1 *Why (or why not), when, and how to replicate research*

Alison Mackey

The importance and prevalence of replication research varies greatly depending on the discipline and research area. In the so-called hard or pure sciences, for example, replication studies are common, and play an integral role in the process of testing and demonstrating the generalizability of crucial findings. Gross (1997) notes two issues that support the need for replication studies in scientific disciplines. First, replication studies check the probability of error in the testing of null hypotheses, or the likelihood of a Type I or Type II error having been made. For instance, the probability for error in rejecting or accepting null hypotheses might have been affected by unrepresentative sampling or low numbers of participants. Thus, testing additional samples of the target population with the same methods provides supporting or contradictory evidence regarding the existence of a phenomenon. Second, replication studies are necessary to more effectively control for extraneous variables that might have confounded the original findings. As a result, replication contributes to increasing the explanatory power and generalizability of previous findings in the "pure" sciences.

In social sciences, such as sociology, psychology, and economics, as well as linguistics, conducting replication research contributes to "the essence of the scientific method" involving "observations that can be repeated and verified by others" (American Psychological Association, 2010: 4). Within social science research, some scholars believe a study is not complete until it has been replicated (Muma, 1993: 927), yet results often prove difficult to reproduce. For example, according to Schneider, "a major problem in educational research is that investigators find it difficult or are unable to replicate their work or that of their peers" (2004: 1472). This scarcity of replication and re-analysis of previous findings undermines "the community's ability to accumulate knowledge" (ibid.: 1473). Although calls have been made for more replications in many areas of the social sciences, including areas related to applied linguistics, such as speech and hearing research (Muma, 1993), research into how second languages are learned has only recently begun to be incorporated in replication studies.

1 Replication in L2 research and other fields

1.1 The interdisciplinary nature of SLA research

Whereas the research area of SLA borrows certain methodologies and research principles from social sciences research, the role and, accordingly, the value of replication research in SLA has not been clearly defined to date for a number of reasons. To begin with, SLA is a relatively young field that has come into its own only in the past 40–50 years. It is clearly interdisciplinary in that it "draws from and impacts many other areas of study, among them linguistics, psychology, psycholinguistics, sociology, sociolinguistics, discourse analysis, conversational analysis, and education, to name a few" (Gass and Selinker, 2008: 2). With such a variety of contributing fields comes a corresponding variety of approaches to studying and analyzing aspects of SLA, some of which rely more heavily on replication than others. Research in linguistics, for example, is not traditionally based on replication, in contrast to research in psychology (Polio and Gass, 1997). L2 research, however, is informed by linguistics, psychology, education, and even sociology. This interdisciplinary nature of SLA research has made it difficult to emphasize the need for conducting replication studies over the need to keep up with other methodological trends from all the associated sub-areas and fields. However, although this status quo may have been acceptable in the earlier years of the establishment and development of the field of SLA, it is increasingly the case that there are sufficient studies present in SLA that need replication, and there is a growing understanding of the importance of replication research (Santos, 1989; Ortega, 2008; Porte, 2010).

1.2 Categorizing replication research

Replication research in the field of SLA has so far been categorized in accordance with the degree of its closeness to, or difference from, the original study. Polio and Gass (1997) outlined a "continuum of replication," which was recast as different replication types in a *Language Teaching* Review Panel article (2008) on replication (see this volume, Introduction).

Exact replication is almost nonexistent in the field of SLA due to the fact that it is usually impossible to get exactly the same type of subjects and exact stimuli as would be found in the original study (see Polio, Chapter 2 this volume). As noted earlier, exact replications are more common in other fields – in the field of bioelectromagnetics, for example, Krause et al. (2004) conducted an approximate replication

of Krause et al. (2000) on the effects of electromagnetic fields emitted by cellular phones on varying EEG frequency bands in participants performing auditory memory tasks. Although this replication resulted in disconfirmatory findings, the same memory tasks, data analyses, and methodologies, as well as comparable subjects, were used as in the original study. It is also worth mentioning that since some of the authors were the same as for the original study, they were replicating their own work, and thus they would not have to "'prove' that they did things the same way" (*Language Teaching* Review Panel, 2008: 6). However, even within the hard sciences, exact replication possibilities are affected by subject and condition variation, particularly in the environmental and ecological sciences. Exact replications in fields like the physical sciences are more common due to the potential for less experimental variation in physiological and psychological factors (see Nassaji's discussion of "Internal replications" in Chapter 3 this volume).

In some areas of the social sciences, such as sociology, it is possible to carry out exact replications; for example, the often-cited Kessler and Stipp study (1984), which replicated Phillips' (1982) study investigating the impact of fictional TV suicide stories on fatalities in the United States. Phillips' original findings suggested a causal link between fictional suicides on daytime television serials, or soap operas, and subsequent real-life suicides and fatal or nonfatal single-vehicle crashes. From the increasing trends of suicides and single-vehicle accidents, both fatal and nonfatal, after soap opera suicide stories, Phillips concluded that "soap opera suicide stories trigger some overt suicides and some covert suicides disguised as motor vehicle deaths" (1982: 1354). However, Kessler and Stipp's exact replication contradicted these results. They found that in using newspaper summaries as sources for the soap opera suicide stories, Phillips assigned an incorrect date range to eight out of 13 stories. Kessler and Stipp corrected this problem by investigating the exact date each story was aired and disaggregating the time series from weekly to daily information, allowing for a more precise before–after analysis. Their analysis also included important stories and controls that had not been included in the original study, and their findings found no substantial and statistically significant relationship between fictional and real-life suicides. Rather, they found "an average decrease of one-half of a suicide and a decrease of seven single vehicle motor fatalities" (Kessler and Stipp, 1984: 166).

Another widely known replication study from the field of psychology involves the investigation of the "bystander effect." Darley and Latané (1968) examined the murder of Kitty Genovese in 1964,

which was supposedly witnessed by 38 people who did nothing to intervene. They (and colleagues) conducted a series of studies (Latané and Darely, 1968; Latané and Rodin, 1969; Latané and Darley, 1970; Latané and Nida, 1981) investigating how participants reacted to dangerous situations, with the overall finding that the presence of bystanders (i.e., other people in the situation) hinders a person's helping behavior. Further replications simulating dangerous emergencies (e.g., Schwartz and Gottlieb, 1976; Harari et al., 1985) found that the bystander effect is inversely affected by the apparent danger of the situation, such that when the costs of not helping are greater than the costs of helping, people are more likely to intervene in a dangerous situation. To observe the bystander effect in a naturalistic setting, Harari et al.'s (1985) replication study simulated a violent crime – rape – on a college campus. In their study, the male subjects observed the crime under either an individual condition or a group condition, and their intervention rate was measured. Unlike previous laboratory studies on the bystander effect (e.g., Borofsky et al., 1971; Field, 1978), Harari et al.'s replication in a realistic setting enhanced "realism, demand characteristics, social desirability, and generalizability" (1985: 654) in explaining the bystander effect.

1.2.1 REPLICATION IN LINGUISTICS

In formal linguistics (e.g., syntax and semantics), one linguist's introspective judgments about meaning and grammaticality of a certain language might be disagreed with by another linguist, with such a disagreement calling into question the reliability and generalizability of the theoretical work. In order to avoid this, studies in formal linguistics often utilize tasks asking native speakers about the plausibility or acceptability of sentence meaning. These judgments contribute to testing linguistic theory. For instance, the use of evaluation tasks (e.g., asking about acceptability or preference on a three- or four-point scale) with the same or different populations allows formalists to replicate the previous study, which might eventually contribute to testing or enhancing the explanatory power of theories. Some research in formal linguistics, then, lends itself to both approximate replication and conceptual replication.

In SLA research, conceptual replications are generally viewed as the easiest to realistically carry out. Leow (1995), for example, replicated his own (1993) original study with a different, but comparable, population and in a different modality (aural versus written). His original study investigated the effects of the complexity of written input (simplified versus unsimplified texts), linguistic item (present

perfect versus present subjunctive verbs), and language experience (first semester versus fourth semester students) on learner intake. Participants at both learning levels were assigned written input that was either simplified with present perfect or present subjunctive verbs, or unsimplified with present perfect or present subjunctive verbs. In an analysis of pre-tests and post-tests, Leow (1993) found no differences in intake due to complexity of input or linguistic items, whereas language experience was a significant factor in determining the number of linguistic items that learners take in. Leow's (1995) replication using aural data corroborated his original results using written data with regard to complexity of input and language experience, but not linguistic item. In the aural modality, learners took in significantly more present perfect forms than present subjunctive forms in the input. This difference in results between aural and written input stresses the importance of considering the role of modality when investigating cognitive processes in SLA.

While Leow's (1993, 1995) work raises awareness of the potential contributions of replication studies to L2 research, there are many subfields within SLA where studies are rarely replicated. For example, there is only one study that is explicitly labeled as a replication in the abstract in the *Journal of Second Language Writing*. This study, by Allison et al. (1999), was a contextualized critique and approximate replication of Reid's (1996) exploratory work investigating the prediction of L2 sentences by native and nonnative English speakers, and resulted in different findings from the original study. Given the degree of controversy surrounding many of the key questions in the field of L2 writing, for instance the efficacy of feedback for the development of grammatical accuracy (e.g., Ferris and Roberts, 2001; Chandler, 2003; Hyland, 2003) and varying operationalizations of errors and their type classifications (e.g., Casanave, 1994; Ishikawa, 1995; Polio, 2003; among many others), a number of areas of L2 writing research, like SLA in general, would significantly benefit from replication research (see Polio, Chapter 2 this volume; Porte and Richards, 2012).

1.2.2 INSUFFICIENTLY DETAILED METHODS IN SLA ARTICLES

Researchers who intend to replicate a study need first to establish the rationale for its replication. One way to begin this process is by explaining the significance for the field of the original study and establishing its worthiness of replication (*Language Teaching* Review Panel, 2008). Sometimes, it may not be feasible to replicate a study if there are methodological issues that cannot be addressed

without making multiple methodological changes which render the new study too different from the original. Even conceptual replication research can be difficult to carry out. Insufficient reporting of the kind of details that would allow replication is a problem with many studies. First, the language proficiency of subjects is not always stated in exact terms, so the equivalence of sample populations is difficult to determine in replication studies. For instance, Thomas (1994: 314) was one of the first to describe the problem of impressionistic judgments of L2 proficiency in SLA studies, noting that some publications did not provide enough information about participants' proficiency (e.g., the subjects "were more or less beginners" – Ellis, 1988: 260; "have some degree of oral reading ability" – Carlisle, 1991: 83; "spoke English with a noticeable foreign accent in the authors' opinion" – Flege and Bohn, 1989: 41). In other cases, assessment criteria, such as placement tests, what each program/assessment level represents, and so on, are often reported vaguely, if at all. Even when researchers use standardized proficiency measures, there is often considerable variation between researchers, for example, in what constitutes an "advanced" learner, making it difficult to directly compare subjects across studies. Second, in many methodology sections in journal articles there is not enough space for detailed information regarding the settings and contexts of experimental conditions. Also, coding systems vary widely and are not always represented in sufficient detail. To address the latter concern, Mackey and Gass suggest making more "use of existing coding schemes, because this would facilitate comparison between studies" (2005: 230). However, sometimes existing coding schemes are refined to capture new knowledge or they may need to be revised to address the research questions, at least if the prevalence of research using new or custom-made coding systems is anything to go by. Third, the variability of operationalizations applied to concepts of the same name is a problem. For example, according to Polio (2003), it is notoriously difficult to find a common denominator for the concept of "linguistic accuracy" in L2 writing research, which makes establishing a starting point problematic in a replication study. Finally, lack of direct access to examples of materials used in the original study is often a serious barrier to successful replication although the establishment of an SLA database of instruments may go some way toward mitigating this problem (Marsden and Mackey, 2010: www.iris-database.org).

Another problem with replication studies is the uncertainty that arises when the replication results are different from the results in the original

study. Such situations raise the question of whether the results of the original or the replication study are correct. For example, DeKeyser et al. (2002) tackle the issue of operationalizations of learning conditions in Input Processing (IP) research, pointing out that replication studies carried out by VanPatten and colleagues (VanPatten, 1990; VanPatten and Cadierno, 1993; VanPatten and Oikkenon, 1996; Wong, 2001) confirm VanPatten's theoretical claims about IP, whereas replications carried out in other contexts (Collentine, 1998; Benati, 2001; Farley, 2001; Cheng, 2002) resulted in alternative interpretations. DeKeyser et al. (2002) pointed out that vaguely defined constructs can cause operationalizational issues, and can produce overgeneralization and overinterpretation of results.

1.3 The catch-22 of replication in the field

Porte, in his Introduction, suggested an additional factor contributing to the general paucity of replication research is its relatively unglamorous status in professional journals and in the academic community in general. According to Valdman, "in replication one loses the aura of glamour and the exhilaration of innovation" associated with original research (1993: 505). Original research is often more valued by tenure/promotion/reward committees and journal editors, and major universities in the United States require that dissertations should be original work. For example, the Department of Linguistics at the University of Hawai'i at Manoa encourages PhD students as follows: "The third and final part of the PhD program involves preparing and defending a dissertation that makes a significant original contribution to knowledge in the candidate's chosen field." The Department of Linguistics and Germanic, Slavic, Asian, and African languages at Michigan State University has similar originality requirements: "The dissertation is based on original research that makes a significant contribution to knowledge in some area of theoretical and/or applied linguistic." Although those requirements do not include any explicit indication that replication studies are not allowed, students might be dissuaded from replicating a previous study in choosing their dissertation topic by the requirement of originality.

This sort of value judgment might be passed on from faculty to graduate students, leading to a preference for original research (*Language Teaching* Review Panel, 2008). Despite replication being considered a basic tenet of scientific advancement (Smith, 1975), replication

research in the field of SLA finds itself in the proverbial "catch 22" situation. On the one hand, replication is essential to verify important findings:

- Replication is an important step in validating research and is considered a criterion for the acceptance of new theories and knowledge (*Language Teaching* Review Panel, 2008).
- Replication is, of course, crucial in order to distinguish the spurious from the real … (Polio and Gass, 1997: 500).

On the other hand, current practices in the field prevent replication research from gaining more acceptance as a useful procedure. This is unfortunate for many reasons. For example, in addition to verifying existing findings, replication research can serve as a learning tool, providing valuable experience for novice researchers or graduate students, as noted elsewhere in this volume (see Abbuhl, Chapter 5, and also Fitzpatrick, Chapter 6). Polio and Gass (1997) also suggested that faculty should encourage graduate students to conduct a replication study, by including it as a requirement in a course syllabus, such as research methods.

1.4 Identifying studies for replication

To qualify as a candidate for replication, a study should address appropriate, theoretically interesting, and currently relevant research questions. Or, it should address studies that are generally accepted in the field, but might have been insufficiently investigated in the original studies. It is not uncommon to find gaps in existing research. If an original study failed to control for important variables or discovered a variable post hoc that was not controlled for in the original research but should have been, then a replication study that takes those variables into account may be in order. In other cases, a study may be selected for replication because it would be interesting to assess whether its results would hold in different settings (e.g., laboratory results extending to the classroom) or different languages, or with learners of different ages (children versus adults). Authors often provide suggestions for replications in the "limitations" section of their papers, which open the door for many more replication opportunities. Issues typically mentioned include things like a limited number of participants/tokens, only one L1 background or setting being considered, as well as potentially intervening variables such as learners' diverse backgrounds. Although replication studies often make changes to the original research design, it is important to stick to the previously established constructs of the research objects (e.g., L2 writing, L2 attention, L2 anxiety). However,

these are often redefined for local contexts rather than further tested as originally operationalized (*Language Teaching* Review Panel, 2008). Replication studies using the original research design are useful in reexamining the theoretical relationship among constructs. For instance, in a situation where the original research identified multiple explanations for the results, a conceptual replication can manipulate nonsignificant variables and operationalizations of the original study to examine the strength of the causal relationship among variables.

1.5 SLA: Ripe for replication

Areas of research within SLA that are ripe for replication are relatively easy to identify. Inconsistent findings across different studies are a good starting point. One currently hot topic is research on the effects of implicit and explicit conditions on language learning (including the relationship between explicitness and awareness). While many studies have attempted to look at the relationship between awareness and L2 learning (e.g., Leow, 1997, 2000; Rebuschat and Williams, 2006), the researchers have reported different results, sometimes possibly due to the methodological differences in measuring awareness and learning (see Section 1.5.1).

Testing findings in different instructional contexts is also an interesting area for replication studies. Currently, most studies on the effects of feedback are "laboratory based," where learners and native speakers typically (although not always) interact in dyads, which is why it is important to repeat the studies in classroom contexts, with multiple participants and usually only one instructor. Since laboratory environments, where intervening variables are controlled, are very different from regular learning situations, such as classrooms, caution is necessary before assuming research from one context applies to another (Hulstijn, 1997). Also, research investigating the effects of instructional interventions on learning outcomes would benefit greatly from careful verification of its tools and practices, as well as from assessing the value of other methods (such as online measures, concurrent or retrospective protocols, etc.) and their applications. The following section further describes replication studies in two particular areas: SLA in the classroom and interactive SLA.

1.5.1 POSSIBLE REPLICATION AREAS

Explicit and implicit learning

Research on explicit and implicit learning has focused largely on evidence from qualitative research (e.g., Leow, 1997, 2000) and

psycholinguistics (e.g., Rebuschat and Williams, 2006). These studies investigate the role of learners' awareness in adults' L2 learning, but different findings are reported. Leow (1997, 2000) found that awareness and attention play a significant role in L2 processing and accuracy whereas Rebuschat and Williams (2006) reported that adult learners were able to learn some syntactic regularities incidentally without conscious awareness of the forms. Different methodologies used to measure awareness seem to result in the contrastive findings: think-aloud protocol data and two tasks (a multiple-choice recognition task and a written production task) were used in Leow's (2000) study, while a grammaticality judgment task on artificial grammar was used in Rebuschat and Williams's (2006) study. Since the studies in this area use different methods (e.g., subjective measure, online/offline verbal report) and coding systems (dichotomous, continuous) to measure awareness, consciousness, and learning, the reported results are hard to compare. Additional evidence from a series of replication studies using systematically unified coding systems and perhaps also incorporating new technologies in psycholinguistics and neurolinguistics that measure the original constructs in new ways (e.g., eye-tracking for measuring attention, event-related potential [ERP] for measuring products) may provide converging evidence related to the differences in these types of learning.

Individual differences

Measuring individual differences in learners before carrying out experimental research can be crucial for understanding data obtained from studies. Factors, such as working memory span, anxiety, motivation, personality, language aptitude, willingness to communicate, and drift, could all be profitably assessed in replication research in order to dissociate, for example, the benefits of instructional methodologies from individual differences. Trofimovich et al. (2007) asked whether individual differences in factors such as learners' phonological memory, working memory, attention control, and analytical ability could determine their ability to notice and benefit from recasts. Unlike Mackey et al. (2010), which found learners with larger working memory and phonological memory spans are more likely to notice the error targeted by recasts than learners with smaller spans, Trofimovich et al. (2007) found no association of working and phonological memory with noticing rates. Trofimovich et al. (2007) attributed these contradictory results to different measures for noticing employed in the two studies. As the authors noted in their limitations, their participants