Studies in Second Language Acquisition (2026), 1–8
doi:10.1017/S027226312610179X

EDITORIAL

A call for greater attention to the external validity of studies of second language acquisition

Stuart Webb

Faculty of Education, University of Western Ontario, London, Ontario, Canada
Email: swebb27@uwo.ca

(Received 16 April 2026 UTC; Revised 16 April 2026 UTC; Accepted 22 April 2026 UTC)

Abstract

Internal and external validity are key criteria for evaluating the quality and usefulness of research on second language acquisition (SLA). Although considerable attention is given to internal validity, external validity—the extent to which findings can be generalized beyond the specific study conditions—may be neglected during peer review and subsequent interpretations of findings. This editorial argues that a lack of attention paid to external validity reduces the interpretability and responsible use of SLA research through overgeneralization and misinterpretation of results. Current initiatives aimed at improving generalizability, including replication, meta-analysis, and multisite studies, are discussed, and a framework for evaluating external validity is proposed. The editorial concludes by calling for more transparent reporting practices and more cautious interpretation of research findings, with the aim of promoting accurate interpretation, replication, and theoretically motivated follow-up research in SLA.

Keywords: external validity; internal validity; multisite studies; replication; meta-analysis

Internal and external validity are foundational constructs that serve as critical criteria for evaluating research quality (Shadish, Cook, & Campbell, 2002). Internal validity refers to the extent to which the design, measurement, and analytic procedures employed within a study support a causal interpretation of findings by ruling out alternative explanations. Readers evaluate the internal validity of studies by assessing the strength of research designs; studies with strong internal validity allow readers to be confident that the findings are credible. In other words, internal validity concerns whether findings accurately reflect the variables and relationships of interest. External validity, by contrast, refers to the degree to which research findings can be generalized beyond the study at hand. Because there is a great deal of precision involved in SLA research designs, it is important to question the degree to which results may be valid beyond the boundaries of the methodology. Therefore, when readers evaluate the external validity of a study, they are assessing the degree to which the research findings

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S027226312610179X Published online by Cambridge University Press

are transferable across different learner populations, target items, materials, learning conditions, tests, and procedures. This editorial argues that insufficient attention to external validity in SLA research limits the interpretability and responsible use of research findings and that more systematic reporting and design practices are needed to address this issue.

One way to assess external validity is to consider whether and to what extent changes in key design variables are likely to influence findings. These variables include learner, material, learning condition, target item, measurement, and procedural factors.

1. Learner variables
   To what extent can the findings be generalized to learners who differ in age, proficiency, L1 background, amount of formal instruction, level of exposure to the L2, metalinguistic awareness, learning strategies, motivation, and so forth?
2. Material variables
   To what extent can the findings be generalized to learning with materials that differ in length, intended audience, difficulty, lexical complexity, lexical sophistication, topic familiarity, and mode of input (e.g., spoken, written, multimodal)?
3. Learning-condition variables
   To what extent can the findings be generalized to spoken and written conditions that differ in length, frequency, learning context, dyadic or group configuration, mode (speaking, listening, reading, writing), degree of meaningful engagement, and feedback?
4. Target item variables
   To what extent can the findings be generalized to target items that differ in form, meaning, frequency of exposure, difficulty, salience, and degree of L1–L2 formal and semantic congruency?
5. Measurement variables
   To what extent can the findings be generalized across assessment formats that differ in length, task type, sensitivity to knowledge, mode of input and output, scoring criteria, and time allocated per item?
6. Contextual and procedural variables
   To what extent can the findings be generalized across instructional contexts (e.g., classroom vs. out-of-class learning; EFL, ESL, ESP, EAP) and procedures with variations in instructional time, task instructions, level of instructional support, and number of exemplars?

The peer review process is well-designed to evaluate the internal validity of studies submitted for publication. Reviews must place emphasis on whether research designs meet the requirements necessary to effectively answer the research questions. Most studies are rejected for publication due to shortcomings in their research design. Although reviewers seldom label these issues explicitly as threats to internal validity, doing so would be useful because it would make their feedback clearer and easier for authors to apply. The external validity of studies of SLA is likely to receive less attention in the review process. This pattern is to some extent understandable. If research does not have strong internal validity, the results have little value and so there is little need to assess external validity. Moreover, any single SLA study will necessarily have limited external validity, because its findings cannot encompass the wide range of learner, target-item, material, learning-condition, test, and procedural variables that may influence outcomes. There is an expectation that authors will address limitations in

https://doi.org/10.1017/S027226312610179X Published online by Cambridge University Press

external validity within discussion, limitation, and conclusion sections of a manuscript. However, because authors must often prioritize addressing multiple reviewer concerns within strict word limits, discussion of external validity may be neglected. Moreover, there is good reason why authors may include relatively little detail about the external validity of their research. The review process can be very negative and carefully describing limits on a study's external validity, while useful, may reduce the potential for reviewers to recommend the manuscript be accepted for publication. Together, these factors may discourage authors from engaging in detailed discussion of external validity.

A lack of attention to external validity in published SLA research can have significant consequences, particularly through the misinterpretation or overgeneralization of findings. Such limitations are often compounded when claims are restated in subsequent publications. Research on lexical coverage and reading comprehension provides a clear example. Lexical coverage—the proportion of known words in a text—is among the most frequently discussed constructs in SLA; the five most cited studies of the topic (Hu & Nation, 2000; Laufer, 1989, 1992; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011) have together accrued over 7,000 citations (Google Scholar, March 22, 2026). It is therefore unsurprising that the claim that readers can understand a text when they know approximately 98% of its words has become widely accepted in the literature. However, empirical support for this claim is considerably weaker than is often assumed, particularly when evaluated across diverse contexts.

Early studies suggested that readers might achieve comprehension at coverage levels of 95% (Laufer, 1989) or 98% (Hu & Nation, 2000), but subsequent research has shown that lexical coverage alone is a weak predictor of reading comprehension. Schmitt et al. (2011) concluded that while high coverage is necessary, it is insufficient, and that individual comprehension remains difficult to predict. More recent partial replications have found little evidence for a clear lexical coverage threshold and indicate that results are sensitive to learner characteristics, text features, and the measures used to assess comprehension (Kremmel et al., 2023; Webb et al., 2025). More importantly, the external validity of this body of research is highly constrained. Confidence in the generalizability of these findings would require evidence that the texts, learners, lexical items, assessments, and procedures employed were representative of the wide range of conditions under which L2 reading occurs. This includes variation in text length, genre, difficulty, and topic; learner proficiency, background knowledge, and reading experience; lexical characteristics and densities of unknown words; the quality and sensitivity of comprehension measures; and the ecological validity of research procedures. Even if consistent results were observed across these studies, generalizability would remain limited. Collectively, this body of research has examined comprehension using only five short texts, despite the many text-level variables known to influence reading comprehension. Although these studies clearly demonstrate the importance of vocabulary knowledge, they also underscore the need for substantially more research before lexical coverage can be considered a reliable predictor of reading comprehension across contexts. Moreover, these limitations in external validity raise wider concerns about how findings are interpreted and reported beyond the original studies.

More broadly, limited attention to external validity is evident in how research findings are reported in the SLA literature. Reviewers frequently suggest expanding the limitations sections of submitted manuscripts. However, very few published studies comprehensively address limitations in external validity. Limitations that are often provided tend to note that the findings should not be generalized to a wider and more diverse learner population. This is less likely to negatively influence the potential for

https://doi.org/10.1017/S027226312610179X Published online by Cambridge University Press

publication because individual studies are unlikely to extend to diverse learner populations and address the large number of participant variables (e.g., age, proficiency, background knowledge, L1 background, L2 instruction). In contrast, if authors acknowledge limited generalizability across target items, materials, tests, learning conditions, and procedures, reviewers may recommend improving the research design, additional data collection, or rejecting the study for publication.

Researchers who have been trained in programs that place strong emphasis on research design and methodological rigor may be less influenced by the limited attention given to external validity in published studies. However, access to such training is unevenly distributed, with most leading programs concentrated in a relatively small number of WEIRD (Western, Educated, Industrialized, Rich, and Democratic) contexts. As a result, research communities that may already have fewer institutional resources or mentoring opportunities may be disproportionately affected. Novice researchers may have difficulty identifying limitations in external validity and consequently misinterpret research findings. They may also be less aware of opportunities to design well-justified follow-up studies that systematically test the generalizability of earlier results. Increased attention to external validity in research reporting would support more accurate interpretation of findings and help identify clear and methodologically motivated pathways for further investigation. Moreover, explicitly framing follow-up studies in terms of extending external validity, whether through formal replication studies or otherwise, may reduce barriers to conducting new research by providing a transparent and compelling rationale grounded in both the original findings and their limitations.

Established researchers often use limitations in the external validity of their findings as the motivation for subsequent studies. A clear example of this approach can be seen in a series of studies on the comprehensibility of L2 speech conducted by Kazuya Saito, Pavel Trofimovich, Talia Isaacs, and colleagues. In an early study, Saito, Trofimovich, and Isaacs (2016) examined how native speakers evaluated the comprehensibility and accentedness of L2 learner speech. Recognizing both the substantial societal and pedagogical relevance of L2 speech comprehensibility and the possibility that findings based on native speaker ratings might not generalize to other populations, the researchers conducted a set of theoretically motivated follow-up studies. These studies systematically examined whether the original findings extended across different rater groups, including L2 learners (Saito et al., 2019), expert and non-expert raters (Saito, Trofimovich, Isaacs, & Webb, 2016), and native speakers of North American English and Singapore English (Saito & Shintani, 2016). This line of research illustrates the importance of not assuming that findings related to L2 speech comprehensibility generalize across populations. Similar research programs aimed at extending external validity can be seen in studies of interactional feedback (e.g., Nassaji, 2007, 2009, 2017; Sarandi & Nassaji, 2024), errorless learning (e.g., Boers, Yu, & Wang, 2025; Elgort, Baliaeva, & Boers, 2020; Stengers & Boers, 2015), and lexical profiling (e.g., Dang & Webb, 2014; Paribakht & Webb, 2016; Webb, 2011). Thus, recognizing limits in the external validity of prior research can help novice researchers formulate clearly motivated replication or follow-up studies that extend findings beyond the constraints of the original designs.

In recent years, three initiatives have been promoted as ways of increasing the external validity of studies in SLA: replication studies, meta-analytic studies, and multisite studies. Calls for replication emphasize the need to examine the extent to which research findings are consistent across repeated investigations (e.g., Marsden et al., 2018; McManus, 2024; Porte, 2012; Porte & McManus, 2019). Meta-analyses

https://doi.org/10.1017/S027226312610179X Published online by Cambridge University Press

reveal the extent to which external validity holds across learner, material, learning condition, measurement, context, and procedure variables. Notably, perhaps the most common conclusion across meta-analyses is the need for further research to increase external validity across diverse variables. Multisite studies, in contrast, focus primarily on determining whether findings generalize across different learner samples (e.g., Morgan-Short et al., 2018; Vitta et al., 2022). Increasing generalizability across participant groups is important because it raises the likelihood that findings with one group of learners also occur in other populations and reduces the extent to which findings are driven by characteristics of a single participant sample. While this represents an important advance, multisite studies typically address only learner variables and do not, on their own, rule out the influence of other design features such as target item variables, material variables, dependent measures, or procedural variables. Consequently, although replication, meta-analyses, and multisite studies make valuable contributions to understanding generalizability, they do not on their own resolve the broader issue of external validity.

Moving towards increased external validity in SLA research

There are several ways in which the external validity of studies of SLA could be made more transparent. First and most importantly, there is a need for authors to include comprehensive discussion of the external validity of their findings. This is easier said than done because authors are understandably wary of describing how their findings may not extend beyond the constraints of their research design; those who do this could be penalized by reviewers while those who do not might find an easier pathway to publication. In order to achieve a fair and transparent process that benefits authors and the research community, it may be necessary to have required limitations sections that include focus on learner, material, target item, learning condition, test, and procedural variables. If this were a requirement in submission guidelines, then both authors and reviewers could follow a process benefitting the wider research community that was fair and transparent. Table 1 provides an initial framework that could be used, with modifications as needed, to guide evaluations of external validity.

A second way to improve the external validity of studies of SLA would be to continue to advocate for more replication studies. The journal *Language Teaching* has long included replication as a category for submission and *Language Learning* and *Studies in Second Language Acquisition* also publish replication studies. *Studies in Second Language Acquisition* recently put out a call for registered report replications of five studies with an aim to publish multiple replications of the same study. This useful initiative may reveal both the need to further examine the robustness and credibility of earlier findings, as well as the many ways in which one or two variables from one study can be modified to test the external validity of earlier findings.

A third way in which external validity can be increased is to continue to advocate for multisite studies. It is important to note, however, that while multisite studies do help to extend the generalizability of findings to different populations, generalizability across learner populations is still likely to be limited. This is because it would be difficult to account for variation across all learner variables (e.g., age, proficiency, L1 background, years of formal language instruction, previous exposure to the L2, vocabulary knowledge, working memory, metalinguistic awareness, learning strategies, motivation) in a single study. In addition, multisite studies do not account for variation across target items, materials, learning conditions, tests, and procedures. For this reason, we should

https://doi.org/10.1017/S027226312610179X Published online by Cambridge University Press

Table 1. Framework for Evaluating the External Validity of SLA Research Designs

Design Variable	Key Considerations for External Validity
Learner variables	Transferability across learners differing in age, proficiency, L1 background, amount of formal instruction, exposure to the L2, metalinguistic awareness, learning strategies, and motivation.
Material variables	Transferability across instructional materials differing in length, topic, genre, intended audience, difficulty level, topic familiarity, and mode of input (e.g., spoken, written, multimodal).
Interactional and learning-condition variables	Transferability across interactional conditions differing in length, frequency, learning context, dyadic or group configuration, mode (speaking, listening, reading, writing), and degree of meaningful engagement.
Target item variables	Transferability across target items differing in form, meaning, frequency of exposure, difficulty, salience, and degree of L1–L2 formal or semantic congruency.
Measurement variables	Transferability across assessment formats differing in length, task type, sensitivity to knowledge, modes of input and output, scoring criteria, and time allocated per item.
Contextual and procedural variables	Transferability across instructional contexts (e.g., classroom vs. out-of-class learning; EFL, ESL, ESP, EAP) and procedures, including instructional time, task instructions, level of instructional support, and number of exemplars.

also advocate for research that increases external validity through research designs that include more than one material (multimaterial studies), more than one type of target item (multitarget item type studies), and more than one learning procedure (variation could occur across procedural variables such as amount of L2 exposure/use, number of sessions, time on task, learning support provided: multiprocedure studies). Although studies often include multiple tests and learning conditions, greater emphasis on a wider range of learning conditions and measurement across different test formats would be beneficial. For example, Morgan and Rinvolucri (2004) described 118 activities for vocabulary learning, and Webb and Nation (2017) outlined 23 activities that they reported were effective for learning. Nevertheless, a meta-analysis on word-focused vocabulary found that there were only four activities that met inclusion requirements (Webb et al., 2020). One of the key values of meta-analyses is their ability to highlight such gaps. Taken together, this underscores the need for cautious interpretation of the external validity of research findings; while there are typically many ways to learn an L2, researchers have investigated relatively few learning conditions with most studies including the optimal conditions for learning with an aim towards finding significant results.

Conclusion

The aim of this editorial was to raise awareness of the extent to which studies of SLA are externally valid. External validity tends to be lacking in a single study or across a small number of studies because the validity of results may not extend beyond the constraints of research designs. Because there are many learner, material, target item, test, and procedural variables that may affect results, readers should always be cautious in their interpretation of findings. Moreover, there should be awareness of the need for further investigation to extend the findings beyond the parameters of original research designs.

https://doi.org/10.1017/S027226312610179X Published online by Cambridge University Press

For this reason, externally valid answers to research questions typically require evidence accumulated across many studies.

Acknowledgments. I am grateful to Luke Plonsky and Pavel Trofimovich for their helpful feedback on this topic.

References

Boers, F., Yu, X., & Wang, X. (2025). Meaning-inferencing versus meaning-given procedures: The case of idioms. *Language Learning*, *75*(4), 1019–1050.

Dang, T.N.Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*(1), 66–76.

Elgort, I., Beliaeva, N., & Boers, F. (2020). Contextual word learning in the first and second language: Definition placement and inference error effects on declarative and nondeclarative knowledge. *Studies in Second Language Acquisition*, *42*(1), 7–32.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403–430. http://hdl.handle.net/10125/66973

Kremmel, B., Indrarathne, B., Kormos, J., & Suzuki, S. (2023). Unknown vocabulary density and reading comprehension: Replicating Hu and Nation (2000). *Language Learning*, *73*(4), 1127–1163. https://doi.org/10.1111/lang.12622

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P.J.L. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics*. Palgrave Macmillan. https://doi.org/10.1007/978-1-349-12396-4_12

Laufer, B., & Ravenhorst-Kalovski, G.C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30. https://doi.org/10.64152/10125/66648

Marsden, E., 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*(2), 321–391.

McManus, K. (2024). The future of replication in applied linguistics: Toward a standard for replication studies. *Annual Review of Applied Linguistics*, *44*, 133–149. doi:10.1017/S0267190524000011.

Morgan, J., & Rinvolucri, M. (2004). *Vocabulary*. Oxford University Press.

Morgan-Short, K., Marsden, E., Heil, J., Issa Ii, B. I., Leow, R. P., Mikhaylova, A., Mikołajczak, S., Moreno, N., Slabakova, R., & Szudarski, P. (2018). Multisite replication in second language acquisition research: Attention to form during listening and reading comprehension. *Language Learning*, *68*(2), 392–437.

Nassaji, H. (2007). Elicitation and reformulation and their relationship with learner repair in dyadic interaction. *Language Learning*, *57*(4), 511–548.

Nassaji, H. (2009). Effects of recasts and elicitations in dyadic interaction and the role of feedback explicitness. *Language Learning*, *59*(2), 411–452.

Nassaji, H. (2017). The effectiveness of extensive versus intensive recasts for learning L2 grammar. *The Modern Language Journal*, *101*(2), 353–368.

Paribakht, T.S., & Webb, S. (2016). The relationship between academic vocabulary coverage and scores on a standardized English proficiency test. *Journal of English for Academic Purposes*, *21*, 121–132.

Porte, G. K. (Ed.) (2012). *Replication research in applied linguistics*. Cambridge University Press.

Porte, G. K., & McManus K. (2019). *Doing replication research in applied linguistics*. Routledge.

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive comprehensibility in second language speech? *TESOL Quarterly*, *50*(2), 421–446.

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. *Studies in Second Language Acquisition*, *41*(5), 1133–1149.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217–240.

https://doi.org/10.1017/S027226312610179X Published online by Cambridge University Press

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2016). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 141–156). Multilingual Matters.

Sarandi, H., & Nassaji, H. (2024). Recast frequency and the acquisition of English articles in a computer-mediated context. *Studies in Second Language Learning and Teaching*, *14*(3), 571–602.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Stengers, H., & Boers, F. (2015). Exercises on collocations: A comparison of trial-and-error and exemplar-guided procedures. *Journal of Spanish Language Teaching*, *2*(2), 152–164.

Vitta, J. P., Nicklin, C., & McLean, S. (2022). Effect size–driven sample-size planning, randomization, and multisite use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition*, *44*(5), 1424–1448.

Webb, S. (2011). Selecting television programs for language learning: Investigating television programs from the same genre. *International Journal of English Studies*, *11*(1), 117–136.

Webb, S., & Nation, P. (2017). *How Vocabulary is Learned*. Oxford University Press.

Webb, S., Pellicer-Sánchez, A., & Wang, A. (2025). Lexical coverage and reading comprehension revisited. *Reading in a Foreign Language*, *37*(1), 1–21. https://doi.org/10.64152/10125/67517

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary learning activities? A meta-analysis. *Modern Language Journal*, *104*(4), 715–738.

Cite this article: Webb, S. (2026). A call for greater attention to the external validity of studies of second language acquisition. *Studies in Second Language Acquisition,* 1–8. https://doi.org/10.1017/S027226312610179X

https://doi.org/10.1017/S027226312610179X Published online by Cambridge University Press