SSLA, **19**, 499–508. Printed in the United States of America.

REPLICATION AND REPORTING

A Commentary

Charlene Polio and Susan Gass
Michigan State University

In this paper we address the need for replication studies in the field of second language acquisition and discuss the problems surrounding standards of reporting research. A particular focus is on the level of detail necessary for replication to take place. The lack of uniform standards is noted with specific examples of problematic reporting taken from descriptions of second language learners' proficiency levels and measures of linguistic accuracy in second language writing research. We propose ways that we believe will lead to more thorough reporting of research and that will therefore allow others to engage in replication.

One of the issues that researchers, editors, and referees must face in addition to standards of quality in research is standards of reporting such research. This latter issue is central to any discussion of replication because one cannot replicate a study without a certain amount of information about the original study. Reporting standards and expectations, however, are not uniform. For example, one of the authors, when writing an article from her dissertation, referred readers to the dissertation for details about the study, but she was told by a reviewer that this was not appropriate; all information necessary to replicate her study had to be included in

This is a revised version of a paper presented at TESOL 1996 in a colloquium organized by Margaret van Naerssen titled "Professional ethics in TESOL-related research."

We are grateful to Cathy Fleck, Nevin Leder, and Ildiko Svetics for perceptive comments on an earlier draft of this paper. Many of their comments have been incorporated into the body of this paper, but even more important were the cogent discussions we had with them on the contents of this paper. We owe them a debt of gratitude for helping us to clarify our thinking on certain aspects of this paper. Alison Mackey also read an earlier version and provided us with comments that allowed us to amplify some points. Unfortunately, we cannot blame any of these individuals for any conceptual or content errors that remain. They belong uniquely to us.

Address correspondence to: Charlene Polio and Susan Gass, English Language Center, Michigan State University, A714 Wells Hall, E. Lansing, MI 48824-1027. e-mail: polio@pilot.msu.edu or gass@pilot.msu.edu.

© 1997 Cambridge University Press 0272-2631/97 $7.50 + .10

https://doi.org/10.1017/S027226319700404X Published online by Cambridge University Press

an article. In a review for a different article in which the same author lamented the lack of information available about a particular measure, a reviewer suggested that it might be unrealistic to expect an author to include enough information for replication in the article; journal space, the reviewer pointed out, is limited. Given these contradictory comments, as well as our own observations, we believe this topic to be worthy of discussion, timely, and to some extent, controversial.

Discussions of replication have received little attention in the field of second language acquisition and, one might venture to say, have been considered relatively unimportant. To some extent, this may have been an acceptable state of affairs in the early years when our field was attempting to identify itself as a valid discipline and valid area of study. In fact, some "chaos" in the form of nongrounded speculation is often typical of early disciplinary work (e.g., alchemy led to chemistry, theosophy to philosophy). However, the search for a discipline, characteristic of earlier years, is now behind us. The scope of inquiry is now reasonably well defined; research questions have been identified and accepted; accepted methodologies exist. Hence, we are now at the point where it seems appropriate and necessary to consider issues surrounding replication more carefully. Replication is, of course, crucial in order to distinguish the spurious from the real and will, undoubtedly, become even more central to our field as we attempt to push our theories forward. Although we can safely say that there are many findings that are part of the current orthodoxy, there are others that are still in debate. Further, in retrospect, it seems that some of the so-called facts of our field have been determined on the basis of scanty evidence.

Within the field of psychology, a discipline that relies heavily on empirical work, the necessity of replication is clearly stated in the *Publication Manual of the American Psychological Association* (1994, p. 292): "The essence of the scientific method involves observations that can be repeated and verified by others." Because SLA is a multidisciplinary field that incorporates research methodologies from different disciplines, the necessity of replication is not uniformly recognized. For example, as an anonymous reviewer of this article pointed out, linguistic research is typically not based on replication in the same way as is conceived of in psychology. As a result, SLA researchers trained primarily as linguists may not have the same concern for replication as those trained in psychology, a discipline that has a tradition of replication. This leads to different views on the relative value of and need for replication studies.

However, because there have not been many replication studies does not mean that there has been no interest or recognition of the need. Santos (1989), in an early article on the need for replication in applied linguistic research, pointed out that "[w]hat is considered standard procedure in other disciplines that hypothesize, quantify, and generalize is ignored in ours" (p. 699). In the field of SLA, a lack of attention to replication is particularly striking when one considers that many published studies base generalizations on relatively few subjects.

Recognizing the significance of replication to a discipline, Albert Valdman pioneered the way toward placing greater value on such studies with the introduction in 1993 of this section in *Studies in Second Language Acquisition*.[2] He stated that

https://doi.org/10.1017/S027226319700404X Published online by Cambridge University Press

"the way to more valid and reliable SLA research is through replication" (p. 505), and importantly, he also added, and perhaps this is the major reason for the lack of replication studies, "to be sure, in replication one loses the aura of glamour and the exhilaration of innovation" (p. 505), for original research is what pushes the boundaries of a field. It is likely that candidates who present themselves for tenure with a record of replication studies will find considerable opposition from colleagues wondering if they can produce "original research." As van der Veer, van Ijzendoorn, and Valsiner (1994, p. 6) noted: "As these replication studies do not yield novelty, but rather check the reliability of the original results, they are less valued in a community where (limited) originality is highly valued." Furthermore, Dewald, Thursby, and Anderson (1986) argued that researchers may avoid replications for fear that such studies may be interpreted as a personal attack on or a lack of trust in a colleague.

Others, particularly in the field of psychology, have noted the bias against publication of studies that are replications (Neulip & Crandall, 1990, 1993a). However, Neulip and Crandall (1993b) also found that many more studies than claim to be are actually some type of replication. And, clearly, a fine line exists between replication and extending research.[3]

Most scholars involved in empirical, quantitative research will not dispute that replication is necessary but will acknowledge that researchers do not often engage in replication studies for the reasons given above. But there is a pragmatic issue to which we turn and which is the focus of this paper: What information should be reported in a published article to allow others to replicate the study?

The most appropriate place to begin to answer this question is the *Publication Manual of the American Psychological Association.* Unfortunately, it simply states that one should report "only the information essential to comprehend and replicate the study" (p. 13). Similarly, if we look to applied linguistics books on research design, we find little information. For example, Brown (1988) states only, "if you are not given enough information to replicate a study, the author has not explained it adequately" (p. 43).

Before we continue, we should clarify what it means to replicate a study. Within the hard sciences (e.g., physics, chemistry), one can imagine what it means to replicate a study: All of the physical conditions are identical to those of the original study. And, provided that all of the information is available, this seems reasonably straightforward. For the social sciences, the situation is less clear because, due to variability from individual to individual, one cannot create identical conditions, making exact repetition virtually impossible. Replication in SLA is even more difficult than replication in, for example, psycholinguistics, because in the latter one is dealing with issues of native language knowledge and processing, the assumption being that adult native language knowledge does not vary much from individual to individual. For SLA studies, on the other hand, researchers must pay particular attention to the distinct knowledge that each individual has. We return to this point below.

van der Veer et al. (1994) and Hendrick (1990) discussed a continuum of replication types from virtual to conceptual. Virtual replications, in which one tries to copy the original research as closely as possible, serve the purpose of confirming that the

https://doi.org/10.1017/S027226319700404X Published online by Cambridge University Press

original results were reliable. Within the field of SLA, for reasons stated above, even virtual replications are difficult, if not impossible. A virtual replication would include, *inter alia,* identical stimuli, identical conditions, and identical subjects (however defined). If we assume that experimental stimuli can remain constant, and if we further assume that experimental conditions can be duplicated, we are still left with the most serious difficulty: Is it possible to have comparable subjects? The answer is clearly a negative one because, as discussed earlier, we have no adequate measure for determining language knowledge with any degree of precision. Although we have valid and reliable tests, as we will discuss later, it is probable that no two randomly selected second language learners, even those with the same L1 and roughly the same L2 proficiency, have knowledge that can be considered as close as two randomly selected adult native speakers.

But there are other ways of viewing replication; one such way is through conceptual replication. Conceptual replications alter various features of the original study and serve the purpose of confirming the generalizability or external validity of the research. Researchers will attempt replication to see if the results hold for a different population, in a different setting, or for a different modality (e.g., Leow, 1995).

To attempt a virtual replication, a great deal of information on a study is needed. However, to attempt a conceptual replication, one needs only to consider the claims of the research and how one might want to make changes to the original study to see if the claims hold up in a new situation. In principle, a conceptual replication may require less information. However, a problem may arise in the interpretation of the results: If the results are not the same in the replication as those in the original, one needs detailed information on the original study to determine why.[4] Were the original results merely spurious or is there something in the methodology or subject population that differed significantly?

The issue of proficiency and the need to be able to determine the equivalence of subject populations is of pressing concern when considering issues of replication. In a review article on proficiency level in SLA research, Thomas (1994) discussed the inadequate assessment of L2 proficiency in several SLA studies. She actually went a step further saying that researchers do not even collect enough information, let alone report enough information. In some cases, the assessment is impressionistic with researchers reporting something like: The subjects "were more or less beginners" (p. 314). In other cases, according to Thomas, subjects' proficiency is reported in terms of placement into a certain class level within the context of a particular intensive language program; and in others, we may learn that subjects are placed in the third level of a four-level system, but we know nothing about what level four is like: Is it prematriculation (i.e., a TOEFL[5] score of approximately 530)? Or are the students already matriculated (perhaps a TOEFL score greater than 550)? In such instances, it may be the case that a reliable and valid placement test has been given, yet one rarely sees mention of such tests in research articles. Thomas argued that even when such information may be available, researchers are not consistent in the degree to which they provide descriptions of their subjects.[6] Sorace (personal communication, January 13, 1997) suggested that researchers provide specific and accurate information about the particular measure used to determine degree of

https://doi.org/10.1017/S027226319700404X Published online by Cambridge University Press

second language knowledge so that other researchers can avail themselves of that same measure. This may or may not be an option. In some cases measures of proficiency lack precision. For example, two individuals with a TOEFL score of 520 may have very different sorts of knowledge, not only across the skill levels tested, but also within each component. Thus, these general proficiency measures constitute a rough equivalence at best and may or may not incorporate information about specific targeted grammatical structures. Additionally, many of the measures are "in-house" placement tests, unique to a particular setting and, hence, for reasons of security they cannot be made publicly available.

When, on the other hand, one is focusing on a particular structure, it is easier to imagine how information could be provided so as to allow for future replication. Schachter (1989, 1990), for example, in her studies on the principle of Subjacency, gave pretests to subjects on just the prerequisite information needed for her study. When such pretests are made available, one then has a measure with which to determine relevant equivalence even though overall equivalence might be lacking. As Thomas pointed out, we need to be particularly alert to these difficulties and, as a result, be vigilant in our reporting requirements.

With regard to research in other fields, King (1995) described an unusual study funded by the National Science Foundation (Dewald et al., 1986) in which the grant recipients attempted to replicate every study published in an economics journal in the space of a year. They found that "Replication was often impossible even with the help of the original author—and help from the authors was often not provided" (King, 1995, p. 445). King (p. 444) also lamented the lack of adherence to the replication standard that he claims "holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information."

We have argued for the centrality of replication studies to validate research findings. We have also pointed out some of the difficulties involved in actually conducting replication studies. We now turn to a detailed discussion of reporting. If, as a field, we are to adequately conduct replication studies, original research must be sufficiently detailed so as to allow others to determine with precision what has been done. So the question is: How much information should one present? Although we have generally noted that too little is reported, there are also occasional instances in which too much is presented. In a recent paper we observed that the researcher included the type of tape recorder used. Although this may be crucial to a study on speech perception or production, for which it must be ascertained that the recording equipment is sufficient for the task at hand, it is probably unnecessary for a study in which subjects have to respond with simple utterances to specific prompts. But such reporting of seemingly irrelevant details is rare, and the irony is that although the *APA Manual* warns against reporting too much detail, we would argue that it is too little rather than too much information that is presented. What follows is a detailed discussion of a particular construct and how various researchers have reported measuring such a construct in the literature.

Polio (1997) reviewed several published studies in the applied linguistics literature that had linguistic accuracy in L2 writing as a dependent variable. After reviewing

https://doi.org/10.1017/S027226319700404X Published online by Cambridge University Press

16 articles, she found that they could be grouped into three categories: (a) a holistic measure of accuracy, (b) an error-free unit (such as a clause or T-unit), and (c) the number of errors (with or without classification of errors). She examined the articles to see if sufficient information existed to enable her to use the same measure in another study.

All the studies in the holistic measure category simply had a copy of the rating scale included in their appendices, thus enabling anyone to use or modify the instrument. No ambiguity about how much information to report was apparent. There were nine studies that counted the number of errors. We will not go into detail because of the large number; we will, however, focus specifically on three studies that used error-free T-units, or error-free clauses.

Defining T-units and clauses can be done in a few sentences, and few disagreements about these categories arose in Polio's (1997) study. Defining "error" is much more problematic. When attempting to count "error-free T-units," Polio and her assistant came up with two single-spaced pages of guidelines. To what extent should these guidelines be reported? We turn now to the three studies that used error-free units to see how reporting was done.

Robb, Ross, and Shortreed (1986) examined the effect of four different kinds of feedback on EFL students' essays. There is no detailed discussion of how they identified an error-free unit.

Casanave (1994) documented changes in students' writing over three semesters of an intensive English program. She included a footnote regarding what constituted an error: "I did not count spelling or typing mistakes as error, but did count word endings, articles, prepositions, word usage, and tense. In a few cases it was difficult to determine whether the writer had made an error or not" (pp. 199–200).

Finally, Ishikawa (1995) was concerned with the effect of two different types of writing practice tasks on the writing of low-proficiency EFL students. Although she acknowledged that determining correctness was difficult, she gave far more detail than most researchers on how she coded her data. For example, she specifically stated that she did not count punctuation except at sentence boundaries and that spelling was disregarded unless it involved a grammatical marker (cf. Casanave's statement above). She explained that when more than one tense was used, the most common was considered correct and that in cases of ambiguity, students were given the benefit of the doubt. Furthermore, she stated that correctness was determined "with respect to discourse, vocabulary, grammar, and style, and strictly interpreted" (p. 59). She also noted that a sentence or clause was considered in context; its correctness was not considered in isolation, but rather as part of the discourse.

Ishikawa went into further detail, and although one may not agree with all of her decisions, the important point here is that anyone reading her study has a good sense of how she handled correctness. A relevant issue is how much journal space such detail covers. In Ishikawa's article it was half of a journal page. Thus, the amount of detail in the three studies ranged from no information to one-half page. One needs to consider which is appropriate. From the perspective of someone who is attempting to conduct research, it is clear that the more information available, the better. A great amount of detail in the literature helps other researchers save time

https://doi.org/10.1017/S027226319700404X Published online by Cambridge University Press

by bringing up problems that might not occur to them until later. In other words, it helps prevent others from reinventing the wheel. In this particular case, it was helpful to learn from Ishikawa that others were grappling with problems of identifying an error, something not evident in most of the other articles.

From an epistemological point of view one needs to consider the extent to which this is important. For example, Ishikawa even described how she counted a word in calculating words per error-free T-unit. (Is a contraction one word or two? If a student writes "grandmother" as two words, how do we count that?) Anyone conducting research has to come up with a solution to these problems. But in the end, does it matter if study A counts contractions as one word and study B counts them as two words? In terms of the internal validity of a study, reliability is what is important. If two researchers can obtain reliability on a measure, then that should be sufficient and more detailed guidelines are not needed. We should note, however, that of the 16 studies reviewed, more than half did not report interrater reliability, and a lack of reliability can mask significant group differences.

If one attempts to use a measure to replicate a study in another setting, slightly different definitions of error may be important in some cases, but not in others. For example, assume two studies examining the effects of peer-editing on the linguistic accuracy of ESL students' essays: One group of students receives no feedback, another receives correction from peers. If one study counts spelling errors and another does not, there could be different results because spelling errors may be easier errors for peers to correct. Spelling errors are also more prevalent among some L1 groups and within particular proficiency levels. Thus, a change in subjects could affect the results. On the other hand, in Polio's study (1997), the coders were struggling with sentences that were ungrammatical in American English, but not in British English. Even though this came up in several essays, it is unlikely that the difference between counting one way or the other would have been significant.

What we have done in this brief paper is identify issues of reporting surrounding replication. Given these problems, we will attempt to articulate some solutions to the problems of standards and consistency in describing research.

First, journal editors and referees should pay closer attention to this issue. We suspect, however, that one can probably not truly judge the appropriate level of detail until one tries to replicate a study oneself. As we have learned in our own research, it is often difficult to know how much to report. This leads to our second suggestion.

Second, faculty should assume the responsibility of encouraging graduate students to conduct replication studies. This could quite easily be done by building such a requirement into course syllabi, particularly in courses dealing with research methods. More replication is an important step toward the creation of standards for reporting research that is replicable. Neulip and Crandall (1993a) suggested reserving space in journals, as is done in *Studies in Second Language Acquisition*, for replications. An increase in the number of such studies should make scholars more aware of some of the problems related to the reporting of research and replication. And, should such replication studies have theoretical implications for the field and should they ultimately lead to ground-breaking changes in our thinking, we predict that

https://doi.org/10.1017/S027226319700404X Published online by Cambridge University Press

there would be a positive effect for the overall value of such studies with regard to fundamental issues relating to tenure and promotion.

Third, in general, we would like to see more detail reported in journal articles. Editors and publishers, on the other hand, are concerned about journal space. However, as noted earlier, detailed guidelines on error in Ishikawa's (1995) article took up half of a page—not an overwhelming amount of space. An alternative is to add details in footnotes and appendices, which take up less space because they are in a smaller point size. If none of these options is selected, authors can add that further information can be obtained by contacting the authors. (This was not done in any of the 16 studies reviewed in Polio [1997].)

Finally, a more radical solution is to require detailed appendices from researchers who submit articles for publication. Such appendices would be reviewed as part of the review process. This information would then be kept by the journal (perhaps on a web site, but also available in hard copy for those who do not have Internet access). King (1995) went a step further by suggesting that political scientists make use of already established digital archives in the field to allow others access not only to detailed procedures regarding data collection and analysis, but also to the data themselves. Similarly, Dewald et al. (1986) suggested that economists submit data and computer programs used to analyze the data.

If one were to capitalize on this suggestion and submit detailed appendices, what might be included? We list here examples of what might ultimately be useful to future researchers: (a) Detailed guidelines and examples of coding categories, (b) A listing of examples that were excluded from consideration, (c) Measures of proficiency (descriptions of tests where test security is a problem), (d) Instruments for data elicitation, including pre-tests and post-tests, (e) Experimental protocols and instructions to subjects, and (f) Demographic background of subjects.

We believe that three benefits of requiring such information are apparent: First, more replication would take place. Second, researchers would save time because measures, elicitation tasks, and coding guidelines would already be available. And third, research might be conducted more carefully. Dewald et al. (1986) pointed out that the very act of preparing one's data and programs for others leads to more careful research and the discovery of inadvertent errors.

We hope that this discussion of replication will lead the field to consider these issues more seriously and, in particular, will lead others involved in replication to flesh out more thoroughly and precisely just what needs to be reported to allow others to attempt replication of existing studies.

(*Received 17 July 1996*)

NOTES

1. As an example, consider the case of Schumann's study of Alberto (1975) and Schmidt's study of Wes (1983). The conclusions of the two studies are contradictory, but, notably, each is based only on one subject. Certainly one cannot replicate these studies in the traditional sense of experimental replication; nevertheless one needs to consider the implications of serious conclusions, if not theories, being based on limited data. Clearly, this is also an issue of types of data and methodologies that are used in the fields and the conclusions that one can legitimately draw from different data types.

https://doi.org/10.1017/S027226319700404X Published online by Cambridge University Press

2. In Valdman's introduction to this new section he noted that the idea went back to 15 years earlier when Stephen Krashen had made a similar suggestion to him. But, as we said earlier, perhaps the field had to develop its own body of knowledge and its own methodological uniqueness (e.g., our own understanding of acceptable techniques for gathering data) before this possibility would be viable. That is, until we know enough, it is not particularly useful to begin replication at the expense of pushing our knowledge of the field forward.

3. A case in point is a 1982 article published by Gass and a 1988 study by Eckman, Bell, and Nelson. The original study was a study of the acquisition of relative clauses by learners in a classroom context. The subject pool was small. The second study tested the same principles, but it did so with a larger database and in a slightly different way. The studies are often cited together with the second one being pointed to as corroborating the results of the first. Depending on one's definition of replication (e.g., conceptual, exact, or virtual), the Eckman et al. study may or may not be a replication of the Gass study.

4. Hendrick (1990) listed several factors that can influence the outcome of a conceptual replication. A partial list of such factors in SLA research could include the following: subject characteristics (L1 and other L2s, L2 proficiency, L2 exposure, L2 instruction, gender, education level, and age); the L2 itself; social-historical-cultural context; elicitation techniques (modality, setting, instrument), and so on.

5. We are sidestepping the important issue of the "North Americancentricity" of the TOEFL. Nonetheless, readily available information on that particular standard test exists.

6. In looking through the literature on replication in other fields, there is still little written about how much to present. We did locate two articles from the area of learning disabilities (Wolery & Ezell, 1993; Rosenberg, Bott, Majsterek, Chiang, Gartland, Wesson, Graham, Smith-Myles, Miller, Swanson, Bender, Rivera, & Wilson, 1992) in which it was argued that without standards for describing subjects, the source of failure in replication will be unknown.

REFERENCES

American Psychological Association. (1994). *Publication Manual of the American Psychological Association* (4th ed.). Washington, DC: American Psychological Association.

Brown, J. D. (1988). *Understanding research in second language learning.* Cambridge: Cambridge University Press.

Casanave, C. (1994). Language development in students' journals. *Journal of Second Language Writing, 3,* 179–201.

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: *The Journal of Money, Credit, and Banking* project. *American Economic Review, 76,* 587–603.

Eckman, F., Bell, L., & Nelson, D. (1988). On the generalization of relative clause instruction in the acquisition of English as a second language. *Applied Linguistics, 9,* 1–20.

Gass, S. (1982). From theory to practice. In M. Hines & W. Rutherford (Eds.), *On TESOL '81* (pp. 129–139). Washington, DC: Teachers of English to Speakers of Other Languages.

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: Are they important? *Journal of Social Behavior and Personality, 5,* 41–49.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4,* 51–70.

King, G. (1995). Replication, replication. *PS: Political Science and Politics, 28,* 444–452.

Leow, R. (1995). Modality and intake and second language acquisition. *Studies in Second Language Acquisition, 17,* 79–90.

Neulip, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5,* 85–90.

Neulip, J. W., & Crandall, R. (1993a). Reviewer bias against replication research. *Journal of Social Behavior and Personality, 8,* 21–29.

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

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47,* 101–143.

Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly, 20,* 83–96.

Rosenberg, M. S., Bott, D., Majsterek, D., Chiang, B., Gartland, D., Wesson, C., Graham, S., Smith-Myles, B., Miller, M., Swanson, H. L., Bender, W., Rivera, D., & Wilson, R. (1992). Minimum standard for the description of participants in learning disabilities research. *Learning Disabilities Quarterly, 15,* 65–70.

Santos, T. (1989). Replication in applied linguistics research. *TESOL Quarterly, 23,* 699–702.

https://doi.org/10.1017/S027226319700404X Published online by Cambridge University Press

Schachter, J. (1989). Testing a proposed universal. In S. Gass & J. Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 73–88). Cambridge: Cambridge University Press.

Schachter, J. (1990). On the issue of completeness in second language acquisition. *Second Language Research, 6,* 93–124.

Schmidt, R. (1983). Interaction, acculturation, and acquisition of communicative competence. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and second language acquisition* (pp. 137–174). Rowley, MA: Newbury House.

Schumann, J. (1975). Affective factors and the problem of age in second language acquisition. *Language Learning, 25,* 209–235.

Thomas, M. (1994). Review article: Assessment of L2 proficiency in second language acquisition research. *Language Learning, 44,* 307–336.

Valdman, A. (1993). Replication study. *Studies in Second Language Acquisition, 15,* 505.

van der Veer, R., van Ijzendorn, M., & Valsiner, J. (1994). General introduction. In R. van der Veer, M. van Ijzendorn, & J. Valsiner (Eds.), *Reconstructing the mind: Replicability in research on human development.* Norwood, NJ: Ablex.

Wolery, M., & Ezell, H. K. (1993). Subject descriptions and single-subject research. *Journal of Learning Disabilities, 26,* 642–647.

https://doi.org/10.1017/S027226319700404X Published online by Cambridge University Press