This is a repository copy of *Data, open science, and methodological reform in second language acquisition research.*.

Data, open science, and methodological reform in second language acquisition research

Emma Marsden & Luke Plonsky

University of York and Northern Arizona University

The chapters in this volume attest to the fact that second language acquisition (SLA) is, undeniably, a field of diverse interests. Since the field's inception in the latter half of the 20th century (Gass, 2009), the substantive landscape of SLA has been in a state of continuous expansion. For the most part, this growth has been constructive, and our understanding of second-language (L2) development has advanced in large part due to massive efforts in the theoretical and empirical realms.

Just as critical to the advancement of this or any scientific discipline, however, is methodological rigor and transparency. Put another way, knowledge about L2 development cannot increase in the absence of high-quality research. Historically, though, there has been very little evidence of concern over research methods in the published SLA literature. The need for researcher training and methodological standards have largely been assumed to be addressed and met through institutional mechanisms, such as graduate training/supervision/assessment, and peer review (see Gonulal, Loewen, & Plonsky, 2014). This is all changing in unprecedented ways.

In recent years, applied linguists have come to explicitly take methodological issues more seriously. We interpret the move toward greater methodological awareness, the "methodological turn" (Byrnes, 2013, p. 825), as critical to our understanding of L2 acquisition and a sign of the field's maturity. In this wider context, illustrated in broad strokes in this chapter, it is not by

chance that the present volume would be published at this moment. We see this book and the reflective, methodologically-oriented discussions therein as comprising both an indicator of and a contributor to this movement.

Other indicators of methodological reform can be found in a range of outlets, many with parallels to this volume. Many of the calls for change have stemmed from the growing body of methodological syntheses, which systematically code and analyze research and reporting practices in a given (ideally comprehensive or at least representative) sample of primary studies. Some methodological syntheses have targeted substantive domains such as task-based language teaching (Plonsky & Kim, 2016) or computer-mediated interaction (Ziegler, 2016), the findings of which speak to the designs, tasks, and other features particular to those domains. In this volume, synthesizing methodological features across 34 studies of the effects of oral corrective feedback, Li noted high levels of idiosyncrasy in the way that certain oral feedback treatments have been operationalized, posing a potential threat to their construct validity. Taguchi (this volume) also presented a systematic review ($K = 22$). As with Li, Taguchi speaks directly to one particular type of study: longitudinal (non-instructed) development of L2 pragmatics. Despite this narrow focus, many of her findings and recommendations resonate with other domains of SLA research. Taguchi describes, for example, the types of evidence to be gained from spoken versus written discourse-completion tasks (see also Plonsky, Marsden, Crowther, Gass, & Spinner, under review, for meta-analytic evidence about the effects of modality in judgement tasks). Based on her findings, Taguchi also notes that role play data are not often fully utilized, with utterance-level analyses lacking consideration of the larger interactional context at work.

Other methodological syntheses have been concerned with research techniques that span broader theoretical terrain and larger swaths of research. For example, Plonsky (2013) described

and evaluated the designs, analyses, and reporting practices in 606 quantitative studies published in the journals *Language Learning* and *Studies in Second Language Acquisition*; and Marsden, Thompson, and Plonsky (in press) provide an in-depth systematic examination of the implementation of self-paced reading tests in L2 research ($K = 64$). Systematic review is also fruitfully applied by Leal (this volume), who questions L2 researchers' use of what is possibly the most frequently observed independent, and also very often dependent, variable in SLA: proficiency. Critically, though, the discussion is not limited to the author's arguments, persuasive as they may be, which build naturally on recent empirical, synthetic, and theoretical contributions in this area (e.g., Hulstijn, 2012; Thomas, 2006; Tremblay, 2011). Her application of methodological synthesis allows for strong claims regarding the use (and misuse) of proficiency on multiple levels — from design to definition to measurement.

Beyond evidence generated by systematic reviews, issues related to research design figure prominently throughout the present volume. One design feature common to all L2 research is sampling. Leal argues not only for larger samples as a means to achieve lower sampling error and greater internal validity but for samples consisting of greater diversity as well, which will yield more generalizable findings (see Plonsky, 2017). Many of the authors in this volume also share a common concern over the need for longitudinal designs and data as a means to understand L2 development in a richer, more dynamic fashion (e.g., Bardovi-Harlig; Tracy-Ventura & Huensch; see also Ortega & Byrnes, 2008).

Yet another design feature that must be considered is context. Multiple chapters describe the tension between the sterile control of lab-based production and the 'noisy' ecological validity of data collected in classrooms or elsewhere 'in the wild' (Bardovi-Harlig; Li; Taguchi; Shively; Solon; Tracy-Ventura & Huensch). Attention to this issue is rightly heightened in light of

numerous meta-analyses of instructed SLA showing larger effects for lab- over classroom-based treatments (see Li, 2010, this volume); differences in effects for second and foreign-language settings are also commonly observed. In this case, findings vary across different subdomains. Whereas some meta-analyses have observed larger effects among participants in L2 environments (e.g., Plonsky, 2011), others have found the opposite (e.g., Boulton & Cobb, 2017). Shively's chapter, in which contextual effects feature most prominently, takes the argument a step further, stating that "naturalistic data are crucial for diversifying the object of study […] and, in fact, can be a catalyst for expanding into new areas of inquiry in pragmatics".

The benefits of considering alternative techniques to data elicitation and coding surely apply in other linguistic domains as well. For example, although Solon was interested in L2 phonological development of Spanish (/l/), re-analyzing participants' samples led to a much more varied and, arguably, more informative interpretation of participants' L2 production data. The bottom-up approach in Solon's chapter, which is not tied to the binary coding scheme or to a single acoustic correlate as in her previous analysis (i.e., F2 values; Solon, 2017), availed the identification of multiple lateral types as well as interlanguage tokens that could not be attributed to the L1 or L2. In a similar vein, Edmonds and Gudmestad re-analyzed data from a previous study. Perhaps more critical than their findings (on temporal distance and adverbial specification as predictors of future time expression) is their call for greater explicitness and transparency regarding coding decisions. Similar positions are also expressed in Li's and Tracy-Ventura and Huensch's calls for piloting and validating of tasks/protocols (see also Derrick, 2016).

Another focal point common to the methodological reform movement and to the chapters in this volume is, of course, data analysis. Statistically speaking, L2 researchers have been introduced in recent years to a number of procedures such as bootstrapping (Larson-Hall &

Herrington, 2010; Plonsky, Egbert, & LaFlair, 2015), mixed effects modeling (Cunnings & Finlayson, 2015; Linck & Cunnings, 2015), and Bayesian inference (Gudmestad, House, & Geeslin, 2013; Norouzian, de Miranda, & Plonsky, under review). Similarly, discussions of data analysis run throughout the current volume. Leal, for example, warns against the loss of data variance, a statistical property central to most quantitative analyses used in SLA, that results from reducing continuous phenomena such as proficiency to nominal level data (see Plonsky & Oswald, 2017). Edmonds and Gudmestad (this volume) make the case both here and elsewhere (e.g., 2015) for multinomial regression in place of binomial (logistic) regression. Also worth noting in the context of data analysis are Tracy-Ventura and Huensch's comments on attrition and missing data, both common in L2 research and especially in longitudinal designs but rarely dealt with in a systematic or open manner.

In sum, a theme across all the chapters is the justification behind methodological choices, but with a strong emphasis on the *transparency* of these choices and the resulting data. We turn now to the importance of methodological transparency and its place in the wider open science movement.

Methodological advancement and its relationship with open science

Some of the broadest challenges facing not only SLA but many research domains are the closely related principles of open science, replication, and reproducibility, all of which rely heavily on methodological transparency. Some groups within the SLA community have been at the cutting edge of open science practices. For example, we have seen large publicly available corpora of L2 data, such as FLLOC (Marsden, Myles, Rule, & Mitchell, 2003) and SPLLOC (Mitchell, Dominguez, Arche, Myles, & Marsden, 2008) as discussed by Tracy-Ventura and Huensch (this volume) and openly available vocabulary research materials on Paul Meara's

Lognostics site (http://www.lognostics.co.uk/). In 2011, IRIS, the open repository for materials

and data in language learning research was established (https://www.iris-database.org/) (see

Marsden & Mackey, 2014; Marsden, Mackey & Plonsky, 2016).[1] Since then, the exponential

growth of its content and usage attest to increasing levels of awareness among the SLA

community of the importance of methodological transparency and a more collaborative ethic (see

this volume: Li; Tracy-Ventura & Huensch; Edmonds & Gudmestad). IRIS now holds

approximately 4,000 files covering a vast spectrum of theoretical, methodological, and

epistemological perspectives. It makes instruments, stimuli, coding and analysis protocols, and

data openly accessible and has drawn over 22,000 downloads (41,000 total site hits) to date from

researchers, teachers and students. IRIS has the support of a large network of journals and

research and teaching associations.

However, overall, the SLA community has been somewhat slow to avail itself of the

benefits of open science, and more efforts are still needed for a range of reasons. First, as

observed by several chapters in the current volume, there is still a distinct need for more clarity

about the operationalization of constructs, stimuli design and, critically, coding and analysis

decisions. For example, Li (this volume) examined the amount and duration of corrective

feedback and found that more than half of his study sample did not report this information.

Edmonds and Gudmestad (this volume) neatly demonstrated the need for increased transparency

in coding temporal distance and adverbial specification, as this coding had tangible

consequences for findings and, in turn, the consistency and standardization of constructs that are

used for theory building. Second, non-transparent methods have consequences for the quality and

---

quantity of replication research, as observed by several methodological syntheses (such as Derrick, 2016; Marsden, Morgan-Short, Thompson & Abugaber, 2018; Marsden et al., in press; Plonsky et al., under review). Poor methodological transparency is no doubt both a cause and a consequence of a culture in which replication research has little chance of being published, even though replication is widely agreed to be fundamental for the progress of science (see for example Li's and Edmonds & Gudmestad's calls for replication, this volume). Indeed, Marsden, Morgan-Short, Thompson, et al. (2018) found that fewer than one in every 400 journal articles has been a self-labelled replication study. This paucity probably reduces the perceived need to make a study's methods accessible in the first place, as there is a low chance of anyone replicating it. Thus, low rates of published replication research probably both partially account for and have led to inadequate methodological transparency, in terms of both materials and data.

In terms of materials, their lack of availability means that researchers need to create their own by extrapolating from the often-cursory descriptions and examples provided in published articles, as observed by Edmonds & Gudmestad (this volume). This (often unacknowledged) heterogeneity between studies adversely affects the inter-connectedness and replicability within agendas. Indeed, Marsden, Morgan-Short, Thompson, et al. (2018) found a positive association between the extent of materials availability and the extent to which replication studies supported the initial studies. Also, poor materials availability means that collaboration with the author(s) of the initial study becomes necessary in order to access materials, which is not ideal if we aspire to objective, independent validation of previous findings. Illustrating the potential concerns surrounding non-independent replication, Marsden, Morgan-Short, Thompson, et al. (2018) found a positive, statistically significant association between author overlap and the likelihood of

a replication study supporting the initial study (90% of replication studies), compared to when there was no author overlap (59% of replication studies).

In terms of the availability of data in SLA research, there has generally been very little raw data that is widely accessible (Larson-Hall & Plonsky, 2015; Larson-Hall, 2017; Plonsky et al., 2015). However, as Tracy-Ventura & Huensch (this volume) note, open data have many benefits. First, full datasets allow researchers to ascertain parity of samples, via checking data against previous datasets. This can be especially important given that differences between participants, such as proficiency, represent important variables in SLA (Leal, this volume). A second benefit of transparent data is in allowing secondary analysis / re-analysis of data, as executed in this volume by Solon and by Edmonds and Gudmestad. Indeed, re-analysis of data is deemed to be a type of replication research for determining reproducibility, according to some taxonomies (National Science Foundation, 2015). However, such activity is rare, or at least rarely published. Marsden, Morgan-Short, Thompson, et al. (2018) found no examples of this as a type of replication, and just 6 percent of replication studies that had combined the initial data with the replication data, of which only one had undertaken a meta-analysis of both sets of data. Morgan-Short, Marsden, Heil, et al. (2018) demonstrated the potential for multi-site replication studies to facilitate such combined analyses. A final reason for sharing data is the possibility that open data are positively associated with better reporting and stronger evidence (Wicherts, Bakker, & Molenaar, 2011).

Nevertheless, data sharing is not without challenges. Researchers need to think ahead about gaining informed consent from participants in the early stages of ethics approval/IRB. Shively (this volume) provides a useful account of the ethical considerations in collecting naturalistic audio-recorded data 'in the wild' (see also De Costa, 2016). Additionally challenging

is that, in order to be useful to others, data need to be structured in a user-friendly format, with codes and data cleaning procedures made explicit. Corpus researchers have perhaps been the quickest in the SLA community to capitalize on the affordances of technology for data-sharing, yet even these more experienced colleagues are realizing the importance of full transparency for achieving a level of standardization to facilitate re-analysis and inter-study comparability, as illustrated by Tracy-Ventura & Huensch's (this volume) discussion of the choices about what information to annotate in a corpus and which metadata to include. (See also Bell, Collins, & Marsden, in press).

On a brighter note, signs of the vitality of open science are emerging. Both journals and professional associations are in a prime position to effectuate change that can address, on a field-wide scale, many of the challenges raised in this volume. For example, several journals (such as *Language Learning, Studies in Second Language Acquisition,* and *The Modern Language Journal*) require the reporting of effect sizes and also recognize open science practice with Center for Open Science 'badges' that clearly signpost studies that are open for scrutiny and replication (e.g. Trofimovich & Ellis, 2015) (see Figure 1).

Figure 1. Badges indicating open science practices in participating journals.

More widespread engagement can be seen in the approximately 36 journals that encourage their authors to upload their materials to IRIS, and, in 2017, this practice was also flagged in the American Association for Applied Linguistics (AAAL)'s publication guidelines. Further

evidence of the role of professional associations is seen in AAAL's recent amendments to their promotion and tenure guidelines, which now value replication studies as part of a scholar's portfolio, and their recent 'Research Methods' strand at the annual conference. Other signs of open science practices include a growing number of researchers producing and using open-source software, such as R data analysis scripts (see this volume Leal; Mizumoto & Plonsky, 2015) and calls for the large-scale data collection via online platforms (MacWhinney, 2017). Many materials for investigating the substantive areas covered by the current volume are now openly accessible on IRIS: 147 search hits are retrieved for the research area 'proficiency', 71 for 'pragmatics', 106 for 'corrective feedback', 147 for 'phonology', and 141 for 'tense'. The repository also has two special collections of research materials reviewed in methodological syntheses: 62 self-paced reading tests (Marsden, Morgan-Short, Thompson, et al. 2018) and 110 acceptability judgement tests (Plonsky et al., under review). Over the last two years, since IRIS began accepting data, 43 datasets have been made available, further testament to a changing culture. It is hoped that the availability of materials on IRIS engenders more replication research, further incentivized by the IRIS Replication Award (see iris-database.org/iris/app/home/replication_award). Beyond SLA, other disciplines, including SLA's sister disciplines of social and cognitive psychology and education (see Marsden, Morgan-Short, Thompson, et al., 2018) are going through similar cultural changes propelled by large infrastructure such as the Centre for Open Science (https://cos.io/) and its open research platform, the Open Science Framework. IRIS complements and is distinct from such developments as it offers a discipline-specific — and therefore highly searchable — platform, hosting only peer-reviewed and accepted materials and data (rather than also holding works-in-progress).

Highly relevant to our focus on methodology is the new article type Registered Reports, just launched at *Language Learning* (Marsden, Morgan-Short, Trofimovich, & Ellis, 2018). Originally launched at *Cortex* in 2013 (Chambers, 2013), about 90 journals now offer this article type. RRs undergo peer review *before* data have been collected, thus shining a spotlight on the *methodological choices* made as they pertain to the substantive relationships under investigation. First, manuscripts are peer reviewed for just their rationale, methods, materials, and coding and analysis 'pipelines' (i.e., critical decision points are identified and justified, allowing appropriate flexibility over which of these pre-determined choices will be taken once the data are collected). If approved, the manuscripts are given 'In Principle Acceptance' – critically, they cannot now be rejected on the grounds of methodological flaws or results that are unexpected or perceived as difficult to explain. As well as reducing publication bias, RRs focus our attention on the importance of methods, rather than the nature of the findings, for determining confidence in the validity and reliability of science.

Figure 2: Center for Open Science badge indicating studies have been registered prior to data collection.

As shown in Figure 2, such studies are eligible for additional recognition for having practiced open science, by participating journals such as *Language Learning* (Trofimovich & Ellis, 2015).

Conclusion

This volume has alluded to the plethora of decision points involved in research methodology, from participant demographics, research design, instrument design, protocol administration, data categorization and coding choices, through to choices about what to report and share. The emphasis has been on the importance of justifying and being transparent about these choices. The authors all observe, or even empirically demonstrate (e.g., Solon; Edmonds & Gudmestad), that there are no one-size-fits-all methodological decisions. However, the need for transparency does 'fit all'. The messages and evidence communicated in this volume indicate that we can look forward to more methodological reflection, innovation, and open collaboration to improve the health of the science in the SLA community.

References

Bell, P., Collins, L., & Marsden, E. (in press). Building a learner corpus to address the instructional needs of L2 language learners in compulsory school programmes. In M. Paquot & B. Le Bruyn (Eds.), *Learner corpora and second language acquisition*. Cambridge: Cambridge University Press.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*, 348-393.

Byrnes, H. (2013). Notes from the editor. *Modern Language Journal, 97*, 825-827.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*, 609-610.

Cunnings, I., & Finlayson, I. (2015). Mixed effects modelling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159-181). New York: Routledge.

De Costa, P. I. (Ed.) (2016). *Ethics in applied linguistics research: Language researcher narratives*. New York: Routledge.

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly, 50,* 132-153.

Edmonds, A., & Gudmestad, A. (2015). What the present can tell us about the future: A variationist analysis of future-time expression in native and nonnative French. *Language, Interaction, Acquisition, 6,* 15-41.

Gass, S. (2009). A historical survey of SLA research. In T. K. Bhatia & W. C. Ritchie (Eds.), *The new handbook of second language acquisition* (pp. 3-27). Bingley: Emerald.

Gonulal, T., Loewen, S., & Plonsky, L. (2017). The development of statistical literacy in applied linguistics graduate students. *International Journal of Applied Linguistics, 168,* 4-32.

Gudmestad, A., House, L., & Geeslin, K. L. (2013). What a Bayesian analysis can do for SLA: New tools for the sociolinguistic study of subject expression in L2 Spanish. *Language Learning, 63,* 371-399.

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition, 15,* 422-433.

Larson–Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics. *The Modern Language Journal, 101,* 244-270.

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language Acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31,* 368-390.

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.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*, 309-365.

Linck, J., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185-207.

MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning*, *67*, 254-275.

Marsden, E., & Mackey, A. (2014). IRIS: A new resource for second language research. *Linguistic Approaches to Bilingualism*, *4*, 125-130.

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

Marsden, E. J., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews, and recommendations for the field. *Language Learning*, *68*.

Marsden, E., Morgan-Short, K., Trofimovich, P., & Ellis, N. (2018). Introducing Registered Reports at Language Learning: Promoting transparency, replication, and a synthetic ethic in the language sciences [Editorial]. *Language Learning*, *68*.

Marsden, E., Myles, F., Rule, S., & Mitchell, R. (2003). Using CHILDES tools for researching second language acquisition. In S. Sarangi & T. van Leeuwen (Eds.), *Applied Linguistics and Communities of Practice* (Vol. 18, pp. 98-113). London: BAAL/Continuum.

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

Mitchell, R., Dominguez, L., Arche, M., Myles, F., & Marsden, E. (2008). SPLLOC: A new database for Spanish second language acquisition research. *EUROSLA Yearbook, 8*, 287-304.

Mizumoto, A., & Plonsky, L. (2016). R as a lingua franca: Advantages of using R for quantitative research in applied linguistics. *Applied Linguistics, 37,* 284-291.

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

National Science Foundation. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science. Report of the subcommittee on replicability in science.* Retrieved 1 November 2017 from http://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

Norouzian, R., de Miranda, M. A., & Plonsky, L. (under review). The Bayesian revolution in L2 research: An applied approach. Manuscript under review.

Ortega, L., & Byrnes, H. (2008). *The longitudinal study of advanced L2 capacities*. New York: Routledge.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning, 61,* 993-1038.

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,* 655-687.

Plonsky, L. (2017). Quantitative research methods. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 505-521). New York: Routledge.

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

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics, 36,* 73-97.

Plonsky, L., Marsden, E., Crowther, D., Gass, S., & Spinner, P. (under review). A methodological synthesis and meta-analysis of acceptability judgement tests.

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition, 39*, 579-592.

Solon, M. (2017). Do learners lighten up? Phonetic and allophonic acquisition of Spanish /l/ by English-speaking learners. *Studies in Second Language Acquisition*, *39,* 801-832.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment in second language proficiency. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279-298). Amsterdam: Benjamins.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research. *Studies in Second Language Acquisition, 33*, 339-372.

Trofimovich, P., & Ellis, N. (2015). Open science badges [Editorial]. *Language Learning, 65*, v–vi.

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*, e26828.

Ziegler, N. (2016). Methodological practices in interaction in synchronous computer mediated

communication: A synthetic approach. In A. Mackey & E. Marsden (Eds.), *Instruments for

research into second languages: Empirical studies advancing methodology* (pp. 197-223).

New York: Routledge.