles and the 70 initial studies they had replicated.

Coding the studies.

Our initial scheme, containing 61 categories for coding characteristics of these studies, was based on the narrative review above, including literature on replication in L2 research (e.g., Norris & Ortega, 2000; Polio & Gass, 1997; Porte, 2012). After 12 iterations during development, 42 of the original categories were maintained (marked ^ in the coding sheet in Supplementary Materials 2; 19 of these 42 were modified slightly during development) and 94 categories were added such that the final coding scheme had 136 categories (marked # in the coding sheet in Supplementary Materials 2). Of these, 80 were categorical (27 dichotomous, 53 of which had 3+ codes), 36 continuous, and 20 open-text. These categories captured information relating to seven clusters of characteristics consisting of:

- journal, article and author information (25 categories),
- study design and participant characteristics (81 categories, including differences between initial and replication studies),

- analysis procedures (4 categories),
- findings (16 categories, including 14 relating to the nature of analysis and discussion of the two sets of findings),
- materials availability (4 categories),
- response/commentary from the initial author(s) (1 category), and
- additional notes (5 categories).

The first pilot coding involved the four authors coding two articles, discussing these initial decisions and changing the scheme accordingly. In the second pilot, two authors both coded the same 14 'replication-initial' pairs of studies (21% of the total sample of studies). These were done over several weeks, and included meetings with all four authors in which some aspects of the coding scheme were clarified and disagreements were addressed. Interrater reliability was calculated for the coding of these 14 pairs of studies. Fifty-seven coding categories allowed Cohen's kappa (κ) reliability coefficient to be calculated, whereas other coding categories such as bibliographic information, long text answers, and entirely constant codes could not yield a κ value. The mean percent agreement between the two raters was 89%, and the mean κ was 0.80. To set this in the context of other methodological syntheses, Plonsky and Derrick (2016) reported κ = .74, Plonsky (2013) κ = .56, and Marsden et al. (accepted) κ = .86.

To further enhance reliability of coding for the remaining studies, categories for which the percent agreement fell below 80% (13 columns) were re-examined by the two coders, who either amended or confirmed their initial codes. After this, the percent agreement for every category was at least 80%, and the mean interrater reliability was 94%, κ = 0.88.

Using this finalized coding scheme, the two researchers individually coded the remaining 101 studies.

Analysis.

Our analysis of the codes almost exclusively draws on descriptive statistics, such as percentages and measures of central tendency and dispersion, as we sought to identify potential trends and formulate plausible accounts for them. During the analysis phase, 9 columns were added to the coding sheet, including article and journal citation data.

The final coding sheet, including percent agreement rates, κ values, and the data, is provided in 'Supplementary Materials 2: Coding sheet with data, IRR, analysis' and is openly available on IRIS (iris-database.org).

Results of the Systematic Review

Results are presented for each research question. Given the number and range of research questions, we provide some discussion with each set of results, to help readability. Further discussion and recommendations are provided in the final section (part three).

RQ1. How much self-labelled replication has been published, and in which journals?

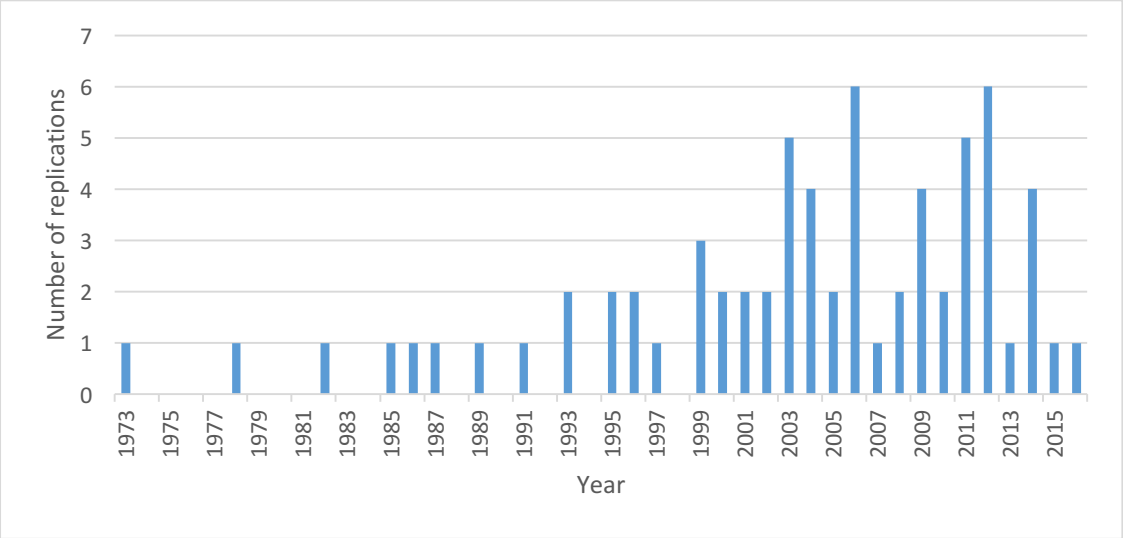
Our search found 67 self-labelled replications of 70 initial studies for a total of 129 study reports that were coded for further analysis.[1] All studies were published as journal articles except for five book chapters that were initial studies. (Three of these chapters were replicated in one replication study, Cobb, 2003). During our search, we also found 50 articles and chapters that were commentaries on or calls for replication in L2 research. This is just over two-thirds of the number of empirical self-labelled replications. See Supplementary 'Materials 1: Included studies and exclusion criteria,' in which replications are marked with *, initial studies with †, and commentaries and calls with °.

The earliest replication study was 1973, with a fairly steady increase beginning in the late 1990s until the most recent published replication in 2016, at the close of our search in October 2016 (Figure 1). In that time period, there was a mean of 1.55 (*SD* = 1.69) replications per year. The steady increase probably mainly reflects an increase in volume of research rather than in the proportion of replications itself. However, some of the increase

may be due to seminal papers promoting a synthetic approach, such as Polio and Gass (1997) and Norris and Ortega (2000), as well as dedicated replication article types in certain prominent L2 journals (e.g., launched in 1993 and refreshed in 2015 by *Studies in Second Language Acquisition*, and started in 2014 by *Language Teaching*). The mean time between a study and its published replication was 6.64 years (*SD* = 6.16, median = 5, mode = 1 (*k* = 11), range = 0 to 37). This time delay demonstrates to us the need for a sustained infrastructure to help replications to be performed and published more quickly, as ascertaining the generalisability and reliability of study findings can reduce the chance of self-perpetuating misinformed agendas and of drawing implications for practice too hastily (see also Koole & Lakens, 2012).

Figure 1*:*

Self-labelled replications published in journals



RQ1a Which replication labels have been used?

Examining the nomenclature used for replications, we found that after the single term "replication" (used in combination with "extension" in 25% of studies), the next most common label was "partial replication" (21%) (Table 1). However, a wide variety of terms

were used, including "strict replication," "replication design," "modified replication," and "follow-up study." Many (k = 24) used multiple terms for the same study. Certain terms were never used despite having been used or recommended in commentaries on replication: 'true,' 'direct,' 'exact,' 'quasi,' and 'ceteris paribus.' Overall, nomenclature was not precisely defined or consistent across studies, reflecting the confusion mentioned by Polio (2012b). We revisit nomenclature in our analysis of the extent to which labels reflected the amount of heterogeneity between initial and replication studies.

Table 1

Terms used to label replications

	percent of studies (k = 67) self-labelling as replication in title or abstract
Close replication	1%
Approximate replication	3%
Partial replication	21%
Conceptual replication	4%
Replicat* (without a qualifier)	67%
Other	3%

RQ1b Which journals have published replications?

Replication articles were found in 26 different journals. Five journals published four or more replications, giving a replication rate of 0.66% across these journals (number of replication articles / total number of research articles (excluding editorials etc.); see Table 2). Of the 26 journals that have published replications, the great majority of journals (k = 21) published three or fewer replications (Supplementary Materials 3: Table 1). Across all 26

journals, we estimated the replication rate as 0.26%. This was calculated as follows: the number of replication articles / estimated total number of articles. The estimated total number of articles used the mean total of 996.8 articles produced by each of the top five journals in the time period found by the synthesis (see Table 2), multiplied by 26 journals for a total of 25,917 articles. Expressed differently, the formula estimates that one in every 400 journal articles was a self-labelled replication. This is a generous[2] estimation of the rate of self-labelled L2 replication, but still falls below the mean rate in psychology in 2012, which would now be higher since the recent surge in replications (as discussed above). We estimate that the field of L2 research may have a similar rate as education (calculated in 2014 at 0.13% by Makel & Plucker, 2014), or perhaps lower given that the denominator for education used a much larger number of journals whereas we used only those journals that *have* published a self-labelled replication.

Table 2

Rates of replications in the five journals publishing the most replications between 1973 – 2015[a]

	Studies in Second Language Acquisition	*The Modern Language Journal*	*Language Learning*	*Foreign Language Annals*	*Applied Psycho-linguistics*	Mean	Total across five journals
Number of replications (of which, initial study is in same journal)	11 (6)	8 (4)	5 (0)	5 (2)	4 (1)	6.6	33
Total number of articles[b]	562	1009	855	1528	1030	996.8	4984
Percent replication	1.96	0.79	0.58	0.33	0.39	0.81	0.66

Note. [a]The year 2015 was the last complete year captured by our synthesis. [b]To calculate the denominator (total articles published), we used a start date of either 1973 (the date of our first replication) or the start of the journal if that fell after 1973.

The low replication rate may be partially due to journals' (dis)encouragement of replications. Of the 26 journals that had published replications, only four explicitly stated that they accepted replications (*Studies in Second Language Acquisition (SSLA), Second Language Research (SLR), Language Teaching (LT),* and *Language Testing*). Interestingly, however, only one of these (*SSLA*) was in our 'top five' of journals publishing self-labelled replications, the others having published three, two, and two respectively. Two of the four journals that stated they accepted replications emphasized originality in the first sentence of their aims/scope sections. Three of these four journals reserved specific strands for replications (*SSLA*, *LT*, *SLR*). Two of these were shorter article types, which might make overt comparability with the initial study difficult (we refer to this issue further in our recommendations about peer reviewing of replications). Ten of the 26 journals implicitly discouraged replications, and nine of these emphasised originality, novelty or innovation in the first or second sentence of the aims/scope sections. Although three journals specified that methods should be clear enough to allow others to replicate the study, two of these did not explicitly state that they accepted replications in their own journal. Finally, two journals explicitly mentioned that null findings would not be grounds for rejection *per se* (*SLR* and *Language Testing*, both journals that encouraged originality and explicitly accepted replications).[3]

Note that beyond this analysis of the number of replications published by journals and the explicit and implicit messages that journals send to authors, we cannot know how the replication rate of L2 research reflects the extent to which authors *submit* replications that are

ultimately rejected versus the extent to which replications are simply not submitted. For this, surveys of editors and reviewers are necessary. Martin & Clarke (2017)'s review of such research shows that none has yet been done specific to language learning or education; data on this is central to improving our understanding the causes of low rates of published replications.

RQ1c) What are the citation counts of replications, of their initial studies and of the journals in which the replication and initial studies have been published?

With insight into the numbers and places of publication, we turn to examining the impact of self-labelled replications. Journal Impact Factors (IF) from the Web of Science (Thompson Reuters) and the replication and initial studies' total citations (according to Google Scholar) were recorded in May 2017[4].

Table 3 shows that article citations were higher for initial compared to replication studies, unsurprising given that high citation often motivates replication and that initial studies had been citable for more years.[5] To take 'years since publication' into account, we divided total citations by the number of years between publication and 2017, to provide mean citations per year. In terms of the relationship between median citations of replications and their initial studies, we found a ratio of 0.25 for L2 research, which aligned very closely with psychology (0.27) and was a little higher than in education (0.16) (calculated from Makel and colleagues' data).

Table 3

Article citation counts and journal impact factors for replication and initial studies

	Total article cites	*Annual* article cites	Journal five-year IF

	Mean (SD)	Median (range)	Mean (SD)	Median (range)	Mean (SD)	Median (range)
Initial studies[a]	364.03 (678.14)	173 (1-4445)	17.65 (24.30)	8.68 (0.03-118.8)	2.39 (1.22)	1.95 (0.24-6.29)
Replication studies	92.91 (113.41)	44 (3-618)	7.26 (6.58)	4.89 (0.33-38.63)	2.00 (0.97)	1.88 (0.24-4.36)

Note: [a]In cases where two initial studies were replicated by one replication study, the citation count of both initial studies was recorded.

Although the total and annual citations were higher for initial studies, the citations of replications were far from low, despite this being a frequent concern about replication work. The mean *annual* citation of replication articles (7.26) was well above the mean IF of the journals publishing replications (2.00) and initial studies (2.39). It was also above even the highest journal IF in the SSCI for linguistics (*Journal of Memory and Language*, 5.22) and education (*Educational Psychologist,* 5.69). This is compelling evidence that replications, at least those published to date, do not have low impact.[6]

We estimated that the mean number of citations of a study before its replication was published was 117.20 (based on the mean 6.64 years between an initial study and its replication, and the mean 17.65 annual citations of an initial study). We acknowledge that this is an estimation based on an *average* evenly spread over time. For L2 research (where citation counts are generally much lower than, for example, psychology), we consider this to be a high number of citations before a study's reliability and generalisability are investigated via replication, especially given the large standard deviations in our data that indicate that some studies received many hundreds of citations before they were replicated.

In terms of the IF of journals that publish replications (Table 3), journals with both high and low five-year IF published replications, with no discernible association between IF

and number of replications published (r_s (26) = .157, p = .443). On average replications were published in journals with slightly lower IF than the initial studies, though with a very small effect size whose lower 95% confidence interval almost reached zero (t(128) = 2.059, p = 0.042, Cohen's d = .36, CIs 0.01-0.70). This small to negligible difference would partly be due to the fact that just over a third (38.8%) of replications were published in the same journal as the initial study (compared with 30.6% in education and 19% in psychology).

RQ1d) To what extent have closely inter-connected series of initial and replication studies been conducted?

Our search identified 67 replications based on 70 initial studies. The mismatch in these numbers reveals some inter-connectedness between groups of studies, where four studies replicated more than one initial study: DeKeyser and Sokalski (1996) replicated VanPatten and Cadierno (1993a & 1993b); Liu (1985) replicated Au (1983 & 1984); Walters (2012) replicated Fitzpatrick and Meara (2004) and Fitzpatrick and Clenton (2010); and Ellis et al. (2014) replicated both Ellis and Sagarra (2010) and Ellis and Sagarra (2011, experiment 1). In these cases, the replications were conceived of (both by the authors and by us) as one replication. Further interconnectedness was found in two lines of research as follows. First, Ellis and Sagarra (2011) served both as an initial study for the Ellis et al. (2014) replication and was itself a replication of Ellis and Sagarra (2010), and thus was coded as both a replication and an initial study. Second, VanPatten and Cadierno (1993a) served as an initial study for DeKeyser and Sokalski (1996) and for VanPatten and Oikkenon (1996), and so was coded twice in its capacity as an initial study. Overall though, the interconnectedness of groups of studies was minimal, given that from the 67 self-labelled replications, only four were associated with more than one initial study, only one continuing line of replications was identified, and only one study was replicated more than once.

It is of course highly likely that more interconnectedness in L2 research exists than was evidenced in our search, due to unwillingness to self-label as 'replication.' Indeed, several of the initial studies were closely related (close enough to be replicated simultaneously by one study), but did not self-label as replications themselves. However, it remains worrying that our sample only provided two clusters of studies that self-labelled as an overt sequence of an agenda that extended beyond two studies (the VanPatten/Cadierno/DeKeyser cluster and the Ellis/Sagarra et al. cluster). Among other concerns, it suggests that the many syntheses and meta-analyses in the field (e.g., Plonsky & Brown, 2015 examined 81 meta-analyses) are bringing together studies that did not self-identify as replications of any kind. This issue is frequently observed by meta-analysts as they comment on the less-than-ideal comparability between studies in the domain under investigation (due to inconsistency of materials and measures etc.), and is one cause of low *k* in meta-analyses (e.g., Oswald & Plonsky, 2010, found a median of 16 studies reviewed in 27 meta-analyses in L2 research).

RQ2 What kinds of studies have been replicated?

RQ2a) Have the findings from initial studies tended to be 'statistically significant' or 'null'?

First, we checked the nature of analyses reported in the initial studies, and found that, as expected, statistical procedures largely reflected null hypothesis significance tests (mainly ANOVAs and *t* tests) that are normally used in L2 research (Plonsky, 2013) (for details, see Supplementary Materials 3: Table 2).

Next, we coded how the initial studies' findings were reported by the authors, on a four-point scale, as follows: (a) rejecting a null hypothesis[es]; this was usually reported as a finding of statistically significant difference/association between the variables under investigation, with alpha=0.05); (b) failing to reject the null hypothesis(es); this was usually

reported as no statistically significant difference/association between the variables under investigation; (c) trend/borderline differences/associations, as interpreted by the authors; or (d) 'other' (usually indicating that statistical significance was not applicable to the research design). This coding was necessarily broad-brush, but the overwhelming finding was that researchers replicated studies that had a statistically significant finding (87%), with only 3% of studies replicating a study with 'null' findings, 3% with a trend towards an effect, and 7% other. This suggests an influence of publication bias and/or the file drawer problem (note, though, that we *did* include replications of initial studies that were not from journals). It is also possibly a consequence of (perceived or real) difficulties in interpreting 'null findings' *without* ascribing methodological flaws to the study, which probably decreases the impetus to replicate such studies.

In our view, these data fuel compelling arguments to (a) investigate the extent of publication bias generally by increasing overall replication effort (among other approaches), (b) increase all types of replication (exact, partial, and conceptual) of studies with null findings, to inform theory and ascertain the extent to which initial null findings were indeed 'due to methodological flaws,' and (c) undertake peer-review prior to data collection to reduce publication bias. We revisit these issues in the Discussion and Recommendations section.

RQ2b) What were the designs and contexts of the initial studies?

We examined characteristics of the initial studies to explore whether particular design features seemed to have a propensity to be replicated. The majority of replicated studies were one-shot, cross-sectional designs. However, more complex designs were also replicated, such as longitudinal (40%) and intervention (37%) studies.[7] In terms of context, 50% were laboratory-based and 39% had collected data in a classroom (see Table 4).

Table 4

Study types/contexts in initial studies

	Percent initial studies (*k*=70)
Laboratory	50%
Experimental/manipulated classroom	20%
Intact/ecologically valid classroom	13%
Lab plus intact or experimental class	6%
Not reported	9%
N/A	3%

In terms of the measures used in the studies, the majority examined morphosyntax and used measures that were linguistic, written and administered offline (see Table 5). However, overall, a very wide range of linguistic forms and assessments appeared in the initial studies, as shown in Table 5.[8]

Table 5

Measure and instrument types used in the initial studies

Language feature	Percent initial studies[a]	Measure focus/type	Percent initial studies[d]	Measure modality/mode	Percent initial studies
Morphosyntax	40%	Linguistic	87%	Oral	26%
Lexicon[b]	23%	Non-ling.	9%	Written	49%
Pragmatics	10%	Both	4%	Both	17%
Speech[c]	9%	Offline	83%	n/a	9%

Multiple features	19%	Online	9%	Comprehension	23%
Not reported, n/a	7%	Both	0%	Production	23%
		n/a	9%	Both	44%
				n/a	10%

Note. [a]Adds up to more than 100% as some studies had more than one. [b]Including collocation and figurative language. [c]Including phonology, prosody, pronunciation, fluency. [d]Throughout, unless otherwise stated, where a column (or row where applicable) does not add up to 100%, this is due to rounding error.

This variation in design characteristics and the findings that 67% of studies included a production measure and 43% had oral measures (which are usually more difficult to administer and score) suggest that L2 replication efforts have *not* tended to replicate only 'easier' studies. Interesting, although one might think that highly controlled, laboratory-based research would be more conducive to replication, studies with an online measure (such as self-paced reading or eye-tracking) were rarely replicated in our sample ($k = 6$). This may reflect the relatively recent adoption of such techniques in main stream L2 research (as found by Marsden et al., accepted), but also the challenges posed by accessing and using expensive hardware and software that is also comparable across sites and studies (as noted by Laws, 2016, experienced by Morgan-Short et al. 2018, and discussed, with practical advice, by Schmid et al., 2015). Infrastructure for collection of data via the Internet, such as that proposed by MacWhinney (2017), would help to alleviate this problem.

RQ2c) What were the participant characteristics in the initial studies?

Participant characteristics, such as age, language background, and proficiency, also provide critical insight into the kinds of studies that tend to be replicated. In terms of language proficiency, we found that of the 62 initial studies with language learners,[9] 17 gave some indication of whether participants were beginner, intermediate, advanced, or a

combination. However, 25 did not specify the proficiency level and 20 studies were coded as 'other' for a range of reasons (e.g., gave number of years of learning experience). In terms of ages, 29 studies used university students without specifying ages, which in reality vary enormously but typically range between 18 and 30. Of the 22 studies that did report participants' age, we calculated a mean of 22.18 years (*SD* = 11.68).[10] Finally, most initial studies involved English as the target language. There was a little more variation seen in participants' first language, and seven studies did not report the participants' first language (see Supplementary Materials 3: Table 3). In all, replications have been largely of initial studies with young adult learners of English, in line with previous observations about participant demographics in L2 research (Plonsky, 2013). Most critically for the current study, our data (or lack of it) clearly demonstrates how unclear reporting practices have adverse consequences for replicability, as replicators cannot know what sample population to target, which characteristics they may wish to intentionally change, or which they should acknowledge as being different from the initial study).

RQ2d) To what extent were the initial study's materials accessible?

A final feature related to the kind of studies that have been replicated is the degree to which initial studies are transparent in terms of materials. We found that 17% of initial studies did not provide any materials at all, and that 41% provided only partial examples in the article's text. Although 37% did provide at least one full instrument, these did not provide all of the instruments used to collect the data that were ultimately analyzed in the study. Only three of the studies in our sample provided a full set of materials (see Table 6).

Table 6

Availability of materials in initial studies

Material availability	*k* initial studies	Percent behind journal paywall[a]	Percent open access	Percent n/a or other
No materials	12	-	-	-
Partial examples	29	90%	3%	7%
One full instrument (not all materials)	26	73%	8%	19%
Full materials used for analysis	2	100%	0%	0%
All full materials	1	100%	0%	0%

Note. [a] When materials are available behind a journal paywall, this does not make replication easy as not everyone has access to all journals (e.g., researchers in certain socio-economic contexts or practitioners without journal subscriptions). Additionally, it is possible to acquire some articles via open access portals, and so know about a study but not have access to its materials which can remain behind journal paywalls in supplementary materials.

Our data regarding availability of materials begs the question of how replicating researchers acquired the materials needed to replicate the study. In our sample of replication and initial studies, it was often unclear how materials were obtained or whether they had been 'recreated,' especially in cases where no materials or just examples were available (see Tables 6 and 7). Thus, replications seemed to have been carried out even when materials were not available or were only described. As with gaps in reporting about participant characteristics, poor availability of materials reduces the replicability of studies and also weakens claims that can be made by replications (as the extent of parity with them is difficult to ascertain).

Table 7

How materials were made available to the replicators

Availability in initial study (for *k* replications)	Percent in article	Percent passed on in private[a]	Percent shared authorship[b]	Percent unclear
No materials (12)	0%	25%	33%	42%
Partial examples of an instrument (26)	54%	12%	31%	4%
One full instrument (26)[c]	85%	4%	19%	0%
Full materials used in analysis (2)	50%	50%	0%	0%
All full materials used in entire study (1)	100%	0%	0%	0%
Total (67)	54%	12%	25%	9%

Note [a]Acknowledgement sections were searched to determine whether researchers were thanked for materials. [b]Materials were not available with the initial article or open access, so we assumed materials were passed on via the author(s) common to the initial and replication studies. [c]Adds up to more than 100% because two studies had an instrument in the article and had shared authorship.

RQ3. To what extent and how did researchers change the initial studies?

In the narrative review, we noted that a limited number of motivated changes between an initial study and its replication (such as those suggested by the initial study authors as future directions) can be desirable for systematic research agendas, but that too many changes, or changes that are unmotivated or unacknowledged, impede one's ability to account for differences in the findings between studies. To gain insight into the types and numbers of changes between initial and replication studies, we coded and counted each change between pairs of studies. We distinguished between three types of changes: changes that were overtly reported as intentional alterations that explicitly motivated the replication, as expected in partial and conceptual replications (henceforth referred to as 'motivated

changes' or a 'motivation for replication'); changes that were acknowledged by the authors but were not explicitly articulated as principled motivations for the replication (henceforth 'acknowledged changes but not motivations for the replication'); and changes that were noted by the coders but were not acknowledged by the authors (henceforth 'unacknowledged changes').

In terms of changes to participant characteristics (see Table 8), the participants' L1 was the most common, often as an intentional change motivating the replication (k=21) or an acknowledged but unmotivated change (k=6). There were a few instances of motivated changes to participants' L2 or level of L2 proficiency. Reassuringly, there were no instances where authors overtly claimed that participant characteristics were constant between studies but the coder thought there had been a change. However, there were several instances of unmotivated or unacknowledged changes. For example, six studies changed the L1, six the proficiency and 19 the ages of the participants without explicitly acknowledging these differences.

Table 8

Percent of replication studies with changes to participant demographics

Participant characteristic	No change	Claimed constant, but coder identified change	Change motivation for replication	Change acknowledged, but not motivation for replication	Change not acknowledged	Unclear/ not reported/ n/a
L1	43%	0%	31%	9%	9%	7%
L2	76%	0%	13%	7%	3%	0%
Proficiency	39%	0%	15%	10%	6%	30%

Age	58%	0%	1%	3%	19%	18%

In terms of linguistic features, mode (production/comprehension) and modality (written/oral), we observed surprisingly few changes, with only about one in five of the replications amending one or more of these characteristics (Supplementary Materials 3: Table 4). However, about half of the replications changed the outcome measures in various ways, such as using different items, tasks, stimuli, or proficiency measures, or manipulating whether a test was done in a pair or a group. A quarter of studies made such changes to the measures that were either not motivated or not acknowledged (Supplementary Materials 3: Table 4, final row).

Changes to measures were often justified as improvements to the data elicitation techniques used in the initial study. Thus, one reason might have been poor instrument or coder reliability found in the initial studies. However, indices of reliability (such as Cohen's alpha, percent agreement, or Cohen's kappa coefficients) were reported in only 17% (k=12) of the initial studies. Thus, changes to instruments appeared to be largely based on the replicating researchers' subjective evaluation of the instrument.

The extent and purpose of these changes is concerning. For example, changing the data elicitation instrument is a significant change, best conceived of as an intentional alteration that motivated a replication. Such changes can, if they are not an intentional design feature (which was the case in a quarter of our replication studies), constitute a major threat to interpretability, particularly in cases where findings are different between the studies. Of course, there is a tension between changing a measure for perceived improved internal validity and compromising the initial study's characteristics and, therefore, one's capacity to determine the cause of differences in findings. To us, these findings underscore the need to continue refining and sharing the field's measurement toolkit in order to reduce the need to

change measures between inter-connected studies and thus increase parity between those studies. Indeed, this goal was one of the main purposes behind establishing the IRIS database of research materials (Marsden et al., 2016).

In terms of study design more generally, we observed very few changes (Supplementary Materials 3: Table 5). However, 23% of replications made changes to the study's context, i.e., a second versus a foreign language context (though this change was motivated for only 10% of replication studies). Researchers largely maintained the longitudinal or cross-sectional designs of the initial studies, with just three exceptions. There were, again, instances where changes were not acknowledged, the most concerning of these being in the domain of the statistical analysis, with over a third of studies using different statistical procedures without clearly justifying this change. Although some of these changes were appropriate, given the other changes made by the replication, the explicit flagging to the reader was inconsistent.

Another change that may occur between initial and replication studies involves the sample size. We found that the replications' subgroup sample size was a mean of 4.4 (median = 1.4) smaller than that of the initial studies, with a very large standard deviation (51.4) and a wide range from -304.0 to 108.5.[11] As noted earlier, smaller sample sizes in replication studies compared to the initial studies can be problematic if effects observed in the initial study are not observed in the replication, as this could be accounted for both by lower power and/or a genuinely different finding. Despite this concern and variation in sampling practices, sample sizes in replication research generally seemed to be higher than the averages found in other, broader syntheses of L2 research: mean study sample size $N = 114.4$ for initial studies and $N = 88.1$ for replications; mean subsample size $n = 41.1$ for initial studies, and $n = 36.4$ for replications (compare with the median[12] of $n = 19$ reported by Plonsky, 2013; the mean of

n = 22 reported by Plonsky & Gass, 2011; the medians per condition of 26 [within-subject designs] and 20 [between-subject and mixed designs] reported by Lindstromberg, 2016).

Collapsing across the types of changes (see Table 9, last row), there was, per replication, a mode of (a) two motivated changes, (b) one acknowledged but not motivated change, and (c) two changes that were not acknowledged by the authors. Overall our findings suggested that in much L2 replication work to date, there have been about as many or more unmotivated and unacknowledged changes per study as motivated changes. As such, it would currently be difficult to make any general evaluation of the reproducibility of L2 research.

Table 9

Percentage of replications making different types of changes to the initial studies

Number of changes	Claimed constant, but coder identified change (% replications)	Change motivation for replication (% replications)	Change acknowledged, but not motivation for replication (% replications)	Change not acknowledged (% replications)
0	94	33	45	46
1	4	28	31	25
2	1	21	13	15
3	0	9	9	9
4	0	7	1	4
5	0	1	0	0

Mean per study	0.07	1.34	0.91	1.00
(SD)	(0.32)	(1.31)	(1.04)	(1.18)
Mode per study	0	2	1	2

RQ3a) To what extent did the amount of change between initial and replication studies relate to nomenclature of replications?

Given that there was such variability in nomenclature (Table 1) and that the majority of studies are simply self-labeled as 'replication' (with no further qualification), we were unable to statistically examine the numbers of changes as a function of the sub-label of replication. Descriptively, we were not able to find any clear discernible patterns. For example, in the three studies that called themselves 'conceptual replications,' where one could expect several and all types of change, we found very different patterns. Specifically, two 'conceptual replications' had no 'motivated changes,' whereas the other had three; regarding 'changes acknowledged but not a motivation for the study,' one had none, one had one, and the other had three; and, finally, regarding 'unacknowledged changes,' two conceptual replications had two, one had four. Our one self-labelled 'close replication' (Waring, 1997) perhaps fitted the expected profile, having one change that motivated the replication and no other changes to key variables. We provide, in Supplementary Materials 3: Table 6, the two sets of self-labels that had the largest *k* in our sample: 'partial' (*k*=14) and 'replicat*' (without a qualifier) (*k*=45). The data show that the amount of change seems to be similar regardless of the label. We acknowledge that the low number of 'partial replication' studies precludes firm conclusions, but at the very least the data demonstrate little systematicity of nomenclature. This replication self-identity crisis is arguably one cause of the lack of self-labelled replication published in the field, as authors, reviewers and editors

vary in their understanding of what does and does not constitute (different types of) replication. We return to this issue in our Further Discussion and Recommendations.

RQ4) To what extent did replications support the initial studies' findings?

In order to examine this question, we coded on a four-point scale the extent to which the initial study's findings were supported by the replication as claimed by the authors of the replication (see Table 10): 0 = not supported (results did not support the initial findings at all); 1 = partially not supported (the majority of results did not support the initial findings); 2 = partially supported (the majority of the results supported the initial findings); 3 = very supported (results supported the initial findings).[13] We found that most studies (68%) presented findings that were generally supportive of the initial studies, which aligns closely with Makel and colleagues' findings of 67.4% for education and, more loosely, to the 78.9% found for psychology. That is, just under a third of our replication studies produced findings that were divergent from the initial study, arguably demonstrating the basic need for replication research to corroborate the validity of findings in L2 research generally. However, as noted above and below, supportive or non-supportive findings from studies that were not *direct* replications (as in the current synthesis) cannot provide a meaningful indication of reproducibility in the field, as many of the replication studies introduced substantial heterogeneity into their design, either intentionally or otherwise.

Table 10

Extent to which replications supported the initial studies' findings

	Percent replication studies (k = 67)
Not supported	15%

Partially not supported	13%
Partially supported	34%
Very supported	34%
Unclear	3%

RQ4a) How did authors compare their findings with the initial findings?

We explored how replicating authors compared their findings with the initial study's findings by coding for two main issues. First, we coded how the initial study's data were presented by the replicators (Table 11) and found that only about a quarter presented descriptive statistics from the initial study, and that even fewer studies used other types of statistics or data from the initial study. Whatever this is due to (space constraints, an assumption that reviewers and readers will access the initial article, or lack of incentive to report fully), it renders basic comparisons between studies difficult.

Table 11

How replications presented and used the results from the initial study

	Provided descriptive statistics	Provided inferential statistics	Extracted reported data and analysed with replication data	Provided effect size	Used raw data in a new statistical analysis
Percent replications	28%	13%	12%	6%	6%

Second, we coded for how the data from both studies were compared (Table 12) and found that comparisons between the studies were generally narrative or based on a

dichotomous interpretation of an NHST, e.g., findings were 'significant' or not. These two observations (that comparisons were almost exclusively narrative or based on NHST and that so few analyses used the initial study's data) are hardly surprising given the lack of availability of effect sizes and raw data in the initial studies.

Table 12

How replications drew comparisons with the initial studies

	Narrative comparison	Mentioned initial study's findings	Based on dichotomous interpretation from NHST	Compared descriptive statistics	Unclear	Compared effect sizes
Percent replications	93%	90%	84%	34%	6%	1%

Effect sizes, as noted many times (e.g., Norris et al., 2015), are useful as they enable comparisons using standardized units across studies to interpret the magnitude of difference or association in meaningful paired comparisons. Morgan-Short et al. (2018) provide an example of a study giving independent effect sizes for inter-site comparisons and aggregated effect sizes in an intra-study meta-analysis of the direct replications (see also Ellis et al. 2014). In our sample of 70 initial studies, Cohen's *d* was provided in 7 studies and *r* by one study, whereas 81% did not provide any effect sizes[14]. We also did not find instances of replicators extracting effect sizes from the initial studies (e.g., Cohen's *d* can be calculated from *t* and *F* statistics when comparing two groups). We found it surprising that the use of effect sizes had not become more embedded by the time of this review, given that many of the initial and most of the replications happened after Norris and Ortega's (2000) influential

meta-analysis emphasizing the importance of effect sizes, and after several journals started requiring the provision of effect sizes.

As noted above, there were similarly small numbers of studies that used the raw data from the initial study in the replication's analysis. Such access to data was possible because all four studies had author overlap (final column, Table 11). Interestingly, three of these found 'very supportive' evidence for the initial study. The fourth study, Ellis and Sagarra (2011) found evidence partially not supportive of Ellis and Sagarra (2010) (and was the only study to use Cohen's *d* to draw comparisons). This brings us to the question of the factors—including that of author overlap—that may be associated with replication studies producing findings that were supportive of the initial findings.

RQ4b) Which factors might have been associated with the extent to which the replications were supportive of the initial findings: i) author overlap

We first quantified the amount of author overlap in our sample of studies and found that 6% ($k = 4$) of the replications had the same authorship as the initial study; 25% ($k = 17$) had some authorship overlap (one or more authors in both the initial and replication studies); and 69% ($k = 46$) were carried out by entirely new authorship teams.[15] This could imply a degree of independence in the replication research in our sample. We explored various effects that overlap in authorship may have had on the extent to which replications supported the initial findings.

First, as seen in Table 13, authorship overlap seemed to be associated with supportive findings. When there was no author overlap between the initial and replication studies, 37% ($k = 17$) of replication studies were generally not supportive and 59% ($k = 27$) were generally supportive, whereas with some author overlap, approximately just 10% ($k = 2$) tended not to be supportive and 90% ($k = 98$) were supportive. This pattern was statistically significant (Pearson χ^2 (1) = 5.824, $p = 0.016$; Likelihood ratio 6.634, $p = 0.01$). It also aligns with the

ratios found by Makel and colleagues in psychology (91.7% supportive with author overlap; 64.6% supportive without) and education (88.7% with; 54% without).

Table 13

Percent of replications that were supportive/not supportive of initial findings, as a function of author overlap

Author overlap (*k* studies)	Not supportive	Partially not supportive	Partially supportive	Very supportive	Not reported/clear
No overlap (46)	20%	17%	30%	28%	4%
Some overlap (21)	5%	5%	43%	48%	0%
Total percent of replications (67)	15%	13%	34%	34%	3%

There are several explanations for this data about author overlap. They could reflect QRPs, which may be more likely if an initial study author is biased towards finding a particular outcome in the replication. They could (also) be a consequence of greater fidelity to the initial study because materials were available and protocols were more strictly adhered to. This might be because author overlap could incur fewer 'researcher degrees of freedom' (Simmons, Nelson, & Simonsohn, 2011), i.e., a reduced likelihood of divergence at the many decision points in any study. There may (also) be a possibility that replications with author overlap might be more likely to have a confirmatory aim (and therefore be closer, with fewer changes), rather than to test generalizability by intentionally manipulating several variables.

To further investigate this last possibility, we compared the number and type of changes between replication and initial studies as a function of author overlap. Although author overlap did not seem to be associated with the proportion of studies that changed just

one feature as a specific motivation for the replication, we found that, overall, replications with author overlap tended to make fewer changes to the initial studies (Table 14). First, there were slightly more replications with author overlap than without overlap that made *zero* changes of all types (motivated, acknowledged, and unacknowledged). Second, there were more studies without author overlap than with overlap that made several unmotivated or unacknowledged changes. This indicates closer replications (i.e., involving less heterogeneity) with author overlap, which, intuitively at least, seem more likely to produce findings that are more in line with the initial studies. Thus, the extent to which replications supported the initial findings could, at least partially, be accounted for by the trend that replications with author overlap were closer than those without author overlap.[16]

Table 14

Percentages of the studies with total number of changes, as a function of author overlap

	Changes were motivation for replication			Changes acknowledged but not motivation for replication		Changes not acknowledged by authors	
Number of changes	0	1	2+	0	1+	0	1+
No overlap (*k*=46)	30%	28%	42%	41%	59%	39%	61%
Some overlap (*k*=21)	38%	29%	34%	52%	48%	62%	39%
k replications	22	19	26	30	37	31	36

RQ4b) Which factors might have been associated with the extent to which the replications were supportive of the initial findings: ii) the amount of change from the initial study

It may be that increased heterogeneity—quantified as the number of changes between the replication and initial studies—independently from authorship overlap, could be linked to a higher likelihood of replications producing findings that were supportive of the initial studies. However, the data in Supplementary Materials 3: Table 8 suggested no strong or interpretable patterns in this matter. This broadly aligns with the lack of evidence in psychology and education research that direct replications were not any more (or less) likely to support initial findings than conceptual replications (Makel et al., 2012; Makel & Plucker, 2014). It also chimes with the negligible to small effects of heterogeneity found in the Many Labs project (Klein et al. 2014). These findings suggest that other issues may be more strongly linked to the extent of supportiveness, such as the nature of the effect under investigation (as argued by Klein et al., 2014), the theorized intention of the heterogeneity, or perhaps, as examined below, the transparency of the initial study's materials.

RQ4b) Which factors might be associated with the extent to which the replications are supportive of the initial findings: iii) transparency of the initial study's materials

Finally, it may be that without access to full materials from the initial study, replicating researchers need to create their own materials; this would introduce unintentional and unacknowledged heterogeneity between studies, which could in turn account for less supportive findings. Thus, we examined whether the availability of the initial study's materials was associated with supportiveness of findings (Table 15). Of the 65 studies that could be included in such an analysis, we observed that instrument transparency was associated with an increased likelihood of replications producing supportive findings, a

pattern that was statistically significant (Pearson χ^2 (1) = 11.489, p = .003).[17] We think that this provides some evidence for one benefit of making materials transparent.

Table 15

Supportiveness of replications as a function of the availability of the data collection instruments

	None	Examples	One full
% (partially) not supportive (*k*=18)	42%	37%	16%
% partially/very supportive (*k*=44)	9%	41%	48%
k	12	26	24

Overall, regarding the factors that are associated with reproducibility, our results seem to suggest that author overlap and the availability of materials was associated with supportive findings, whereas the number of changes between initial and replication studies was not.

Part 3: Further Discussion and Recommendations

In light of these narrative and systematic reviews and our own experiences with replication work, we summarize key findings and propose a set of recommendations. Our discussion and recommendations align with the four main themes addressed by both the narrative and systematic reviews above, though these themes are fragmented into seven sub-sections here, and we adopted a slightly different linear order, as follows: the quantity and nomenclature of replication (see recommendations 1, 2, 3, and 4); changes between initial and replication studies (see recommendations 4, 5, and 7); the warranting of what research gets replicated (see recommendations 6, 7, and 10); the extent of reproducibility and its relations with author overlap, materials transparency, and heterogeneity between replication and initial studies (see recommendations 7, 8, 9, 10, 11). In line with our aim to consider infrastructural challenges

to replication research, recommendations 11 – 16, along with recommendations 2 and 9, allude to infrastructural and cultural needs in publishing, funding, and training. All recommendations are united by the aims of increasing the quantity and improving the quality of replication research in the field of L2 and multilingualism research.

1) Increasing the amount and speed of replication.

Although we cannot determine an optimum rate of replication or an ideal balance between replication and innovation, our data certainly demonstrate an extremely low rate: replications have constituted approximately 1 out of 400 of articles in those journals that have published at least one self-labelled replication in L2 research, since the first published L2 replication in 1973. Critically, this rate would be *much* lower if it could be calculated using the whole, larger set of journals that ever publish L2 research, and from the start of their history. Makel and colleagues were able to calculate this broader denominator easily and objectively by using the set of journals delineated by the 'discipline categories' of Education and Psychology in the ISI Web of Knowledge Journal Citation Reports, whereas there is no such discipline specific list for 'L2 and multilingual journals'. Even more worryingly, despite our more generous calculation, the rate we found was much lower than that in psychology, the key parent discipline for L2 research that adopts quantitative hypothesis-driven approaches and a discipline that is itself concerned that its own replication rate is too low. Our data also demonstrate a slow *speed* of replication. The observed mean gap of 6.4 years is not likely to expedite the checking and refining of theories before implications for academic and practitioner communities take root. As argued by Makel and Plucker (2014), "science may be self-correcting, but the often glacial pace of that correction does not match the speed of dissemination when results enter the public consciousness" (p. 313). We are unequivocal in our first and overarching recommendation. ***Recommendation #1: Increase the amount of and rate at which replications are performed and published.***

We also emphasise that data is needed about the *causes* of low published replication rates, in order to inform our efforts, including those recommended in the following sections, in empirically-grounded ways. For example, the publication of replication studies that have null findings or that did not support the initial findings may be adversely affected by publication bias, and so be one cause of the overall low rate of published replications. Initiatives such as "Positively Negative" (PLOS, 2015), an open collection of studies with null or inconclusive findings, which includes studies entitled 'failure to replicate', may be useful and worth evaluating. There are many other potential causes of the low rate of replication research, such as low prestige and a related unwillingness to self-label as a replication. ***Recommendation #2: Systematic enquiry into the causes of low rates of published replication studies, including (more) empirical evidence about the extent and causes of publication bias in the field.***

2) The importance of nomenclature

The low rate of replication is likely in part due to a lack of willingness to self-label as 'replication' (Neulip & Crandall, 1993; Polio, 2012b). This reticence is complex. Anecdotally, we observed during colloquia discussing the current study and Morgan-Short et al. (2018) that some researchers reported actively undertaking and promoting replication, with students and in their own work, yet were less enthusiastic about labelling this as replication. Here we illustrate, with three relatively recent examples, what we think is fairly standard practice. This is, we stress, not to criticise these studies (and a good proportion of our own research certainly has aligned with this practice). We aim, rather, (a) to acknowledge that our synthesis is not a fully comprehensive reflection of the amount and nature of replication effort in the field, and (b) to recognise the complexities that our arguments and recommendations about nomenclature entail. First, Kim and Nam (2017) had closely related aims, used the same tests and similar analyses procedures as Ellis (2005). They did not self-

label as a replication (their title used the term 'revisited') and yet referred to three other studies that used the same materials in the same agenda as "replications" though none of those studies were self-labelled replications. Second, Trenkic, Mirkovic and Altmann (2014) did not self-label their study as a replication, but their aims, design, and stimuli were closely informed by Chambers, Tanenhaus, Eberhard, Filip and Carlson (2002)'s study (who were acknowledged for sharing stimuli) and they reported that their findings "replicated" the findings of Chambers et al. (2002). Third, several large, coordinated studies have used the same (or very similar) shared materials across different sites (e.g., Bergmann, Meulman, Stowe, Sprenger, & Schmid, 2015; Dimroth, Starren, Rast, & Watorek, 2013; Meulman, Wieling, Sprenger, Stowe, & Schmid, 2015; Schmid, 2011). None of these studies were yielded in our search, even though they are likely examples of partial or conceptual replications as they sought to make claims about replicating previous aims and findings and used or adapted materials from earlier studies. Thus, arguably, our estimation of the amount of published replication research in the field under-represents the total wider replication effort, when more broadly defined.

Frank debate is required about the advantages of three broad approaches to nomenclature: i) using a replication label in title or abstract, ii) maintaining a range of other terms for alluding to replication effort (e.g., 'exten* by McManus & Marsden, 2018, and Nakamura, 2012; or 'revisit*' by Au, 1983, and Kanno, 2000), iii) alluding to closely related theoretical and methodological precedents more covertly within study reports. Here, we present arguments that a reticence to label with the term replication is detrimental for the field. First, it hinders the general tracking of intellectual connections and hides theoretical and methodological precedents under an invisibility cloak (or a "cloaking device," Makel et al., 2012, p. 541). Second, without 'replication' labels, heterogeneity from one study to the next can pass largely unchecked. We found that despite many suggestions in the

limitations/further research sections of articles regarding necessary replications with different language combinations, participant demographics, or design features, very little such specific variation is undertaken systematically in self-labelled replications, and variation was often accompanied by other, potentially confounding, changes. In contrast, using the label 'replication' establishes a need, for both the researchers and reviewers, to monitor inter-study variation and identify precise relations with one or more specific study/studies. This in turn increases the field's ability to confirm and reject theories across studies. Finally, the lack of self-labelling adversely affects efforts to synthesize and meta-analyze research. Among other purposes of synthesis, better self-labelling would facilitate future efforts to examine reproducibility in the field of SLA to ascertain the reliability and generalisability of findings (as done by the large-scale replication efforts in psychology). In sum, we argue that explicit identification, via self-labelling with some replication nomenclature in titles or abstracts, clarifies the relations between studies and this would help the quality and scope of research agendas. For example, it would: render theoretical and methodological precedents more visible; facilitate reviewers' evaluation of the extent of changes to previous research procedures and materials; encourage more tightly-knit series of inter-connected studies by requiring researchers to explicitly operationalise and articulate changes to earlier research; improve the quality of syntheses and meta-analyses by, for example, facilitating the comparability of studies. ***Recommendation #3: More self-labelling with the term 'replication' wherever appropriate.***

In terms of sub-types of replication, we found a very wide range of labels and negligible relations between these labels and the amount or type of change between the initial and replication study. We thus propose a simple distinction based on the principle that 'direct replications' aim to test *data* and *analysis* (i.e., to confirm previous findings via a study with, as far as possible, the same conditions), whereas 'partial replications' test a *construct* by

manipulating one of the initial conditions or study characteristics to test generalisability to one new context/condition (see Lykken, 1968). Thus, 'direct replication' would describe a study in which there was no intention to change any variables deemed likely (according to current knowledge) to affect results. As minor deviations from the initial study can be unavoidable, especially with human participants, any such heterogeneity would be reported in full. 'Partial replication,' on the other hand, would describe a study that intentionally changes only one significant component of the initial study to check, *a priori,* for one well defined 'boundary condition' or 'moderator' of the initial findings. This could include a principled change in instrumentation, analysis, linguistic form, or a participant characteristic. In our study sample, 'partial replication' was the most frequently used sub-label, so although we are confident that the term is already in our nomenclature, we recommend that its usage/function become more consistent. 'Conceptual replications' introduce more than one significant change to the initial study and can extend agendas in multifaceted ways, but are in a weaker position to ascribe different findings to the adaptations made to the initial study. However, retaining this label is, we think, helpful for authors and reviewers seeking to identify the extent of relations between studies. ***Recommendation #4: A principled, standard nomenclature as follows: 'direct replications' make no intentional change to the initial study and seek to confirm methods, data, and analysis; 'partial replications' introduce one principled change to a key variable in the initial study to test generalisability in a clearly pre-defined way; 'conceptual replications' introduce more than one change to one or more significant variables. In all cases, potential heterogeneity and contextual details are documented as fully as possible.***

Ascertaining the extent and nature of change between initial and replication studies can be severely hampered by unclear reporting. For example, we found that some of the changes that replications introduced were not acknowledged by the authors. We recommend

that authors of replications clarify the relationship with the initial study (including descriptive statistics and effect sizes) and combine analyses where possible. However, even with better reporting of the methods, data, and findings of the initial studies, it is unlikely that the replication articles can do full justice to the initial report, especially given that some journals assign replication studies to a shorter article type. In view of this, reviewers need to be familiar with the initial study and read it alongside the replication to be able to corroborate the claimed relationships. This will have implications for authorship blinding practices, in cases where there is author overlap between the initial and replication studies.

Recommendation #5: Reviewers are asked to read the initial study that is being replicated.

3) Warranting what should be subject to a replication study

Various propositions exist to set benchmarks or define rationales for when a study merits replication, such as a citations metric or the co-occurrence of specific characteristics (e.g., "the troubling trio" of low *n*, large effect sizes, and unexpected findings [Lindsay, 2015]). We, however, do not propose a set of benchmarks, as these may become over-interpreted (as with the alpha level '0.05'; and small/medium/large effect sizes) and could exacerbate the image of replication as an unoriginal, mechanistic undertaking. Part of the skill in replication work is surely choosing studies worth replicating and justifying this to reviewers and editors. These justifications are likely to include citation counts, low sample size, surprising results, and theoretical, methodological or practical issues; but a rigid formula based on a fixed composite of these is likely, we think, to be cumbersome and unreliable. Thus, we suggest that there should be little or no top-down (e.g., journal or professional association) control, and researchers' agendas should drive what is replicated.

Recommendation #6: Warrants for replication studies are provided, and peer-reviewed, on a case by case basis, with rationales including, but not restricted to, one or more of the following characteristics of the initial study: surprising findings, one or more concerning

methodological features, high (potential) impact such as theoretical or practical significance.

A related phenomenon that may, however, require top-down influence, is the rate of publishing studies that replicate initial studies with null or borderline findings (as the current synthesis found a paucity of such replications, just 4 of 67). We do not suggest that this should be addressed by a blanket recommendation such as 'increase attempts to replicate initially null findings', as the phenomenon is tightly related to the low rate of publication of initial studies with null findings in the first place, which is in turn influenced by publication bias. However, given that replication can increase the interpretability and, therefore, value of initial null findings, we recommend that these issues are certainly worthy of empirical investigation (see Recommendation 2).

4) Collaborative ethic to sustain an independent replication effort

Several issues will determine the extent to which and speed with which we can adopt more collaborative approaches to facilitate replication.

*Transparency of materials and data***.** We found that changes to stimuli, instruments and measures (such as elicitation tests) were relatively frequent between an initial and a replication study. Although these changes were sometimes intentional as a motivation for the replication, often this was not the case. This is a key concern because measures often constitute the key dependent variables and changes to them reduces comparability with previous research (Marsden et al. 2016; Thomas, 1994 & 2006). For example, several meta-analyses have shown that effects of instruction vary as a function of measurement type (e.g., Norris & Ortega, 2000; Lee, Jang, & Plonsky, 2015). Another problem we found was that the extent of change could not be ascertained due to omissions in the initial study's report and lack of availability of materials and data. Methodological transparency can improve these problems, facilitating replication and improving its quality and reliability (Marsden et al.,

2016). Another motivation to make materials fully available is that our findings suggested the more that materials were available, the more likely a replication was to find support for the initial study. Transparency may also influence replication in other ways that require further investigation, such as on the *quantity* of replication. There is emerging evidence (Plonsky et al., in progress) that positive correlations exist between materials transparency (number of entries on IRIS) and citation counts. With high citation being one factor that can trigger a replication, it seems that materials transparency could be associated with increased replication research (though of course other factors may cause methodological transparency itself). ***Recommendation #7: Increased open availability of materials, including proficiency measures, for L2 research.***

In addition, sharing *data* is essential for cumulative analyses that join datasets and examine moderator effects of inter-study variation, especially important given the well documented lack of power in L2 research (e.g., Plonsky, 2013; 2015). We found only one bundle of self-labelled replications that conducted an internal meta-analysis, which was possible as they used the same materials and had the data from the initial study fully available (Ellis et al., 2014; and see Morgan-Short et al., 2018). (See also Lindsay, 2017). Making data available entails ethical (IRB) considerations early in the research process, and is not possible in all situations, but is increasingly a requirement of funders. ***Recommendation #8: More research is made 'fully transparent and open for replication' by making data available.***

Researchers, reviewers and editors all have the responsibility to improve our collaborative ethic. Trofimovich and Ellis (2015) adopted the Open Science Badges for *Language Learning*, and several other journals now also value open materials and data in this way (e.g., *Studies in Second Language Acquisition*, 2017*; The Modern Language Journal*, 2017). Kidwell et al. (2016) and Giofrè, Cumming, Fresc, Boedker and Tressoldi (2017) provide quantitative evidence for the effectiveness of this initiative for the transparency of

materials and data. Indeed, partly as a result of these initiatives and push from journal editors, IRIS now holds 24 sets of L2 data (in addition to approximately 3,600 files of materials and analysis protocols). Although we have high expectations that transparency via materials and data sharing will improve the quality and quantity of replication efforts, there is still much work to be done in these endeavours. For example, Marsden et al. (accepted) found only 4% of self-paced reading studies had openly available materials and 77% had only a brief example of stimuli available in their article. ***Recommendation #9: More journals give more and stronger incentives to their authors to make materials and data systematically openly available.***

5) Independence, combined with professional practice and collegiality

Authorship practices. Our observation that supportive findings from a replication study were significantly more likely when authorship overlapped between the initial and replication studies, compared to independent replication, aligned very well with those of Makel and colleagues from two related disciplines. We do not make conclusive claims about *why* author overlap tended to be linked to more supportive replications, as this could be accounted for by increased QRPs, and/or by reduced heterogeneity due to access and fidelity to materials and fewer researcher degrees of freedom. However, we argue that each of these explanations is concerning, as reduced heterogeneity should be possible without overlapping authorship so that reproducibility of findings is unrelated to author overlap. We recommend that replication carried out *independently* from the initial studies is desirable, to reduce any influence that author overlap may have on our insight into the reproducibility of L2 research findings. Thus, when materials and data for initial studies are available, author overlap would become a matter of collegiality rather than necessity.

However, independent replications can be perceived negatively as bullying, as discussed by Bohannon (2014). Inviting the initial author to review replication studies can

help reduce this, and, in the case of a Registered Report, they can be invited to provide a Stage 1 review before data collection (see Marsden, Morgan-Short, Trofimovich, & Ellis, 2018, and below). Even more transparent practices that may promote more and higher quality replication, reduce publication bias, and reduce perceptions of bullying, include: publishing 'open reviews' and authors' responses to reviews (Laws, 2016, *BMC Psychology*); giving initial authors an automatic right to a peer-reviewed published commentary (in our sample we found one such example, Kanno, 2000) (and see *Perspectives in Psychological Science*); and 'adversarial collaborations' (Coyne, 2016; Koole & Lakens, 2012; Kahneman, 2014; Mellers, Hertwig, & Kahneman, 2001), wherein researchers who account for phenomena differently agree to work together following a single protocol. We raise awareness of the existence of these more extreme measures, but hope that the other mechanisms that we recommend, such as transparent materials and data and the reviewing of methods prior to data collection, serve to reduce any perception of bullying that independent replication may engender.

Recommendation #10: When possible, replication studies are done by researchers independently of the initial study's authors, but the initial authors are invited to be involved at some stage of the review process, preferably prior to data collection (see Recommendation #12 about Registered Reports).

For multi-site replications, authorship practices may be required that are relatively rare to date in L2 research. The large multi-site efforts have, thus far, been in fields where large authorship teams are the norm. In line with these practices, Morgan-Short et al. (2018) offered authorship to those running studies, with lead authorship for those convening the multi-site replication, providing the materials, and formally documenting the results. Even with this co-authorship agreement, we were fortunate in securing collaborators, and a reciprocal ethic is needed to support such large-scale multi-site replication efforts (such as 'I collect data for others; others collect it for me'). Formal infrastructure is likely to help here,

such as the 'Call for Replication Collaborators' button on IRIS and the Centre for Open Science's 'Study Swap' (https://osf.io/view/StudySwap/) whereby researchers seek collaborators or offer participant availability. ***Recommendation #11: More multi-site collaborative replication efforts are desirable.***

6) Cultural and procedural changes in publishing

Various initiatives are available to increase the amount and quality of replication, the most obvious of which is perhaps journals' Author Guidelines explicitly encouraging replications. However, our data suggested that this alone was not a reliable or necessary mechanism: although the journal which had published the most replications had a statement inviting replications, the other three journals with such a statement actually published fewer replications than journals without such a statement. Indeed, journals that simply state they publish replications reach only Level 1 of the Transparency and Openness Promotion Guidelines (TOP) about replication (Nosek et al., 2015).

Another mechanism might be the idea of an "Accountable Replication Policy" (proposed by Chambers, 2016, due to launch at *Royal Society Open Science* in January 2018), whereby a journal would guarantee to publish replications of studies that they have published (unless there is a demonstrated significant methodological flaw with the initial study). This could incur a large commitment from journals, but if publishers are no longer bound by printed page limitations, such initiatives become more feasible (e.g., Wiley has removed page limitations for many of their journals). Another step is for more journals to explicitly comment on the acceptability of null findings, as one hindrance to replication is that not reproducing initial 'statistically significant' findings may leave authors vulnerable to negative reviews from the authors of the initial study or from general bias against null findings. One direct way of reducing such bias is via a 'results-free peer-review' (Button, Bal, Clark, & Shipley, 2016), where authors seek reviews on the basis of rationale, methods and planned

analyses alone and, once approved, the full manuscript with results is submitted for a second stage review (e.g., *BMC Psychology*). Although mitigating against bias at review, such a mechanism cannot reduce problems earlier in the research process, as the data are already known to the researcher and so QRPs (e.g., as Hypothesising After Results are Known, *p*-hacking) could still have happened, prior to the results-free review. Thus, journals that encourage submission of replication studies and carry out a results-free review attain only Level 2 of the TOP guidelines on replication.

A mechanism that aims to address these problems, as well as increase the amount and quality of replication, is the article type referred to as 'Registered Reports' (RRs) (see Marsden et al., 2018). RRs were pioneered by the journal *Cortex* in 2013 and have been adopted by over 52 journals (https://cos.io/rr/) at the time of writing. For RRs, a manuscript receives an initial (stage 1) review of the study purpose, aims, materials, data collection and analysis protocols. Crucially, the stage 1 review occurs before the data is collected. If approved, the materials and procedures are time-stamped as a pre-registration and given formal 'In Principle Acceptance' (IPA) by the editor (Nosek & Lakens, 2014), and data collection can then begin. Later, data collection, analysis, and report writing proceed and are submitted for a stage 2 review. At this stage, as the design and methods were approved beforehand, studies cannot be 'reviewed out' due to assertions relating to methodological flaws. Thus, IPA incentivises researchers to undertake a replication by reassuring them with a pledge of publication *prior* to investing in the data collection. It is unsurprising, therefore, that to date RRs include a high proportion of replication studies (OSF 2017), relative to the proportions found in standard publication routes as observed in the current and previous studies. Indeed, journals that offer Registered Reports as a route to publishing replication research meet the highest level (Level 3) of the TOP guidelines on replication. Thus, RRs have the potential to address many of the observations in our synthesis, such as: (a) few

replications of studies with 'null findings,' (b) the low rate of publication of replications overall, (c) the lack of direct replications, (d) extensive and unacknowledged heterogeneity between initial and replication studies, and (e) potential associations between supportiveness of replications' findings and author overlap or materials availability. RRs also carry other benefits such as: peer review informs the study at the design stage (rather than when it can be too late); they reduce QRPs; they accommodate any methods where data collection, coding, and analyses can be pre-determined (including, for example, observations and interviews). RRs do not preclude additional exploratory data collection or analysis as authors can report these in addition to the registered protocol and analysis, although such exploratory endeavours would be subject to review at stage 2. Importantly, there is also potential to adapt the procedure to fit uniquely to exploratory designs and associated epistemologies in 'Exploratory Reports' (McIntosh, 2017). ***Recommendation #12: Journal editorial boards consider accepting Registered Report article types, and where this is not possible, they consider undertaking results-free reviews.***

Another barrier to replication is that it is difficult to include both a replication and an extension study within one published article given normal space limitations, yet this study structure may alleviate stigma attached to doing replications. We found few examples of such article types (e.g., Barcroft & Sommers 2005; Marsden, Liu, & Williams, 2013; neither included in our current synthesis as we investigated replications of studies in different publications). Current limitations on article length are probably one reason why this was rare, but we are hopeful that this situation will change as publishers remove formal word limits as publication moves online (though to the best of our knowledge, only Wiley have yet done this). There are at least two models: one is a study that begins with a direct replication of a study published previously, followed by a partial or conceptual replication (to ascertain generalisability, 'boundary conditions' etc.); another is an initial study followed by a

confirmatory direct replication, to test the robustness of the original data and methods. We recommend both of these routes. ***Recommendation #13: Publishers lift word limits or provide online capacity to encourage more replication work within individual study reports.***

Our data demonstrate that the perceived low prestige of replication research is unfounded in, at least, two respects: perceived ease and perceived low impact. Carrying out well-justified, carefully administered replications, which are rigorously analysed in relation to their initial study, is no trivial task and very rare in self-labelled replications to date. Our data also showed that replications have been relatively highly cited and have been published in some of the highest impact journals. Further, the three journals that we found to have published the highest number of replications were found by Plonsky et al. (in progress) to have the highest perceived prestige. As a community, we can further enhance the impact and prestige of replications by co-citing them along with their initial studies (see Koole & Lakens's, 2012 proposals for incentivising replication). Beyond enhancing the impact of replications, such a practice would reflect a valid and comprehensive reporting of the state of the literature, as readers would know the extent to which the results of the initial study are reliable or generalizable. ***Recommendation #14: When the initial study is cited, citation should also be made to (at least any direct and partial) replication studies of it.***

7) Wider cultural changes in academia

Changing the incentives in our wider academic culture is even more challenging than changing editorial, review, and citation practices discussed above. Of course, a driving force to shape behaviour is funding (as noted by Baker, 2015). Although we found that replication studies to date have not uniquely been of 'cheap and easy' studies (as a reasonable proportion had relatively 'costly' characteristics such as oral measures, classroom environments, and longitudinal and intervention designs), we found few replications using expensive equipment,

corroborating Laws's (2016) concerns. Combined with the low rates of replication research overall, this indicates that funding mechanisms are indeed critical for improving replication effort. However, incentivisation in academia tends to be entrenched in rewarding originality (Chambers, 2017). For example, approximately 60% of UK universities' centrally distributed funding is allocated on the basis of the three criteria of 'originality,' 'rigour,' and 'significance' of research (Research Excellence Framework, 2011, [REF]). Although replication studies could score highly on rigour and significance, as these are arguably inherent in good replication work, they are likely to score lower on originality. Nevertheless, changes to new REF criteria (REF, 2017) (such as a reduced *number* of outputs and reward for open science practices) could incentivise large multi-site pre-registered replication projects. We further note five recent funding initiatives that should help replication effort. Two directly promote replication research: the 'IRIS Replication Award,' for published replications that used materials from IRIS; and The Netherlands' Organisation for Scientific Research scheme dedicated to funding replication studies (NWO, 2017). The other three indirectly promote replication efforts by adopting an RR approach to review: *Language Learning*'s Early Career grant scheme, which prioritises one award for a RR (Marsden et al., 2018); and funder collaborations with journals that integrate the RR model of peer review into the grant funding process (*The Children's Tumour Foundation* with *PLOS ONE* (PLOS ONE, 2017), and the charity funder *Cancer Research UK* with the journal *Nicotine and Tobacco Research* (Munafò, 2017). ***Recommendation #15: We recommend more funding, from institutional through to international levels, to promote replication as an integral part of the research process.***

Professional associations could also incorporate replication strands into their conference programmes and endorse replication as a valued part of tenure applications (see, for example, the *American Educational Research Association*'s Standards (2006) and Code

of Ethics (2011)). The *American Association of Applied Linguistics* recently amended their guidelines to recommend that "high quality replication studies, which are critical in many domains of scientific inquiry within applied linguistics, be valued on par with non-replication-oriented studies" (AAAL, 2017). Engaging students with conducting replication studies has also been discussed (see Frank & Saxe, 2012; Porte 2012) and we are aware of several graduate programmes where replication is an integral part of training and assessment.

Recommendation #16: Efforts should be made (e.g., via teaching and training infrastructures, institutional recognition, national funding mechanisms, and professional association conferences and promotion guidance) to reward those who include replication research in their work.

Concluding Remarks

We conclude by considering a few key implications for the future meta-science and production of replication research. This includes acknowledgements of some of the limitations of our study and arguments.

First, somewhat inevitably, we hope this study will stimulate replications and extensions of the systematic review itself. Also, when more direct replications are available, future syntheses will be able to investigate the extent of reproducibility in the field quantitatively. That is, rather than using author interpretations and subjective ratings as was appropriate and necessary in the current study, meta-analytic techniques would be appropriate for examining reproducibility in direct replications (where high reproducibility is clearly expected) and assessing the effects of any operational heterogeneity (recalling that *intentional* heterogeneity is sometimes designed and predicted to yield non-reproduced findings).

Second, we do not suggest that increased replication *alone* will improve the reliability and validity of *all* L2 research. To some extent, we agree with Schmidt and Oh (2016)'s

argument that rather than increasing replication, other issues such as publication bias and QRPs need to be tackled first and then meta-analyses could address the lack of direct replication (see Coyne, 2016 for related arguments; and Schimmack, 2016, on the value of replicability indices to detect likely publication bias in lieu of actual replication studies for investigating reproducibility). Whilst we agree that these other issues require attention, we argue that meta-analysis could not address the lack of replication. As a retrospective mechanism, meta-analysis cannot address some problems that can be addressed by replication, such as lack of parity between studies, which reduces the critical mass of adequately powered comparable studies that answer sufficiently similar questions to be included in any meta-analysis (Laws, 2016).

Third, understanding the causes of low levels of published self-labelled replication requires data about the experiences and opinions of editors, reviewers and researchers, using questionnaires and interviews. This would reveal the extent to which the observed lack of replication originates in low levels of execution, self-labelling, article submission, and/or actual publication. That is, we do not have a good understanding of the extent to which replications are in fact submitted to journals but rejected, and, if so, why.

Finally, our key finding is perhaps the very low number (67) of self-labelled replication studies, especially striking when set against the 50 calls and commentaries on replication in the same field. All four anonymous reviewers requested we express in stronger terms the perturbing amount and quality of self-labelled replication research. Using the words of one such reviewer, we could sum up our data as "providing an unequivocal view of the state of replication research in the field: It is disparate, loose, rare, flawed, inconsistent, and opaque. If a foundation of high quality replication studies is a pre-requisite for a healthy discipline, the field of second language research occupies very hazardous terrain." We have identified many factors that must work together to change production of and attitudes towards

replications, including increased transparency of materials and data, multi-site collaboration, more consistent self-labelling of replications, fewer and more transparent alterations of features from one study to the next, and increased publication via article types such as Registered Reports. Recommending that these and other practices are incorporated more systematically into our communities is intended to precipitate us towards a more mature field, whose terrain embraces replication research that is more convergent, tighter, more frequent, less flawed, more consistent, and more transparent.

Notes

[1] Where a study replicated more than one initial study (k=7), we coded according to the replicators' aims and analyses: if the two initial studies were replicated separately (as each initial study had different aims and designs and analyses in the replication were presented separately), then these were coded as unique initial-replication pairs (Chen 2011; Cobb, 2003; Robinson, 2005); if, on the other hand, two initial studies were replicated because the initial studies had very similar designs and aims and the replication was presented as if replicating one 'collapsed' study, then this was coded as a single initial-replication pair (e.g., DeKeyser & Sokalski, 1996; Ellis et al., 2014; Liu, 1985; Walters, 2012). Five studies included in Polio's (2012b) review were not included in our study because they did not self-label clearly as a replication study in the title or abstract or because they were not a replication of a study reported in a separate publication.

[2] This is generous for two main reasons. First, the calculation is based only on journals that *have* published self-labelled replications, rather than all journals that have ever published L2 research. The latter would be very difficult to estimate, and probably provide an unfair representation of replication rate, as it would include an extremely wide range of journals across multiple disciplines (the field of L2 research does not have an SSCI discipline-specific list such as the ones used by psychology or education by Makel and colleagues). Second, our start date is from the earliest replication published rather than the start date of each journal (e.g., the *MLJ* began publishing in 1916).

[3] "We do not discourage contributions that present null results" (*SLR*) and "Lack of statistically significant results, or difficulty in drawing clear conclusions, will not necessarily rule out publication of interesting contributions" (*Language Testing*).

[4] Google Scholar includes citations in many types of publications, including books (unlike the Web of Science used by Makel & Plucker (2014)).

[5] Replications were published a mean of 13.1 (mode 11) years ago, and the initial studies a mean of 20.5 (mode 21) years ago.

[6] Nor did this seem to be affected by whether the replication's findings tended to support the initial study's findings or not (see RQ4 for more details): mean citations 'not/partially not supportive' = 6.35 (SD 8.609, k=19), 'partially/very supportive' = 7.62 (SD 5.742, k=46) (Mann-Whitney U 330.5, Z -1.536, p=.124) (d = 0.190, CI -0.3456 - 0.7254).

[7] With 'intervention' defined for coding purposes as 'an experimental manipulation to cause learning, beyond normal practice.'

[8] We coded our sample studies for their research areas, and the data are available in Supplementary Materials 2 and at www.iris-database.org. We found a very wide range of subdomains of research and the coding was subjective involving multi-layered coding categories. We could not discern any patterns in terms of particular areas that had more or fewer replication studies.

[9] Studies that did not use language learners collected data from, for example, teachers, corpora or textbooks

[10] Where age ranges were given, the median was used.

[11] This was calculated by dividing the study sample size by the number of groups (or conditions) in each study. The calculation excluded two pairs of studies that gathered large scale data from formal tests – these replications increased the *n* of the initial studies by 44,612 and 1,415.

[12] We were unable to locate the mean, for a direct comparison, though a personal communication indicated this was 35 (SD 64, CI 30-40).

[13] We acknowledge that these ratings are subjective, but the technique is very similar to, though slightly more fine-grained than, Makel et al. (2012) and Makel & Plucker's (2014) three-point scale: "success, failure, mixed." This approach was fit for our purpose, as unlike the recent large-scale replication efforts in psychology that set out to statistically assess reproducibility across multiple studies by conducting a *direct* replication of each of the initial studies, we aimed to provide a review of replications - of all kinds - that had already been conducted.

[14] Other studies provided (partial) eta squared on omnibus tests, or were coded other/unclear/not applicable. Ellis et al. (2014) used regression coefficient '*beta,'* another standardized effect size of magnitude.

[15] In cases where there was no authorship overlap, several replications ($k = 14$) included the initial authors in the acknowledgements (which sometimes indicates academic lineage/collaboration). Combining overlap in authorship with a mention in the acknowledgements yielded almost equal numbers of studies with authorship commonalities ($k = 33$) and those with none ($k = 34$).

[16] In terms of association between supportiveness and replications being in the same journal as the initial study, our data did not suggest a strong trend (Supplementary Materials 3: Table 7), broadly in line with Makel and Plucker (2014). However, a dichotomous coding of supportiveness (same journal: 24% not supportive *versus* 76% supportive; different journal: 32% not supportive *versus* 64% supportive) suggests this may be worth pursuing once the field has a larger body of more direct replications.

[17] Likelihood ratio for small samples: 11.052, 2, p=.004; Fisher's Exact Test because one cell (16.7%) had cell count < 5 (10.569, p=.005). Five studies were excluded because cell counts were too small: only two studies provided all the instruments used to collect all data used in the analysis (one supportive and one not); one study provided *all* instruments (supportive); two studies could not be coded as to whether they were supportive or not.

References

American Association for Applied Linguistics, (2017). *Promotion and tenure guidelines*. Retrieved from: http://www.aaal.org/?page=PT

Au, T. K. (1983). Chinese and English counterfactuals: The Sapir-Whorf hypothesis revisited. *Cognition, 15*(1), 155-187. https://doi.org/10.1016/0010-0277(83)90038-0

Au, T. K. (1984). Counterfactuals: In reply to Alfred Bloom. *Cognition, 17*(3), 289-302. https://doi.org/10.1016/0010-0277(84)90012-X

Bakan, D. (1967). *On method*. San Francisco: Jossey-Bass. https://doi.org/10.1177/001316446802800431

Baker, M. (2015). Over half of psychology studies fail reproducibility test. *Nature News and Comment.* https://doi.org/10.1038/nature.2015.18248.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543-554. https://doi.org/10.1177/1745691612459060

Barcroft, J. & Sommers, M. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27,* 387–414. https://doi.org/10.10170/S0272263105050175

Basturkmen, H. (2014). Replication research in comparative genre analysis in English for Academic Purposes. *Language Teaching, 47*(3), 377-386. https://doi.org/10.1017/S0261444814000081

Bergmann, C., N. Meulman, L. A. Stowe, S. A. Sprenger and M. S. Schmid. (2015). Prolonged L2 immersion engenders little change in morphosyntactic processing of bilingual natives. *Neuroreport, 26*, 1065-1070. https://doi.org/10.1097/WNR.0000000000000469

Bohannon, J. (2014). Replication effort provokes praise and 'bullying' charges. *Science, 344*(6186), 788-789. https://doi.org/10.1126/science.344.6186.788

Branco, A., Cohen, K.-B., Vossen, P., Ide, N., Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: introducing an LRE special section. *Language Resources and Evaluation 51*(1), 1–5. http://dx.doi.org/10.1007/s10579-017-9386-7

Button, K. S., Bal, L., Clark, A., & Shipley, T. (2016). Preventing the ends from justifying the means: Withholding results to address publication bias in peer-review [Editorial]. *BMC Psychology, 4*(59). https://doi.org/10.1186/s40359-016-0167-7

Chambers, C. (2016, November 9). *Accountable replication policy at Royal Society Open Science* [Blog post]. Retrieved from: http://neurochambers.blogspot.co.uk/2016/11/an-accountable-replication-policy-at.html

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice.* Princeton, New Jersey: Princeton University Press. http://dx.doi.org/10.1177/1745691613514450

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of memory and language, 47*(1), 30-49. https://doi.org/10.1006/jmla.2001.2832

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review, 59*(3), 393-424. https://doi.org/10.3138/cmlr.59.3.393

Collins, H. M. (1985). The possibilities of science policy. *Social Studies of Science*, *15*(3), 554-558. Retrieved from: http://www.jstor.org/stable/285370

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology, 4*, 28. https://doi.org/10.1186/s40359-016-0134-3

DeKeyser, R. M., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning, 46*(4), 613-642. https://doi.org/10.1111/j.1467-1770.1996.tb01354.x

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly, 50,* 132-153. https://doi.org/10.1002/tesq.217

Devlin, H. (2016, September 21). Cut-throat academia leads to 'natural selection of bad science', claims study. *Science Guardian.* Retrieved from: https://www.theguardian.com/science/2016/sep/21/cut-throat-academia-leads-to-natural-selection-of-bad-science-claims-study

Dimroth, C., Rast, R., Starren, M., & Watore, M. (2013). Methods for studying a new language under controlled input conditions: The VILLA project. *Eurosla Yearbook, 13*, 109-138. https://doi.org/10.1075/eurosla.13.07dim

Earp, B. D. (2016). What did the OSC replication initiative reveal about the crisis in psychology? An open review of the draft paper entitled "Replication initiatives will not salvage the trustworthiness of psychology" by James C. Coyne. *BMC Psychology, 4*(28), 1-19. Retrieved from: https://www.researchgate.net/publication/293651901_What_did_the_OSC_replication_initiative_reveal_about_the_crisis_in_psychology

Earp, B. D., Everett, J. A., Madva, E. N., & Hamlin, J. K. (2014). Out, damned spot: can the "Macbeth Effect" be replicated?. *Basic and Applied Social Psychology, 36(1),* 91-98. https://doi.org/10.1080/01973533.2013.856792

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6*, 1-11. https://doi.org/10.3389/fpsyg.2015.00621

Editorial (2006). Let's replicate. *Nature*, *442*(330). https://doi.org/10.1038/442330b

Ellis, N. C., & Sagarra, N. (2010). The bounds of adult language acquisition. *Studies in Second Language Acquisition*, *32*(4), 553-580. https://doi.org/10.1017/S0272263110000264

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition, 33*(4), 589-624. https://doi.org/10.1017/S0272263111000325

Ellis, N. C., Hafeez, K., Martin, K. I., Chen, L., Boland, J., & Sagarra, N. (2014). An eye-tracking study of learned attention in second language acquisition. *Applied Psycholinguistics*, 35(03), 547-579. https://doi.org/10.1017/S0142716412000501

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*, 141–172. https://doi.org/10.1017/S0272263105050096

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication research's disturbing trend. *Journal of Business Research*, *60*(4), 411-415. https://doi.org/10.1016/j.jbusres.2006.12.003

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891-904. https://doi.org/10.1007/s11192-011-0494-7

Faretta–Stutenberg, M., & Morgan–Short, K. (2011). Learning without awareness reconsidered: A replication of Williams (2005). In G. Granena et al. (Eds.), S*elected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions* (pp. 18–28). Somerville, MA: Cascadilla Proceedings Project.

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLoS ONE, 10*(2), e0118053. https://doi.org/10.1371/journal.pone.0118053

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing 27*(4), 537-554. https://doi.org/10.1177/0265532209354771

Fitzpatrick, T., & Meara, P. (2004). Exploring the validity of a test of productive vocabulary. VIAL, *Vigo International Journal of Applied Linguistics*, *1*, 55-74. Retrieved from: webs.uvigo.es/ vialjournal

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19,* 975-991. https://doi.org/10.3758/s13423-012-0322-y

Frank, M. C. & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science, 7*(6)*,* 595–599. https://doi.org/10.1177/1745691612460686

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & Tressoldi, P. (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PLoS ONE, 12*(4), e0175583. https://doi.org/10.1371/journal.pone.0175583

Han, C. (2016). Reporting practices of rater reliability in interpreting research: A mixed-methods review of 14 journals (2004-2014). *Journal of Research Design and Statistics in Linguistics and Communication Science, 3*(1), 49-75. https://doi.org/10.1558/jrds.29622

Hartshorne, J., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Front Comput Neurosci*, *6*(1), 1-14. https://doi.org/10.3389/fncom.2012.00008

Hubbard, R., & Armstrong, J. S. (1994). Replications and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing*, *11*(3), 233-248. https://doi.org/10.1016/0167-8116(94)90003-5

Ioannidis, J. (2005) Why most published research findings are false. *PLoS Med, 2*(8). https://doi.org/10.1371/journal.pmed.0020124

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again…. *Science*, *334*(6060), 1225-1225. https://doi.org/10.1126/science.334.6060.1225

Kahneman, D. (2014). A new etiquette for replication. *Social Psychology, 45*(4), 310.

Kanno, K. (2000). Case and the ECP revisited: Reply to Kellerman and Yoshioka (1999). *Second Language Research, 16*(3), 267-80. https://doi.org/10.1191/026765800672956803

Kelly, C. W., Chase, L. J., & Tucker, R. K. (1979). Replication in experimental communication research: An analysis. *Human Communication Research*, *5*(4), 338-342. https://doi.org/10.1111/j.1468-2958.1979.tb00646.x

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., ... & Errington, T. M. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biol, 14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Kim, J., & Nam, H. (2017). Measures of implicit knowledge revisited: Processing modes, time pressure, and modality. *Studies in Second Language Acquisition 39*(3), 431-457. doi:10.1017/S0272263115000510

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142-152. https://doi.org/10.1027/1864-9335/a000178

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. Perspectives in Psychological Science, *7*(6), 608-14. https://doi.org/ 10.1177/1745691612462586

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*, Supp. 1, 127-159. https://doi.org/ 10.1111/lang.12115

Larson-Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics. *The Modern Language Journal, 101*(1), 244-270. https://doi.org/ 10.1111/modl.12386

Laws, K. R. (2016). Psychology, replication and beyond [Editorial]. *BMC Psychology, 4*(1), 30. https://doi.org/ 10.1186/s40359-016-0135-2

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*(3), 345–366. https://doi.org/10.1093/applin/amu040

Lee, S-K and Huang, H-T (2008). Visual input enhancement and grammar learning. *Studies in Second Language Acquisition, 30,* 307–331. https://doi.org/10.10170/S0272263108080479

Lindsay, S. (2015). Replication in psychological science. *Psychological Science, 26*(12), 1827-1832. https://doi.org/10.1177/0956797615616374

Lindsay, S. (2017). Sharing data and materials in psychological science [Editorial]. *Psychological Science,* 1-4. https://doi.org/10.1177/0956797617704015

Lindstromberg, S. (2016). Inferential statistics in Language Teaching Research: A review and ways forward. *Language Teaching Research*, *20*(6), 741-768. https://doi.org/10.1177/1362168816649979

Liu, L. (1985). Reasoning counterfactually in Chinese: Are there any obstacles? *Cognition, 21*(3), 239-270. https://doi.org/10.1016/0010-0277(85)90026-5

Luijendijk, H., and Koolman, X. (2012). The incentive to publish negative studies: how beta-blockers and depression got stuck in the publication cycle. *Journal of Clinical Epidemiology, 65*(5), 488-492. https://doi.org/10.1016/j.jclinepi.2011.06.022

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological bulletin, 70*(3), 151-159. Retrieved from: www.psy.cmu.edu/~siegler/lykken68.pdf

MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning, 67*(S1), 254-275. https://doi.org/10.1111/lang.12220

Makel, M., & Plucker, J. (2014). Facts are more important than novelty: Replication in the Education sciences. *Educational Researcher, 43*(6), 304 -316. https://doi.org/10.3102/0013189X14545513

Makel, M., Plucker, J., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science, 7*(6), 537-542. https://doi.org/10.1177/1745691612460688

Markee, N. (2017). Are replication studies possible in qualitative second/foreign language classroom research? A call for comparative re-production research. *Language Teaching, 50*(3), 367-383. https://doi.org/10.1017/S0261444815000099

Marsden, E. (in press). Open science and methodological transparency in applied linguistics research. In C. Chapelle (Ed.), *Encyclopaedia of Applied Linguistics*. Oxford: Oxford: Blackwell Publishing.

Marsden, E. J., & Mackey, A. (2014). IRIS: A new resource for second language research. Linguistic Approaches to Bilingualism, 4(1), 125-130. https://doi.org/10.1075/lab.4.1.05mar

Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS repository: Advancing research practice and methodology. In Mackey, A. & Marsden, E. (Eds.) *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1-21). New York: Routledge. https://doi.org/10.4324/9780203489666

Marsden, E., Morgan-Short, K., Trofimovich, P., & Ellis, N. (2018). Editorial: Registered reports, replication, and open science. *Language Learning,* 68, 1.

Marsden, E., Thompson, S., & Plonsky, L. (accepted). A methodological synthesis of self-paced reading tests in second language research. *Applied Psycholinguistics*.

Marsden, E., Williams, J., & Liu, X. (2013). Learning novel morphology: The role of meaning and orientation of attention at initial exposure. *Studies in Second Language Acquisition, 35*(4), 619-654. https://doi.org/10.1017/S0272263113000296

Marsman, M., Schönbrodt, F., Morey, R., Yao, Y., Gelman, A., & Wagenmakers, E. (2017). A Bayesian bird's eye view of 'Replications of important results in social psychology'. *Royal Society Open Science* 4: 160426. http://dx.doi.org/10.1098/rsos.160426

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Front. Psychol, 8,* 523. https://doi.org/10.3389/fpsyg.2017.00523

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487-498. https://doi.org/10.1037/a0039400

McIntosh, R. (2017) Exploratory reports: A new article type for Cortex. *Cortex, 97.* https://doi.org/10.1016/j.cortex.2017.07.014

McManus, K. & Marsden, E. (2018). Online and offline effects of L1 practice in L2 grammar learning: a partial replication. *Studies in Second Language Acquisition.* https://doi.org/10.1017/S0272263117000171

Mellers, B.A., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269-275. https://doi.org/10.1111/1467-9280.00350

Meulman, N., Wieling, M., Sprenger, S.A., Stowe, L., & Schmid, M. (2015). Age effects in L2 grammar processing as revealed by ERPs and how (not) to study them. *PLoS ONE 10*(12), https://doi.org/10.1371/journal.pone.0143328

Morgan-Short, K., Heil, J., Botero-Moriaty, A., & Ebert, S. (2012). Allocation of attention to second language form and meaning: Revisiting the use of think aloud protocols. *Studies in Second Language Acquisition, 34*(4)*,* 659-685. https://doi.org/10.1017/S027226311200037X

Morgan-Short, K. Marsden, E., Heil, J., with Issa. B., Leow, R., Mikhaylova, A., Mikołajczak, S., Moreno, N., Slabakova, R., & Szudarski, P. (2018). Effects of attending to form whilst comprehending: A multi-site replication study. *Language Learning, 68*(1).

Munafò, M. (2017). Improving the efficiency of grant and journal peer review: Registered Reports funding. *Nicotine & Tobacco Research, 19*(7), 773. https://doi.org/10.1093/ntr/ntx081

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(21). https://doi.org/10.1038/s41562-016-0021

Nakamura, D. (2012). Input skewedness, consistency, and order of frequent verbs in frequency-driven second language construction learning: A replication and extension of Casenhiser and Goldberg (2005) to adult second language acquisition. *IRAL: International Review of Applied Linguistics in Language Teaching, 50*, 1-37. https://doi.org/10.1515/iral-2012-0001

National Academies of Sciences, Engineering, and Medicine. (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop.* Washington, DC: The National Academies Press. https://doi.org/10.17226/21915.

Neulip, J. W., & Crandall, R. (1993). Everyone was wrong: There are lots of replications out there. *Journal of Social Behavior and Personality, 8*, 1-8.

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, *65*, 470-476. https://doi.org/10.1111/lang.12104

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

Nosek, B., & Lakens, D. (2013). Call for proposals special issue of social psychology on "Replications of important results in social psychology." *Social Psychology, 44*(1), 59-60. https://doi.org/10.1027/1864-9335/a000143

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 43,* 137-141. https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, D., Breckler, S. J.,…& Yarkoni, T. (2015). *TOP Guidelines*. Retrieved from: https://cos.io/top/

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

NSF. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science. Report of the subcommittee on replicability in science. Advisory committee to The National Science Foundation directorate for social, behavioral, and economic sciences.* Retrieved from the National Science Foundation Web site: www .nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_ Research_Report.pdf

NWO (2017). *Replication studies*. Retrieved from: https://www.nwo.nl/en/research-and-results/programmes/replication+studies

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Oswald, F., & 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

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science, 11*(4), 539-544. https://doi.org/10.1177/1745691616646366

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning 61*(4), 993–1038. https://doi.org/10.1111/j.1467-9922.2011.00663.x

Plonsky, L. (2012). Replication, meta-analysis, and generalizability. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 116-132). New York: Cambridge University Press.

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. https://doi.org/10.1017/S0272263113000399

Plonsky, L. (2015). Statistical power, p values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23-45). New York: Routledge.

Plonsky, L., & Derrick, D. J. (2016). A Meta- Analysis of Reliability Coefficients in Second Language Research. *The Modern Language Journal*, *100*(2), 538-553. https://doi.org/10.1111/modl.12335

Plonsky, L., Egbert, J., & LaFlair, G. T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, *36*, 591-610. https://doi.org/10.1093/applin/amu001

Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*(2), 267-278. https://doi.org/10.1177/0267658314536436

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

Plonsky, L., Blair, R., Boyce, K., Kim, A., Li, F., Qi, D., … & Zhuang, J. (in progress). *Quality, prestige, and impact in L2 research journals*. Manuscript in preparation.

PLoS (2015). Positively negative: A new PLoS One collection focusing on negative, null and inconclusive results. [Web log article]. *The Missing Pieces.* Retrieved from: http://blogs.plos.org/collections/collections/the-missing-pieces/

PLoS ONE (2017). The Children's Tumour Foundation and PLoS ONE announce a new funder-publisher partnership. [Press release]. Retrieved from: http://www.ctf.org/news/ctf-plos-one-funder-publisher-partnership

Polio, C. (2012a). No paradigm wars please! *Journal of Second Language Writing*, *21*(3), 294-295. https://doi.org/10.1016/j.jslw.2012.05.008

Polio, C. (2012b). Replication in published applied linguistics research: A historical perspective. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 47 – 91). New York, NY: Cambridge University Press.

Polio, C., & Gass, S. (1997). Replication and reporting: A commentary. *Studies in Second Language Acquisition, 19*(4), 499-508.

Porte, G. (2012). *Replication research in applied linguistics*. New York, NY: Cambridge University Press.

Porte, G., & Richards, K. (2012). Focus article: Replication in second language writing research. *Journal of Second Language Writing, 21*(3), 284-293. https://doi.org/10.1016/j.jslw.2012.05.002

Research Excellence Framework. (2011). *Assessment framework and guidance on submissions.* Retrieved from: http://www.ref.ac.uk/pubs/2011-02/

Research Excellence Framework. (2017). *Initial decisions on REF 2021.* Retrieved from: http://www.hefce.ac.uk/pubs/year/2017/CL,332017/

Rohrer, D., Pashler, H., & Harris, C. (2015). Do subtle reminders of money change people's political views?. *Journal of Experimental Psychology: General, 144*(4), e73- e85. https://doi.org/10.1037/xge0000058

Rosenthal, R. (1979). The 'File Drawer' Problem and Tolerance for Null Results. *Psychological Bulletin 86*(3), 638-641. https://doi.org/10.1037/0033-2909.86.3.638

Santos, T. (1989). Replication in applied linguistics research. *TESOL Quarterly*, *23*(4), 699-702. https://doi.org/10.2307/3587548

Schmid, M. S. (2011). *Language attrition*. Cambridge: Cambridge University Press.

Schmid, M., Berends, S.M., Bergmann, C., Brouwer, S. M., Meulman, N., Seton, B. J., Sprenger, S. A., Stowe, L. (2015) *Designing Research on Bilingual Development: Behavioral and Neurolinguistic Experiments*. London: Springer.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol, 13*, 90-100. https://doi.org/10.1037/a0015108

Schmidt, F. &, Oh, I. S. (2016) The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology, 4,* 32-37. https://doi.org/10.1037/arc0000029

Schimmack, U. (2016, January 31). *The replicability-index: Quantifying statistical research integrity* [Blog post]. Retrieved from https://replicationindex.wordpress.com/2016/01/31/a-revised-introduction-to-the-r-index/

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., & Srinivasan, M. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66*, 55-67. https://doi.org/10.1016/j.jesp.2015.10.001

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U. (2016). Each reader decides if a replication counts: Reply to Schwarz and Clore (2016). *Psychological Science, 27*(10), 1410-1412. https://doi.org/10.1177/0956797616665220

Smith, B., & Lafford, B. A. (2009). The evaluation of scholarly activity in computer-assisted language learning. *The Modern Language Journal*, *93*, 868 - 883. https://doi.org/10.1111/j.1540-4781.2009.00978.x

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice-versa. [Editorial]. *American Statistician, 49*(1), 108-112. https://doi.org/10.1080/00031305.1995.10476125

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59-71. https://doi.org/10.1177/1745691613514450

Studies in Second Language Acquisition (2017). *Instructions for Contributors*. Retrieved from: https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/information/instructions-contributors

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 435-452). New York: Russell Sage Foundation.

The Modern Language Journal (2017). *Author guidelines for contributors*. Retrieved from: http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1540-4781/homepage/ForAuthors.html.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning, 44,* 307-336. https://doi.org/10.1111/j.1467-1770.1994.tb01104.x

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. M. Norris & L. Ortega (Eds.), *Synthesizing Research on Language Learning and Teaching* (pp. 279-301), Amsterdam, NE: John Benjamins Publishing Company.

Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology. *Theory & Psychology, 26*(4), 540-548. https://doi.org/10.1177/0959354316637136

Trafimow, D., & Earp, B. D. (2017). Null hypothesis significance testing and Type 1 error: The domain problem. *New Ideas in Psychology, 45*(1), 19-27. https://doi.org/10.1016/j.newideapsych.2017.01.002

Trenkic, D., Mirkovic, J., & Altmann, G. (2014). Real-time grammar processing by native and non-native speakers: constructions unique to the second language. *Bilingualism: Language and Cognition, 17*(2), 237-257. https://doi.org/10.1017/S1366728913000321

Trofimovich, P., & Ellis, N. (2015). Open Science Badges [Editorial]. *Language Learning, 65*(3), v-vi. https://doi.org/10.1111/lang.12134

Tversky, A., & Kahneman, D. (1971) Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105-110. http://dx.doi.org/10.1037/h0031322

Vandergrift, L., & Cross, J. (2017). Replication research in L2 listening comprehension: A conceptual replication of Graham & Macaro (2008) and an approximate replication of Vandergrift & Tafaghodtari (2010) and Brett (1997). *Language Teaching*, *50*(1), 80-89. https://doi.org/10.1017/S026144481500004X

VanPatten, B. (2002a) Processing Instruction: An update. *Language Learning, 52*(4), 755-803. https://doi.org/10.1111/1467-9922.00203

VanPatten, B. (2002b) Processing the content of input-processing and processing instruction research: A response to DeKeyser, Salaberry, Robinson, and Harrington. *Language Learning, 52*(4), 825-831. https://doi.org/10.1111/1467-9922.00205

VanPatten, B., & Cadierno, T. (1993a). Input processing and second language acquisition: A role for instruction. *The Modern Language Journal, 77*(1), 45-57. https://doi.org/10.1111/j.1540-4781.1993.tb01944.x.

VanPatten, B., & Cadierno, T. (1993b). Explicit instruction and input processing. *Studies in Second Language Acquisition, 15*(2), 225-243. https://doi.org/10.1017/S0272263100011979

VanPatten, B., & Oikkenon, S. (1996). Explanation versus structured input in processing instruction. *Studies in Second Language Acquisition, 18*(4), 495-510. https://doi.org/10.1017/S0272263100015394

VanPatten, B., & Williams, J. (2002). *Research criteria for tenure in second language acquisition: Results from a survey of the field*. Unpublished manuscript, University of Illinois at Chicago.

Walters, J. (2012). Aspects of validity of a test of productive vocabulary: Lex30. *Language Assessment Quarterly, 9*(2), 172-185. http://dx.doi.org/10.1080/15434303.2011.625579

Waring**,** R. (1997). The negative effects of learning words in semantic sets: A replication. *System, 25*(2), 261-274. https://doi.org/10.1016/S0346-251X(97)00013-4

Wicherts, J., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One, 6*(11), e26828. https://doi.org/10.1371/journal.pone.0026828

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726-728. https://doi.org/10.1037/0003-066X.61.7.726

Wong, W. (2001). Modality and attention to meaning and form in the input. *Studies in Second Language Acquisition, 23*(3), 345-368. https://doi.org/10.1017/S0272263101003023