



UvA-DARE (Digital Academic Repository)

Sampling bias and the problem of generalizability in Applied Linguistics

Andringa, S.; Godfroid, A.

DOI

[10.1017/S0267190520000033](https://doi.org/10.1017/S0267190520000033)

Publication date

2020

Document Version

Final published version

Published in

Annual Review of Applied Linguistics

License

CC BY-NC-SA

[Link to publication](#)

Citation for published version (APA):

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in Applied Linguistics. *Annual Review of Applied Linguistics*, 40, 134-142.
<https://doi.org/10.1017/S0267190520000033>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

SHORT PAPER

Sampling Bias and the Problem of Generalizability in Applied Linguistics

Sible Andringa¹ and Aline Godfroid^{2*}

¹Universiteit van Amsterdam, the Netherlands and ²Michigan State University, USA

*Corresponding author. E-mail: godfroid@msu.edu

Abstract

In this final contribution to the issue, we discuss the important concept of generalizability and how it relates to applied linguists' ability to serve language learners of all shades and grades. We provide insight into how biased sampling in Applied Linguistics currently is and how such bias may skew the knowledge that we, applied linguists, are building about second language learning and instruction. For example, our conclusions are often framed as universally-applying, even though the samples that have given rise to them are highly specific and Western, Educated, Industrialized, Rich, and Democratic (WEIRD; Henrich, Heine, & Norenzayan, 2010). We end with a call for research and replication in more diverse contexts and with more diverse samples to promote progress in the field of Applied Linguistics as ARAL celebrates its 40th anniversary.

Introduction

Research in the language sciences is generally motivated by a desire to understand the architecture of the language system, language learning, and language use. As language scientists, we tend to strive for knowledge that has universal value, knowledge that is true beyond the particular group of language users we observe in our studies. An important lesson in behavioral research methodology is that such generalization rests on the premise that samples are randomly drawn from the population under study. If the selection of participants is somehow biased, then the reliability of researchers' statements about the behavior under investigation is compromised (Henrich, Heine, & Norenzayan, 2010). In practice, however, participant selection is almost always biased. For the field of psychology, Arnett (2008) has shown that samples are drawn primarily from WEIRD societies—Western, Educated, Industrialized, Rich, and Democratic groups (Henrich et al., 2010). Arnett (2008) estimated that 80% of the samples presented in six major psychology journals consisted of university students drawn from academic psychology programs. Arnett (2008) also clearly demonstrated that sampling is not only biased in terms of educational background: 90% of the participants represented in the journals surveyed came from North American or European populations.

© The Author(s), 2020. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (<http://creativecommons.org/licenses/by-nc-sa/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is included and the original work is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use.

Current psychological insight, then, is based on a very specific, highly educated group drawn from WEIRD societies.

In this position paper, we explore to what extent our own field of Applied Linguistics suffers from biased sampling and what the consequences of such skewed sampling may be. We do this in full recognition that we are not alone in raising concerns about the demographics of the samples used in our field (e.g., Bigelow & Tarone, 2004; Cox, 2019; Kormos & Sáfár, 2008; Mackey & Sachs, 2012; Ortega, 2005, 2019; Pot, Keijzer, & de Bot, 2018; Tarone & Bigelow, 2005), that we cannot do justice to all these voices within the scope of this paper, and in full recognition of the fact that we ourselves are part of the problem in that we rely on academic samples as much as anyone else. We do think it is important to keep this conversation on the front burner in Applied Linguistics, and we intend to do so in this paper by sketching just how skewed sampling currently is and how problematic this can be.

Sampling Practices in Applied Linguistics

There is no reason to assume that the field of Applied Linguistics is doing any better in terms of population representation than its neighboring disciplines. There are many indications (but little hard evidence so far) that certain groups are underrepresented. For example, in a plea for attention to illiteracy in second-language (L2) learning published in *TESOL Quarterly*, Bigelow and Tarone (2004) pointed out that there had been no publications on illiterate L2 learners in *TESOL Quarterly* in the previous ten years, even though millions of people in the US and hundreds of millions of people worldwide are illiterate (currently estimated at 750 million by the UNESCO Institute for Statistics [2019]). Sixteen years later, this situation remains largely unchanged (but see Bigelow, Delmas, Hansen, & Tarone, 2006). Conversely, there is also evidence that some groups are overrepresented: Plonsky (2016) estimated that 67% of research samples in L2 research are university student samples. To date, however, no published research has systematically investigated sampling biases in Applied Linguistics research.

One way to gauge the sampling practices in the field is to examine published meta-analyses. The field has witnessed a plethora of meta-analyses, which collectively cover a large number of published studies on a wide array of topics. In some ways, meta-analyses are the hallmark of researchers' wish to generalize. Meta-analysts typically synthesize a collection of studies that, by the nature of their designs, all aim for generalization, and in doing so, meta-analysts aim to present the state of the art in a particular area of inquiry. Meta-analysis also affords a possibility to evaluate the sampling practices in a given domain, given that many meta-analysts will summarize and present the demographic characteristics of the participant samples included in the primary studies. Table 1 lists sample information provided by meta-analyses published in the last five years that coded participants' background in sufficient detail to allow insight into the field's sampling practices. As can be seen, this yielded a selection of 17 meta-analyses collectively covering 863 studies and a fair (but obviously far from exhaustive) range of topics that applied linguists are concerned with. The table shows that 65% of all samples were university student samples. Samples other than university students tended to consist of elementary or high school pupils. These samples of young learners accounted for 25% of the research database. There is thus a sizeable amount of research with young learners, but it still constitutes a heavy underrepresentation in light of the fact that L2 learning takes place in schools much more frequently than at universities (Kormos & Sáfár, 2008).

Table 1. Characteristics of participant samples (k) in recent meta-analyses (2014-2019)¹

Study ²	Topic	No. articles ³	Elem. school	High school	University	Lang. ins.	Miss.	other	% uni.	% young learners	% uni. students in adult samples
Boulton & Cobb, 2017	Corpus use	64		4	33		6		77	9	100
Bryfonski & McKay, 2019	TBLT implementation	52	18		25	4			53	38	86
E. Y. Kang & Han, 2015	Written corrective feedback	21	1	1	12	7			57	10	63
E. Y. Kang, Sok, & Han, 2019	Form-focused instruction	54	7	11	27	6			53	35	82
J. Lee, Jang, & Plonsky, 2015	Pronunciation instruction	86		10	86	10			81	9	90
Li, 2016	Aptitude	66	2	22	26	6		7	41	38	67
H. Lin, 2014	CMC	59	3	10	45		1		76	22	100
H. Lin, 2015a	Reading comprehension	59	5	9	27			18	46	24	60
J.-J. Lin & Lin, 2019	Mobile-assisted vocabulary learning	33	5	5	19			4	58	30	83
Q. Liu & Brown, 2015	Corrective feedback in L2 writing	44	3	5	33	2	2		73	18	94
J. Liu & Zhang, 2018	Extensive reading	21	1	5	14	1			67	29	93

Plonsky & Zhuang, 2019	Pragmatics instruction	50		4	77	11		84	4	88	
Sharifi, AbuSaeedi, Jafarigohar, & Zandi, 2018	Computer-assisted instruction	140	45 ⁴		113			72	28	100	
Teimouri, Goetze, & Plonsky, 2019	L2 Anxiety and achievement	97	3	35	53	2	3	55	40	96	
Uchihara, Webb, & Yanagisawa, 2019	Incidental vocabulary acquisition	26	4	11	30			67	33	100	
Wang, Lan, Tseng, Lin, & Gupta, 2019	3D virtual worlds	13	2	8	21			68	32	100	
Ziegler, 2016	Interaction in synchronous CMC	14		3	11			79	21	100	
		863	108	143	652	49	12	29	65	25	88

¹Notes: Elem. school = elementary school; Lang. ins. = language institute; Miss. = missing

²To be included in the table, the meta-analyses needed to be published between 2014 and 2019, were required to code for educational level, and were not allowed to exclude studies on the basis of educational level. To avoid overlap, the most recent meta-analysis was included in case studies covered the same topic. Journal selection was based on Al-Hoorie & Vitta (2018); journals were SSCI indexed and international in scope. They were: *Applied Linguistics*, *Applied Psycholinguistics*, *Bilingualism: Language & Cognition*, *Computer Assisted Language Learning*, *ELT Journal*, *English for Specific Purposes*, *Foreign Language Annals*, *International Review of Applied Linguistics in Language Teaching*, *Language Assessment Quarterly*, *Language Learning*, *Language Learning & Technology*, *Language Teaching Research*, *Language Testing*, *The Modern Language Journal*, *ReCALL*, *Second Language Research*, *Studies in Second Language Acquisition*, *System*, and *TESOL Quarterly*.

³Number of articles and the total number of samples don't correspond exactly. Articles included in a meta-analysis often provide more than one sample, but some articles in the meta-analysis don't provide information about educational level.

⁴This number represents elementary and high school samples together.

If we zoom in on the adult groups, the picture is particularly bleak. A staggering 88% of all adult samples are university student samples. The remaining 12% were samples coded as “language institute” or “other,” and, while it is not exactly clear what kind of learners fall in these categories, we might conclude that we know comparatively little about adult language learning that takes place outside the walls of academia.

It is important to acknowledge that an investigation of participant samples on the basis of meta-analyses has limitations. For one thing, descriptive and qualitative work remains under the radar, but such work often foregrounds the contextual and situated nature of language learning and use over its generalizability. Also, particular topics foreground particular populations: reading research is often conducted with younger learners and study abroad with university students. Furthermore, in meta-analyses, a large number of studies are typically dropped because of unfit study designs and insufficient detail in reporting. We cannot rule out (but do not assume either) that studies with nonuniversity learners present greater challenges in study design and are, therefore, excluded more often from meta-analysis. If that were the case, research with nonuniversity learners would be underrepresented in the meta-analytic literature and the proportion of research samples with non-university learners might actually be higher.

Zhang’s (2019) recently published bibliometric analysis of second language acquisition (SLA) research provides yet another view of the field’s sampling practices. This analysis shows clearly that the field is dominated heavily by Western scholarship. Using the 16 most influential Applied Linguistics journals as determined by Al-Hoorie and Vitta (2018), Zhang (2019) analyzed the origins of the publications in these journals. The analysis showed that the field has been heavily dominated by journals published in the U.S., publications produced at prestigious North American research institutes, and scholars working in Anglo-Saxon environments (although Japanese scholarship has been strong, and Chinese scholarship is on the rise). It is possible that the research produced at Anglo-Saxon universities was conducted elsewhere, such that there may not always be a one-to-one relationship between the university location and the study location, but overall, the picture emerging is that insight from Applied Linguistics research is strongly shaped by information gathered from university students, and these tend to be enrolled at prestigious research institutes in North America. The inescapable conclusion is that our participants are truly WEIRD.

It is interesting to consider briefly why sampling biases exist. Notions of convenience and accessibility are easily invoked as an explanation for sampling biases, but probably too simplistic (Ortega, 2005). Of course, many researchers are likely to resort to student samples, because they are easily available. Many programs, for example, maintain a database of potential research participants and sometimes have a credit system in place for student participation in research. But the reasons underlying sampling bias probably run deeper. One important reason could be a lack of engagement with language-learning contexts outside the walls of the university (Ortega, 2005). Many applied linguists teach languages at their own research institutes, and it is natural that their classroom observations generate hypotheses that inspire their research. Moreover, the tendency to look for participants within the walls of academia may be exacerbated by the “publish or perish” culture that governs much of the production of knowledge in academia (Chambers, 2019). Making it in academia tends to depend on the number of publications one produces and the perceived quality of the journals in which they appear, and this makes it attractive to go for the more easily accessible student populations.

What are the consequences for generalizability?

The ways in which existing sampling biases may have affected the state of our knowledge must remain a matter of speculation until we systematically gather knowledge from a wider array of language learning contexts. As Henrich et al. (2010) have demonstrated for several areas within psychology, there is every reason to be concerned about this. One consequence of sampling bias could be that behavior attested in particular groups is simply absent in other groups. This is illustrated by Tarone and Bigelow (2005), who argued that the influential noticing hypothesis (Schmidt, 1990, 1994) may need reconsideration in light of evidence showing that illiterate learners typically do not exhibit noticing where literate learners do. Another consequence is that we may underestimate the range of variation in language learners and language learning outcomes. The Applied Linguistics field is replete with studies that try to establish links between language learning or use and particular individual characteristics. Such endeavors are likely to be compromised due to range restriction if observations are sampled from just a narrow band of the population distribution. How conclusive are findings on the role of working memory or aptitude in foreign language learning, if they are based on university students exclusively? Correlations are hard to detect in the absence of variation and may well be deflated or gone if one zooms in on a restricted range of variation. For this reason, at least one psychology lab at a northeastern U.S. elite institution no longer recruits study participants among its university students, although such a measure tends to be the exception rather than the rule.

Very few studies have directly tested how the nature of the sample affects the behavior under scrutiny. One exception is Andringa (2014), who investigated how the characteristics of native-speaker comparison groups affect decisions about the incidence of native-like performance by second language learners. Second language learners were compared against a nonrepresentative (academic-only) native speaker norm group and a representative norm group, which approximately followed the education demographics of the society from which the native speakers were drawn, including participants with an academic background. Andringa showed that L1 proficiency ranges were consistently larger in the representative norm group, suggesting that it is both unrealistic and unrepresentative to hold L2 speakers' language abilities up against academically trained L1 speakers only.

In reading research, the heavy reliance on participant samples learning to read in English (with the English language being an extreme outlier in terms of the inconsistency of the sound-spelling link) illustrates how sampling biases can negatively affect the state of our knowledge. Share (2008) offered a fascinating and somewhat disturbing discussion of the consequences, which include, among others, an undue focus on accuracy in learning to read and in defining and diagnosing dyslexia, and the adoption of pedagogies that are misplaced when learning to read in languages with more regular orthographies. In Share's words, the overwhelming presence of English in reading research has led to "distorted theorizing with regard to many issues—including phonological awareness, early reading instruction, the architecture of stage models of reading development, the definition and remediation of reading disability, and the role of lexical-semantic and supralexical information in word recognition" (p. 584).

As Ortega (2005) has poignantly argued, our sampling choices not only create problems for generalizability, but also pose ethical dilemmas. Our science may not provide answers to questions for a vast majority of language learners, even within WEIRD societies. Ultimately, the challenges we face as a discipline may lie elsewhere. They may lie in how to sustain linguistic diversity, in how to sustain multilingual development, in dealing

with the language-learning needs of refugees and migrants, who may have had limited or interrupted formal education, in the relationship between (second) language skills and healthy aging, and in the dropping numbers of students learning languages other than English (e.g., Cox, 2019; King, Schilling, Fogle, Lou, & Soukup, 2008; Ortega, 2019; Pot et al., 2018; Reynolds, 2019). Does our research and the models that we generate represent the groups that face these challenges, and does it speak to their experiences? Based on the evidence presented above, we cannot be confident they do. Ultimately, for Applied Linguistics research to be considered relevant by stakeholders, researchers need to demonstrate and make a case for the impact of their work beyond the walls of the academy, in a society that faces many real linguistic needs and questions.

Making Applied Linguistics More Open and Less WEIRD: Some Suggestions

Changing current sampling practices will not be easy. It will require applied linguists to broaden the scope of their research. For this to happen, they may need to step outside their comfort zone, to take note of language learning contexts outside the university, to collaborate with scholars working in different settings, and to actively engage in more dialogue with stakeholders, such as teachers, policymakers, and funding agencies. Current publication practices and the emphasis on publication metrics in researcher quality evaluation may also need to change.

Some of these changes are already underway. The OASIS project, which stands for Open Accessible Summaries In Language Studies, is one cross-journal initiative that may help foster interaction between researchers and stakeholders (“OASIS,” n.d.), and has been embraced by a number of journals in our field already (e.g., Marsden, Trofimovich, & Ellis, 2019). Changes are also visible in the fact that many funding agencies now consider the broader impacts or societal impacts of research grant applications as a required and integral part of the grant evaluation process (e.g., NSF, n.d.; NWO, n.d.). And in Europe, the Plan S initiative, which mandates that publicly funded research from 12 European countries be published in open access journals by 2021, aims not only to make science freely available to everyone, but also tries to change the way scholarly achievement is evaluated by putting in place a system that acknowledges that an academic career is also shaped by a scholar’s ability to engage with society and to teach (Science Europe, 2018). All these initiatives are meant to break down walls between the academy and the wider society, so researchers may expand their research, relevance, and reach.

And of course, we should not take the current state of our knowledge for granted, but work directly to diversify participant samples. One powerful way to do this is to replicate existing research with different samples and evaluate to what extent the same conclusions still hold. This is the essence of the *SLA for All?* project (Andringa & Godfroid, 2018), in which we invited scholars to replicate influential studies in different, nonacademic samples. The goals of the project are to (1) raise awareness of the existing sampling biases in Applied Linguistics and the problems associated with them; (2) to get a sense of the extent to which the current state of knowledge in Applied Linguistics is indeed distorted by these sampling practices; (3) to gain experience doing research in different contexts and with learners who currently often remain invisible in research; and (4) to learn to what extent the tools applied linguists have developed are adequate for doing so, as they may well not be (Ortega, 2019). We hope that such an initiative will promote further efforts to make Applied Linguistics less WEIRD and, with that, a bit more fair and inclusive.

References

- Al-Hoorie, A., & Vitta, J. P. (2018). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR impact factors. *Language Teaching Research*, 23(6), 727–744. doi:10.1177/1362168818767191
- Andringa, S. J. (2014). The use of native speaker norms in critical period hypothesis research. *Studies in Second Language Acquisition*, 36, 565–596. doi:10.1017/S0272263113000600
- Andringa, S. J., & Godfroid, A. (2018). *SLA for all? Reproducing SLA research in non-academic samples*. OSF. Retrieved from <https://doi.org/10.17605/OSF.IO/MP47B>
- Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, 63(7), 602–614.
- Bigelow, M., Delmas, R., Hansen, K., & Tarone, E. (2006). Literacy and the processing of oral recasts in SLA. *TESOL Quarterly*, 40(4), 665–689.
- Bigelow, M., & Tarone, E. (2004). The role of literacy level in second language acquisition: Doesn't who we study determine what we know? *TESOL Quarterly*, 38(4), 689–700.
- Chambers, C. (2019). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.
- Cox, J. G. (2019). Multilingualism in older age: A research agenda from the cognitive perspective. *Language Teaching*, 52(3), 360–373.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61–83.
- King, K. A., Schilling, N., Fogle, L. W., Lou, J. J., & Soukup, B. (2008). *Sustaining linguistic diversity: Endangered and minority languages and language varieties*. Georgetown University Press.
- Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, 11(2), 261–271.
- Mackey, A., & Sachs, R. (2012). Older learners in SLA research: A first look at working memory, feedback, and L2 development. *Language Learning*, 62(3), 704–740.
- Marsden, E., Trofimovich, P., & Ellis, N. (2019). Extending the reach of research: Introducing open accessible summaries at *Language Learning*. *Language Learning*, 69(1), 11–17.
- NSF (n.d.). Perspectives on broader impacts. Retrieved January 2020 from https://www.nsf.gov/od/oa/publications/Broader_Impacts.pdf
- NWO (n.d.). Knowledge utilisation to create impact. Retrieved January 2020 from <https://www.nwo.nl/en/about-nwo/organisation/nwo-domains/sgw/knowledge+utilisation>
- OASIS: Open accessible summaries in language studies (n.d.). Retrieved January 2020 from <https://oasis-database.org/>
- Ortega, L. (2005). For what and for whom is our research? The ethical as transformative lens in instructed SLA. *The Modern Language Journal*, 89(3), 427–443. doi:10.1111/j.1540-4781.2005.00315.x
- Ortega, L. (2019). SLA and the study of equitable multilingualism. *The Modern Language Journal*, 103, 23–38.
- Plonsky, L. (2016). *The N crowd: Sampling practices, internal validity, and generalizability in L2 research*. Presentation at University College London.
- Pot, A., Keijzer, M., & de Bot, K. (2018). Do low L2 abilities impede healthy aging for migrant older adults in the Netherlands? *Dutch Journal of Applied Linguistics*, 7(1), 109–120.
- Reynolds, D. (2019). Language policy in globalized contexts. Report RR.3.2019. *Report for the World Innovation Summit for Education, Qatar Foundation*. Retrieved from: <https://www.wise-qatar.org/app/uploads/2019/08/language-policy-in-globalized-contexts.pdf>
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129–158.
- Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review*, 11, 11–26.
- Science Europe. (2018). Plan S: Making full and immediate open access a reality. Retrieved from <https://www.coalition-s.org/>
- Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of over-reliance on an “outlier” orthography. *Psychological Bulletin*, 134(4), 584.

Tarone, E., & Bigelow, M. (2005). Impact of literacy on oral language processing: Implications for second language acquisition research. *Annual Review of Applied Linguistics*, 25, 77–97.

UNESCO Institute for Statistics. (2019). Literacy. Retrieved from <http://uis.unesco.org/en/topic/literacy>

Zhang, X. (2019). A bibliometric analysis of second language acquisition between 1997 and 2018. *Studies in Second Language Acquisition*, 1–24. doi:10.1017/S0272263119000573

Cite this article: Andringa, S., & Godfroid, A. (2020). Sampling Bias and the Problem of Generalizability in Applied Linguistics. *Annual Review of Applied Linguistics*, 40, 134–142. <https://doi.org/10.1017/S0267190520000033>