STUDY QUALITY IN SLA: A CUMULATIVE AND DEVELOPMENTAL ASSESSMENT OF DESIGNS, ANALYSES, REPORTING PRACTICES, AND OUTCOMES IN QUANTITATIVE L2 RESEARCH

By

Luke Plonsky

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Second Language Studies

2011

ABSTRACT

STUDY QUALITY IN SLA: A CUMULATIVE AND DEVELOPMENTAL ASSESSMENT OF DESIGNS, ANALYSES, REPORTING PRACTICES, AND OUTCOMES IN QUANTITATIVE L2 RESEARCH

By

Luke Plonsky

I began this study with two assumptions. Assumption 1: Study quality matters. If the means by which researchers design, carry out, and report on their studies lack in rigor or transparency, theory and practice are likely to be misguided or at least decelerated. Assumption 2 is an implication of Assumption 1: Quality should be measured rather than assumed. Although peer reviews and researcher training in second language acquisition (SLA) are generally considered to be acceptable and thorough, there is very little evidence of the extent of methodological rigor, consistency, and transparency across second language (L2) research.

Beyond these assumptions, this study drew from previous research in several different fields. Central to this paper was the research on study quality which has received considerable attention in the context of research synthetic and meta-analytic methods. Because of the shared historical and methodological tradition of SLA and psychology (e.g., Gass, 1993), I consulted as well the American Psychological Association's guidelines for research when developing the instrument used in this study. Much of the empirical motivation for this project also came from within the field of SLA. Previous reviews have raised concerns about methodological practices across a number of subdomains, warranting further and more comprehensive investigation.

The first two of four questions posed by the present study asked about the use of study designs and analyses (RQ1) and reporting practices (RQ2) in L2 research. My third question sought to measure the relationship between research practices and outcomes (i.e., effect sizes)

found in L2 research (see Plonsky & Gass, 2011). And research question four asked whether and to what extent research practices and outcomes in SLA have changed in recent years. The purpose underlying these questions and the study more generally was not only to better understand conventions in the field but to inform future research practices as well.

In order to answer these questions and meet the study's larger purpose, a representative sample of L2 research published in two L2 journals from 1990 to 2010 was collected. Using research synthetic techniques, I surveyed the sample of studies, 606 in total, using a modified version of the instrument used by Plonsky and Gass (2011) in their investigation of study quality in the interactionist tradition of SLA. The coding scheme was designed to extract information related to study identification, design features, analyses, reporting practices, and effect sizes. Descriptive statistics were then calculated for study features to answer the research questions.

The overall results of this study point to a number of systematic strengths as well as many flaws across the corpus of L2 research. Of particular concern are incomplete and inconsistent reporting practices (e.g., means without standard deviations) and low statistical power, among other issues. Somewhat surprisingly and in contrast to previous findings (e.g., Plonsky, in press), there was very little evidence of a relationship between study quality and effect sizes, a finding that may reflect the broad substantive scope of the study and field. Finally, comparing research practices over the 1990s and 2000s, I found substantial improvements in almost all categories.

The discussion situates the results in terms of reviews from SLA and other fields (e.g., education; Skidmore & Thompson, 2010), shedding light on the methodological and analytical trends and trajectories observed. Based on the findings of the study, I make pointed suggestions for methodological reforms to be enacted by institutions such as the American Association for Applied Linguistics and individuals (e.g., independent researchers, journal editors).

Copyright by
Luke Plonsky
2011

To Pamela, Mateo, and Ruby

ACKNOWLEDGMENTS

I am enormously grateful to those individuals who invested their time, energy, and intellect in me, not only during the realization of this dissertation but throughout my graduate studies in the Second Language Studies program at Michigan State University. Although it will not do them or their efforts justice, I will attempt to express my appreciation.

I thank first my two advisors, Susan Gass and Shawn Loewen, who have provided me with unending support since entering the program. Sue's abundant ideas and insights were instrumental in guiding my progress on this project and my understanding of research in SLA more generally. Her perspective is unique to the field, and I am extremely fortunate to have been her student. The direction and support I have received from Shawn have, likewise, been unwavering, and his challenges invigorating. His expertise and guidance have vastly improved both the quality and rigor of my dissertation and my ability to contribute to the scholarly community. Both of my advisors have always had my best interests in mind, and I will never cease to thank them for that.

The three remaining members of my committee—Frederick L. Oswald, Paula Winke, and Spyros Konstantopoulos—were also instrumental in the successful completion of this project. Fred was integral to every phase of this dissertation and to the entire process leading up to it. He has been and remains a constant source of encouragement and inspiration. Paula's organization and attention to detail is unmatched, and I have benefited generously from both. And Spyros, the last of my committee members, provided me with a broad, outsider perspective in carrying out this project. His understanding of research synthesis as conducted across the social sciences was of great value in improving the strength and arguments of this paper.

Many thanks are also due to my adoring wife, Pamela, who never doubted me or this project when I was unsure of one or both. In addition to her steadfast support and encouragement, I appreciate her patience during seemingly endless days, weeks, and months of studying and writing. (Plus, she never complained when I went on and on about meta-analysis and effect sizes!) And although they won't remember it, I'd like to thank my two beautiful children, Mateo and Ruby. Not only are they a constant and refreshing source of enjoyment in my life, but they are also the perfect reminder that things like meta-analysis and effect sizes don't really matter.

Another individual I am indebted to is Kaytlin Moore. She worked tirelessly on the laborious and largely invisible task of coding. Many thanks to her for her assistance.

Finally, I thank God for helping me to find myself in second language research and for giving me the strength and ability to complete this work.

The dissertation presented here was supported financially by the Second Language Studies Ph.D. program, the College of Arts and Letters, and the Graduate School of Michigan State University in the form of a Dissertation Completion Fellowship, two Summer Support Fellowships, and a Research Enhancement Award. I am very grateful for these awards.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 REVIEW OF THE LITERATURE

The field of second language acquisition (SLA) has made significant progress, both substantive and methodological, since its inception in the second half of the 20th century (e.g., Lightbown, 2000; Pica, 1997). During this time, theoretical and empirical advances have been summarized and discussed regularly in the scholarly literature. In contrast, relatively little attention has been paid to the state or development of SLA research methods, an unfortunate circumstance given the empirical rigor needed to reliably and accurately inform second language (L2) theory and practice. Reflecting on the state of knowledge construction in SLA, Selinker and Lakshmanan (2001) explained that it may be "the case with all developing fields and especially with a field that draws from so many other fields, that very few papers tend to question their basic assumptions" (p. 324). Perhaps also related to its youth and its historical reliance on methodologies of sister-disciplines, SLA lacks the field-wide and field-specific standards for carrying out and reporting on research that help to maintain quality and consistency in other fields such as the *Publication Manual* of the American Psychological Association (APA) in psychology.

Although largely absent from the meta-discourse in SLA, some L2 researchers have hinted at the need for greater reflection on methodological rigor in L2 research in the context of broader conversations encompassing issues such as social utility and research ethics (e.g., Kubanyiova, 2008; Ortega, 2005). More pointed calls for reform of L2 methods and reporting practices in particular have surfaced in meta-analyses and other types of reviews. These studies have found somewhat widespread evidence among other trends for a lack of pretesting in quasi-experimental studies (e.g., Norris & Ortega, 2000), low or unreported estimates of instrument reliability (e.g., Chaudron, 2001), and missing data including basic descriptive statistics needed

to calculate an effect size (i.e., *d* value) (Oswald & Plonsky, 2010; Plonsky, in press-a; Wa-Mbaleka, 2006). Complementary to these findings, surveys of L2 researchers have found a lack of perceived importance of methodological rigor in different subdomains of the field such as computer-assisted language learning (Egbert, 2007; Smith & Lafford, 2009).

Beyond any methodological, statistical, and psychometric motivations for adhering to rigorous research and reporting practices, "respect for the field of SLA can come only through sound scientific progress" (Gass, Fleck, Leder, & Svetics, 1998; see also Henning, 1986). In other words, methodological infirmity hinders not only progress in the development of theory, but it may also negatively affect our reputation and legitimacy as a discipline and limit our potential to contribute to parent fields such as linguistics, education, and psychology from which SLA was conceived and has long-since borrowed research traditions.

To conclude this introduction, there is no controversy over the necessity of rigorous methods to advance the field of SLA. Like any social science, progress in this field depends on sound research designs, principled data analyses, and transparent reporting practices. However, as mentioned above, very little scholarly activity in the field has sought to describe the "how" of SLA and much less has addressed explicitly the quality of its empirical efforts. To be clear, inactivity in this area does not necessarily indicate a lack of concern or any field-wide incapacity to carry out scientifically rigorous studies. There are numerous book-length treatments that illustrate, describe, and prescribe research methods commonly used in SLA (e.g., Hatch & Lazaraton, 1991; Mackey & Gass, 2005, forthcoming; Porte, 2010), and the peer-review process in this field is generally regarded as maintaining rigorous control over published research (see Loewen & Gass, 2009; Valdman, 1998). However, whether and to what extent studies in SLA have been carried out in adherence to standards of quality is an empirical question. The study

reported here represents a step toward answering that question by systematically describing and assessing the research and reporting practices in SLA both cumulatively and over time. The crux of my purpose in carrying out this study is simple: By looking to the past of SLA, I hope to contribute to its future. In order to do so, I examine study quality in L2 research, defined for the purposes of this study as adherence to standards of empirical rigor, appropriateness, and transparency in study design, analysis, and reporting practices (see section on study quality below).

The remainder of this chapter is divided into three parts. The first is an introduction to meta-analysis and study quality, an area obscure to many SLA researchers but central to the present study. Next I describe SLA research and reporting practices and concerns about study quality as described in meta-analyses and previous reviews of L2 research. These papers are generally more concerned with the substantive issues of the subdomains they review, focusing only peripherally on study quality, but their findings and comments related to research practices are useful and informative to this discussion nonetheless. The literature review then summarizes Plonsky and Gass (2011), the first large-scale empirical investigation of methodological quality in SLA, and ends with the study's research questions.

Meta-analysis and the Assessment of Study Quality[1]

Meta-analysis is a procedure for quantitatively synthesizing primary research. In many ways, the steps involved in conducting a meta-analysis parallel those of primary research. Studies (the "participants") each contribute data (usually a standardized effect size index such as Cohen's *d* or a correlation coefficient) which are combined or averaged to answer a particular question. This is the core of meta-analysis. As one might expect, there are numerous decisions along the way leading to ancillary steps and techniques described at great length in the

methodological literature related to meta-analysis (see Oswald & Plonsky, 2010). One major decision point the meta-analyst must cross involves handling the varying degrees of quality among primary studies being collected and synthesized.

Study quality, an interdisciplinary domain in its own right but most often discussed in the context of meta-analysis, is the source for much of the orientation and conceptual motivation for this study. To date, as many as 300 measures have been proposed to assess the quality of quantitative empirical research (Wells & Littell, 2009). These measures have been used to weight effect sizes from primary studies based on the quality or appropriateness of their design, analyses, and reporting practices. Some systems for scoring and weighting primary research are very simple: exclude or include (essentially a weight of 0 or 1). The argument behind a dichotomous quality rating that also functions as part of the inclusion/exclusion criteria is that only those studies deemed to be of sufficiently high quality should be included in the meta-analysis (i.e., garbage in, garbage out). Other systems involve much more sophisticated procedures that attempt to approximate ultimate levels of psychometric precision by accounting for statistical artifacts such as range restriction, measurement reliability, and so forth (see Borenstein, Hedges, Higgins, & Rothstein, 2009, and Hunter & Schmidt, 2004). Unfortunately, this brand of study quality measurement and effect size weighting often sacrifices interpretability for the sake of accuracy only noted several places to the right of the decimal point of the meta-analytic mean. Regardless of the complexity of the approach, however, the assumption underlying instruments of this type is that studies of higher methodological quality should contribute more to the meta-analytic average than those of lower quality.

Although it is not my intention to assign a weighted value to the research reports investigated in this study, the criteria included in previous instruments of this type (e.g., sample

size, random assignment to experimental conditions) constitute an important source and point of departure for an instrument designed to describe and assess L2 research and reporting practices. However, as we might expect, there is little consensus among research synthesists over how to define and measure study quality (see Moja et al., 2005; Wells & Littell, 2009).

A number of studies have compared tools designed for this purpose, finding differences in some cases large enough to alter or reverse meta-analytic outcomes if applied to primary study effects (e.g., Jüni, Witschi, Bloch, & Egger, 1999). That there are differences between these instruments is not entirely surprising, and surely the availability of multiple instruments is valuable to would-be synthesists who are able to select the most appropriate set of measures for their particular domain. For example, one could easily make a case for employing a unique set of measures for assessing research from different fields (e.g., SLA and educational psychology), different subfields (e.g., universal grammar and interactionist traditions within SLA), and even different design types within the same subfield (e.g., observational vs. experimental studies of L2 interaction). In other words, the variety of available instruments may be both overwhelming and beneficial to meta-analysts, assuming care is taken to select measures appropriate for the domain in question. In the case of this study with its broad, field-wide scope and the variety of subfields and designs therein, I defined study quality rather broadly: adherence to standards empirical rigor, appropriateness, and transparency in study design, analysis, and reporting practices.

One could argue that the third component of this definition, reporting practices, should not be considered a component of study quality because it does not reflect the quality of the study itself. Furthermore, what gets (or does not get) reported in a research report is subject to constraints such as the page limits of different journals and the preferences of reviewers and editors. These arguments are worth considering and addressing. The decision to include reporting

practices in the definition and operationalization of study quality was prompted by several considerations. First, reporting practices, especially when examined at the field-wide level, are an indication of transparency and informational richness in the available research, which are certainly preferred to opaqueness. In addition, from a meta-analytic perspective, thorough reporting at the primary level enables more complete analyses at the secondary level and limits the potential for bias created by any relationships between reporting practices and study outcomes (see Plonsky, in press-a; see also discussion below related to problems associated with unreported data). And third, the inclusion of reporting practices follows existing definitions of and tools for measuring study quality (e.g., Downs & Black, 1998). One final clarification regarding this issue: I do not argue in this paper that a poorly reported study is necessarily of poor overall quality or vice versa. My position throughout this paper and reflected in the above definition is that study quality is a multidimensional construct; the extent to which data are reported thoroughly comprises one of those dimensions.

Also in line with the above definition of study quality, the range of items included in the instrument for this study was intentionally broad so as to be relevant to as many studies in the sample as possible. Inevitably some measures/items (e.g., whether or not a study included a delayed posttest) were only applicable to a certain type of design (e.g., experimental studies).

Returning briefly to inconsistencies among study quality instruments, one possible source for these discrepancies is the limited capacity of a single value (i.e., an overall quality rating or score) to express a notion as multifaceted as study quality (Valentine & Cooper, 2008). For this reason, in this study I analyze each aspect of study quality individually in the aggregate rather than assigning overall quality scores to individual studies in the sample.

Also related to the number and variety of existing measures of study quality, I will now introduce different facets of study quality which are operationalized in study quality instruments within the meta-analysis literature. I have labeled these facets according to the categories common to instruments of this type, and each is accompanied by a brief description or rationale for their role in determining study quality. Each category also includes sample measures or items from previous instruments. Of course the many existing instruments include not only different items and different numbers of items but different categories of items as well; and although the categorization here drawn on conventions from the domain of study quality, I do not want to give the impression that it is necessary uniform. They are, however, representative. It should also be noted that the categories found below are not mutually exclusive. Some items or measures could be placed in multiple categories. Reporting the extent and possible causes for attrition, for example, might be relevant to both the external validity and reporting practice categories.

Finally, not all measures listed below were used in the present study. I chose to exclude items from previous instruments for several reasons. First, many items were domain specific and therefore lacked relevance to L2 research (e.g., items particular to medical interventions). Second, in order to avoid necessarily subjective decisions which might pose a threat to the internal validity of my own study, I tended to prefer more objective measures and phrasing of items (see Chalmers et al., 1981). And third, the methodological and substantive scopes of SLA are quite broad and I therefore chose items that would apply to as many studies as possible. (See Chapter 2 for the complete instrument including the list of included items, definitions for each item, and scoring procedures.) I also note that these lists are not exhaustive due to the number of existing/possible measures and to avoid repeating what is presented in Chapter 2.

Internal validity: When the internal validity of a particular study or a body of studies is high (or when threats are minimal), we can have greater confidence in reported outcomes and findings. Some questions to think about when considering internal validity:

- In the case of (quasi-)experimental studies, were participants aware of their group membership?
- Were treatment administrators blinded?
- Were raters or scorers blinded to which groups' data they were handling?
- Were all analyses planned at the outset of the study?
- Were statistical tests appropriate for the type of data collected and for the research questions asked?
- Were outcome measures valid and reliable?
- Were participants (or classes) assigned randomly to conditions?
- If participants were not assigned randomly to conditions, was a pretest carried out to ensure comparability of groups?
- For (quasi-)experiments, was a delayed posttest included in the design?
- Were data compared from more than one group or condition?
- Was statistical power sufficient for the anticipated effects?
- Were alternate explanations for results provided?

External validity: The counterpart to internal validity, external validity, is concerned with the generalizability of the findings of a particular study beyond the sample to a larger population, a goal of much of SLA and quantitative social science research more generally (see Plonsky, in press-b). With this goal in mind, it is valuable for secondary researchers and consumers of secondary research to be aware of the extent to which the results in a body of literature can be

generalized (or the extent to which existing threats to external validity may limit or constrain generalizations). Measures of study quality in this category examine those threats.

- Was the sample representative of the population of interest?
- Was the context or setting of the study appropriate for the procedures and goals of the study?
- Were the parameters of the population of interest defined?
- Were claims and interpretations of the data appropriate to the design and results as presented?

Reporting practices: As described above, the thoroughness with which data and other study characteristics are present in a study report is associated with quality in that greater transparency facilitates more accurate and complete interpretation and reinterpretation at both primary and secondary levels.

- Was a hypothesis or objective stated?
- Were sample demographics and other descriptive information related to the sample provided?
- Were the procedures explained clearly?
- Were main findings explained clearly?
- Were appropriate quantitative indicators of sample variability provided (e.g., standard deviations, confidence intervals, interquartile ranges)?
- Were any potentially adverse effects of or reactions to the procedures documented and explained?
- Were actual rather than relative *p* values reported for all inferential statistics?
- Were estimates of instrument validity and/or reliability reported?
- Was the sample size reported?
- Were effect sizes reported when appropriate?
- Were effect sizes interpreted appropriately?

- Were the assumptions of statistical tests checked and met?

Previous Studies of L2 Research and Reporting Practices[2]

As I mentioned in the Introduction, study quality as a domain of inquiry has just begun to attract attention in the SLA literature. Nevertheless, descriptions and critiques of methodological and reporting practices in SLA have surfaced occasionally in research syntheses, meta-analyses, and historical reviews. In this section I review a number of such studies in three subsections, relating findings to discussions from related disciplines when appropriate: (a) reviews of L2 research, (b) reviews and guidelines of L2 journals, and (c) Plonsky and Gass (2011).

Study Quality as Found in Reviews of L2 Research

The comments and concerns raised in reviews of L2 research have centered mainly around three related issues—study designs, statistical analyses, and reporting practices. In one of the first meta-analyses of L2 research, Norris and Ortega (2000) synthesized research on the effectiveness of instruction. Based on their review, the authors included suggestions for improving future research on L2 instruction as well as for the field more generally such as the inclusion of control groups and increased pretesting in (quasi-)experimental studies to more accurately measure the effects of interventions and verify comparability of groups. Plonsky (in press-a), likewise, called attention to the lack of delayed posttests among studies of L2 strategy instruction. In some cases, the problems found by both Norris and Ortega and Plonsky pose threats to the external validity of previous findings. In other cases, additional findings (e.g., the longevity of treatment effects) were not examined thus leaving unnecessary gaps in the empirical literature. Many such problems are easily remedied with proper planning at the design stage of carrying out a study.

In his historical review of nine decades of classroom-based research in *The Modern Language Journal*, Chaudron (2001) included a thoughtful and at times critical discussion of methodological shortcomings. He lamented, for example, measures of low reliability, generally poor design, and the fact that "intact groups [as opposed to random assignment to experimental conditions] are the norm" (pp. 66-67; see also Henning, 1986; Lazaraton, 2000, 2005; Nunan, 1991, 1996). Yet Nunan (1991), in a similar review of 50 selected reports published in *Studies in Second Language Acquisition*, noted the relatively small portion of classroom-oriented research actually carried out in classrooms. Nunan went on to prescribe more "classroom-based" as opposed to "classroom-oriented" research in order to increase the relevance of SLA research to language practitioners. He also observed that the majority of the studies he examined were non-experimental in nature. Finally, a number of L2 scholars have warned that the small samples typical of L2 research may be problematic for their debilitating effect on statistical power (Chaudron, 1988; Crookes, 1991; Flahive & Ehlers-Zavala, 2010; Hauser, 2001; Henning, 1986; Larson-Hall, 2010; Lazaraton, 1991; Norris & Ortega, 2006; Oswald & Plonsky, 2010; Plonsky & Oswald, in press; Plonsky & Gass, 2011). Addressing this problem may be difficult because it is unclear exactly why sample sizes tend to be small in SLA. Besides the obvious logistical challenge of recruiting large numbers of L2 learner-participants, other factors may be at play such as unfamiliarity with the implications of statistical power (Lazaraton, Riggenbach, & Ediger, 1987) and a lack of previous research reporting effect sizes to facilitate power analyses and determine appropriate sample sizes (Plonsky & Gass, 2011). However, an assessment of sample sizes ought to consider as well the size of L2 classrooms where studies are often carried out. It is not uncommon, depending on the instructional setting and L2 in question, for L2 classrooms to be made up of 20 or less learners. That is, the ecological validity of small samples

found in L2 research must also be weighed against criticisms related to the relatively low power they carry for statistical analyses

With respect to statistical analyses and reporting practices, two major themes are evident from the synthetic literature describing L2 research. The first is the prevalence of means-based analyses (Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka, 2006; Nunan, 1991; Teleni and Baldauf, 1989). That L2 research depends mostly on means-based analyses is not in and of itself problematic, assuming (a) statistical tests are chosen based on their appropriateness for the available data and for the questions being posed rather than out of habit, convention, or convenience, (b) certain conditions and assumptions are met (e.g., normal distributions; see Wells & Hintze, 2007), and (c) the data are reported thoroughly and faithfully, which leads us to the second theme.

Simply put, data often go missing in reports of L2 research. More specifically and perhaps surprisingly, the problem is most acute among simple descriptive statistics such as means and especially standard deviations. In their review of the use of meta-analysis in SLA, Oswald and Plonsky (2010) commented on and cited the number of studies six L2 meta-analyses excluded because insufficient data was reported to calculate an effect size, which ranged from 16 in Russell and Spada (2006) to 35 in Plonsky (in press-a). In three of these meta-analyses (Dinsmore, 2006; Nekrasova & Becker, 2009; Russell & Spada, 2006), the number of primary studies excluded for this reason actually exceeded the total number of studies included!

These figures should give us pause because, depending on their pervasiveness, missing descriptive statistics have the potential to weaken progress in the field in several ways. First and most immediately, unreported data restricts our ability to interpret the findings of primary studies. Second, as meta-analysts regularly point out (see Oswald & Plonsky, 2010), it is often

impossible to calculate certain effect sizes such as Cohen's *d* when descriptive statistics such as standard deviations are missing. In other words, because meta-analyses depend on the data reported in primary studies, missing data at the primary level necessarily yields missing data at the secondary or meta-analytic level, which lowers power for moderator analyses, renders potentially vast amounts of research unmeta-analyzable and therefore inefficient and unaccounted for, and may produce a biased sample of primary studies. The potential for bias due to missing data is greatest when data are more likely to be left out of studies (a) from a particular theoretical or empirical tradition, (b) with nonstatistically significant findings, and/or (c) published in journals or during a time of more lenient requirements for reporting data (see Pigott, 2009; Sutton, 2009; Lipsey & Wilson, 1993; and for a book-length treatment of publication bias, see Rothstein, Sutton, & Borenstein, 2005). Third, the practice of incomplete data reporting may perpetuate itself by implicitly teaching readers and authors of research reports that thorough reporting and transparency are not prerequisites for publication.

Descriptive statistics are not the only aspect of reporting conventions found to be in need of improvement. Concerns have also been expressed over the reporting of reliability estimates and other types of data. Nekrasova and Becker (2009) and Norris and Ortega (2000), for example, found that only 6% and 16% of the primary studies in their meta-analyses of the effects of practice and instruction, respectively, reported any estimate of reliability for their dependent measures. Likewise, only two studies in Mackey and Goo's (2007) meta-analysis of L2 interaction reported test reliability. Of course, whether or not instruments used in L2 research are highly reliable is of greater importance than the presence or absence of a quantitative indication of that reliability. But the availability of reliability estimates is also critical to interpreting research findings. Without an estimate of reliability, it is not clear whether small and/or

nonstatistical effects in experimental studies, for example, should be attributed to an ineffective treatment or a measure with low reliability, a distinction with critical implications for the model being tested as well as future studies of the phenomenon in question (Oswald & Plonsky, 2010; Polio, in press). Reporting of reliability coefficients is also useful in the context of meta-analysis, where reliability estimates are often used to weight study effects as described earlier. (See Norris & Ortega [2003] and Read [2007] for discussions of issues related to reliability in L2 research in general, and Polio [1997, in press] for a discussion of reliability related to studies of L2 writing.) Other practices found to be absent or problematic among areas covered by previous reviews of L2 research include reporting of effect sizes and confidence intervals (Norris & Ortega, 2000; Mackey & Goo, 2007; Plonsky, in press-a), stating research questions and/or hypotheses to be tested (Henning, 1986), checking the assumptions of statistical tests (Nunan, 1991), graphic displays of data (Nekrasova & Becker, 2009), reporting a pre-determined level of statistical significance (Henning, 1986), and reporting whether participants were assigned to conditions randomly (Lazaraton, 2005; Nunan, 1996; Polio, in press).

This review may seem overly critical toward the field of SLA. To be fair, and to situate this discussion in a broader context, it is worth noting that many of the same issues present in L2 research have also been observed in other fields. Bangert and Baumberger (2005) tallied statistical analyses and select reporting practices in 256 studies published over 11 years in the *Journal of Counseling & Development*. They found that less than half of the studies in their sample that conducted tests of statistical significance reported effect sizes and only two studies reported a power analysis (see comparably bleak findings with respect to the use and perception of power in organizational research in Cashen & Geiger, 2004; Mone, Mueller, & Mauland, 1996; and Sedlmeier & Gigerenzer, 1989). And similarly low or lower rates of effect sizes and

confidence intervals were obtained by Kieffer, Reese, and Thompson's (2001) and Keselman et al.'s (1998) studies of research published in several prominent educational journals. Keselman et al. also found studies reporting to have checked statistical assumptions to be scarce.

Reporting of reliability is another area of weakness apparently shared by both L2 research and other social sciences. Three decades ago, Willson (1980) found that only 37% of the quantitative studies surveyed and published in the *American Educational Research Journal* reported a reliability coefficient. He went on to criticize the field for this "unexcusable" practice "at this late date" (p. 9). More recent studies as well have found reliability to be reported in as few as half of the studies published in different journals (Meier & Davis, 1990; Thompson & Snyder, 1998; Vacha-Haase, Ness, Nilsson, & Reetz, 1999). (For a unique vantage on methodological weaknesses, see Brutus, Gill, & Duniewicz's [2010] study of limitations as reported by researchers themselves in industrial and organizational psychology.)

The Relationship between Methods and Outcomes

In addition to summarizing substantive findings and measuring certain methodological and reporting conventions, several meta-analyses have also hypothesized about and undertaken empirical examinations of the relationship between research practices and study outcomes. A starting point for this type of analysis is the assumption that "study results are determined conjointly by the nature of the substantive phenomenon under investigation and the nature of the methods used to study it" (Lipsey, 2009, p. 150). Put another way, "effect sizes are not magically independent of the designs that created them" (Vacha-Hasse & Thompson, 2004, p. 478).

Research setting is a design feature hypothesized to moderate effect sizes. As we might expect intuitively, lab studies generally produce larger effects than classroom studies according to several L2 meta-analyses (Li, 2010; Mackey & Goo, 2007; Plonsky, in press-a; see Gass,

Mackey, & Ross-Feldman, 2005, for a discussion and one of the few empirical studies comparing treatments across research settings). Other meta-analyses have looked at measures of methodological quality in relation to outcomes. Russell and Spada (2006), for instance, calculated the average effect of error correction based on whether studies reported reliability and validity of dependent measures. Likewise, Plonsky (in press-a) formed and compared subgroups of studies of L2 strategy instruction based on three aspects of methodological quality, finding substantially larger effects for studies that (a) pretested (d = 0.54) versus those that did not pre-test (d =0.39), (b) employed random group assignment (d = 0.65 vs. d = 0.42), and (c) reported reliability (d = 0.65 vs. d = 0.42) (see also Adesope, Lavin, Thompson, & Ungerleider, 2010).

These findings along with other suggestions for reform described above point not only to the presence of flaws in SLA research but also to a possible relationship between different methodological/reporting practices and study outcomes (see Lipsey &Wilson, 1993; Prentice & Miller, 1992; Wilson & Lipsey, 2001). To be clear, I am not suggesting that methodological quality might affect outcomes. No one would claim that the act of randomly assigning participants to experimental conditions or reporting reliability *causes* larger or smaller effects (see Lipsey, 2009). (Although it is entirely possible that studies with random group assignment are more likely to be carried out in lab contexts where researchers exercise greater experimental control and are thus able to obtain larger effects. Similarly, studies using highly reliable instruments might be more likely to both report reliability and to obtain larger effects because their results will not be attenuated by low reliability.) Rather, I intend to explore possible relationships between study features and outcomes in order to better understand how SLA research has been carried out and to determine if/how these variables interact.

Changes over Time

In order to more clearly illustrate the state and development of SLA research practices, it will be useful to consider different study features not only cumulatively but over time. This process began recently within one subarea of SLA, interaction-based research, in Plonsky and Gass (2011; see section below) and is being expanded to include the field as a whole in this study. If SLA methods continue to pattern similar to those of related disciplines by attempting to follow the slow but dynamic trajectory of what is considered best practice by quantitative methodologists, we should find changes and improvements in the designs, analyses, and reporting practices in L2 research across decades. Skidmore and Thompson (2010) found data analyses in education and psychology, for example, to move from ANOVA-type techniques to regression analyses. The authors traced the impetus for this gradual shift to Cohen's (1968) discussion of the related and hierarchical nature of statistical analyses within the General Linear Model (GLM).

In SLA we might also expect to find the settings of L2 research to shift toward more classroom-based research as researchers seek to generalize findings from lab studies, and observational/correlational findings, once established, may give way to experimental designs (Oswald & Plonsky, 2010). In addition to the findings in Plonsky and Gass' study (see below), there is already some evidence to demonstrate changes and improvements taking place over time particularly in reporting practices. Russell and Spada (2006), for instance, observed an "evolution of studies on CF [corrective feedback]" in that "most of the *recently* published studies met the criteria for inclusion in a meta-analysis" by reporting enough data to calculate an effect size (p. 156, emphasis added). They then argue that "over time, research on CF is adhering to higher research standards" (p. 156). Further evidence is found in the portion of studies reporting effect sizes as found in two meta-analyses of research on L2 interaction. Whereas none of the

studies in Keck et al. (2006; with inclusion dates 1980-2003) reported effect sizes, four did so in Mackey and Goo's meta-analytic replication (2007; 1990-2007).

Like methods, evidence is never static (see Trikalinos et al., 2004). It is not hard to imagine a trajectory of research for a particular domain in which methodological adjustments or improvements lead to larger effect sizes over time (Fern & Monroe, 1996; Oswald & Plonsky, 2010). In experimental SLA research, such changes may result from the realization that longer or stronger treatments produce larger differences between groups. Plonsky (in press-a), for instance, found that the length of strategy interventions correlated positively with their effectiveness. An increase in effect sizes might also be found when the psychometric properties of instruments, the standards for which are generally lower in an emerging research area (Brutus et al., 2010), are refined over time.

In contrast, an alternate scenario may also play out in a body of empirical literature: Early research in a given area is often characterized by strong manipulations that set out to determine whether an effect exists and thereby determine whether the claims of a particular and usually novel hypothesis merit further attention. Such experiments would tend to yield large effect sizes (Kline, 2004). Subsequently, after an effect is found, research efforts may shift to the generalizability of an effect across samples, settings, tasks, and so forth (see Plonsky, in press-b). In domains where this scenario is observed, theoretical maturity would be inversely correlated with outcomes and thus a decrease in effect sizes would be obtained over time (Plonsky & Oswald, 2010, in press).

Combs (2010) provided an alternate perspective on the latter scenario. In his review of effect sizes (correlation coefficients, in this case) reported in the *Academy of Management Journal*, he found the magnitude of effects to be decreasing over time. However, instead of

attributing that change to empirical demonstrations of theoretical nuance, he argued that the inverse relationship between date of publication and effect size was related to the increase in average sample size that took place during the period in question. More precisely, he claimed that, as reviewers and editors began to recognize the importance of statistical power and require larger samples, contributing authors were able to obtain and publish $p < .05$ for smaller correlations. Statistically speaking, it is hard to counter this argument. (Holding a correlation constant, regardless of how small it is, $p < .05$ can always be attained given a large enough sample [see, for example, Tukey, 1991]). Combs' approach, however, deemphasizes two important factors: First, we should expect that more mature domains will ask more subtle questions that produce smaller effect sizes overall whether or not their sample sizes have increased. And second, meta-research at the field-wide level, including my study, is often blind to the variance across multiple subdomains in the sample. Although I would urge Combs to consider an explanation for the change in effects that accounts for both statistical and substantive developments, his study presents a worthwhile example of an exploration of effect sizes over time from another discipline.

Unlike meta-analyses that examined findings within a single subdomain of L2 research, the field-wide scope of the current study, like Combs (2010), may blur my ability to detect whether one, both, or neither of these patterns has occurred. In other words, there is no doubt that the last few decades of SLA research have seen numerous domains rise, fall, mature, improve, and decline, but these patterns may not be observable in the aggregate.

Reviews and Guidelines of L2 Journals

Earlier I summarized occasional and problematic trends in research and reporting practices as found in reviews of L2 research. Complementary to those findings, two studies of

scholars' perceptions of applied linguistics journals have found a relatively low priority given to methodological rigor. Egbert (2007) surveyed members of the research interest section of TESOL on their journal-reading preferences, and only 2 of 31 respondents cited sound research design in articles as a factor in determining the value of TESOL-related journals. However, seven did mention the much broader trait of "quality of article", which may include methodological rigor. In a similar study, Smith and Lafford (2009) asked established researchers of computer assisted language learning (CALL) to cite their criteria for evaluating CALL journals. None cited quality in design or methodology employed (see also Magnan, 2007).

Contrary to and perhaps because of the perceived value of methodological rigor in L2 journals, editors of some journals have taken steps to improve the research and reporting practices found in the journals they oversee. In contrast to the bottom-up approach to reform seen in recommendations by individual researchers cited above such as Norris and Ortega (2000, 2006) and Oswald and Plonsky (2010), this channel for reform takes a top-down approach. That is, editors have at their disposal and can exploit the unequal footing between themselves and submitting authors to incite change. In an editorial statement (Ellis, 2000), *Language Learning* joined at least 23 other academic journals (see Vacha-Haase & Thompson, 2004) in the social sciences requiring that submitting authors "always present effect sizes and their confidence intervals for primary outcomes" (p. xii; see a reiteration of this policy by DeKeyser & Schoonen, 2007). Since then, *Language Learning & Technology*, *The Modern Language Journal*, and *TESOL Quarterly* have released similar guidelines with respect to effect sizes and other reporting practices including stating the hypotheses to be tested, describing whether assumptions for statistical tests were met, and using graphs and tables to complement in-text presentations and explanations of data. To my knowledge, no other L2 journals have policies stating that

submitting authors are required to adhere to any of these conventions (see DeVaney's, 2001, survey of editorial policies and preferences of non-L2 journals regarding the reporting of effect sizes and related practices; see also reviews of reporting practices from other fields in Sun, Pan, & Wang, 2010, and Matthews, Gentry, McCoach, Worrell, Matthews, & Dixon, 2008).

There is currently little empirical evidence to suggest whether these policies and guidelines have had an effect on reports of L2 research, one of the gaps this study seeks to fill. Reviews of similar efforts from other fields, however, paint a mixed picture of what we might find. Fidler and Cumming (2007), for example, reviewed attempts by the APA and several non-L2 journals and journal editors to reform the use of null hypothesis significance testing (NHST), a ubiquitous but controversial practice regularly condemned for decades by quantitative methodologists (e.g., Cohen, 1994; Lykken, 1968; Schmidt, 1996; Thompson, 2001), cautioned against sporadically by L2 researchers (Crookes, 1991; Larson-Hall, 2010; Lazaraton, 1991; Nassaji, in press; Norris & Ortega, 2000, 2006; Plonsky, 2009, in press-a; Oswald & Plonsky, 2010), and brought to justice recently by the US Supreme Court (*Matrixx Initiatives Inc. v. Siracusano*, 2010). Fidler and Cumming argue that despite the massive potential of the APA *Publication Manual* to improve data analysis and reporting practices, and despite the recommendations of other APA publications (e.g., Wilkinson & the Task Force on Statistical Inference, 1999), the de facto message from the APA and the status of NHST in social science research is essentially "business as usual" (p. 443; for one of many other reviews of NHST policies and practices, see Finch, Thomason, & Cumming, 2002; see also Keselman et al., 1998; Kirk, 1996; Thompson & Snyder, 1998). The development of data reporting practices and the use of NHST in particular in fields such as education and psychology is relevant to this study and to SLA more generally because, first, L2 research has traditionally followed in the

methodological footsteps of these fields (Felser, 2005; Gass, 1993, Pica, 1997) and, second, patterns in other social sciences can serve as a point of reference for informing and comparing the rate and route of progress in SLA. That is not to say that all aspects of the previous reforms should be emulated. As Fidler and Cumming (2007) put it, "the statistical reform 'debate' has been, very largely, the sound of one hand clapping" (p. 441), and the pace of change has been glacial at best.

In order to avoid the dual inefficiency of relying on NHST and waving the anti-NHST flag while marching in circles, Fidler and Cumming (2007) and others (e.g., Huck, 2007; Kieffer et al., 2001) suggested improvements to statistical education in graduate programs. In our own field, Lazaraton et al. (1987) and Teleni and Baldauf (1989) have also suggested a shift in the quantity and curricular focus of training for graduate students.

Plonsky & Gass (2011)

Plonsky and Gass' recent study was motivated by much of the same literature reviewed here. Much like the present study, we used meta-analytic techniques to examine study designs, statistical procedures, data reporting practices, and the relationship between these variables and effect sizes both cumulatively and over time. In fact, the only major difference between ours and the present study is in scope. Whereas the present study examines the field of SLA as a whole, Plonsky and Gass focused on one particularly long-standing and influential line of inquiry in SLA: the interactionist tradition.

To carry out the 2011 study we designed an instrument for describing and assessing research and reporting practices—a slightly modified version of which is used in the present study—and surveyed 174 published reports of research on L2 interaction. Our results revealed several strengths in study designs and reporting practices such as the use of delayed posttests in

80% of (quasi-)experimental studies and the reporting of reliability in 64% of the sample. We also found several design-related weaknesses such as a lack of pretesting (21% of quasi-experiments did not pretest), small samples and low statistical power (estimated at .56 based according to a post hoc analysis), and the lack of random group assignment even among experimental lab-based studies. We also observed inconsistencies in data reporting practices. Means were often reported without standard deviations, *t* tests and ANOVAs without corresponding *t* and *f* values, and several reporting practices associated with rigor and quality (e.g., power analyses, checking statistical assumptions, confidence intervals) were generally absent.

Like my current study, we were also interested in the relationship between study outcomes (i.e., *d* values) and both design types/features and reporting practices. Somewhat larger effects were found for experimental over observational studies ($d = 0.72$ vs. 0.51). Although this difference was not statistically significant due to overlapping confidence intervals, it is worth noting the relatively narrow confidence intervals around the mean for experimental studies, which indicate a relatively precise estimate of that population of studies. There was virtually no difference between studies carried out in classrooms ($d = 0.64$) and laboratories ($d = 0.65$). Using four design features associated with experimental quality to form subgroups of studies and meta-analyze their effects, there was a statistically significant and reliable advantage only for studies that included a delayed posttest in their designs (compared to those that did not). Unlike Plonsky (in press-a), subgroups of studies with other quality-related features did not produce larger effects that those without.

Four reporting practices associated with quality were also investigated in relation to study outcomes. Somewhat surprisingly, the first two comparisons revealed larger effects for studies

with nonpreferred reporting practices. There was a large and consistent difference, indicated by nonoverlapping confidence intervals, between studies that reported (d = 0.42) and did not report reliability coefficients (d = 0.96). And studies that do not report a predetermined level of statistical significance (d = 0.71) produced larger effects than the minority of studies that did (0.53), although this difference may not be reliable due to overlapping confidence intervals. Also related to statistical significance, effects were compared for studies that reported exact p values and those that reported p values as greater or less than a particular Type I error rate such as .05. There was virtually no difference between these two types of studies or between studies that did and did not report an effect size.

Although the overall study quality in interactionist research was less than ideal, there was reason to be optimistic among the findings for changes across the 1980s, 1990s, and 2000s. Increases were observed in several features associated with quality including random group assignment and delayed posttests. A similar pattern was found for reporting practices as well. Consistent and occasionally dramatic increases were found over time for the percentage of studies reporting exact p values, checking of statistical assumptions, and effect sizes among others. The reporting of basic descriptive statistics such as means and standard deviations and test statistics such as t and f values also increased, but none of these data types matched the frequency of the means-based tests that used them, thus indicating that the interpretability and meta-analyzability of this body of research remains limited by researcher oversight.

Our study also examined changes in the magnitude of effects across the three decades of interest. The unambiguous trend was for effect sizes to decrease over time, with average d values of 1.62, 0.82, and 0.52 from the 1980s, 1990s, and 2000s, respectively. Along with a decrease in effect sizes, we also observed an increase in the number of studies published over time and in the

precision of their aggregated effects, as indicated by increasingly narrow confidence intervals. We attributed the decrease in effects over time to the maturation of the interactionist research agenda and increases in theoretical nuance as described earlier (see Plonsky & Oswald, in press). We also proposed that a relationship may exist between the decrease in effect sizes and the increase in the number of statistical tests carried out over time.

Although the Plonsky and Gass study provided us with a rich data set regarding the state and development of research and reporting practices in the interactionist tradition, we cannot assume that it is representative of the rest of the field of SLA. I have already pointed out examples of several practices that vary from one subdomain of L2 research to another (e.g., reporting of reliability, frequency of missing data). What this study aims to produce, then, is a description of designs, analyses, reporting practices, and outcomes found in the field of SLA as a whole, and the relationships between them. In doing so, the L2 community will be provided with a source of empirical data to inform reflections on and perhaps reforms of research practices.

Research Questions

The following research questions were addressed:

RQ1) To what extent has L2 research employed various study designs and statistical procedures?

RQ2) To what extent have data in L2 research been reported thoroughly?

RQ3) Is there a relationship between methodological features and quality as measured in RQ1 and 2 and effect sizes in L2 research?

RQ4) How have different aspects of study quality including designs (as addressed in RQ1), analyses (as addressed in RQ1), reporting practices (as addressed in RQ2), and outcomes (as addressed in RQ3) in L2 research changed over time?

Chapter 2 METHOD

The research questions given in Chapter 1 were addressed by surveying a representative body of quantitative L2 research. Although not strictly a research synthesis or a meta-analysis per se, many of the techniques I used to retrieve, code, and analyze this body of primary research are characteristic of research syntheses and meta-analyses. This study differs from those brands of synthetic research in that the focus of this study is almost exclusively methodological (i.e., the *how* of L2 research) rather than substantive (i.e., the *what*). The following points serve to further clarify the difference between research synthesis, meta-analysis, and the present study.

- Research synthesis is a type of secondary study that comprehensively and systematically reviews the available research on a given topic. The question addressed is almost always substantive rather than methodological in nature, and the analysis of previous studies and their findings may or may not include a quantitative aggregation of results such as in meta-analysis.
- Meta-analysis is, like research synthesis, a procedure for analyzing primary studies that address a common question. Unlike research synthesis, however, meta-analysis necessarily employs a set of fairly well-defined procedures for obtaining findings from previous studies (effect sizes, usually) and combining them to determine the mean and variance of those findings.
- The present study resembles both research synthesis and meta-analysis in its systematic approach to reviewing previous research. Like meta-analysis in particular, this study also combines effect sizes. Unlike both secondary tools, however, the intention of this study is not to summarize the accumulated findings related to any particular question found among primary studies. Rather, the objective of this study is to synthesize and analyze

research and reporting practices found across the field of SLA both cumulatively and over time. One might refer to this type of study as a "methodological synthesis".

Study Identification and Retrieval

The first step in carrying out this study was to determine a principled, representative, and accessible domain of empirical research within the field of SLA. I defined the domain along three dimensions: location (i.e., sources of L2 research), time (i.e., the dates to be included), and substance/content. Following Plonsky and Gass (2011), I began with the assumption that journals (as opposed to books or other publication formats) constitute the primary means by which SLA research is disseminated (Smith & Lafford, 2009; VanPatten & Williams, 2002). Beyond the fact that journals are the medium of choice for publishing primary L2 research, journals are generally accessible through hard-copy and electronic library resources. Moreover, primary studies published in books were excluded on the grounds that the number of possible sources would preclude the collection of an even and representative sample. I initially considered the 15 journals identified by 45 associate and full professors of SLA as regularly publishing L2 research (VanPatten and Williams, 2002): *Applied Language Learning, Applied Linguistics, Applied Psycholinguistics, Bilingualism: Language and Cognition, Canadian Modern Language Review, Foreign Language Annals, Journal of Second Language Writing, Language Awareness, Language Learning, Language Teaching Research, The Modern Language Journal, Second Language Research, Studies in Second Language Acquisition, System,* and *TESOL Quarterly*. After consulting descriptions on journal websites, the list of potential sources was reduced to four journals that focus primarily or exclusively on second language learning (as opposed to language teaching, technology, or other L2-related issues): *Language Learning, The Modern Language Journal, Second Language Research,* and *Studies in Second Language Acquisition*.

The scopes of these journals were then examined more closely, and *The Modern Language Journal* was excluded because of its slightly broader interest in issues such as L2 pedagogy and the profession of language teaching. I also decided that *Second Language Research* would not be adequately representative of the field, given that the studies it publishes almost exclusively employ psycholinguistic methods and are related to morphosyntactic elements of the L2. The two remaining journals were included in the study: *Language Learning* and *Studies in Second Language Acquisition*. Although neither journal quality or any other index for journal prominence such as impact factor was considered in selecting sources of publication—my position in this paper is that quality is an empirical issue and one that has yet to be examined—*Language Learning* and *Studies in Second Language Acquisition* were the two most highly rated SLA journals according to VanPatten and Williams' (2002) survey.

With respect to the temporal dimension, all studies published 1990-2010 (inclusive) in the two journals were eligible for inclusion. A two-decade span of research was chosen to examine the current state of research and reporting practices in the field as well as to examine any recent changes by comparing results across the 1990s and the first decade of the 2000s.

The substantive dimension was very inclusive. All primary, quantitative studies related to L2 learning that fell within the other two dimensions were included. Because of the broad nature of the study and the narrow scope employed when selecting source-journals (Lazaraton [2005] referred to *Language Learning* and *Studies in Second Language Acquisition* as "niche" [p. 218] journals within the parent field of applied linguistics), the default so to speak was to include candidate studies. However, a very small number of reports was excluded based on the following criteria: (a) the substantive focus was studied and discussed only in relation to L1 acquisition (e.g., Nicoladis & Krott, 2007), (b) the exact same study was published elsewhere in the pool of

candidate studies such as in the *Best of Language Learning Monograph Series* (e.g., Pulido, 2004), (c) the study was a literature review or meta-analysis (e.g., Li, 2010), and (d) the report presented only qualitative data (e.g., Waring, 2009). (This final exclusion criterion was determined based on the interest of this study in quantitative research and reporting practices.) All theoretical (i.e., non-empirical), methodological, and editorial articles were also excluded.

To summarize, I included all primary L2 studies published in *Language Learning* or *Studies in Second Language Acquisition* and dated 1990-2010. A total of 606 primary reports met these criteria; 327 were published in *Language Learning* and 279 in *Studies in Second Language Acquisition*. See Appendix for a complete listing of reports included in the study.

Coding

Each study was surveyed using a modified version of the protocol developed and first used by Plonsky and Gass (2011). In designing the instrument, Plonsky and Gass drew from a number of sources, all reviewed in Chapter 1: (a) previous instruments for assessing methodological quality from the meta-analysis literature, (b) surveys of methodological and analytic practices from other fields (e.g., Goodwin & Goodwin, 1985; Kieffer et al., 2001), (c) recommendations found in the *Publication Manual* and other APA publications (i.e., Journal Article Reporting Standards Working Group, 2008; Wilkinson & the Task Force on Statistical Inference, 1999), (d) recommendations and findings of previous reviews of L2 research (e.g., Lazaraton, 2000, Norris & Ortega, 2000), and (e) reviews and guidelines of L2 journals. For (c) and (e), only the most recent guidelines were used.

Based on the findings and recommendations of these sources, five different categories of data were coded: (a) study identification (e.g., year of publication, journal), (b) design (e.g., random group assignment, pretesting), (c) analyses (e.g., correlation, *t* test), (d) reporting of data

(e.g., reliability coefficients, means), and (e) outcomes (i.e., effect sizes). Table 1 lists these categories, the variables on which each study was coded, possible values for each variable, and a brief definition. For most variables, the definition is worded as the question I asked of each study; a zero indicates that the answer was "no" and a one "yes". In order to limit the number of potential items/variables in categories (c) and (d), I focused especially on means-based analyses (e.g., *t* tests, ANOVAs) and related data reporting practices such as means and standard deviations. This constraint was motivated by the predominance of means-based analyses in L2 research (Gass, 2009; Lazaraton, 2005) and the problems associated with missing data related to these analyses (e.g., Plonsky, in press-a; Wa-Mbaleka, 2006). In contrast to similar reviews from other fields (e.g., Bangert & Baumberger, 2005; Goodwin & Goodwin, 1985), I did not assign an a priori evaluation of sophistication to statistical techniques such as basic, intermediate, advanced. Rather, in the absence of any recent surveys of L2 researcher knowledge of statistics, each type of analysis was treated equally.

I coded all 606 articles. An additional trained rater recoded a subset of ten studies chosen at random from the sample. I calculated agreement between myself and the other rater as both a simple percentage and Cohen's kappa to account for agreement due to chance among categorical variables. Overall interrater agreement was 82% with a kappa of .56 (SE = .04; 95% CIs = .48-.63), which is fair; Cohen's kappa is known to be overly conservative, especially when estimating reliability among a large number of categories (e.g., Aguinis, Pierce, Bosco, & Muslin, 2009; Brutus et al., 2010; Fleiss, 1981; Landis & Koch, 1977). Nevertheless, I was concerned about the error rate, so I reviewed all interrater disagreements by consulting study reports. I found the original coding to be accurate 96% of the time (essentially a measure intrarater reliability; kappa = .90; SE = .04; 95% CIs = .86-.94). No additional studies were

recoded because (a) the rate of agreement was considered acceptably high, (b) most items in the coding scheme were low-inference in nature, and (c) high agreement was found in Plonsky and Gass (2011) which used a version of the same instrument to code reports drawn from the same population as the current study (99.5% and 96.3% agreement for the first and second additional raters, respectively, who coded ten studies each or 12% of the sample of 174 studies; overall percentage agreement was 97.9%).

Following Plonsky and Gass (2011), among those features in Table 1, four were considered most important when exploring methodological quality in experimental studies. They were as follows and in order of preference for empirical control: (a) random assignment to experimental conditions; (b) inclusion of a control or comparison group; (c) pretesting; and (d) delayed posttesting (see Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002).

Table 1
Coding Scheme Categories, Variables, Values, and Definitions

Variable	Values	Definition/operationalization
Identification		
Author(s)	Open	Author or authors of each study report
Journal	LL, SSLA	Journal in which the study was published
Year	1990-2010	Year in which the article was published
Volume-Issue	Open	In what volume and issue was the study published?
Design		
Pretest	0, 1	Was a pretest used to determine equivalence of groups, to measure participants' knowledge of a particular structure, and/or as an indication of participants' general L2 proficiency?
Delayed posttest	0, 1	Was a test given to measure the effects of a treatment at a time later than an immediate or initial posttest?
Observational	0, 1	The study did not test a causal relationship using one or more posttests. This includes a variety of data sources such as transcripts of naturalistic and elicited production, questionnaires and surveys, response time measures and so forth.
Experimental	0, 1	Did the study test a causal relationship using one or more posttests?

Table 1 (cont'd)

Comparison group: observational	0, 1	Among observational studies, were data collected and compared from multiple groups formed on the basis of one or more controlled variables?
Control group: experimental	0, 1	Among experimental studies, were data collected and compared using a (true) control group
Comparison group: experimental	0, 1	Among experimental studies, were data collected and compared using one or more comparison groups?
Setting: lab	0, 1	The study was not carried out with intact classes.
Setting: classroom	0, 1	The study was carried out using intact classes.
Random assignment: individual	0, 1	Were individual participants assigned to conditions randomly?
Random assignment: group	0, 1	Were classes or groups assigned to conditions randomly?
Sample size	Open	The number of participants in each sample in the study
Conditions	Open	The number of unique groups or conditions included in the study
Analyses		
Correlation	0, 1	Any type of correlation carried out, not including correlations used to estimate instrument reliability
Chi-square	0, 1	Were frequency data analyzed using a chi-square test?
t test	0, 1	Was a *t* test (or a nonparametric equivalent) used?
ANOVA	0, 1	Was an ANOVA (or a nonparametric equivalent) used?
ANCOVA	0, 1	Was an ANCOVA (or a nonparametric equivalent) used?
MANOVA	0, 1	Was an MANOVA (or a nonparametric equivalent) used?
MANCOVA	0, 1	Was an ANCOVA (or a nonparametric equivalent) used?
Factor analysis	0, 1	Was a factor analysis carried out?
Regression	0, 1	Was a regression carried out?
SEM	0, 1	Was structural equation modeling used?
DFA	0, 1	Was a discriminant function analysis carried out?
Rasch	0, 1	Was a Rasch analysis carried out?
Nonparametrics	0, 1	Was a nonparametric equivalent of a means-based test used?
Other	0, 1 + open	Were any other types of quantitative analyses (e.g., cluster analysis) carried out?
Reporting of Data		
Sample size	0, 1	Was the sample size reported?
Percentage	0, 1	Was a percentage reported?
Frequency	0, 1	Were frequency data reported?
Correlation	0, 1	Was a correlation reported?
Mean	0, 1	Was a mean reported?
Standard deviation	0, 1	Was a standard deviation reported?
Mean without *SD*	0, 1	Was a mean reported without a standard deviation?
Effect size	0, 1	Was an effect size reported, including Cohen's *d*, Hedges *g*, Cohen's f^2 Cramer's *V*, eta-squared, omega squared, and excluding percentages, *r* values, and r-squared values?

Table 1 (cont'd)

Confidence interval	0, 1	Was a confidence interval reported?
t value	0, 1	Was a *t* value reported?
f value	0, 1	Was an *f* value (resulting from a test comparing means) reported?
Chi-square value	0, 1	Was a chi-square value (resulting from a comparison of frequency data) reported?
t test or ANOVA without *M*	0, 1	Was a test comparing means (e.g., *t* test) carried out without reporting one or more means being compared?
t test or ANOVA without *SD*	0, 1	Was a test comparing means (e.g., *t* test) carried out without reporting the standard deviation of one or more means being compared?
t test or ANOVA without *t* or *f* value	0, 1	Was a test comparing means (e.g., *t* test) carried out without reporting the resulting *t* or *f* value?
p values	Open	How many *p* values were reported?
$p =$	0, 1	Was an exact *p* value reported?
$p <$ or $>$	0, 1	Was a "relative" *p* value reported (i.e., reported as greater or less than a particular value such as .05)
Research questions	0, 1	Were research questions or hypotheses stated as such?
A priori alpha	0, 1	Was an a priori level of statistical significance reported?
Power analysis	0, 1	Were the results of a power analysis (a priori or post hoc) reported?
Assumptions checked	0, 1	Did the author report checking the assumptions of statistical tests used, whether or not they were met?
Reliability-interrater	0, 1	Was a measure of interrater reliability reported?
Reliability-instrument	0, 1	Was a measure of instrument reliability (excluding indices of interrater or intrarater agreement) reported?
Visual	0, 1	Were data presented visually (e.g., bar graph, scatter plot)
Effect Size		
Effect size	Open	The effect sizes as reported or as calculated based on reported data

Analysis

The analyses required to answer my four research questions were relatively straightforward and similar to reviews of research practices in other fields (e.g., Goodwin & Goodwin; 1985; Kieffer et al., 2001; Willson, 1980). Research questions 1 and 2 were addressed by calculating frequencies and percentages for the different designs, statistical analyses, and data reporting practices found in the sample. Following Plonsky and Gass (2011) I cross-tabulated frequencies and percentages for four design types as well: (a) observational studies carried out in

classrooms (O+C), (b) observational studies carried out in laboratories (O+L), (c) experimental studies carried out in classrooms (E+C), and (d) experimental studies carried out in laboratories (E+L). Frequencies and percentages were also calculated for the four features associated with experimental control (random group assignment; inclusion of a control or comparison group; pretesting; and delayed posttesting) across the latter two design/setting types, E+C and E+L.

To answer research question 3, I examined effect sizes (*d* values) within subgroups of studies based on different research and reporting practices associated with methodological quality (i.e., using variables used to answer research questions 1 and 2 as grouping or independent variables and using *d* values as the dependent variable). Although several different types of effect sizes were extracted from primary reports, this phase of the analysis included only *d* values based on between-groups contrasts and, in the case of experimental studies, immediate posttests. This decision was based on three related considerations. First, more *d* values of this type were available for analysis than any other type of effect size. Second, although SLA research also collects and analyzes other types of data, the primary type of analysis involves comparisons of group means, a condition suggested by previous reviews and indicated empirically by the number of *d* values available in the Results (see also Gass, 2009). And third, although effects based on pre-post and repeated measures contrasts were calculated, these were not combined or analyzed with effects from between-groups contrasts because the former has been shown to create an upward bias in effects relative to the latter (Lipsey & Wilson, 1993; Morris, 2008; Oswald & Plonsky, 2010; Norris & Ortega, 2000; Wilson & Lipsey, 2001). Moreover, I calculated *d* values for within-group contrasts using the same formula for between-groups contrasts because pre-post correlations were almost never reported; this calculation also overestimates effects (Cheung & Chan, 2004; Gleser & Olkin, 2009; Plonsky & Oswald, in

press). Before combining effect sizes across studies, I converted negative effects to their inverse or absolute value. This choice was motivated first by an interest in determining the magnitude as opposed to direction of effects in the field. Second, in a large portion of the studies in the sample, and especially in observational studies, there was not a clear prediction for the direction of differences between groups. That is, in many cases, researchers tested two-tailed hypotheses, and I chose to include the data from these studies rather than exclude them on the basis of having to arbitrarily assign a direction to their effects. And third, many studies with directional hypotheses included multiple measures, for which some a higher score was predicted for the treatment group and for others a higher score was predicted for a comparison group. This condition was exceedingly common among studies with both self-paced reading or response time instruments (i.e., tests where the assumption is that a *lower* score indicates greater knowledge of a particular structure) as well as more typical language assessments (i.e., tests where the assumption is that a *higher* score indicates greater knowledge).

A sensitivity analysis was also carried out to determine whether outliers or other irregularities were present in the set of *d* values. Specifically, I examined the descriptive statistics as well as a histogram (see Figure 1) and a stem-and-leaf plot, and decided not to exclude any cases from subsequent analyses. (For a recent treatment of graphic techniques for displaying meta-analytic data and for assessing publication bias in particular, see Anzures-Cabrera & Higgins, 2010.) Despite a somewhat wide dispersion of effects greater than the average *d* value (M = .88; SD = .73; skewness = 2.65), graphic portrayals of the data revealed a continuity of values extending evenly away from the mean towards the largest values. Furthermore, using the more typical criterion for identifying outliers beyond three standard deviations from the mean (see Lipsey & Wilson, 2001) would have resulted in the removal of 7 of 236 total values. Based

on this analysis and the comprehensive nature and volume of studies in this sample I chose not to exclude these values.

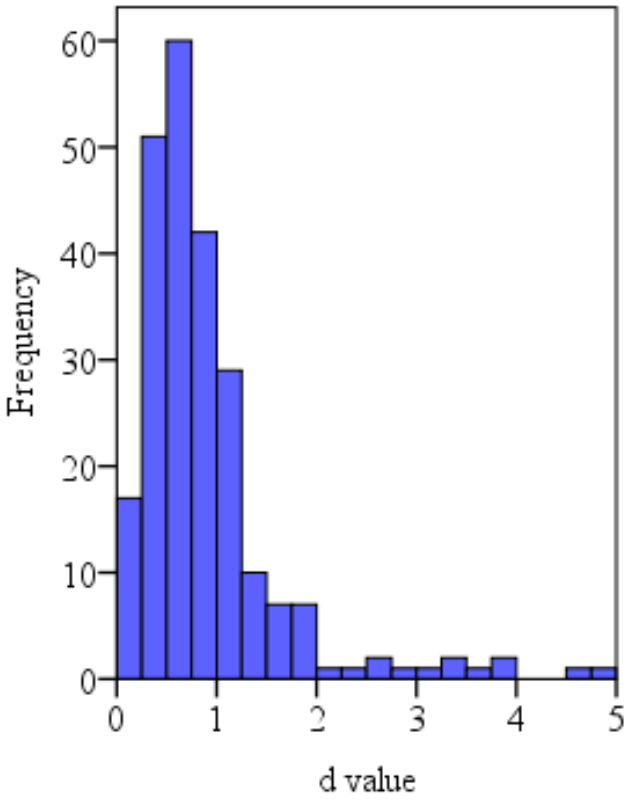

Figure 1. Histogram of *d* values included in the present study. (For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.)

Following these minor preparations of the data, the effects from each study were averaged so that each study would contribute a maximum of one *d* value to the analysis used to answer research question 3 (see Wilson & Lipsey, 2001). I combined effects in this manner to avoid statistical nonindependence (due, for example, to a common standard deviation in experimental-control contrasts) and bias in favor of studies with a large number of samples

(Cameron & Pierce, 1996; Gleser & Olkin, 2009; Hedges & Olkin, 1985; Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998). Mean *d* values, their standard deviations, and 95% confidence intervals were then used to compare effects (a) between the two major types of designs (observational and experimental) and settings (classroom and laboratory), (b) across studies with and without the four features associated with methodological quality in experimental designs (e.g., pretesting) overall and between E+C and E+L studies, and (c) across studies with and without different reporting practices associated with quality both overall and across the four design/setting categories (O+C, O+L, E+C, E+L).

Research question four involved taking a developmental approach. I compared the data obtained to answer research questions 1, 2, and 3 for studies published 1990-1999 and 2000-2010. Percentages of each design type/feature, statistical analysis, and type of data reported were calculated for each decade. Average *d* values were also calculated for the respective ten- and eleven-year intervals overall and across the four setting/design categories in order to examine any possible changes over time occurring as a function of these study characteristics.

Chapter 3 RESULTS

Results for Research Question 1: Designs and Analyses in L2 Research

The first research question was motivated by an interest in several facets of study design and methodological quality. Specifically, it asked the extent to which a number of methodological features and statistical/analytical procedures had been present in quantitative L2 research. Tables 2 and 3 show the frequency and percentage of studies for which different designs and settings were employed. Overall, we see that the vast majority of L2 research is lab-based as opposed to classroom-based and observational as opposed to experimental. There also appears to be a relationship between designs and settings. Whereas experimental research was carried out nearly equally in classrooms and labs, observational studies were 80% lab-based and only 20% classroom-based.

Table 2

Major Designs across Research Settings in L2 Research[a]

	Classroom		Laboratory		Total	
	K	%	*K*	%	*K*	%
Observational	91	20	359	80	450	100
Experimental	78	45	94	55	172	100

Note. [a]A small number of studies were both observational and experimental (e.g., Loewen, 2005), or were carried out in both laboratories and classrooms (e.g., Lyster & Izquierdo, 2009). Thus, the total number of studies across all cells, and the total numbers of studies that percentages are based on, is greater than 606.

Table 3

Research Settings across Major Designs in L2 Research[a]

	Observational		Experimental		Total	
	K	%	*K*	%	*K*	%
Classroom	91	54	78	46	169	100

Table 3 (cont'd)

Laboratory	359	79	94	21	453	100

Note. [a]A small number of studies were both observational and experimental (e.g., Loewen, 2005), or were carried out in both laboratories and classrooms (e.g., Lyster & Izquierdo, 2009). Thus, the total number of studies across all cells, and the total numbers of studies that percentages are based on, is greater than 606.

Looking across the four design features associated with quality in experimental designs (i.e., random assignment to experimental conditions, inclusion of a control or comparison group, inclusion of a pretest, inclusion of a delayed posttest), the results are somewhat mixed (see Table 4). Overall, about half of the experimental studies reported to have assigned participants to experimental conditions randomly whether by group or individually. And as we might expect, lab studies were more likely than classroom-based studies to randomize conditions. Use of comparison groups was very common. Approximately 90% of experimental L2 research carried out in both classrooms and labs compared the effects of interventions to a control or comparison group. A somewhat smaller number of experimental studies included pretests: 78% of classroom- and 59% of lab-based studies, respectively. And among experimental studies in this sample, approximately 38% examined the durability of treatments using one or more delayed posttests, but the rate of inclusion of delayed posttests is not consistent across research settings: Classroom studies were substantially more likely to measure effects at a delayed interval (50%) than lab studies (29%).

Table 5 shows the size of samples in L2 research, a factor related to power and precision of findings and thus research quality. Overall the studies in this sample included 1,732 unique

samples with most studies (85%) including 1-4 groups or samples. The samples in a very small number of studies were in the thousands, but the median group/sample size was 19.

Table 4
Design Features Associated with Quality in Experimental L2 Research

		Experimental + Classroom[a]		Experimental + Lab[b]		Total[c]	
Variable	Value	*K*	%	*K*	%	*K*	%
Random Assignment	Individual	18	23	45	48	63	37
	Group	13	17	4	4	17	10
Comparison group	Yes	70	90	79	84	149	87
Pretest	Yes	61	78	55	59	116	67
Delayed posttest	Yes	39	50	27	29	65	38

Note. [a]Out of a total of 78 Experimental + Classroom studies. [b]Out of a total of 94 Experimental + Lab studies. [c]Out of a total of 172 (quasi-)experiments.

Table 5
Sample Sizes in L2 Research

Total *N*	Unique Samples (*K*)[a]	Median *n*	Min-Max
181,255	1,732	19	1-34,069

Note. Modal *k* per study = 1

Research question 1 also asked about the use of different statistical procedures and techniques. In Table 6 we see that the majority of studies in SLA have been interested in testing for differences between group means using (in descending order) ANOVAs, *t* tests, MANOVA, ANCOVA, and MANCOVA. Nonparametric equivalents of these tests were also used occasionally. Approximately 30% and 15% of the studies in this sample used correlations and regressions, respectively, and categorical/frequency data were used to calculate chi-squares in about one-fifth of the sample. Additional analyses carried out sporadically include factor analysis, structural equation modeling, discriminant function analysis, Rasch analysis, and others

including cluster analysis, unidimensional and multidimensional scaling, and growth curve analysis.

SLA research often employs multiple statistical techniques in conjunction with each other. Table 7 shows that although no inferential statistics were reported in 12% of the sample, and 28% report using a single statistical analysis, 60% reported using multiple analyses. Related to the number of unique analyses is the number of tests of statistical significance (see Table 8). Studies of L2 research report 35 *p* values on average (median = 18). However, the distribution of the number of *p* values is positively skewed with several outliers (on the high end) as indicated by the somewhat large standard deviation relative to the mean (64). The intensive and extensive use of statistical testing in SLA also has implications for statistical power (see Discussion chapter). Finally, due to the omission of nonstatistically significant findings (see results below related to reporting of inferential statistics), these numbers may not accurately reflect the practice of statistical testing in SLA; in reality, these data likely underestimate the number of tests performed.

Table 6
Statistical Analyses in L2 Research

Type of analysis	*K*	% of total
ANOVA	341	56
t test	263	43
Correlation	189	31
Chi-square	115	19
Regression	89	15
MANOVA	40	7
ANCOVA	31	5
Factor analysis	30	5
SEM	14	2
DFA	7	1
MANCOVA	4	1
Rasch analysis	4	1
Nonparametrics	30	5
Other	23	4

Table 7
Number of Different Statistical Analyses Used

Number of analyses	*K*	% of total
Zero	74	12
One	170	28
Multiple*	362	60

Note. Maximum = 6

Table 8
Tests of Statistical Significance in Reports of L2 Research

M	*SD*	95% CIs	Median	Min-Max
35	64	30 - 40	18	0 - 975

Results for Research Question 2: Reporting Practices

Research question 2 approaches study quality by examining data reporting practices with respect to several sources, including the guidelines provided by the APA and several SLA journals, measures of methodological quality from the research synthesis literature (e.g., Valentine & Cooper, 2008), recommendations for improving SLA reporting practices (e.g., Chaudron, 2001; Norris & Ortega, 2006), and meta-analyses that have identified missing data in different areas of L2 research (e.g., Mackey & Goo, 2007). I grouped reporting practices into three categories: one for descriptive statistics, one for inferential statistics, and a third for other types of data and reporting practices associated with study quality and recommended by the APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008) but not covered by the first two categories.

Table 9 displays the frequencies and percentages of studies reporting different descriptive statistics. Thorough reporting of data in this category not only enables consumers of primary research to more fully understand and assess study findings but it also avails primary data to

would-be meta-analysts who require such data to calculate an effect size. The most commonly reported descriptive statistic, not surprisingly, was sample size, reported in 99% of the sample. Also regularly reported were percentages (68%), frequency data (48%), and correlations (30%), all of which reflect the use of analyses described earlier such as chi-squares, correlations, regressions, and so forth. Means are reported in 77% of the sample, reflective as well of the largely means-based analyses in SLA, but standard deviations which we would expect or hope to be reported equally as often as means were reported in only 60% of the sample. The difference of 17%, however, obscures the frequency with which means are reported without standard deviations because reporting both statistics does not ensure that all means reported in studies with means were accompanied by their standard deviations. For this reason, studies were also coded for the presence of a mean without a standard deviation, found in 31% of the sample. About a quarter of the studies in the sample reported an effect size, but only 5% reported confidence intervals. Although not coded for in the entire sample, I noticed that at least nine studies reporting effect sizes did so only for statistically significant results. This practice suggests a bias toward statistically significant results in that those authors did not deem it necessary to report effect sizes along with nonstatistical results.

Table 9
Descriptive Statistics in L2 Research

Variable	*K*	%
Sample size	601	99
Percentage	413	68
Frequency	288	48
Correlation	182	30
Mean	464	77
Standard deviation	364	60
Mean without SD	189	31
Effect size	155	26
Confidence interval	27	5

Table 10 displays the frequencies and percentages of studies reporting different types of inferential statistics with particular attention to means-based analyses which are by far the most common type in SLA. Roughly in line with the results for statistical tests in Table 6, *f* and *t* statistics were reported in 61% and 36% of the sample. While high, these figures are somewhat lower than the percentage of studies reporting use of ANOVAs (56%), *t* tests (43%) and other means-based tests. Moreover, the number of studies omitting one or more *f* or *t* values is much higher. Nearly one quarter of the sample (24%) reported a statistical test comparing means without reporting the appropriate *f* or *t* statistic. In many cases, these omissions co-occurred with nonstatistically significant *p* values. Similar to the situation described in the previous paragraph where effect sizes were reported for statistical results only, an even larger number of reports (*k*=27) omitted *f* and *t* values when $p > .05$ but did not omit them when $p < .05$. Some of these omissions were implicit such as when the results of nonstatistically significant ANOVAs were reported as "$f < 1$". In many other cases, authors carried out statistical tests but stated explicitly that only those results found to be statistically significant would be reported. In addition to test statistics such as *f* and *t*, the descriptive statistics that accompany those tests were often omitted as well. One or more means and standard deviations being compared in a statistical test were not reported in 20% and 35% of the sample, respectively. Although not necessarily indicative of a bias toward statistically significant results, this finding provides field-wide evidence for the missing data problem in SLA discussed regularly in meta-analyses of L2 research (e.g., Russell & Spada, 2006; Plonsky, in press-a). Chi-square values were reported in 17% of the sample, 2% less frequently than the use of chi-square analyses in SLA (19%).

In an effort to better understand the reporting practices for inferential statistics, this part of the analysis also considered the different ways *p* values resulting from tests of statistical significance are reported. The results, also in Table 10, show that four out of five reports of L2 research report one or more *p* values as greater or less than a particular cutoff for statistical significance such as .05. Approximately half of the studies in this sample reported exact (as opposed to relative) *p* values. Studies were also coded for their (in)consistency in reporting *p* values. 44% of the sample reported *p* values in a consistent manner throughout the report, and 42% of the studies reported both relative and exact *p* values.

Table 10
Inferential Statistics in L2 Research

Variable	*K*	%
f	369	61
t	216	36
x^2	105	17
p (none)	80	13
p =	296	49
p < or >	487	80
p either = or >/<	269	44
p = and < or >	257	42
ANOVA / *t* test without M	121	20
ANOVA / *t* test without SD	213	35
ANOVA / *t* test without *f* or *t* value	144	24

The frequencies and percentages of studies with additional reporting practices associated with quality are shown in Table 11. These results are mixed at best in that only one could be said to be reported consistently in SLA research: research questions or hypotheses (80%). Visual displays of data were present in just over half of the sample. Reliability coefficients were found in almost half, with 31% reporting an estimate of interrater reliability and 21% instrument reliability (e.g., internal consistency). Although almost all of the studies reported using statistical

tests (see Table 7), only 22% reported setting an a priori level of statistical significance, 17% reported checking the assumptions of their statistical tests, and power analyses were reported in only six studies (1%), two of which were carried out post hoc.

Table 11
Other Reporting Practices in L2 Research

Variable	*K*	%
Research questions or hypotheses	485	80
Visual displays of data	318	53
Reliability (either or both)	273	45
Interrater reliability	185	31
Instrument reliability	128	21
Pre-determined alpha	133	22
Statistical assumptions checked	101	17
Power analysis	6	1

Results for Research Question 3: The Relationship between Methods and Outcomes

Research question 3 addresses the outcomes (i.e., *d* values) produced in L2 research in relation to different designs and reporting practices. With respect to major design types, virtually no difference was found between observational and experimental studies (see Table 12). Likewise, the data show no relationship between research setting, classroom vs. laboratory, and study outcome. Mean effects from all four categories are very similar to those of the overall meta-analytic mean of the sample (d = .88). Although it is outside the scope of this study to adequately address or propose standards for interpreting *d* values in L2 research, these data provide a starting point for building on Cohen's benchmarks for the social sciences in general and on Oswald and Plonsky's (2010) tentatively proposed benchmarks for SLA, both of which may underestimate the magnitude of effects for L2 research as found in this study.

Table 12
Subgroup Analysis of Effect Sizes (Cohen's d) across Designs and Settings

Variable	Value	*k*	*M*	*SD*	95% CI
Design	Observational	141	.91	.73	.79-1.03
	Experimental	101	.87	.74	.73-1.02
Setting	Classroom	81	.88	.74	.76-1.00
	Laboratory	157	.89	.74	.77-1.00
Total		236	.88	.73	.79-.98

Note. All effect sizes aggregated here were extracted from between-groups comparisons and, in the case of experimental studies, on immediate posttest results.

Table 13 presents descriptive statistics for effect sizes from studies with and without four design features associated with experimental quality overall and in both major settings. Overall these results show little indication of differences in effects between studies of higher and lower quality. Despite the variation in average effects depending on how participants are assigned to groups, overlapping confidence intervals indicate that these differences are not necessarily trustworthy. It is perhaps worth noting, however, that among these three subgroups (random assignment to individuals, to groups, and no random assignment), standard deviations from studies that randomly assign experimental conditions at the individual level are much smaller and have narrower confidence intervals, indications of consistency among observed effects. Because all *d* values used in this phase of the analysis are based on between-groups contrasts, no comparison can be made between studies that do and do not include a control or comparison group in their design. Effects in this group of studies are similar to the larger sample of studies in that no difference is observed based on research setting (i.e., classroom vs. lab). Studies that pretest, however, have produced somewhat larger effects overall and across settings than those that do not. Although this difference is not statistically significant as indicated by overlapping

confidence intervals, the pattern and size of the difference is worth noting. Finally, the most striking relationship between study features and outcomes is found among experimental studies that do and do not include delayed posttests in their designs. The overall effect obtained on the immediate posttest from studies with delayed posttests (d = 1.19) is almost twice as large as those without (.64), and the confidence intervals around these means do not overlap. This pattern also holds when comparing effects across research settings although the confidence intervals between E + L studies with and without delayed posttests overlap slightly.

Table 13
Subgroup Analysis of Effect Sizes (d) *Based on Design Features Associated with Experimental Study Quality*

Variable	Value	*k*	M (*d*)	SD (*d*)	95% CIs
Random Assignment	Individual (all)	45	.72	.49	.57-.87
	E + C	15	.71	.41	.48-.93
	E + L	31	.73	.52	.54-.92
	Group (all)	11	1.24	1.07	.52-1.96
	E + C	9	.80	.47	.44-1.16
	E + L	2	3.22	.25	1.02-5.42
	None (all)	56	.99	.87	.76-1.22
	E + C	36	.93	.82	.65-1.21
	E + L	20	1.10	.97	.65-1.56
Comparison group	Yes (all)	101	.87	.74	.72-1.02
	E + C	51	.86	.73	.66-1.07
	E + L	51	.88	.75	.67-1.09
	No (all)	0	-	-	-
	E + C	0	-	-	-
	E + L	0	-	-	-
Pre-test	Yes (all)	75	.93	.82	.74-1.12
	E + C	40	.91	.79	.65-1.16
	E + L	36	.95	.85	.66-1.23
	No (all)	26	.71	.37	.56-.85
	E + C	11	.71	.39	.44-.97
	E + L	15	.71	.36	.50-.91
Delayed post-test	Yes (all)	42	1.19	.91	.90-1.47
	E + C	25	1.13	.90	.77-1.51
	E + L	18	1.23	.92	.77-1.69
	No (all)	59	.64	.48	.52-.77
	E + C	26	.60	.37	.45-.75
	E + L	33	.68	.55	.49-.88

In addition to design features, seven reporting practices associated with quality were also examined in relation to study outcomes. Because of the relevance of these reporting practices to observational and experimental studies, both designs were included in the analysis. The results are reported in Table 14. Similar to the comparisons between studies with and without design features associated with quality in experimental research, these data provide little to no evidence in favor of a relationship between study outcomes and reporting practices in SLA. In fact, there are no statistically significant differences between subgroup means in overall or between-settings comparisons. Nevertheless, there are three non-preferred reporting practices that appear to be related to studies with larger effects: not reporting an a priori Type I error rate, not reporting having checked statistical assumptions, and not reporting data visually. But the differences in effects between these groups of studies and those with preferred practices are small and not statistically significant. The one preferred practice associated with a somewhat larger effects is the reporting of research questions or hypotheses (d = .91 vs. .70). However, again, the confidence intervals around these means overlapped slightly. Comparisons across designs and research settings followed the overall patterns and will not be described in detail.

Table 14
Subgroup Analysis of Effect Sizes (d) *Based on Reporting Practices Associated with Study Quality*

Variable	Value		*k*	M (*d*)	SD (*d*)	95% CIs
Reliability reported	Yes		112	.85	.73	.72-.99
		O + C	22	.80	.59	.53-1.06
		O + L	44	.93	.81	.69-1.18
		E + C	24	.74	.47	.54-.94
		E + L	29	.98	.91	.63-1.32
	No		124	.91	.73	.78-1.04
		O + C	10	1.23	.72	.72-1.75
		O + L	65	.89	.73	.71-1.07
		E + C	27	.97	.89	.62-1.32
		E + L	22	.74	.44	.54-.94
Pre-set *p*-value	Yes		72	.81	.65	.66-.96
		O + C	10	.83	.49	.48-1.18
		O + L	35	.80	.55	.61-.99
		E + C	19	.89	.77	.52-1.27
		E + L	14	.96	.98	.39-1.52
	No		164	.91	.76	.80-1.03
		O + C	22	.98	.73	.66-1.30
		O + L	74	.96	.84	.76-1.15
		E + C	32	.85	.71	.59-1.10
		E + L	37	.85	.65	.63-1.06
p reported	Exact		151	.86	.74	.74-.98
		O + C	17	.80	.56	.51-1.08
		O + L	70	.87	.77	.68-1.05
		E + C	35	.87	.83	.58-1.15
		E + L	34	.87	.66	.64-1.10
	< or >		219	.88	.71	.79-.97
		O + C	30	.91	.59	.68-1.13
		O + L	103	.92	.78	.77-1.07
		E + C	45	.81	.59	.64-.99
		E + L	48	.90	.76	.68-1.12
	p either = or </>		98	.90	.73	.76-1.05
		O + C	15	.93	.66	.56-1.30
		O + L	45	.94	.72	.72-1.16
		E + C	20	1.01	.83	.62-1.40
		E + L	20	.83	.87	.42-1.23
	Both		136	.86	.72	.74-.98
		O + C	16	.84	.55	.54-1.13
		O + L	64	.89	.79	.69-1.08
		E + C	30	.78	.65	.54-1.02
		E + L	31	.91	.67	.66-1.15
Effect size reported	Yes		93	.86	.67	.72-1.00

Table 14 (cont'd)

		O + C	11	1.10	.76	.59-1.61
		O + L	44	.87	.71	.66-1.09
		E + C	20	.81	.45	.60-1.02
		E + L	23	.79	.65	.51-1.08
	No		143	.90	.77	.77-1.03
		O + C	21	.85	.60	.57-1.12
		O + L	65	.93	.80	.73-1.13
		E + C	31	.90	.87	.58-1.22
		E + L	28	.944	.82	.63-1.26
Assumptions check	Yes		46	.77	.59	.60-.95
		O + C	8	.89	.65	.35-1.44
		O + L	13	.68	.41	.43-.93
		E + C	18	.85	.79	.46-1.25
		E + L	9	.74	.18	.60-.87
	No		190	.91	.76	.80-1.02
		O + C	24	.95	.67	.66-1.23
		O + L	96	.94	.79	.78-1.10
		E + C	33	.87	.70	.62-1.12
		E + L	42	.91	.82	.65-1.16
Visual displays	Yes		133	.92	.79	.79-1.06
		O + C	21	1.02	.72	.69-1.35
		O + L	65	1.01	.91	.78-1.24
		E + C	25	.79	.48	.59-.99
		E + L	27	.83	.78	.52-1.13
	No		103	.83	.65	.71-.96
		O + C	11	.77	.50	.44-1.11
		O + L	44	.76	.41	.63-.88
		E + C	26	.93	.91	.56-1.30
		E + L	24	.93	.71	.63-1.23
Research Questions	Yes		204	.91	.77	.81-1.02
		O + C	28	.99	.67	.73-1.25
		O + L	91	.95	.81	.78-1.11
		E + C	47	.87	.75	.65-1.09
		E + L	45	.90	.78	.66-1.13
	No		32	.70	.39	.56-.84
		O + C	4	.50	.36	-.07-1.08
		O + L	18	.72	.40	.52-.92
		E + C	4	.73	.52	-.09-1.55
		E + L	6	.73	.36	.35-1.10

Results for Research Question 4: Changes over Time

Research question 4 asked how different designs, analyses, reporting practices, and study outcomes have changed over the last 20 years. This question was addressed by calculating the percentage of different types of studies carried out from 1990–1999 and from 2000–2010 and by calculating average *d* values for these decades.

Designs

A number of design-related changes can be observed in Figures 2 and 3, some of which indicate improvements and/or disciplinary progress and others which do not. We see for example that from the 1990s to the 2000s the portion of SLA research concerned with testing the effects of different treatments in experimental studies has increased somewhat in relation to observational research. A very slight change was also observed in research settings where classroom-based studies increased and, conversely, lab-based studies decreased.

In experimental research, a much larger percentage of studies in the 2000s have included both pretests and delayed posttests compared to the 1990s (see Figure 3). Somewhat surprisingly, the use of random group assignment, another indication of study quality in experimental research, decreased over time and at an interval too large to be attributed exclusively to the increase in studies carried out in classrooms where random assignment is not always possible or ethical. The data also reveal a decrease in the portion of experimental studies with a control or comparison group.

Another aspect of research design examined over time was number of participants and unique samples per study. A shown in Table 15 the median total sample size per study increased approximately 11% from 56 in the 1990s to 62 in the 2000s, which may indicate an increase in statistical power (assuming effect sizes and alpha levels remained constant). But the number of

groups per study also rose slightly (approximately 4%) which may mitigate the potential increase in power.

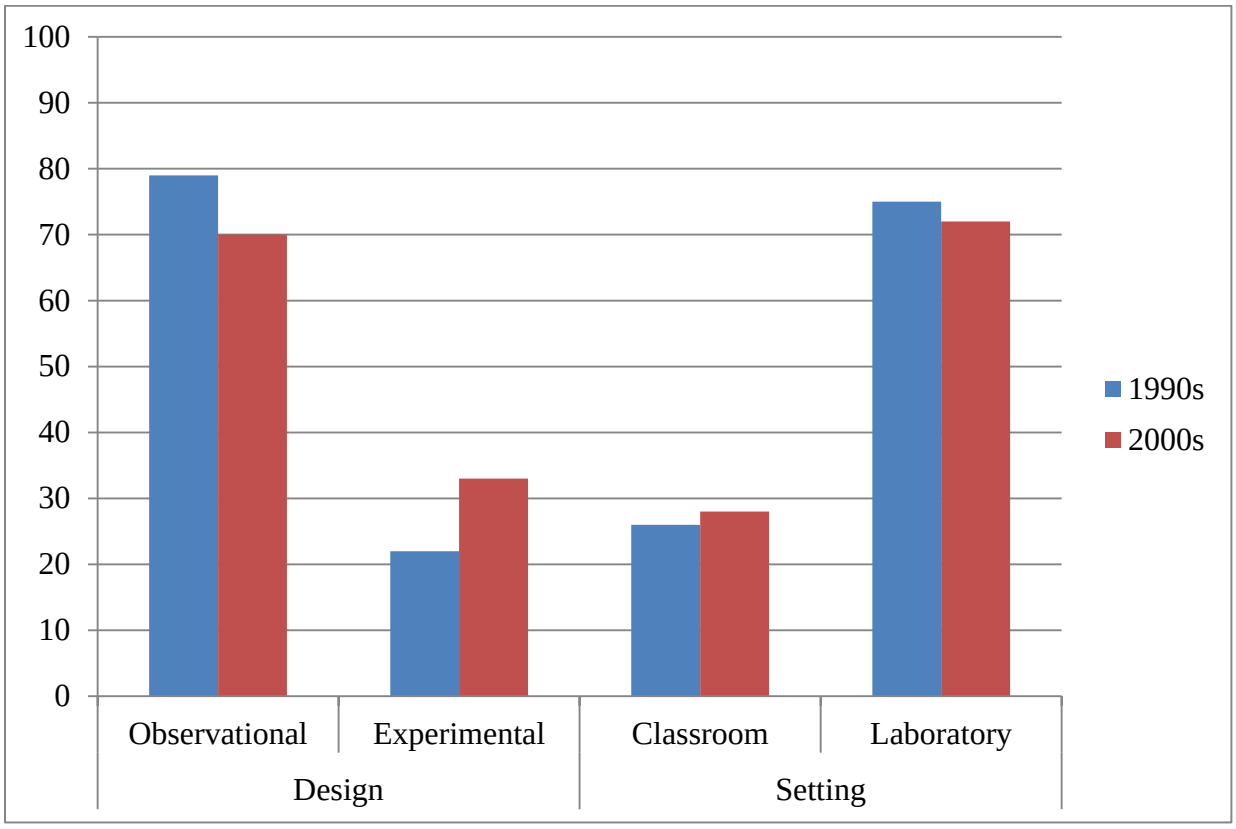

Figure 2. Percentage of major designs and research settings over time.

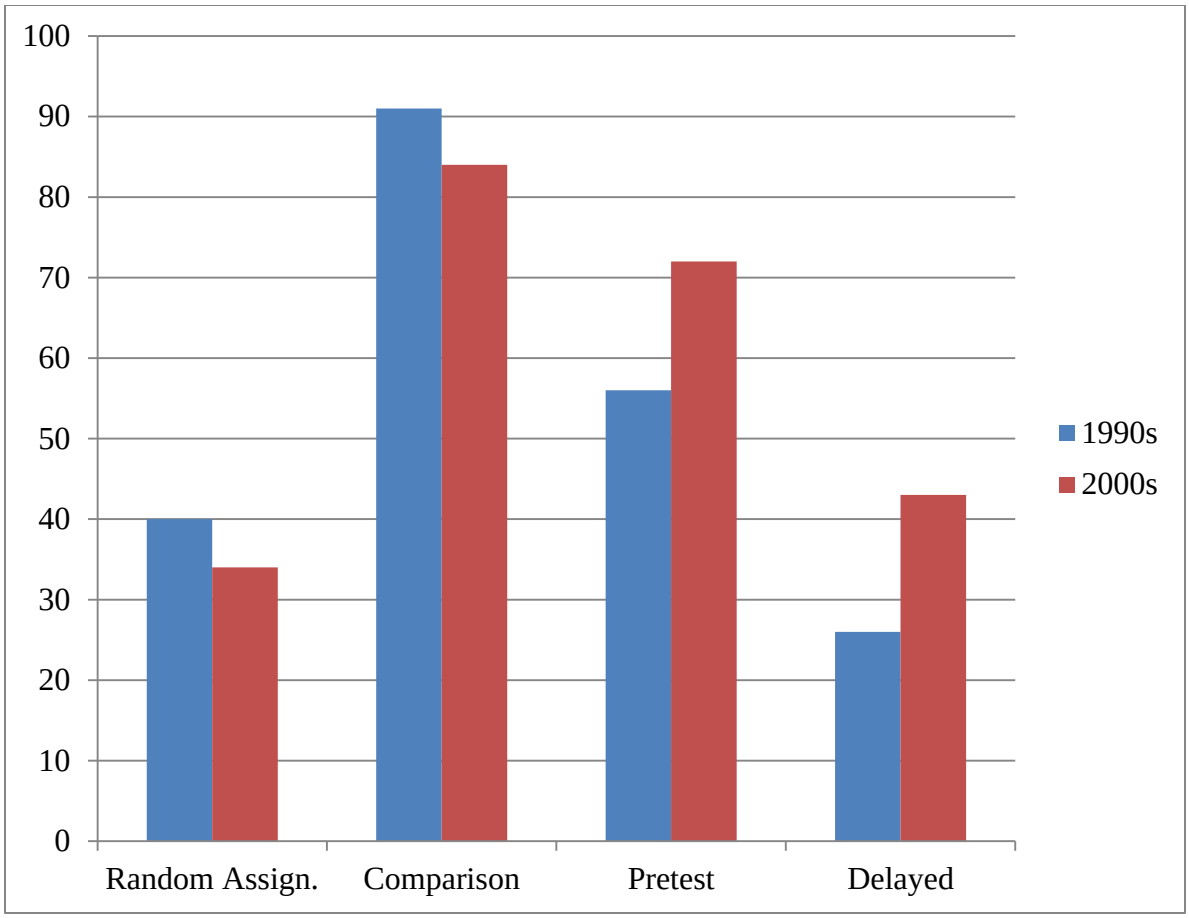

Figure 3. Percentage of design features associated with experimental quality over time.

Table 15
Sample Sizes in L2 Research over Time

Decade	Median *N*	Min-Max *N*	Total *k*	Mean *k*	SD *k*	95% CIs
1990s	56	1-90,789	709	2.78	2.10	2.52-3.04
2000s	62	1-8,593	1,023	2.91	1.90	2.71-3.11

Analyses

Figure 4 compares the percentage of studies employing different statistical analyses during the 1990s and 2000s. Most tests have been used in a greater percentage of studies over time, an indication that studies are now reporting the results of more unique statistical tests than before. ANOVAs and *t* tests, the most frequently used analyses in both decades, both increased

dramatically. Meanwhile, a slight decrease was observed for chi-squares, correlations, factor analyses, and discriminant function analyses.

In the previous paragraph I mentioned an increase over time in the number of unique statistical analyses reported. Figure 5 clearly supports this claim. Here we see the percentage of studies reporting zero or one statistical test decreasing while the percentage of studies reporting multiple tests increases. Finally, the use of *p* values in interpreting the analyses carried out also increased from one decade to the next (see Table 16), which may indicate a loss of statistical power.

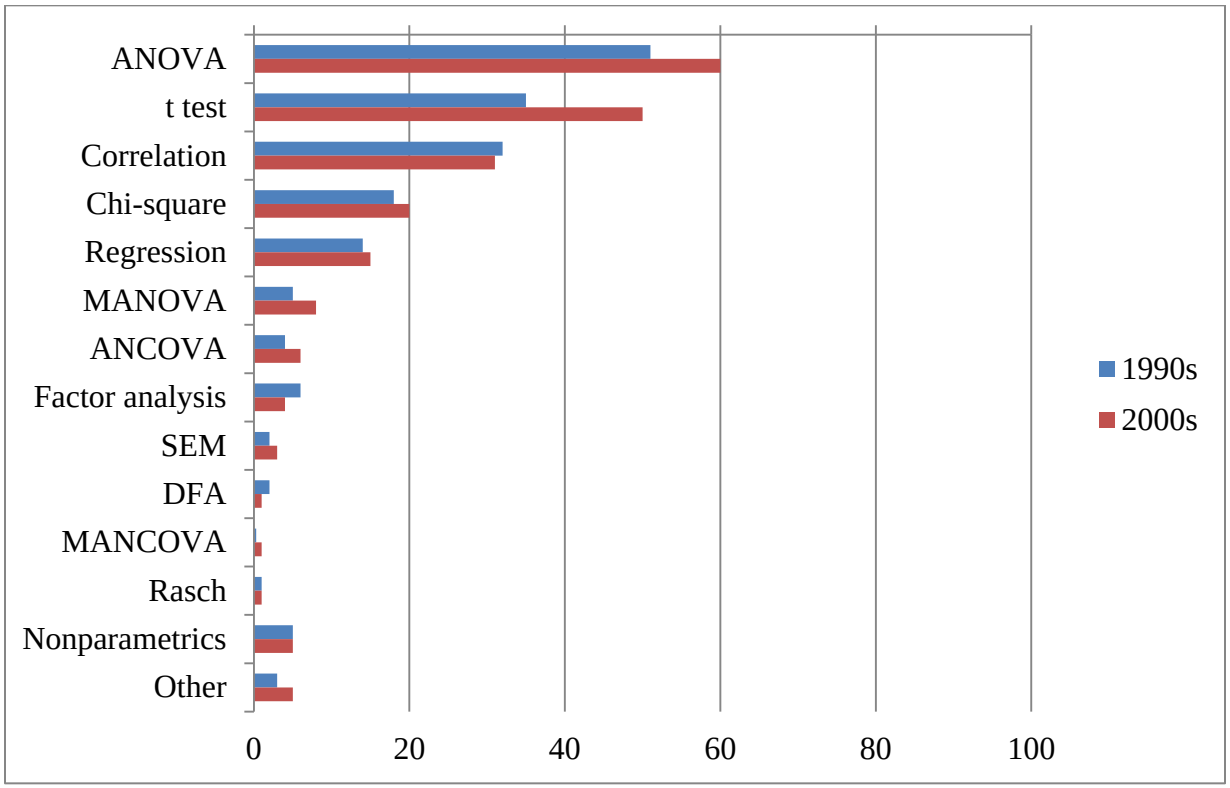

Figure 4. Percentage of studies using different statistical analyses over time.

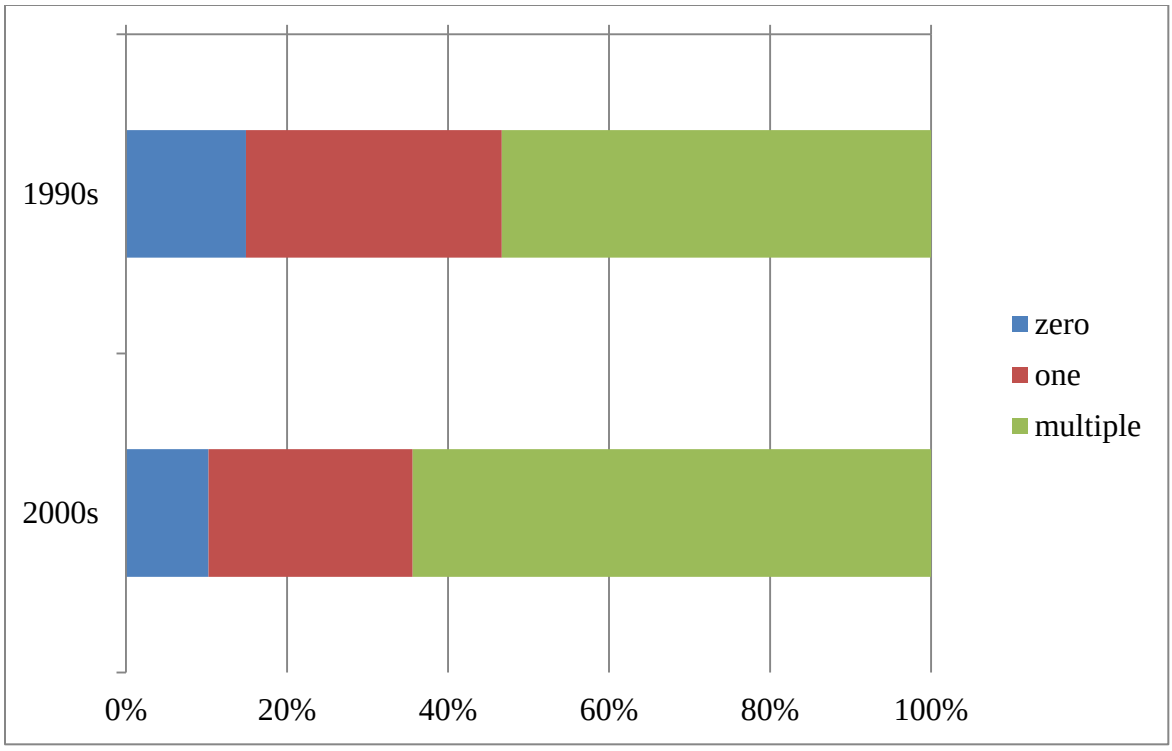

Figure 5. Percentage of studies using different numbers of statistical analyses over time

Table 16
Tests of Statistical Significance in Reports of L2 Research over Time

	M	*SD*	95% CIs	Median	Min-Max
1990s	31	51	25-37	14	0-546
2000s	38	71	30-45	20	0-975

Reporting Practices

Changes in reporting practices were also examined, revealing several important patterns over time. Figure 6 shows that although little change has taken place in the reporting of sample size, frequencies, percentages, and correlations, dramatic improvements have taken place in the reporting of four related statistics: means, standard deviations, confidence intervals, and effect sizes. But reporting in these areas remains far from perfect. There are still studies reporting means without standard deviations, and the number of means and standard deviations in the

2000s is still less than the number of studies carrying out means-based statistical tests during that decade (see Figure 4), but these gains and the benefits they represent for interpreting primary data and analyzing and synthesizing primary data at the secondary level are noteworthy.

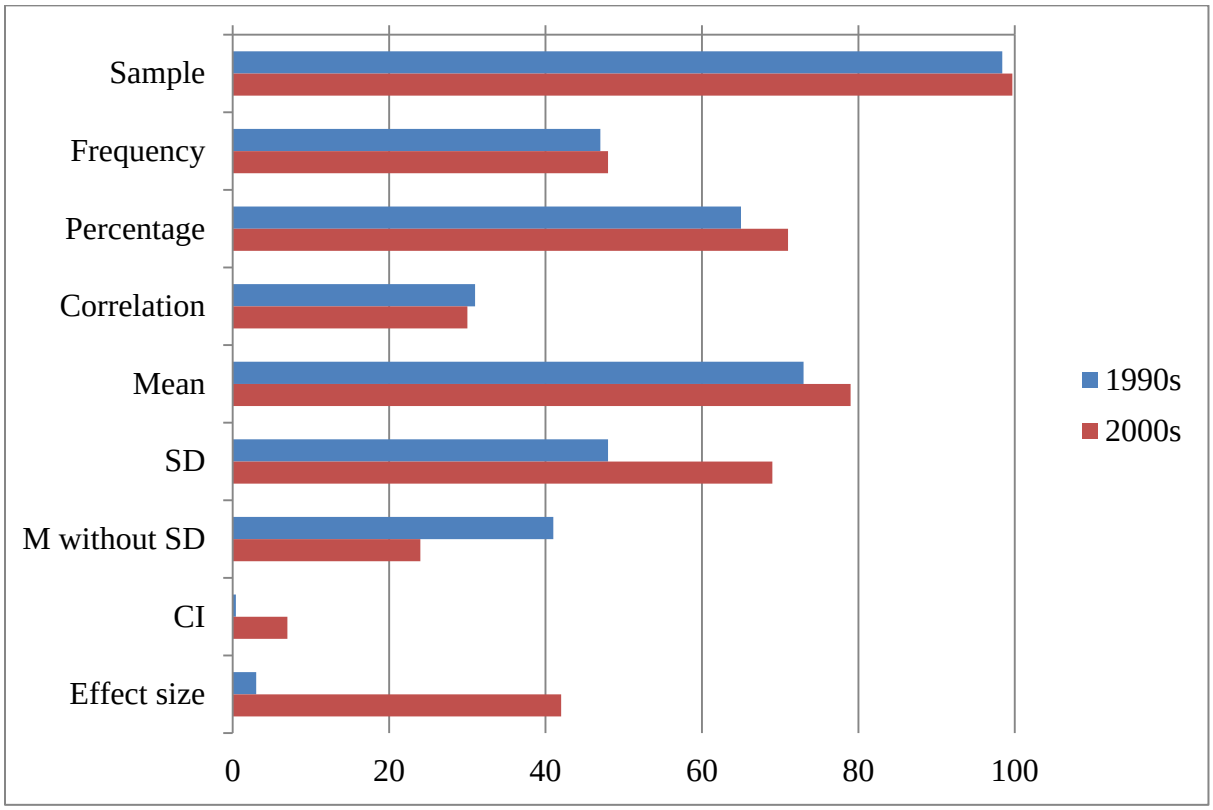

Figure 6. Percentage of studies reporting descriptive statistics over time

Reporting of inferential statistics has also improved (see Figure 7). A much larger portion of L2 research in the 2000s reports exact *p* values. However, at the same time, there was a simultaneous increase in the reporting of relative *p* values, and the number of studies stating *p* in the same way throughout the report (i.e., as either an exact or relative value) decreased. This aspect of data reporting apparently lacks consistency both across and within studies. Related to the increases observed in descriptive statistics, I also found a decrease in the percentage of

studies reporting means-based tests without a mean or standard deviation. A similar change was observed for the reporting of *f* and *t* values for their corresponding statistical tests although it should be noted that neither test statistic reaches the percentage of studies that report analyses producing those statistics.

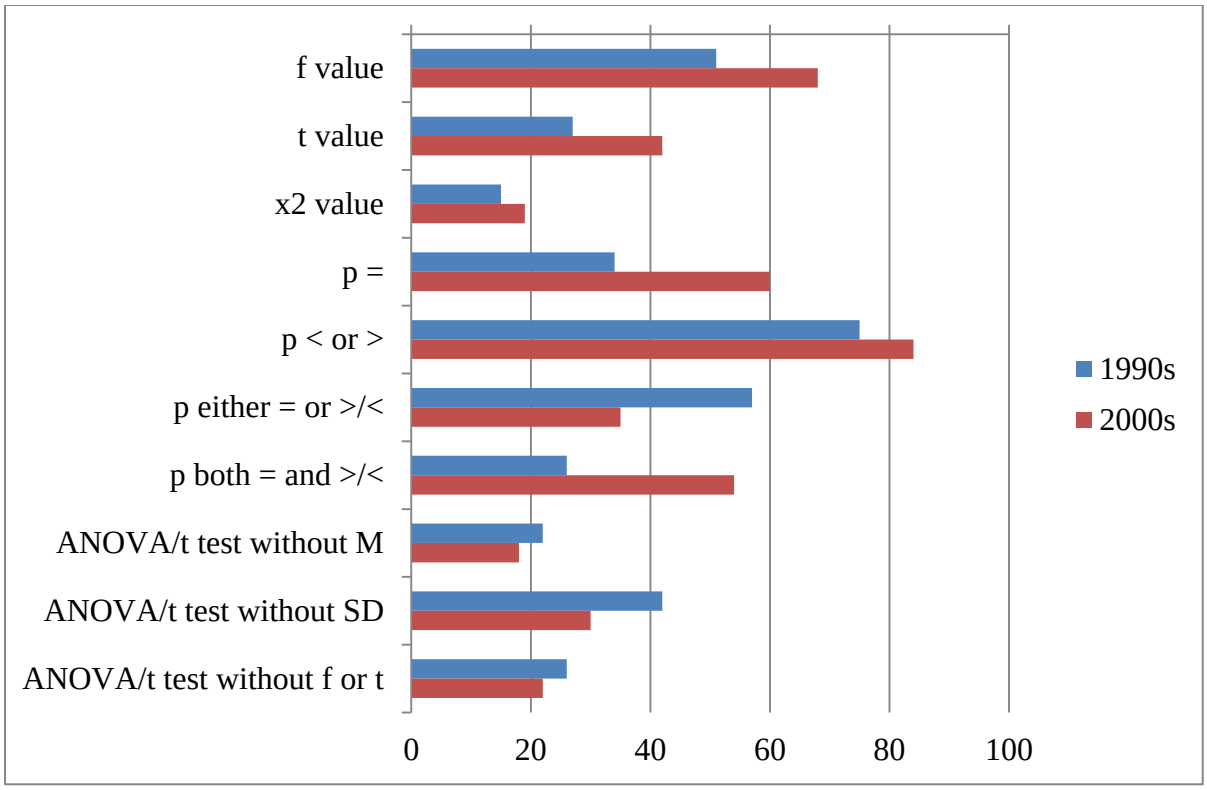

Figure 7. Percentage of studies reporting inferential statistics over time

Changes in the third category of reporting practices improved across the board (see Figure 8). Solid increases were found for the reporting of research questions or hypotheses, visual displays of data, reliability coefficients, a predetermined level of statistical significance, checking of statistical assumptions, and power analysis, the last of which was completely absent in the sample until the 2000s.

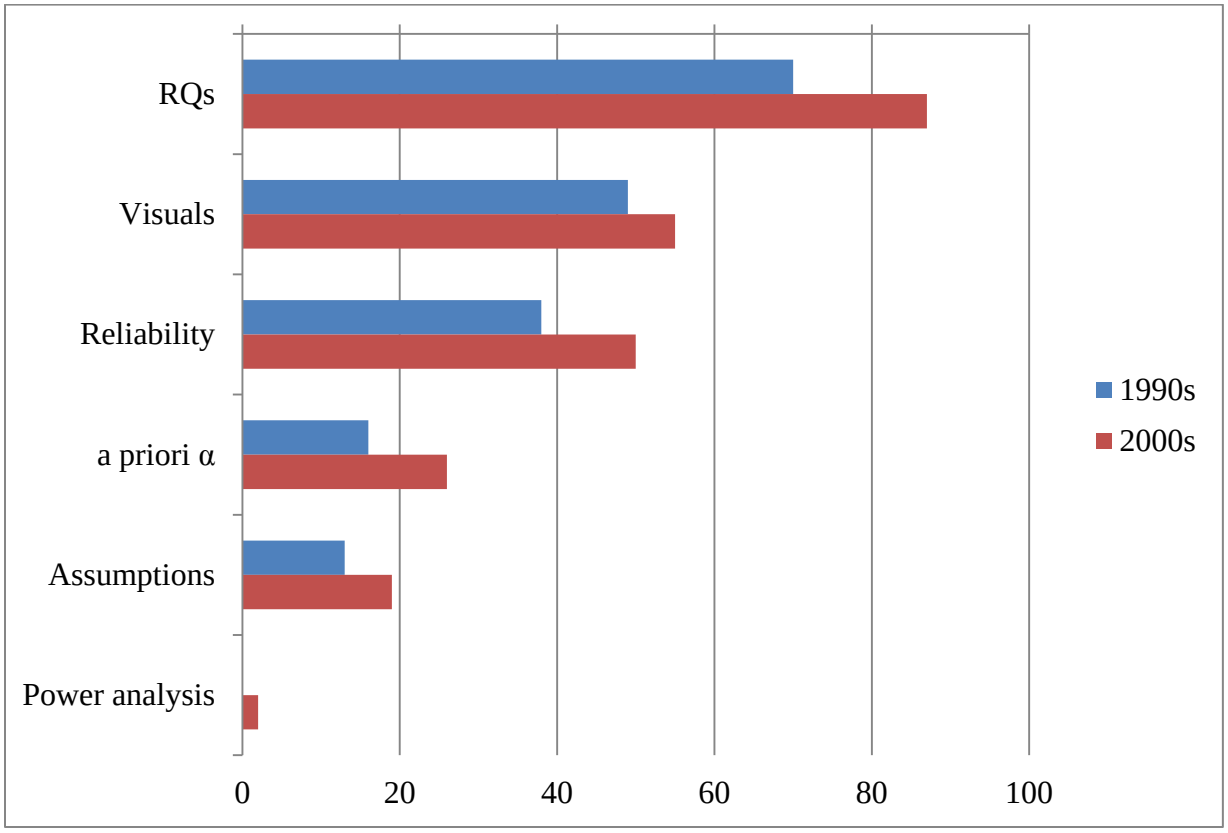

Figure 8. Percentage of studies with other reporting practices associated with quality

Effect Sizes

In addition to probing changes over time in designs, statistics, and reporting practices, research question four asked about changes in the magnitude of effects in SLA. There is no major difference between the 1990s and 2000s overall or across designs and settings (see Table 17). Smaller standard deviations in the 2000s, however, indicate slightly more consistent effects, and a slightly narrower confidence interval indicates increased precision in the estimate of aggregated effects from that decade. Related to the improvements in reporting practices detailed earlier (e.g., standard deviations to accompany means), Table 17 also shows that the number of studies from which a *d* value could be calculated almost doubled from the 1990s to the 2000s.

This change, a proportionately much larger increase than the change in total number of studies in the sample from the 1900s (k = 255) and 2000s (k = 351), demonstrates an increase in the meta-analyzability of the field's research.

Table 17
Effect Sizes over Time across Study Designs and Contexts

Decade		*k*	M (*d*)	SD (*d*)	95% CIs
1990s		81	.87	.75	.71-1.04
	O + C	10	.84	.41	.55-1.13
	O + L	39	.95	.91	.65-1.24
	E + C	17	.74	.47	.50-.98
	E + L	15	.86	.76	.44-1.28
2000s		155	.89	.72	.77-1.00
	O + C	22	.98	.75	.65-1.31
	O + L	70	.89	.67	.73-1.05
	E + C	34	.93	.83	.64-1.21
	E + L	36	.88	.75	.63-1.14

Summary of Findings

- L2 research is predominantly lab- as opposed to classroom-based and observational or non-experimental as opposed to experimental.
- The results are somewhat mixed for design features associated with experimental quality. Random assignment at the individual level was found in 37% of the sample, comparison groups in 87%, pretesting in 67%, and delayed posttesting in 38%.
- The most common inferential statistics were those based on means such as ANOVA (in 56% of the sample) and *t* test (in 43% of the sample). Other common statistics were correlations (31%), chi-squares (19%), and regression (15%). Nearly two-thirds of the sample employed multiple tests. Very few studies (12%) did not use inferential statistics. And the median number of *p* values in the sample was 18.

- The data reported in L2 research parallels the types of analyses conducted, but several significant weaknesses were found including means without standard deviations (in 31% of the sample), ANOVAs or *t* tests without means (20%), ANOVAs or *t* tests without SDs (35%), ANOVAs or *t* tests without *f* or *t* values (24%). Other deficient reporting practices include confidence intervals (found in only 5% of the sample), effect sizes (26%), exact *p* values (49%), reliability estimates (45%), checking of statistical assumptions (17%), and power analyses (1%).
- Little to no evidence was found to support a relationship between study quality and effect sizes at the field-wide level. The only large and statistically significant relationship was found between studies with and without delayed posttests. These results, however, may obscure patterns particular to subdomains within SLA.
- A slight shift has taken place over the last two decades from observational to experimental research and from lab- to classroom-based research.
- Random assignment to experimental conditions and inclusion of a comparison group in experimental research both decreased from the 1990s to the 2000s, but pretesting and delayed posttesting both increased.
- Sample sizes in L2 research have increased slightly over the last two decades, but the number of statistical tests has also increased, mitigating any increase in statistical power.
- The variety of statistical analyses has not changed from the 1990s to the 2000s. The number of unique tests used in reports of L2 research, however, has increased.
- With respect to data reporting practices, several significant improvements were found. Means, standard deviations, and *f* and *t* values were omitted much less frequently. Simultaneously, increases were found over time in the reporting of effect sizes,

confidence intervals, exact *p* values, research questions or hypotheses, visual displays of data, reliability estimates, predetermined levels of statistical significance, checking of assumptions of statistical tests, and power analyses. Substantial improvements are still needed, however, in many of these areas.

- Effect sizes in L2 research did not change across decades.

Chapter 4 DISCUSSION

The primary goal of this study was to better understand previous research and inform future research by examining and quantifying methodological quality in SLA. In light of the multidimensional nature of study quality and the range of study features explored, I did not set out to characterize the breadth of SLA research as "high" or "low" in quality. Rather, this study and the first two research questions in particular asked about the extent to which a number of research and reporting practices were present in L2 research.

In the following two sections I address the results pertinent to those two questions, situating them in the context of previous research and suggesting changes in practice where appropriate. Next I interpret and discuss results of the relationships observed between research practices and outcomes (RQ3), followed by an examination of how both have developed over time (RQ4).

Before concluding, limitations and suggestions for future research on study quality are laid out and I end the chapter by offering specific suggestions for reforming L2 research practices based on the findings of this study. My recommendations are directed to six non-exclusive groups of stakeholders in the field: (a) individual researchers, (b) journal editors, (c) meta-researchers, (d) graduate curriculum committees and researcher trainers in SLA, (e) grant-funding agencies and their reviewers, and (f) the American Association for Applied Linguistics (AAAL), its leadership, and all its constituents with an interest in L2 research. I realize that my recommendations may not reach their audiences, but I feel a responsibility to use the findings from this study to improve the means by which research is carried out and reported on.

Designs and Analyses: Overall

I first examined major design types and settings found in L2 research. These study descriptors are not necessarily associated with quality but are important for understanding the basic approaches taken in SLA. Among the studies surveyed, approximately three-quarters have been lab-based, a somewhat surprising finding given the applied nature of the field and the tendency for researchers to translate their findings into classroom implications. Of course the logistical challenges of classroom research need to be recognized. It is not always practical, pedagogically appropriate, or methodologically viable for L2 researchers to carry out studies in actual classrooms. At the same time, in spite of those who apply SLA research to the classroom, a lab-based majority may indicate L2 researchers' preference/priority in a classic trade-off: Whereas classroom research can preserve ecological validity, lab studies provide the researcher with greater control. I use the term "control" loosely here to refer to design elements associated with quality and statistical validity (e.g., random assignment to experimental conditions; see discussion a few paragraphs down) as well as, more broadly, to the ability to manipulate the procedures, logistics, and environment surrounding a study.

One might interpret these data as evidence that L2 researchers have opted to increase control in these and other forms at the expense of authenticity and ecological validity. Alternately, the relative frequency of lab- versus classroom-based research can be viewed as a rough proxy for different brands of L2 research such as correlational (lab) and experimental (classroom). Of course these generalizations don't hold entirely vis-à-vis the data. Only about half of the classroom research in this sample is experimental. But 80% of lab studies are indeed observational, and although they may not be exclusively "correlational" in the statistical sense—we know that *t* and *f* are the statistics of choice, not *r*—many of these studies are certainly correlational in that they seek to observe how two or more uncontrolled variables (e.g.,

proficiency) co-relate rather than seeking to test the effect of a particular intervention (e.g., instruction). (Later on I will actually argue for increased use of correlation/regression analyses among observational studies.)

With respect to the design types chosen, an issue related to study setting and one I have already begun to discuss, the data show that the vast majority of L2 research is not experimental but observational. This finding contradicts the notion of the typical SLA study where differing instructional interventions are provided to two or more groups who are then compared on posttest measures. That is, experimental research is not so much the rule but the exception.

Also related to and perhaps despite the relative frequency of these two major design types, it is interesting and perhaps telling that nearly all SLA meta-analyses have synthesized corpora of (quasi-)experimental studies (see Oswald & Plonsky, 2010; Plonsky, in press-b). The more immediate relevancy of experimental research for classroom instruction may explain partly this incongruency. Or perhaps experimental research more often addresses the core and/or more intensely-debated questions of SLA—for instance, the effectiveness of instruction, feedback, and CALL—prompting greater attention from meta-analysts.

In either case there appear to be two unique schools of L2 research, each with their own agenda and delimited boundary lines drawn roughly along the two design/setting types described earlier in this paper and much earlier in the context of psychological research (see Cronbach, 1957). One school is populated by observational/correlational/lab studies, the other by experimental/classroom studies. Cronbach (1957) saw these two schools operating separately, much like we find in SLA today. Whereas experimenters exercise great control over an environment (e.g., input, interaction, instruction) to understand its effect (e.g., learning, comprehension, changes in strategies, type and frequency of output), correlators are more

interested in the interplay of learner internal (e.g., L1, proficiency, age of onset) and external variables (e.g., nongrammatical sentences, minimal pairs, foreign vs. second language context) without an eye to posterior consequences.

Cronbach's influential papers (1957, 1975) argued for a convergence of these two schools in a hybrid he called aptitude-interaction treatment (ATI). ATI research has been suggested recently in SLA (DeKeyser, 2009), appearing occasionally in primary studies but especially in meta-analyses which can answer questions not previously addressed explicitly by primary studies (see Oswald & Plonsky, 2010). Spada and Tomita (2010), for example, classified and meta-analyzed effects to measure the types of instruction (explicit or implicit) X linguistic features (complex or simple) interaction. From the primary literature, Li (2010) and Goo (2010) tested the interaction of different types of treatments (feedback types and think-aloud protocols, respectively) and working memory (for other recent examples, see Sheen, 2007, and Mackey & Sachs, in press). The number of possible and theoretically motivated interactions that might bridge correlational and experimental inquiries is nearly endless. However, a thorough exploration of the prospects of ATI for L2 theory, research, and practice is beyond the scope of this paper. My purpose in raising the issue was simply to draw attention to its potential as found by fields such as educational psychology (Cronbach & Snow, 1977; but cf. Tracy, Robins, & Sherman, 2009).

Returning momentarily to design types chosen by L2 researchers, the relative abundance of observational research may also be a function of logistics and "control" as described above in the context of research settings. The one-shot lab-based designs of many non-experimental studies using, for example, grammaticality judgment tests or a battery of aptitude and proficiency measures, may appeal to researchers as a means to acquiring data more immediately than a pre-

post-delayed design with multiple intact classes, multiple instructors, multiple treatment sessions, and considerable potential for contamination and attrition. (Who, among classroom researchers, has not at least once sworn s/he would henceforth use electronic questionnaires as their exclusive means of data collection? I'll be the first to admit that I enjoy the neatness of secondary research.) I am speculating again. Another interpretation of these data—one that gives the benefit of the doubt to researchers—would argue that researcher choices related to designs and settings are simply based on the questions they posed. Of course, conceiving of and conducting a study is a very individual process which is not regularly reported but which might make for an interesting study, as suggested by an audience member at a preliminary presentation of my results (Plonsky, 2011). A case study or a series of case studies could be carried out to get a behind-the-scenes view of the inner workings and decisions otherwise unknown to consumers of primary literature.

Along with basic designs and settings, I also examined the presence of four indicators of methodological quality or rigor in experimental research—random assignment, inclusion of a control or comparison group, pretesting, and delayed posttesting—across both classroom and lab studies. In contrast to what we might expect, classroom-based studies were generally more likely to exhibit these features than lab studies (see Table 4). The only exception was in random group assignment, found in 52% of lab and 40% of classroom studies. Again, the logistics of classroom-based research makes it easy to understand this difference. It is less clear, however, why the figure would be so low in lab studies. (By contrast, 78% of interactionist lab experiments in Plonsky and Gass [2011] assigned individuals randomly to experimental conditions.) It may be that some lab researchers did in fact assign participants randomly but did not report having done so. Other researchers may not be familiar with preferred practices in

experimental research design. Still others may have pretested to verify comparability of groups and thus deemed random assignment unnecessary. The three other features, discussed in the next paragraph, were all found in a greater portion of classroom than lab studies.

Comparison groups were very frequent overall (87%) and in both settings (90% classroom vs. 79% lab), which is similar but in a pattern opposite to that of Plonsky and Gass' (2011) findings (60% vs. 95%, respectively). Pretesting was relatively frequent in classroom studies and somewhat less so in labs (78% vs. 59%). As I mentioned in the previous paragraph, some researchers may have assigned participants randomly to experimental conditions and may have therefore determined a pretest unnecessary. With a relatively small median sample size of 19 (see Table 5) and accordingly low power / large sampling error, however, it may be best to pretest in all experimental research (see Norris & Ortega, 2000). (I return to the issue of statistical power later.) In addition to using pretest scores to verify comparability of groups or as covariates, greater pretesting in experimental research would also enable more precise estimates of gains which can be compared to and meta-analyzed with similar studies. Delayed posttesting was found in relatively few classroom (50%) or lab (29%) studies (38% overall). I find this result somewhat surprising compared to the findings reported by Plonsky and Gass (77% classroom; 81% lab) and in light of the theoretical and practical value of understanding the longevity of instructional effects in SLA.

In addition to designs, settings, and study features, research question 1 also addressed the use of different statistical analyses in L2 research. As found in previous reviews (e.g., Teleni and Baldauf, 1989; Gass, 2009), the most common statistical tests in L2 research are those that compare means (e.g., ANOVA was employed in 56% of the sample; *t* test in 43%). Correlations (31%) and regressions (15%) were also found with some regularity among observed analyses.

Several other types of tests that were found but present in less than 5% of the sample include factor analysis, structural equation modeling, and discriminant function analysis. Skidmore and Thompson (2010) found ANOVA/ANCOVA to be the most frequently used statistic in education and psychology journals. Unlike SLA, however, they also found regression and correlation to be second and third most frequent, and *t* tests were least frequent behind factor analyses, cluster analyses, and nonparametric statistics.

Overall these results show that almost all statistical tests found in L2 research fall under the general linear model (GLM) with the exception of chi-squares (in 19% of the sample) and the nonparametric equivalents of *t* and *f* tests (5%), which are conceptually if not statistically related to the GLM. However, having read 600-some methods sections, I have the sense that the related and hierarchical relationship between the family of GLM statistics is largely unknown to L2 researchers. I mention this not to draw attention to gaps in our knowledge but rather because I wonder if a clear introduction to the GLM and its constituent analyses would prompt a shift in analytical strategies in SLA as was seen in psychology following Cohen (1968; see Skidmore & Thompson, 2010).

Regardless of whether L2 researchers necessarily recognize most of the statistics they use and read about as belonging to the GLM—a revelation obscured or omitted in most quantitative L2 methods textbooks—it might be worth reflecting on the use of ANOVA-type statistics and asking whether SLA is best served by ANOVA and, if not, whether other options might exist. I cannot resolve these issues entirely here, but I will attempt to initiate a conversation by making a few points. First, it appears that the default status of ANOVA has come to shape certain conventions of design and measurement in L2 research, most notably the conversion of continuous-level data to a nominal scale. Researchers in SLA as well as in other social sciences

often convert an intervally-measured variable into a categorical (i.e., between-participants) variable (e.g., motivation → low, high; general proficiency → low, intermediate, high; age → pre- or post-pubescent) to then compare these new groups' means on one or more dependent variables. (Several studies in my sample that took this ill-advised practice one step further in the wrong direction: They formed groups according to the median score on a proficiency test, for example, and then compared the groups' scores using a *t* test.) Several voices from other disciplines (e.g., Humphreys, 1978) and at least one in our own (Lee, 2010) have spoken out against the nominalization of continuous data, arguing instead for correlational/regression analyses. Their point and the one I would like to make here is that "variance is the 'stuff' on which all analyses are based" (Skidmore & Thompson, 2010, p. 791); trading variance for what appears to be a cleaner analytical approach results in a waste of data and a loss of statistical power (Cohen, 1968).

This issue as well was likely brought to the forefront following Cohen's classic paper, now decades old (Cohen, 1968). He demonstrated very convincingly that multiple linear regression presents a more parsimonious and statistically powerful modeling tool than ANOVA, whether the independent variables are intervally or nominally scaled (with dummy coding of the latter). My second point is much the same: L2 theory and research might benefit from greater use of regression analysis. As Cohen (1968) explained, regressions and ANOVA take slightly different analytic paths to reach the same results. To illustrate this concept, consider the fact that (a) ANCOVA is essentially a regression that removes the covariance in the covariate(s), and (b) the well-known eta-squared which is usually calculated from an ANOVA is a type of squared correlation ratio and is mathematically and conceptually identical to the *R*-squared value that results from a regression. But ANOVA and regression part ways in terms of their utility,

parsimony, and statistical power when multiple independent variables are of interest. Whereas multiple ANOVAs would be needed to determine the presence and magnitude of main effects, interactions, and so forth, a regression handles them all in a single test and provides a quantitative indication of their contributions and errors relative to the model. Along these lines multivariate analyses have also been suggested as an alternative to running multiple ANOVAs as is common in L2 research. MANOVA, for example, preserves statistical power and better reflects the multivariate nature of the reality we attempt to measure (Fish, 1988; Raykov & Marcoulides, 2008, chapter 1; see Gelman, Hill, & Yajima, 2009, for an alternate, Bayesian perspective on multiple comparisons). Looking back at Table 8 which shows the use of statistical significance testing in SLA, we might expect these figures to be lower thus preventing inflation of Type I error if more studies had employed regressions and MANOVAs instead of ANOVAs.

Third, if L2 researchers prefer ANOVA, as the data clearly indicate, despite the advantages of regression, why might this be? I think the answer to this question might lie somewhere near the answer to the question of why researchers in SLA and other fields adhere so closely to NHST and *p* values. A commonality between these two analytical practices is the appearance of clarity in results. In the same way that a *p* value is used to indicate that treatment X is / is not categorically more effective than treatment Y or that there is / is not a difference between native and nonnative speakers' judgments of ungrammatical sentences for example, the discrete groups compared using ANOVA provide the researcher with a straightforward means to determining whether groups A, B, and C are / are not different from each other on a particular dependent measure. But these two approaches also share a common weakness. Neither tells us what we really want to know. Instead of asking categorical (yes/no) questions of continuous data, we should be asking and answering questions like "how much more/less effective is treatment A

than B?", "how much more/less accurate are nonnative speakers' judgments?", and "to what extent are groups A, B, and C different from each other on the dependent variable and how much of their difference can be accounted for by one or more independent variables?" To sum up, we trade precision and richness of information for crude clarity.

Before going on I want to be clear about my intention in advocating for analyses such as regression and MANOVA (procedures some would consider somewhat more sophisticated than, say, ANOVA). I have no interest in pushing for state-of-the-art statistics, and I whole-heartedly agree with the APA's recommendation to choose "minimally sufficient analyses" (Wilkinson & the Task Force for Statistical Inference, 1999, p. 598). After all, SLA is not physics or economics, and nor should it try to be. As a consumer and synthesist of primary research, I am generally happiest with a less-is-more approach. Given the opportunity, I recommend thorough reporting of descriptive statistics and graphic displays of data along with as few statistical tests as possible. And it appears at least one physicist and one economist might agree with me despite the reputations of their fields for using very advanced mathematical and statistical procedures. Einstein said "Models should be as simple as possible, but not more so." And Allen Greenspan predicted: "I suspect greater payoffs will come from more data than from more technique." In other words and in another context, accelerated progress in SLA will not come from impressive stats (alone) but from impressive ideas in the realms of both substance and method.

Fortunately, SLA does not suffer from an infatuation with statistical sophistication. This study found a relatively narrow range of analytical strategies. Nevertheless, it is common for studies in SLA to carry out multiple unique statistical procedures within a single study. This finding speaks to the variety and perhaps to the levels and scales of data and questions posed in L2 research. Along these lines, I also found L2 research to rely heavily on NHST (the median

number of *p* values reported per study = 18). This figure may seem high in terms of its effect on power, especially considering the median sample/group size of 19, but the actual number of statistical tests is likely even larger. As I reported in the results, missing data of this type was occasionally declared by authors to be unreported because of nonstatistical results, but it is hard to gauge the number of studies that omit results for the same reason but that do not report the omission. Both cases introduce an upward bias in effects by removing nonstatistical findings from the available literature (Pigott, 2009).

Reporting Practices: Overall

My second research question asked about the thoroughness of data reporting in L2 research. As with the results related to design features, the findings here are mixed at best. Before discussing the results for this phase of the analysis, I want to be clear about my perspective on data reporting practices as a measure of study quality. I do not equate completeness with overall study quality nor incompleteness with low quality. I view the thoroughness with which data are reported as one among several indicators or study quality.

I presented the results for three types of data reporting practices, and each will be discussed in turn. With respect to descriptive statistics, the most notable finding is that approximately one-third (31%) of the studies in the sample reported one or more means without standard deviations. As mentioned above, this practice is problematic because it limits the interpretability in a great number of studies and restricts their data from being included in meta-analytic reviews.

Also related to meta-analytic reviews is the reporting of effect size indices found in 26% of the sample. At the very least, this figure inspires hope in those who see value in alternatives to NHST. Nevertheless, the act of reporting effect sizes such as *d* and eta-squared is in and of itself

of little value because L2 researchers generally do not interpret their meaning. And when they do, they almost always turn to Cohen's (1988) benchmarks (*d* = .2 for small, .5 for medium, and .8 for large) rather than effects from previous and related L2 studies and meta-analyses. Cortina and Landis (2011) and many others before them (e.g., Cohen, 1994) warn that dismissing one arbitrary cutoff (.05) for interpreting findings for another (.2, .5, .8) amounts to nothing more than "being stupid in another metric" (Thompson, 2001, p. 82-83). Oswald and Plonsky (2010) also noted rigidity among interpretations of effect sizes in their review of the methodological choices and outcomes among 27 L2 meta-analyses. Based on this observation and the finding that Cohen's benchmarks generally underestimated the magnitude of effects in L2 research, Oswald and Plonsky (2010) and Plonsky and Oswald (2010) offered ten alternatives to consider when interpreting *d* values including a tentative, field-specific scale (.4 for small, .7 for medium, and 1.00 for large). Future researchers should maximally exploit the data they collect by giving more precise and nuanced interpretations of effect sizes. We owe it to ourselves and to our colleagues to avoid the temptation to boil our results down to a yes/no or to an effect of small, medium, or large.

The pattern of omitted data among descriptive statistics also follows into the realm of inferential statistics. Studies that carried out ANOVAs and *t* tests regularly left out *f* and *t* statistics (24% of the sample), means (20%), and standard deviations (35%). What is particularly troubling is the fact that many of these authors were totally transparent, openly equating nonstatistical significance with insignificance. These findings, along with those regarding the omission of standard deviations, are particularly problematic in light of the potential for bias that they introduce (Pigott, 2009). I'll explain. Chan, Hróbjartsson, Haahr, Gøtzsche, and Altman (2004) found that outcomes reported thoroughly were twice as likely to be statistically

significant than outcomes that were missing data such as *M*s, *SD*s, or *t* values. Based on the stated omission of *f* or *t* values in 27 studies (only) when $p > .05$ and of effect sizes in nine studies (only) when $p > .05$, there is reason to believe that the same practice observed by Chan et al. may also be found in L2 research. My main cause for concern here is not so much for the impact that missing data has on primary studies themselves but for their effect on meta-analytic means which perhaps we should begin to assume to be upwardly biased by nonstatistical findings that go unreported. (The bumper sticker for those who feel strongly about this issue might read something like "*BEWARE: EFFECTS IN REAR VIEW MIRROR ARE SMALLER THAN THEY APPEAR!*") A formal assessment of the presence of publication bias in SLA is certainly needed.

Among other reporting practices related to inferential statistics, I also coded for different uses of *p* values. Overall, reporting of *p* values was widespread (found in 87% of the sample) and exhibited inconsistency both across and within studies. By across-study variability, I refer to the 80% of the sample that reported a relative *p* value whereas 49% reported an exact *p* value (the preferred practice). As for within-study variability or inconsistency, 42% of the sample included both exact and relative *p* values in the same report. On the brighter side, 44% reported all *p* values as only exact or relative throughout the study. Inconsistencies such as this may be dealt with by and the responsibility of journal reviewers and editors, if not individual researchers. However, many reviewers may be burdened by the largely thankless task at hand and they may feel that closely examining the statistical analyses and reporting thereof to be too time-consuming or outside their realm of expertise. Perhaps with these issues in mind, former editor of *The Modern Language Journal*, Magnan (1994) expanded the review process of the journal to include a "specific review for appropriateness of research design, methods, and statistical procedures" (p. 8) for all empirical submissions that advance beyond the first round of reviews. I

do not know what effect the additional round of reviews has had on the manuscripts that passed through that stage, nor do I know if the current editor of *The Modern Language Journal* has maintained Magnan's policy. (I emailed him to inquire about this, but I never received a response). Nevertheless, other editors might consider adopting a similar strategy toward improving analytical and reporting practices in their journals.

The third category of data collected for research question 2 consisted of reporting practices associated with quality and recommended by the APA. Again, in this category, the findings were mixed. In contrast to Henning's (1986) claim that research questions and hypotheses were not often stated, this study found them clearly reported in 80% of the sample. Reliability in one form or another was found to be reported in 45% of the sample. But this practice clearly varies among different subdomains, as previous reviews have found reliability coefficients in as few as 6% of their samples (L2 practice; Nekrasova & Becker, 2009) and as much as 64% (L2 interaction; Plonsky & Gass, 2011). Regardless of variability across SLA, there is clearly room for improvement in estimating and reporting reliability. As discussed in Chapter 1 and similar to other reporting practices discussed already in this chapter, omission of an estimate of reliability not only weakens the interpretability and trustworthiness of individual studies but it also restricts meta-analysts from making the most of existing research and leaves future researchers using the same or a similar instrument without a point of comparison for their instruments and samples.

The last three reporting practices examined were all related to statistical testing in that best practice and journal/societal guidelines often require these steps to be taken prior to conducting the analysis. First, 22% of the sample reported a predetermined level of statistical significance. This finding is approximately the same as found in the interactionist literature

where 25% reported an a priori cutoff for statistical significance (Plonsky & Gass, 2011). Second, 17% of the sample reported having checked statistical assumptions. Whether or not they were met is a different issue, as is any actions taken as a result of having violated one or more assumptions such as bootstrapping. To investigate researchers' responses to having violated assumptions, I cross-tabulated studies that did/did not report checking assumptions with those that did/did not conduct nonparametric equivalents of means-based analyses, and found a strong relationship between the two. Studies that reported and checked assumptions were five times more likely to employ a nonparametric test than those that did not. Third and last is power. Only six studies in the sample (1%) conducted a power analysis. Again, the rarity of power analyses does not necessarily indicate low power. Based on the median sample size of 19, and *d* value of .71, we can, however, roughly gauge post hoc power in the field at .57 or a 57% chance of appropriately detecting statistical significance. The evidence of a "power problem" seems to be accumulating. To review: L2 studies (a) tend to rely on small samples, (b) typically conduct about 18 tests of statistical significance, (c) do not generally produce very large effects (median *d* = .71), (d) occasionally omit results of nonstatistical results, (e) rarely check/report whether statistical assumptions have been met, (f) rarely use multivariate analyses, and (g) almost never conduct power analyses to determine an appropriate sample size despite available effect size estimates in the existing literature.

The Relationship between Study Quality and Outcomes

In Chapter 1 I quoted Lipsey (2009): "Study results are determined conjointly by the nature of the substantive phenomenon under investigation and the nature of the methods used to study it" (p. 150). Taking this assumption as a starting point along with previous reviews investigating effect sizes as a function of research and reporting practices (e.g., Lipsey & Wilson,

1993; Plonsky, in press-a; Plonsky & Gass, 2011), my third research question explored effects in L2 research across different designs and indicators of methodological quality.

In contrast to Plonsky (in press-a) and Russell and Spada (2006), my results showed almost no difference between subgroups of studies based on designs types, settings, or methodological features associated with quality. Some of these results were unexpected and perhaps even counterintuitive. For example, we might expect lab studies to have larger effects than classroom studies, but average *d* values from the two settings (and their respective *SD*s and 95% confidence intervals) were almost identical. No difference between classroom and lab studies was also reported among studies of interaction (Plonsky & Gass, 2011), but a difference was found by three L2 meta-analyses (Li, 2010; Mackey & Goo, 2007; Plonsky, in press-a) thus indicating that study effects may vary across settings in some areas of L2 research but not in others. Despite the finding of no difference in this study, future studies and meta-analyses should consider the setting(s) of their research domains when interpreting results.

Looking across the four design features associated with experimental quality and control, very similar *d* values were generally found for studies with and without each preferred feature. However, there was one exception. As in Plonsky and Gass (2011), studies with one or more delayed posttests produced significantly larger effects than those without. One explanation for this difference/advantage might be that studies are more likely to include a delayed posttest when the initial or immediate effect is expected to be large. In more concrete terms, if a researcher does not expect (and/or does not find) the immediate effect to be significant, s/he may be less likely to assess the permanence of that effect using a delayed posttest. There was no difference between studies that assigned participants randomly to experimental conditions (vs. did not). This finding contradicts what I found among studies of strategy instruction (Plonsky, in press-a)

but replicates Lipsey and Wilson's (1993) finding of no difference between randomized and nonrandomized designs as found across more than 300 meta-analyses from education and psychology. Likewise, no difference was found between studies that did vs. did not pretest, which also differs from Plonsky (in press-a). And no interactions were found between these study features across research settings and effect sizes.

Because I only included *d* values calculated based on between-groups contrasts in this phase of the analysis (see Chapter 2), I could not compare effect sizes from studies that did/did not include a comparison group or between studies with between groups designs vs. pre-post designs. Nevertheless, I was curious whether studies with true control groups (i.e., receiving no treatment) might produce larger effects than studies with comparison groups (i.e., receiving an alternate or traditional treatment). The answer is a qualified yes. The average effects from studies with control and comparison groups differ considerably (d = 1.24, SD = 1.03, k = 14 vs. d = .81, SD = .66, k = 90, respectively), but the 95% confidence intervals around those means overlap somewhat (.64-1.83 vs. 67-.95). To date, L2 meta-analyses have not generally distinguished between these two types of contrasts but future meta-analytic reviews might consider exploring the source of comparisons in assessing the treatment effectiveness as a potential moderating variable.

Findings for the relationship between reporting practices and outcomes were similar to those for design features in that very little evidence of a relationship was found. I partly expected to find a relationship between the reporting of reliability and larger effects. My thinking was that studies that report reliability might be more likely to have piloted and refined their instruments therefore leading to higher reliability and larger effects. But no such relationship was found in this study or in Plonsky and Gass (2011).

To summarize the results of research question 3, the evidence of a relationship between research practices and effect sizes is minimal. Neither higher quality nor lower quality studies were associated with larger (or smaller) effects. However, as discussed earlier, the broad scope of research included in this study may reduce the visibility of patterns occurring among particular subdomains of L2 research. Therefore I hope that the results for this part of my study are not interpreted as conclusive evidence that no relationship exists between methodological practices and study outcomes. As the saying goes, "the absence of proof is not proof of absence". Moreover, given the patterns found in much more localized reviews (e.g., Plonsky, in press-a), I would recommend that future meta-analyses explore effect sizes in relation to the quality of primary studies being synthesized. Syntheses that find larger effects among studies of higher quality might interpret this result as evidence of a link between researchers' substantive knowledge (i.e., knowledge of how to manipulate variables to induce a large effect) and their methodological knowledge (i.e., knowledge of how to appropriately design and report on a study). It is not unreasonable to suppose that those researchers who are most likely to understand the predictions of a particular model (assuming the model is accurate) and are therefore able to exploit differences between variables to generate large and statistically significant effects are the same researchers who are most likely to adhere to preferred practices in designing and reporting the results of their research. Of course, not all predicted relationships are strong/large, so alternate explanations would be needed. We might explain a result of an association between smaller effects and higher quality in reporting practices, for example, by proposing that studies that report data more thoroughly in general may be less likely to suppress or omit nonstatistical findings and other data such as those used to calculate or weight an effect size for meta-analytic

averaging. In this way, thorough reporting might be associated with reduced (but more accurate) effect sizes in a particular study and across a particular subdomain.

Changes over Time

The findings of this study so far have indicated several strengths but mostly weaknesses in L2 research. Looking ahead, however, the trends over time provide reason to be optimistic about the future of our field. I do not want to suggest that a methodological and statistical utopia for SLA is imminent or even inevitable; major strides are still needed on multiple fronts (see recommendations at the end of this chapter). These findings also illustrate changes in the types of questions we have asked and give a general indication of methodological and analytical approaches taken to address and answer those questions.

Designs and Analyses over Time

Regarding changes in major design types and settings over time, experimental studies make up an increasingly large portion of all studies, but observational studies are still in the majority. Although I only considered studies published 1990 and later, according to Henning (1986), this pattern was also taking place in the 1970s and early 1980s. The move towards experimental research may be evidence of a field-wide change in the type of relationships suggested in models of SLA. Alternatively, an increase in experimental studies may also be an indication of the maturity of our domain. Different substantive areas of SLA may have migrated from an early phase of research looking to establish a correlation to a more developed phase testing causation. Cooke and Payne (2002) point out that another force that may influence the types of designs are federal grant-funding agencies which have begun to require evidence of programmatic effectiveness to be based on randomized experimental designs. Pressure of this sort may also be slowing the subtle shift taking place from lab- to classroom-based research in

order to allow for random assignment. Future domain-specific syntheses are needed to determine which patterns are taking place and in which areas of SLA.

The trends for design features associated with experimental quality are mixed. Use of control or comparison groups and random assignment to experimental conditions both decreased, contrary to the patterns observed for both features in interactionist research (Plonsky & Gass, 2011). As I mentioned in the previous paragraph, we may see random assignment increasing in future research due to the requirements of federal grant funding agencies (Cooke & Payne, 2002). Increases, however, were found in both pretesting and delayed posttesting. The rise in pretesting is not entirely surprising because, despite the opposite trend observed by Plonsky and Gass (2011), random assignment decreased. In other words, there may be a kind of tradeoff in experimental research practice between random group assignment and pretesting because of the role both can play in ensuring pretreatment comparability of groups. The increase in delayed posttesting might reflect development occurring in substantive areas during the period studied. Whereas early studies in a particular line of experimental research may have only tested whether or not an immediate effect exists, later studies or studies in more mature areas may be more likely to be interested in testing the longevity of experimental interventions as well.

I also looked at changes in sample sizes and in the number of unique groups or samples in L2 research. Both increased over time (1990s: median $N = 56$, $k = 2.78$; 2000s: median $N = 62$, $k = 2.91$) resulting in a slight net increase in power, assuming equal or larger effects and an equal number of statistical tests across the two decades in question. Although effect sizes remained essentially unchanged (see below), the median number of statistical tests increased by approximately 50%, which diminished overall power. To summarize, several factors that lower power appear to be present (see list above) and increasing in L2 research.

Moving forward, the first step toward solving the power problem is simple: larger samples. Again, the paths taken by other fields can be illustrative. In their review of personnel research, Lent, Aurbach, and Levin (1971) found a median sample size of 68 and noted the debilitating effect that it had on power. But more recently, Salagado (1998) found the median sample in personnel research to have increased to 113. Based on the findings of this and other reviews of L2 research (e.g., Plonsky & Gass, 2011), researchers, reviewers, and editors would increase the accuracy (and therefore the efficiency) of our results by insisting on larger samples. Obtaining larger study-wise and group-wise samples will introduce various logistic and financial constraints not to mention limitations to hypothesis testing. For example, larger subsamples may translate to fewer between-group contrasts and statistical tests. But it is no doubt preferable to sacrifice quantity of analyses for enhanced precision of results. Of course the sample sizes typical of L2 research are often small because the number of available participants is small. Imagine, for instance, the challenge of recruiting participants in studies on the acquisition of less commonly taught languages. Moreover, in classroom-based research there is a kind of tension between obtaining large enough samples (and sufficient statistical power) on one hand and preserving ecological validity on the other. Considering these issues, it may not be fair to hold SLA to the same standard or expectation of large samples as one might in a field such as psychology where researchers often have access to undergraduate participant pools or otherwise larger populations. However, the perils of low power must still be recognized, regardless of the cause. For this reason it may be best for researchers working with necessarily small samples to limit their use of inferential statistics or to avoid them entirely.

There are two results I want to highlight among changes in statistical analyses over time. First, nearly all types of analyses increased from one decade to the next but especially tests

comparing means. An increasingly large portion of L2 research relies on *t* tests and ANOVAs to analyze quantitative data. This trend is worth noting because it is exactly opposite to what has occurred and is occurring in education and psychology. As Skidmore and Thompson (2010) and others have reported (e.g., Goodwin & Goodwin, 1985; Kieffer et al., 2001; Willson, 1980), use of *t* tests and ANOVAS in those fields has dropped dramatically in recent decades, a shift associated with a simultaneous increase in regression and attributed at least in part to Cohen (1968; see above). And second, although the diversity and sophistication of analyses across the field does not appear to be increasing, the diversity within individual studies does. The percentage of studies employing zero or one type of statistical analysis dropped from the 1990s to the 2000s while the percentage of studies using more than one increased substantially. A shift may be taking place in the number of different types of questions and research objectives posed by individual studies. This finding also replicates previous observations of an increase in the use inferential statistics as opposed to using only descriptive statistics (Gass, 2009; Henning; 1986).

Reporting Practices over Time

To begin this section, I repeat what I consider to be the largest, most salient, and most significant changes in reporting practices: increases in means, standard deviations, confidence intervals, effect sizes, exact *p* values, and confidence intervals; and decreases in means without standard deviations, ANOVAs / *t* tests without means, ANOVAS / *t* tests without standard deviations, ANOVAs / *t* tests without *f* or *t* values, and inconsistent reporting of *p* values. These changes indicate that data are overall more thoroughly reported now than before. They may also signify an awareness of the importance of thorough reporting and a move in the direction of synthetic-mindedness at the primary level (see Norris & Ortega, 2006). Whatever the cause,

effect sizes in the 2000s are somewhat regularly reported in and can be calculated from a larger portion of studies than was previously possible.

Before celebrating these small but significant successes, I want to draw attention to two potential sources of bias related to the increase over time in reporting practices, effect sizes, and meta-analyzability of L2 research. Although it is certainly preferable to be able to meta-analyze a larger portion of existing studies, the difference in meta-analyzability between earlier and more recent decades may produce a "top heavy" set of primary effects and therefore constitute a source of bias. As I discussed in Chapter 1, over time effect sizes in a particular research area may swell or shrink (or both, leading to greater overall variance) due to factors such as theoretical maturity, subtlety of analyses, improvements to research design and instrumentation, and so forth. In areas where an increase or decrease has occurred, greater meta-analyzability among more recent studies would bias the overall meta-analytic average upward or downward depending on the pattern of effects. Meta-analysts should explore individual study effects to determine whether any such patterns are present. Another potential source of bias in meta-analyses comes as the result of different reporting practices across journals. For example, studies published in *Language Learning* may be more likely to be included in meta-analytic reviews because of the journals' explicitly stated editorial policy requiring effect sizes to be reported (see Ellis, 2000). To provide a more concrete sense of how this might play out, consider the fact that 34% of the studies in *Language Learning* reported an effect size whereas only 16% did so in *Studies in Second Language Acquisition* which currently has no stated policy regarding the reporting of effect sizes. To be clear, greater meta-analyzability in one journal over another does not necessarily introduce a threat to the validity of meta-analyses unless there are also systematic differences in effect sizes across journals. In the case of the two journals included in this review,

the difference in average *d* values between them was minimal and nonstatistical: .86 (*SD* = .65) in *Language Learning* vs. .91 (*SD* = .82) in *Studies in Second Language Acquisition*. To prevent and make known of any bias of this nature, meta-analysts should examine and compare effects across journals/sources and aim for inclusivity when locating primary studies.

The third category of reporting practices consists of a mix of different types of data and study elements: research questions, visual displays of data, reliability estimates, a predetermined Type I error rate, checking of statistical assumptions, and power analysis. Improvements were found for all six, which demonstrates a larger move in L2 research toward adhering to generally accepted notions of best practice in social science research (see Klingner, Scanlon, & Pressley, 2005) and to specific APA recommendations found the latest/6th edition of the *Publication Manual*, in Wilkinson and the Task Force for Statistical Inference (1999), and in the APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). The results for one particular reporting practice, providing estimates of reliability, support previous anecdotal (e.g., Mackey & Gass, 2006) and empirical (Plonsky & Gass, 2011) reports that L2 researchers have developed in recent years a heightened concern for psychometric issues. The overall percentage of L2 studies reporting estimates of reliability is now roughly equal to the same in Kieffer et al.'s (2001) findings from two respected journals of education and psychology (*American Educational Research Journal* and *Journal of Counseling Psychology*).

The many improvements to L2 research discussed throughout this section raise the question of their source. Recognizing the impetus for change allows credit to be given where it is due and, more importantly, informs future attempts at reform on what has worked. In education and psychology, for example, significant methodological and analytical developments seen over

the last half century have been attributed to preeminent figures and seminal papers they authored such as Cohen (1962, 1968, on power and the GLM, respectively) and Cronbach (1957, on the combined potential of correlational and experimental research). In the case of SLA, one cannot be entirely certain to whom the systematic and field-wide changes observed here can be attributed, though certainly one would be tempted to nominate Norris and Ortega (2000) in light of the nature of changes seen over the last decade.

Finally, along with jubilation and optimism over the reforms in L2 methods and the individuals who championed them, a realistic look at the data show that the glass (and the bar graph in Figure 8) is still at least half empty. In the 2000s, estimates of reliability were reported in just over half of the sampled studies, a priori alphas in about a quarter, assumptions in less than 20%, and Crookes' (1991) claim that power analyses "are almost never used" (p. 762) still applies. Of course it is important to distinguish between practices such checking for statistical assumptions, reporting reliability, and power analyses, on one hand, and violations of statistical assumptions, low reliability, and low power on the other. But the lack of the former may be symptomatic of researchers' unfamiliarity with these issues and their implications in quantitative research. Consequently, improving SLA research practices may be best achieved by not only establishing field-wide and field-specific standards but by reforming the curricula and researcher training in graduate SLA programs (see specific suggestions related to both below).

Effect Sizes over Time

The last feature of L2 research examined over time was the magnitude of effect sizes. The results from this phase of the analyses found almost no change from the 1990s to the 2000s. That is, none of the scenarios in which effect sizes increase or decrease or both was observed. The lack of any measureable change is probably an artifact of the study's wide scope which

diminishes its sensitivity to movements taking place over time in different areas of SLA. Recall that Plonsky and Gass (2011) found clear evidence for a decrease in effect sizes over three decades of interactionist research. It is likely that effects in other areas of L2 research have also increased over time. The presence of domain-specific patterns of effects subsumed by the volume of this sample will have to be determined by future meta-analyses.

Despite any patterns that may be hidden within the average *d* values from the 1990s (.87, median = .72), 2000s (.89, median = .70), and overall (.88, median = .71), these values provide another valuable source of comparison for effect sizes in future primary studies and meta-analyses. To begin with, effects in L2 research are substantially larger on average than effects typically found in education and psychology (see meta-syntheses by Hattie, 1992; Lipsey & Wilson, 1993; Richard, Bond, & Stokes-Zoota, 2003). It is also clear that Cohen's (1988) benchmarks are not appropriate for interpreting effect sizes in SLA. No set of benchmarks is universally appropriate even within a single field (e.g., Thompson, 2007), but the median effects reported here do support Oswald and Plonsky's (2010) tentatively proposed scale for interpreting effect sizes in the absence of other domain-specific effects as a source of more precise, contextualized comparisons.

A closer analysis of the present data set as well as that used by Oswald and Plonsky should be used to produce a more refined scale for interpreting effect sizes in L2 research. Future research should also replicate these and other analyses carried out presently using other commonly reported effect sizes indices such as correlation coefficients and eta-squared.

To close this section on changes over time in SLA, I want to reiterate that the field of SLA has made important strides toward more rigorous research and reporting practices. There is evidence that the state of our science today is more precise and more efficient than it was two

decades ago. Consequently SLA is also more deserving of recognition and respect from our colleagues in other social sciences for its many contributions to understanding human cognition and behavior. Looking forward, the challenge ahead of us will be to continue on the path toward improving the means by which we construct knowledge on how languages are learned. Before outlining specific steps to that end, I will describe some of the limitations of this study and areas within the domain of study quality that I feel to be in need of future research.

Limitations and Areas for Future Research on Study Quality

Study quality as an area of empirical inquiry in SLA is in its infancy. It is exciting to contemplate the potential of this domain to contribute to the field and to the dynamic nature of its research and reporting practices. For the sake of transparency, and in the hopes of encouraging and directing further research in this area, the following points acknowledge limitations of this study and provide corresponding suggestions for future studies.

- Limitation 1: The eligibility criteria of this study were inclusive in one dimension (substantive domain) but more restricted in the other two dimensions (source-journals and time).
    - Suggestion 1: Get a more complete picture of L2 research by replicating and/or expanding on this study using additional journals (e.g., *Applied Linguistics*, *The Modern Language Journal*, *Second Language Research*) and/or earlier and forthcoming studies (back to 1980 or 1970 and beyond 2010).
- Limitation 2: This study did not explicitly assess or compare the quality of research in either journal.
    - Suggestion 2: Compare methodological quality across *Language Learning* and *Studies in Second Language Acquisition* as well as other journals (see Suggestion

1) to provide an empirically-based indication of journal quality that might (a) serve as an alternative to traditional perceptions of journal quality (which probably favor older or more visible journals) and controversial measures such as impact factors and (b) encourage editors to enact stricter publication guidelines in their journals.

- Limitation 3: This study gives us a macro-level view of how L2 research is carried out and reported on but tells us nothing about the process and motivation behind the decisions made by individual researchers.
    - Suggestion 3a: Conduct case studies of individual researchers as they design, carry out, analyze, and report on studies.
    - Suggestion 3b: Replicate Lazaraton et al.'s (1987) survey of applied linguists' familiarity with different study features and statistical concepts.
- Limitation 4: As I mentioned throughout the Discussion, the broad net used for this study may have blurred overall and chronological trends taking place across smaller subdomains of L2 research.
    - Suggestion 4: Carry out additional methodological reviews in different areas of L2 research. Such analyses are probably best conducted along with meta-analyses of substantive findings because of their mutual dependency.
- Limitation 5: Several different effect size indices were coded from primary reports, but only *d* values based on between-groups contrasts were analyzed.
    - Suggestion 5a: Replicate the analyses conducted here using other commonly used effect sizes in L2 research such as correlation coefficients, *d* for pre-post contrasts, and eta-squared.

- Suggestion 5b: Use unanalyzed effect size data to develop field-specific benchmarks for correlation coefficients, *d* for pre-post contrasts, and eta-squared.

Suggestions for the Field of SLA

Beyond providing direction for future studies related to methodological quality, the findings of this study have implications for the field of applied linguistics more generally. The purpose of this study was not only to look back but to look ahead, and the data reported here make a compelling case for reform. With an eye to progress and the future, this section outlines suggestions for reforming L2 research. I direct my comments to six different but non-exclusive groups of stakeholders in the field: individual researchers, journal editors, meta-researchers, graduate curriculum committees and researcher trainers in SLA, grant-funding agencies and their reviewers, and The American Association for Applied Linguistics.

To Individual Researchers

- When planning a study, consider power. More specifically, use an estimate of the anticipated effect size to help determine an appropriate sample size. Also consider sample size and its inverse relationship with sampling error when interpreting results.
- Be skeptical of *p* values. Specifically, remember that (a) a *p* value at less than .05 with a very large sample is meaningless because any size difference between groups (or correlation) will reach statistical significance given a large enough sample, (b) small samples with statistical findings may not be reliable either because they are likely infected with high sampling error, and (c) when small samples are used to study small effects, a finding of statistical significance is probably if not necessarily an overestimate (see Gelman & Weakliem, 2009).

- As an alternative or in addition to *p* values, calculate and report effect sizes. But don't forget to explain what they mean. To do so and do so well, you have to accept that your findings will probably be best understood as a matter degree, not a dichotomous yes or no (or even a trichotomous small, medium, or large). See Oswald and Plonsky (2010) and Plonsky and Oswald (in press) for some ideas on how to interpret effect sizes such as in simple standard deviation units, in relation to previous studies and meta-analyses, and/or relative to the L2-specific scale we proposed.
- Calculate and report an estimate of reliability for your instruments. When relevant, also explain attenuation to effects that may have occurred due to instrument (un)reliability or other psychometric artifacts such as range restriction.
- Be sure to report your data thoroughly. Specifically, report a standard deviation with all means, and report both an exact *t* or *f* value and exact *p* values with all statistical tests used to compare means.
- Consider whether correlational or regression analyses might be an appropriate approach to your data rather than comparing group means.
- Consider whether multivariate analyses might be an appropriate approach to your data rather than multiple univariate tests.
- If you typically conduct observational/correlational studies, try teaming up with an experimental researcher, and vice versa. There is great potential in aptitude-treatment interventions.
- Work towards an in-depth understanding of one or more specialized research techniques or statistical analyses. The range of research practices and especially statistics in SLA is somewhat narrow, and we might benefit from the introduction of new procedures

developed and used in other fields. In order to do this, you might need to take a class outside of your department. But getting familiar with other disciplines and the ways they do research might give you new perspectives on our own field and lead to some exhilarating cross-disciplinary collaborations.

To Journal Editors

- Remember that you have the power to influence and improve research practices. Use it. If your journal has guidelines, uphold them and remind reviewers do the same. If it does not, consult with trusted individuals and sources to compose some. Better yet, work with other editors and societal/organizational leadership to establish a common set of requirements and guidelines for publication across multiple journals (see AAAL below).
- Related to the previous point, it's not enough to simply require that authors report effect sizes and confidence intervals. Effect sizes are only more informative than *p* values if they are interpreted and contextualized; labeling them as small, medium or large and citing Cohen (1988) is inadequate.
- Demand consistency (e.g., in *p* values) across and within papers.
- Consider including a methodological review as part of the review process similar to Magnan (1994). Doing so will improve the quality of studies in your journal and both authors and reviewers will likely develop a more defensible understanding of research designs and statistics.

To Meta-researchers

- As an area expert and accumulator of massive amounts of data, you have a unique perspective that people will listen to. Use this voice, this megaphone (your meta-phone, so to speak) to make known the strengths and weaknesses of your subdomain. Generate

awareness and encourage continued used of effective research practices and expose weak or absent ones.

- Be sure to do more than summarize. Examine relationships not addressed or not addressed sufficiently in the primary literature, and pinpoint deficiencies in the available data to propose specific studies that can later be included in a meta-analytic replication (see Plonsky, in press-b).
- Examine the methods in primary studies not only to review them and assess their adequacy in addressing the research questions but to explain variance in effects as well. Then, use your methodological review to answer and interpret questions related to the sample of effect sizes such as (a) Were most of the samples very small? (If so, this might explain a relatively high amount of variance across studies.); (b) Were studies carried out in labs or classrooms or both?; (c) How were treatment group effects generally measured? As pre-post contrasts? In relation to comparison or true control groups?; (d) Do effects vary according to these and other research and reporting practices?
- Examine changes in effects over time because of the three scenarios discussed earlier but also because of the potential for bias due to a greater portion of meta-analyzable studies in recent years.
- Use your findings to guide future studies in interpreting their effects.
- Cast the net wide when searching for primary studies. They may vary in quality, but whether they do or not is an empirical problem that you can help solve.

To Graduate Curriculum Committees and Researcher Trainers in SLA

- There is no doubt that statistical know-how among L2 researchers has improved immensely since Meara (1995) wrote "[When I was in graduate school], anyone who

could explain the difference between a one-tailed and two-tailed test of significance was regarded as a dangerous intellectual; admitting to a knowledge of one-way analyses of variance was practically the same as admitting to witchcraft in 18th century Massachusetts." (p. 341; see also Lazaraton, 1987). Nevertheless, there is still plenty of room for growth in this area.

- I argued in this paper that regression might be more appropriate than ANOVA in some cases. However, ANOVA's status as the statistical test of choice in L2 research is not likely to change soon and ANOVAs should and will continue to be used, so graduate students need to know how to test the assumptions of, use, report on, and interpret the results and effect sizes of ANOVA perhaps more than any other statistic.
- Emphasize the importance of understanding, interpreting, and reporting descriptive statistics.
- Emphasize the importance of and relationship between power, sampling error, effect sizes, and statistical significance.
- Emphasize that we should not expect the findings of a single study to provide a definite or conclusive answer to any question worth asking. In other words, encourage students to take a synthetic approach to their consumption and production of primary research.
- Graduate curriculum committees should consider advising their students to take more specialized courses in research methods and statistics. Encouraging graduate students to take classes outside their home department will carry benefits field-wide (e.g., by expanding our collective methodological and analytical horizons) and locally (e.g., by students sharing what they learn with classmates and faculty both in class and in other settings such as workshops and brown bags).

To Grant-Funding Agencies and their Reviewers

- Like journal editors, grant-funding agencies and the standards they require can have a major influence on funded activities and research practices (*à la* the carrot or the stick). In addition to other qualities used to determine which proposals will be funded (e.g., theoretical or practical relevance, feasibility), I recommend that grant-funding organizations in applied linguistics such as the TESOL International Research Foundation and The *Language Learning* Grants Program determine and state a clear set of methodological standards for grant proposals. Because grants are generally written before the research has begun, review boards cannot assess all the aspects study quality investigated here (e.g., whether data are reported thoroughly). They can, however, insist that researchers make appropriate and a priori decisions in matters such as study design and statistical power (see Altman, 2004).

To the American Association for Applied Linguistics (AAAL), its Leadership, and All its Constituents with an Interest in L2 Research

- In the past, AAAL has advocated for research and policy relating to substantive matters, but to my knowledge it has been silent with respect to how applied linguistics research is conducted. Based on the findings of this study, the inconsistencies observed within and across studies and journals, and the debilitating effects both of these have on progress in L2 research, it is time for the leadership of AAAL to designate a task force to construct methodological standards for L2 research, which makes up a large if not a majority portion of the research carried out by its members. (Alternately, AAAL may want to put a more permanent committee in place that can regularly discuss and respond to developments in quantitative methodology as they relate to L2 research.) I do not propose

that we reinvent the wheel, but rather that we draw on the experience, expertise, and standards of related disciplines in addition to our collective understanding to establish field-specific norms for conducting and reporting on research. The task force or committee I envision would be comprised of the following: at least one member of the executive committee of AAAL, members from the editorial boards of L2 journals, a small number of both quantitatively- and qualitatively minded researchers, and perhaps (for an outside perspective) one or more methodologists who work in other disciplines but that are at least somewhat familiar with L2 research.

Conclusion

This study set out to accomplish two primary goals. The first was to gain a better understanding of research and reporting practices in SLA. The data collected for this study provide us with a quantitative indication of many aspects of L2 research previously unknown such as the extent to which L2 research is lab- versus classroom-based, the inclusion of delayed posttests in experimental research, and the frequency of NHST. A considerable number of weaknesses were also observed in the sample of primary studies, and the concerns I raise regarding these weaknesses merit serious attention. However, I prefer to look forward, which is why the second major goal of the study was less retrospective and more prospective. The findings of this study should prompt us, at the very least, to reflect on and investigate further the means by which L2 research is carried out and reported. Ideally, though, more concrete action will be taken by the field of SLA at both institutional and individual levels to enact reform. I am hopeful and optimistic that we will do what is needed to improve our field and the means by which we move it forward.

Notes

[1] Parts of this section were borrowed, with the permission of S. Gass, from Plonsky and Gass (2011).

[2] Parts of this section were borrowed, with the permission of S. Gass, from Plonsky and Gass (2011).

APPENDIX

APPENDIX: REPORTS INCLUDED IN THE PRESENT STUDY

Table 18
Reports Included in the Present Study

Author(s)	Article Title
Studies in Second Language Acquisition, 2010	
Conroy, Cupples	We could have loved and lost, or we never could have love at all: Syntactic misanalysis in L2 sentence processing
Ellis, Sagarra	The bounds of adult language acquisition: Blocking and learned attention
Shea, Curtin	Discovering the relationship between context and allophones in a second language: Evidence for distribution-based learning
Shintani, Ellis	The incidental acquisition of English plural – s by Japanese children in comprehension-based and production-based lessons: A process-product Study
Van der Slik	Acquisition of Dutch as a second language
GeesLin, Gudmestad	An exploration of the range and frequency of occurrence of forms in potentially variable structures in second-language Spanish
Hama, Leow	Learning without awareness revisited
Sheen	Differential effects of oral and written corrective feedback in the ESL classroom
Yang, Lyster	Effects of form-focused practice and feedback on Chinese EFL learners' acquisition of regular and irregular past tense forms
Storch, Wigglesworth	Learners' processing, uptake, and retention of corrective feedback on writing
Rothman, Judy, Guijarro-Fuentes, Pires	On the (un)-ambiguity of adjectival modification in Spanish determiner phrases
Rah, Adone	Processing of the reduced relative clause versus main verb ambiguity in L2 learners at different proficiency levels
Qasem, Foote	Crosslanguage lexical activation: A test of the revised hierarchical and morphological decomposition models in Arabic-English bilinguals
Studies in Second Language Acquisition, 2009	
Derwing, Munro, Thomson, Rossiter	The relationship between L1 fluency and L2 fluency development
Henry, Culman, VanPatten	More on the effects of explicit information in instructed SLA: A partial replication and a response to Fernández (2008)
Stæhr	Vocabulary knowledge and advanced listening comprehension in English as a foreign language
Trofimovich, Lightbown, Halter, Song	Comprehension-based practice: The development of L2 pronunciation in a listening and reading program
Gabriele	Transfer and transition in the SLA of aspect: A bidirectional study of learners of English and Japanese

Table 18 (cont'd)

Neubauer, Clahsen	Decomposition of inflected words in a second language: An experimental study of German participles
Révész	Task complexity, focus on form, and second language development
Camponelle, Williams	Learner versus nonlearner patterns of stylistic variation in synchronous computer-mediated French: yes/no questions and nous versus on
Song, Schwartz	Testing the fundamental difference hypothesis: L2 adult, L2 child, and L1 child comparisons in the acquisition of Korean wh-constructions with negative polarity items
Zyzik, Azevedo	Word class distinctions in second language acquisition: An experimental study of L2 Spanish
Pulido	How involved are American L2 learners of Spanish in lexical input processing tasks during reading?
van de Craats	The role of is in the acquisition of finiteness by adult Turkish learners of Dutch
Duffield, Matsuo	Native speakers' versus L2 learners' sensitivity to parallelism in VP-ellipsis
	Studies in Second Language Acquisition, 2008
Taguchi	The role of learning environment in the development of pragmatic comprehension: A comparison of gains between EFL and ESL learners
Dekydtspotter, Donaldson, Edmonds, Liljestrand Fultz, Petrush	Syntactic and prosodic computations in the resolution of relative clause attachment ambiguity by English-French learners
Abrahamsson, Hyltenstam	The robustness of aptitude effects in near-native second language acquisition
Bohnacker, Rosén	The clause-initial position in L2 German declaratives: Transfer of information structure
Fernández	Reexamining the role of explicit information in processing instruction
Roberts, Gullberg, Indefrey	Online pronoun resolution in L2 discourse: L1 influence and general learner effects
Bowles	Task type and reactivity of verbal reports in SLA: A first look at a L2 task other than reading
Brown, Gullberg	Bidirectional crosslinguistic influence in L1-L2 encoding of manner in speech and gesture: A study of Japanese speakers of English
Anderson	Forms of evidence and grammatical development in the acquisition of adjective position in L2 French
McDonough, Mackey	Syntactic priming and ESL question development
Nguyen, Macken	Factors affecting the production of Vietnamese tones: A study of American learners
Webb	Receptive and productive vocabulary sizes of L2 learners

Table 18 (cont'd)

Studies in Second Language Acquisition, 2007	
Egi	Interpreting recasts as linguistic evidence: The roles of linguistic target, length, and degree of change
Major	Identifying a foreign accent in an unfamiliar language
O'Brien, Segalowitz, Freed	Phonological memory predicts second language oral fluency gains in adults
Cuervo	Double objects in Spanish as a second language: Acquisition of morphosyntax and semantics and semantics as a second language: acquisition of morphosyntax and semantics
Colantoni, Steele	Acquiring /alveolar approximant/ in context
Trofimovich, Gatbonton, Segalowitz	A dynamic look at L2 phonological learning: seeking processing explanations for implicational phenomena
Steinel, Hulstijn, Steinel	Second language idiom learning in a paired-associate paradigm: effects of direction of learning, direction of testing, idiom imageability, and idiom transparency
Ozeki, Shirai	Does the noun phrase accessibility hierarchy predict the difficulty order in the acquisition of Japanese relative clauses?
Kanno	Factors affecting the processing of Japanese relative clauses by L2 learners
Yabuki-Soh	Teaching relative clauses in Japanese: Exploring alternative types of instruction and the projection effect
Jeon, Kim	Development of relativization in Korean as a foreign language: The noun phrase accessibility hierarchy in head-internal and head-external relative clauses
Yip, Matthews	Relative clauses in Cantonese-English bilingual children: Typological challenges and processing motivations
Sugaya, Shirai	The acquisition of progressive and resultative meanings of the imperfective aspect marker by L2 learners of Japanese: Transfer, universals, or multiple factors?
Rossomondo	The role of lexical temporal indicators and text interaction format in the incidental acquisition of the Spanish future tense
Sachs, Polio	Learners' uses of two types of written feedback on a L2 writing revision task
Studies in Second Language Acquisition, 2006	
Ammar, Spada	One size fits all?: Recasts, prompts, and L2 learning
Harada	The acquisition of single and geminate stops by English speaking children in a Japanese immersion program
Sunderman, Kroll	First language activation during second language lexical processing: an investigation of lexical form, meaning, and grammatical class
Lieberman, Aoshima, Phillips	Nativelike biases in generation of wh-questions by nonnative speakers of Japanese
Zyzik	Transitivity alternations and sequence learning: Insights from L2 Spanish production data

Table 18 (cont'd)

Lee, Guion, Harada	Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals
McDonough	Interaction and syntactic priming: English L2 speakers' production of dative constructions
Carpenter, Jeon, MacGregor	Learners' interpretations of recasts
Polio, Gass, Chapin	Using stimulated recall to investigate native speaker perceptions in native-nonnative speaker interaction
Lyster, Mori	Interactional feedback and instructional counterbalance
Pica, Kang, Sauro	Information gap tasks: Their multiple roles and contributions to interaction research methodology
Ellis, Loewen, Erlam	Implicit and explicit corrective feedback and the acquisition of L2 grammar
Trofimovich, Baker	Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech
Morgan-Short, Bowden	Processing instruction and meaningful output-based instruction: Effects on second language development
Munro, Derwing, Morton	The mutual intelligibility of L2 speech
	Studies in Second Language Acquisition, 2005
Guion	Knowledge of English word stress patterns in early and late Korean-English bilinguals
Sharma	Language transfer and discourse universals in Indian English article use
Zareva, Schwanenflugel, Nikolova	Relationship between lexical competence and language proficiency: Variable sensitivity
Loewen	Incidental focus on form and second language learning
Barcroft, Sommers	Effects of acoustic variability on second language vocabulary learning
Bowles, Leow	Reactivity and type of verbal report in SLA research methodology: Expanding the scope of investigation
Ellis	Measuring implicit and explicit knowledge of a second language: A psychometric study
Tokowicz, MacWhinney	Implicit and explicit measures of sensitivity to violations in second language grammar: Implicit and explicit measures of sensitivity to violations in second language grammar
De Jong	Can second language grammar be learned through listening? An experimental study
Robinson	Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: replications of Reber, Walkenfeld, and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance for SLA
Willams	Learning without awareness
Gass, Alvarez Torres	Attention when? An investigation of the ordering effect of input

Table 18 (cont'd)

	and interaction
Webb	Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge
Marinis, Roberts, Felser, Clahsen	Gaps in second language sentence processing
McDonough	Identifying the impact of negative feedback and learners' responses on ESL question development
	Studies in Second Language Acquisition, 2004
Escudero, Boersma	Bridging the gap between L2 speech perception research and phonological theory
Leeser	The effects of topic familiarity, mode, and pausing on second language learners' comprehension and focus on form
Smith	Computer-mediated negotiated interaction and lexical acquisition
Lyster	Differential effects of prompts and recasts in form-focused instruction
Murphy	Dissociable systems in second language inflectional morphology
Segalowitz, Freed	Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts
Lafford	The effect of the context of learning on the use of communication strategies by learners of Spanish as a second language
Collentine	The effects of learning contexts on morphosyntactic and lexical development
Diaz-Campos	Context of learning in the acquisition of Spanish second language phonology
Freed, Segalowitz, Dewey	Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs
Dewey	A comparison of reading development by learners of Japanese in intensive domestic immersion and study abroad contexts
Flege, MacKay	Perceiving vowels in a second language
Leow, Morgan-Short	To think aloud or not to think aloud: The issue of reactivity in SLA research methodology
Ellis, Yuan	The effects of planning on fluency, complexity, and accuracy in second language narrative writing
Hansen	Developmental sequences in the acquisition of English L2 syllable codas: A preliminary study
	Studies in Second Language Acquisition, 2003
Mondria	The effects of inferring, verifying, and memorizing on the retention of L2 word meanings: the effects of inferring, verifying, and memorizing on the retention of L2 word meanings
Papadopoulou, Clahsen	Parsing strategies in L1 and L2 sentence processing: A study of relative clause attachment in Greek
Dussias	Syntactic ambiguity resolution in L2 learners: Some effects of

Table 18 (cont'd)

	bilinguality on L1 and L2 processing strategies
Erlam	Evaluating the relative effectiveness of structured-input and output-based instruction in foreign language learning: Results from an experimental study
Abrahamsson	Development and recoverability of L2 codas: A longitudinal study of Chinese-Swedish interphonology
Montrul, Slabakova	Competence similarities between native and near-native speakers: An investigation of the preterite-imperfect contrast in Spanish
Zsiga	Articulatory timing in a second language: evidence from Russian and English
O'Grady, Lee, Choo	A subject-object asymmetry in the acquisition of relative clauses in Korean as a second language
Helms-Park	Transfer in SLA and creoles: The implications of causative serial verbs in the interlanguage of Vietnamese ESL learners
Clements	The tense-aspect system in pidgins and naturalistically learned L2
Iwashita	Negative feedback and positive evidence in task-based interaction: Differential effects on L2 development
Leeman	Recasts and second language development: Beyond negative evidence
Prévost	Truncation and missing inflection in initial child L2 German
Philp	Constraints on "noticing the gap": Nonnative speakers' noticing of recasts in NS-NNS interaction
Rehner, Mougeon, Nadasdi	The learning of sociolinguistic variation by advanced ESL learners: The case of nous versus on in immersion French
	Studies in Second Language Acquisition, 2002
Slabakova	The compounding parameter in second language acquisition
Izumi	output, input enhancement, and the noticing hypothesis: an experimental study on ESL relativization
Whong-Barr, Schwartz	Morphological and syntactic transfer in child L2 acquisition of the English dative alternation
Jiang	Form–meaning mapping in vocabulary acquisition in a second language
Hu	Psychological constraints on the utility of metalinguistic knowledge in second language production
Jarvis	Topic continuity in L2 English article use
GeESLin	The acquisition of Spanish copula choice and its relationship to language change
Butler	Second language learners' theories on the use of English articles: An analysis of the metalinguistic knowledge used by Japanese students in acquiring the English article system
Bardovi-Harlig	A new starting point? Investigating formulaic use and input in future expression
Biber, Reppen	What does frequency have to do with grammar teaching?

Table 18 (cont'd)

Liu, Gleason	Acquisition of the article the by nonnative speakers of English: An analysis of four nongeneric uses
Akiyama	Japanese adult learners' development of the locality condition on English reflexives
	Studies in Second Language Acquisition, 2001
Lee	The incidental acquisition of Spanish: Future tense morphology through reading in a second language
de la Fuente	Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in the receptive and productive acquisition of words
Munro, Derwing	Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate
Upton, Lee-Thompson	The role of the first language in second language reading
Riazantseva	Second language proficiency and pausing: A study of Russian speakers of English
Flege, Liu	The effect of experience on adults' acquisition of a second language
Bogaards	Lexical units and the learning of foreign language vocabulary
Wong	Modality and attention to meaning and form in the input
MacIntyre, Baker, Clément, Conrod	Willingness to communicate, social support, and language-learning orientations of immersion students
Glahn, Hakansson, Hammarberg, Holmen, Hvenekilde, Luund	Processability in Scandinavian second language acquisition
Inagaki	Motion verbs with goal PPS in the L2 acquisition of English and Japanese
Montrul	Agentive verbs of manner of motion in Spanish and English as second languages
Bley-Vroman, Joo	The acquisition and interpretation of English locative constructions by native speakers of Korean
Hirakawa	L2 acquisition of Japanese unaccusative verbs
Sorace, Shomura	Lexical constraints on the acquisition of split intransitivity: Evidence from L2 Japanese
Haznedar	The acquisition of the IP system in child L2 English
Wolter	Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model
Helms-Park	Evidence of lexical transfer in learner syntax: The acquisition of English causatives by speakers of Hindi-Urdu and Vietnamese
	Studies in Second Language Acquisition, 2000
Mackey, Gass, McDonough	How do learners perceive interactional feedback?
DeKeyser	The robustness of critical period effects in second language acquisition
Jarvis, Odlin	Morphological type, spatial reference, and language transfer
Leow	A study of the role of awareness in foreign language behavior:

Table 18 (cont'd)

	Aware versus unaware learners
Dimroth, Watorek	The scope of additive particles in basic learner languages
Ahrenholz	Modality and referential movement in instructional discourse: Comparing the production of Italian learners of German with native German and native Italian production
Hendriks	The acquisition of topic marking in L1 Chinese and L1 and L2 French
Bernini	Negative items and negation strategies in nonnative Italian
Carroll, Murcia-Serra, Watorek, Bendiscioli	The relevance of information organization to second language acquisition studies: The descriptive discourse of advanced adult learners of German
Kormos	The timing of self-repairs in second language speech production
Toth	The interaction of instruction and learner-internal factors in the acquisition of L2 morphosyntax
Montrul	Transitivity alternations in l2 acquisition: Toward a modular view of transfer
Cebrian	Transferability and productivity of L1 rules in Catalan-English interlanguage
Rose	An exploratory cross-sectional study of interlanguage pragmatic development
Allen	Form-meaning connections and the French causative: An experiment in processing instruction
Ju	Overpassivization errors by second language learners: The effect of conceptualizable agents in discourse
	Studies in Second Language Acquisition, 1999
Rosa, O'Neil	Explicitness, intake, and the issue of awareness: another piece to the puzzle
Mackey	Input, interaction, and second language development: An empirical study of question formation in ESL
Rott	The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading
Izumi, Bigelow, Fujiwara, Fearnow	Testing the output hypothesis: Effects of output on noticing and second language acquisition
Paribakht, Wesche	Reading and "incidental" L2 vocabulary acquisition: an introspective study of lexical inferencing
Fraser	Lexical processing strategy use and vocabulary learning through reading
Wode	Incidental vocabulary acquisition in the foreign language classroom
Brown, Sagers, LaPorte	Incidental vocabulary acquisition from oral and written dialogue journals
Ellis, He	The roles of modified input and output in the incidental acquisition of word meanings
Williams	Memory, attention, and inductive learning

Table 18 (cont'd)

Myles, Mitchell, Hooper	Interrogative chunks in French L2: A basis for creative construction?
Moyer	Ultimate attainment in L2 phonology: The critical factors of age, motivation, and instruction
Ortega	Planning and focus on form in L2 oral performance
Studies in Second Language Acquisition, 1998	
Bardovi-Harlig	Narrative structure and lexical aspect: Conspiring factors in second language acquisition of tense-aspect morphology
Rounds, Kanagy	Acquiring linguistic cues to identify agent: Evidence from children learning Japanese as a second language
Kempe, MacWhinney	The acquisition of case marking by adult learners of Russian and German
Carrell, Wise	The relationship between prior knowledge and topic interest in second language reading
Beck	English-speaking learners of German and the local impairment hypothesis
Munro	The effects of noise on the intelligibility of foreign-accented speech
Flege, Frieda, Walley, Randazza	Lexical factors and segmental accuracy in second language speech production
Riney, Flege	Changes over time in global foreign accent and liquid identifiability and accuracy
Carlisle	The acquisition of onsets in a markedness relationship: A longitudinal study
Wennerstrom	Intonation as cohesion in academic discourse: A study of Chinese speakers of English
Scarcella, Zimmerman	academic words and gender: ESL student performance on a test of academic lexicon
Lyster	Recasts, repetition, and ambiguity in L2 classroom discourse
Mehnert	The effects of different lengths of time for planning on second language performance
Studies in Second Language Acquisition, 1997	
Bongaerts, van Summeren, Planken, Schils	Gender as social practice
Yuan	Asymmetry of null subjects and null objects in Chinese speakers' L2 English
Watanabe	Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary
de Bot, Paribakht, Wesche	Toward a lexical processing model for the study of second language vocabulary acquisition: Evidence from ESL reading
Hancin-Bhatt, Bhatt	Optimal L2 syllables: Interactions of transfer and developmental effects
Harley, Hart	Language aptitude and second language proficiency in classroom learners of different starting ages

Table 18 (cont'd)

Ellis, Schmidt	Morphology and longer distance dependencies: Laboratory research illuminating the a in SLA
Yang, Givón	Benefits and drawbacks of controlled laboratory studies of second language acquisition: The keck second language learning project
DeKeyser	Beyond explicit rule learning: Automatizing second language morphosyntax
Robinson	Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions
de Graaff	The Experanto experiment: Effects of explicit instruction on second language acquisition
Derwing, Munro	Accent, intelligibility, and comprehensibility: Evidence from four L1s
Schmitt, Meara	Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes
Lyster, Ranta	Corrective feedback and learner uptake: Negotiation of form in communicative classrooms
	Studies in Second Language Acquisition, 1996
Chikamatsu	The effects of L1 orthography on L2 word recognition: A study of American and Chinese learners of Japanese
Horiba	Comprehension processes in L2 reading: Language competence, textual coherence, and inferences
Davies	Morphological uniformity and the null subject parameter in adult SLA
VanPatten, Oikkenon	Explanation versus structured input in processing instruction
Derwing	Elaborative detail: Help or hindrance to the NNS listener
Foster, Skehan	The influence of planning and task type on second language performance
Takahashi	Pragmatic transferability
House	Developing pragmatic fluency in English as a foreign language
Paradis, Genesee	Syntactic acquisition in bilingual children: Autonomous or interdependent?
Robinson	Learning simple and complex second language rules under implicit, incidental, rule-search, and instructed conditions
Major, Faudree	Markedness universals and the acquisition of voicing contrasts by Korean speakers of English
	Studies in Second Language Acquisition, 1995
Flowerdew, Tauroza	The effect of discourse markers on second language lecture comprehension
Oliver	Negative feedback in child NS-NNS conversation
Juffs, Harrington	Parsing effects in second language sentence processing
Munoz	Markedness and the acquisition of referential forms: The case of zero anaphora
Flanigan	Anaphora and relativization in child second language acquisition
Polio	Acquiring nothing? The use of zero pronouns by nonnative

Table 18 (cont'd)

	speakers of Chinese and the implications for the acquisition of nominal reference
DeKeyser	Learning second language grammar rules: An experiment with a miniature linguistic system
Whyte	Specialist knowledge and interlanguage development
Reynolds	Repetition in nonnative speaker writing: More than quantity
Bouton	A cross-cultural analysis of the structure and content of letters of reference
Hartford	Zero anaphora in nonnative texts: Null-object anaphora in Nepali English
Bardovi-Harlig	A narrative perspective on the development of the tense/aspect system in second language acquisition
Slavoff, Johnson	The effects of age on the rate of learning a second language
Munro	Nonsegmental factors in foreign accent: Rantings of filtered speech
	Studies in Second Language Acquisition, 1994
Flege, Munro	The word unit in second language speech production and perceptions
Meisel	Code-switching in young bilingual children: The acquisition of grammatical constraints
Kern	The role of mental translation in second language reading
Scott	Auditory memory and perception in younger and older adult second language learners
Gass, Varonis	Input, interaction, and second language production
Loschky	Comprehensible input and second language acquisition: What is the relationship?
MacIntyre, Gardner	The effects of induced anxiety on three stages of cognitive processing in computerized vocabulary learning
Matsumura	Japanese learners' acquisition of the locality requirement of English reflexives
Sasaki	Paths of processing strategy transfers in learning Japanese and English as foreign languages
Ioup	Reexamining the critical period hypothesis: A case study of
Boustagui, Tigi, Moselle	successful adult SLA in a naturalistic environment
	Studies in Second Language Acquisition, 1993
Robinson, Ha	Instance theory and second language rule learning under explicit conditions
Towell, Hawkins, Bazergui	Systematic and nonsystematic variability in advanced language learning
Tyler, Bro	Discourse processing effort and perceptions of comprehensibility in nonnative discourse
Harlig, Hartford	Learning the rules of academic talk: A longitudinal study of pragmatic change
Damhuis	Immigrant children in infant-class interactions: Opportunities for

Table 18 (cont'd)

	second language acquisition of young multilingual children in Dutch infant classes
Leow	To simplify or not to simplify: A look at intake
Carroll, Swain	Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations
Trahey, White	Positive evidence and preemption in the second language classroom
Spada, Lightbown	Instruction and the development of question in L2 classrooms
VanPatten, Cadierno	Explicit instruction and input processing
Duff	Syntax, semantics, and SLA: The convergence of possessive and existential constructions
Laufer, Eliasson	What causes avoidance in L2 learning: L1-L2 difference, L1-L2 similarity, or L2 complexity?
Horiba	The role of causal reasoning and language competence in narrative comprehension
Kraemer	Social psychological factors related to the study of Arabic among Israeli high school students: A test of Gardner's socioeducational model
	Studies in Second Language Acquisition, 1992
Young, Milanovic	Discourse variation in oral proficiency interviews
Pfaff	The issue of grammaticalization in early German second language
Skiba, Dittmar	Pragmatic, semantic, and syntactic constraints and grammaticalization
Bohn, Flege	The production of new and similar vowels by adult German learners of English
Ross, Berwick	The discourse of accommodation in oral proficiency interviews
Tang	The effect of graphic representation of knowledge structures on ESL reading comprehension
Gardner, Day, MacIntyre	Integrative motivation, induced anxiety, and language learning in a controlled environment
Ellis	Learning to communicate in the classroom: A study of two language learners' requests
Harrington, Sawyer	L2 working memory capacity and L2 reading skill
Wolfe Quintero	Learnability and the acquisition of extraction in relative clauses and wh-questions
Tyler, Bro	Discourse structure in nonnative English discourse
	Studies in Second Language Acquisition, 1991
Doughty	Second language instruction does make a difference: Evidence from an empirical study of SL relativization
Bond, Fokes	Perception of English voicing by native and nonnative adults
Pica, Holliday, Lewis, Berducci, Newman	Language learning through interaction: What role does gender play?
Edge	The production of word-final voiced obstruents in English by L1 speakers of Japanese and Chinese

Table 18 (cont'd)

Cohen	Feedback on writing: The use of verbal report
Ellis	Grammaticality judgments and second language acquisition
Adamson, Regan	The acquisition of community speech norms by Asian immigrants learning English as a second language
Eckman	The structural conformity hypothesis and the acquisition of consonant clusters in the interlanguage of ESL learners
Huffines	Acquisition strategies in language death
Gardner, Macintyre	An instrumental motivation in language study: Who says it isn't effective
	Studies in Second Language Acquisition, 1990
Schachter, Yip	Grammaticality judgments: Why does anyone object to subject extraction?
Koda	The use of L1 reading strategies in L2 reading: Effects of L1 orthographic structures on L2 phonological recoding strategies
Schneider, Connor	Analyzing topical structure in ESL essays: Not all topics are equal
Lightbown, Spada	Focus-on-form and corrective feedback in communicative language teaching: Effects on second language learning
VanPatten	Attending to form and content in the input: An experiment in consciousness
Derwing	Speech rate is no simple matter: Rate adjustment and NS-NNS communicative success
Robison	The primacy of aspect: Aspectual marking in English interlanguage
Lee, Riley	The effect of prereading, rhetorically-oriented frameworks on the recall of two structural different expository texts
Chaudron, Parker	Discourse markedness and structural markedness: The acquisition of English noun phrases
Reseigh Long	What you don't know can't help you: An exploratory study of background knowledge and second language listening comprehension
	Language Learning, 2010
Goo	Working memory and reactivity
Schmid, Fägersten	Disfluency Markers in L1 Attrition
Tight	Perceptual Learning Style Matching and L2 Vocabulary Acquisition
Peng, Woodrow	Willingness to Communicate in English: A Model in the Chinese EFL Classroom Context
Ionin, Montrul	The Role of L1 Transfer in the Interpretation of Articles with Definite Plurals in L2 English
Rast	The Role of Linguistic Input in the First Hours of Adult Language Learning
Williams	Initial Incidental Acquisition of Word Order Regularities: Is It Just Sequence Learning?
Mackey, Adams,	Exploring the Relationship Between Modified Output and

Table 18 (cont'd)

Stafford, Winke	Working Memory Capacity
Boulton	Data-Driven Learning: Taking the Computer Out of the Equation
Crossley, Salsbury, McNamara	The Development of Polysemy and Frequency Use in English Second Language Speakers
Hsieh, Kang	Attribution and Self-Efficacy and Their Interrelationship in the Korean EFL Context
Hakansson, Norrby	Environmental Influence on Language Acquisition: Comparing Second and Foreign Language Acquisition of Swedish
Paradis	Bilingual Children's Acquisition of English Verb Morphology: Effects of Language Exposure, Structure Complexity, and Task Type
Li	Sociolinguistic Variation in the Speech of Learners of Chinese as a Second Language
Collentine, Asención-Delaney	A Corpus-Based Analysis of the Discourse Functions of Ser/Estar + Adjective in Three Levels of Spanish as FL Learners
Yuan, Woltz, Zheng	Cross-Language Priming of Word Meaning During Second Language Sentence Comprehension
Vandergrift, Tafaghodtari	Teaching L2 Learners How to Listen Does Make a Difference: An Empirical Study
Bowden, Gelfand, Sanz, Ullman	Verbal Inflectional Morphology in L1 and L2 Spanish: A Frequency Effects Study Examining Storage Versus Composition
Gor, Cook	Nonnative Processing of Verbal Morphology: In Search of Regularity
Kempe, Brooks, Kharkhurin	Cognitive Predictors of Generalization of Russian Grammatical Gender Categories
Morgan-Short	Second Language Acquisition of Gender Agreement in Explicit and Implicit Training Conditions: An Event-Related Potential Study
Murphy	Processing English Compounds in the First and Second Language: The Influence of the Middle Morpheme
	Language Learning, 2009
Bardovi-Harlig	Conventional Expressions as a Pragmalinguistic Resource: Recognition and Production of Conventional Expressions in L2 Pragmatics
Kim, Jang	Differential Functioning of Reading Subskills on the OSSLT for L1 and ELL Students: A Multidimensionality Model-Based DBF/DIF Approach
Schwieter, Sunderman	Concept Selection and Developmental Effects in Bilingual Speech Production
Boyd, Gottschalk, Goldberg	Linking Rule Acquisition in Novel Phrasal Constructions
Keating	Sensitivity to Violations of Gender Agreement in Native and

Table 18 (cont'd)

	Nonnative Spanish: An Eye-Movement Investigation
Potowski, Jegerski, Morgan-Short	The Effects of Instruction on Linguistic Development in Spanish Heritage Language Speakers
Rau, Chang, Tarone	Think or Sink: Chinese Learners' Acquisition of the English Voiceless Interdental Fricative
Tonzar, Lotto, Job	L2 Vocabulary Acquisition in Children: Effects of Learning Method and Cognate Status
Nekrasova	English L1 and L2 Speakers' Knowledge of Lexical Bundles
Abrahamsson, Hyltenstam	Age of Onset and Nativelikeness in a Second Language: Listener Perception Versus Linguistic Scrutiny
Crossley, Salsbury, McNamara	Measuring L2 Lexical Growth Using Hypernymic Relationships
Webb, Rodgers	Vocabulary Demands of Television Programs
Chen	Perception of Paralinguistic Intonational Meaning in a Second Language
Nassaji	Effects of Recasts and Elicitations in Dyadic Interaction and the Role of Feedback Explicitness
Lyster, Izquierdo	Prompts Versus Recasts in Dyadic Interaction
De Jong, Silbert, Park	Generalization Across Segments in Second Language Consonant Identification
Sanz, Lin, Lado, Bowden, Stafford	Concurrent Verbalizations, Pedagogical Conditions, and Reactivity: Two CALL Studies
Havik, Roberts, Van Hout, Schreuder, Haverkort	Processing Subject-Object Ambiguities in the L2: A Self-Paced Reading Study With German L2 Learners of Dutch
Peters, Hulstijn, Sercu, Lutjeharms	Learning L2 German Vocabulary Through Reading: The Effect of Three Enhancement Techniques Compared
Hall, Newbrand, Ecke, Sperr, Marchand, Hayes	Learners' Implicit Assumptions About Syntactic Frames in New L3 Words: The Role of Cognates, Typological Proximity, and L2 Status
Sparks, Patton, Ganschow, Humbach	Long-Term Crosslinguistic Transfer of Skills From L1 to L2
Kempe, Brooks	Second Language Learning of Complex Inflectional Systems
	Language Learning, 2008
Graham, Macaro	Strategy Instruction in Listening for Lower-Intermediate Learners of French
Lazarte, Barry	Syntactic Complexity and L2 Academic Immersion Effects on Readers' Recall and Pausing Strategies for English and Spanish Texts
Sheen	Recasts, Language Anxiety, Modified Output, and L2 Learning
Jackson	Proficiency Level and the Interaction of Lexical and Morphosyntactic Information During L2 Sentence Processing
Dewaele, Petrides, Furnham	Effects of Trait Emotional Intelligence and Sociobiographical Variables on Communicative Anxiety and Foreign Language Anxiety Among Adult Multilinguals: A Review and Empirical

Table 18 (cont'd)

	Investigation
Munro, Derwing	Segmental Acquisition in Adult ESL Learners: A Longitudinal Study of Vowel Production
Montrul, Foote, Perpiñán	Gender Agreement in Adult Second Language Learners and Spanish Heritage Speakers: The Effects of Age and Context of Acquisition
Ayoun	Acquisition of English Tense-Aspect Morphology by Advanced
Salaberry	French Instructed Learners
Kovács, Racsmány	Handling L2 Input in Phonological STM: The Effect of Non-L1 Phonetic Segments and Non-L1 Phonotactics on Nonword Repetition
Lee	Argument-Adjunct Asymmetry in the Acquisition of Inversion in Wh-Questions by Korean Learners of English
Leow, Hsieh, Moreno	Attention to Form and Meaning Revisited
Toth	Teacher- and Learner-Led Discourse in Task-Based Grammar Instruction: Providing Procedural Assistance for L2 Morphosyntactic Development
Kim	The Role of Task-Induced Involvement and Learner Proficiency in L2 Vocabulary Acquisition
Kormos, Csizér	Age-Related Differences in the Motivation of Learning English as a Foreign Language: Attitudes, Selves, and Motivated Learning Behavior
Tseng, Schmitt	Toward a Model of Motivated Vocabulary Learning: A Structural Equation Modeling Approach
Tavakoli, Foster	Task Design and Second Language Performance: The Effect of Narrative Type on Learner Output
Hamada, Koda	Influence of First Language Orthographic Experience on Second Language Decoding and Word Learning
Taguchi	Cognition, Language Contact, and the Development of Pragmatic Comprehension in a Study-Abroad Context
Min	EFL Vocabulary Acquisition and Retention: Reading Plus Vocabulary Enhancement Activities and Narrow Reading
Dimroth	Age Effects on the Process of L2 Acquisition? Evidence From the Acquisition of Negation and Finiteness in L2 German
Jansen	Acquisition of German Word Order in Tutored Learners: A Cross-Sectional Study in a Wider Theoretical Context
	Language Learning, 2007
Nassaji	Elicitation and Reformulation and Their Relationship With Learner Repair in Dyadic Interaction
Manchón, Roca de Larios	On the Temporal Nature of Planning in L1 and L2 Composing
Holtgraves	Second Language Learners and Speech Act Comprehension
Davis	Resistance to L2 Pragmatics in the Australian ESL Context
Lee, Kim	On Crosslinguistic Variations in Imperfective Aspect: The Case of L2 Korean

Table 18 (cont'd)

Geyer	Self-qualification in L2 Japanese: An Interface of Pragmatic, Grammatical, and Discourse Competences
Charkova	A Language Without Borders: English Slang and Bulgarian Learners of English
Mills, Pajares, Herron	Self-efficacy of College Intermediate French Students: Relation to Achievement and Motivation
Cheung, Chan, Chong	Use of Orthographic Knowledge in Reading by Chinese-English Bi-scriptal Children
Rott	The Effect of Frequency of Input-Enhancements on Word Learning and Text Comprehension
Leeser	Learner-Based Factors in L2 Reading Comprehension and Processing Grammatical Form: Topic Familiarity and Working Memory
Schiff, Calif	Role of Phonological and Morphological Awareness in L2 Oral Word Reading
Bae	Development of English Skills Need Not Suffer as a Result of Immersion: Grades 1 and 2 Writing Assessment in a Korean/English Two-Way Immersion Program
Stevenson, Schoonen, De Glopper	Inhibition or Compensation? A Multidimensional Comparison of Reading Processes in Dutch and English
Pulido	The Relationship Between Text Comprehension and Second Language Incidental Vocabulary Acquisition: A Matter of Topic Familiarity?
Wang, Koda	Commonalities and Differences in Word Identification Skills Among Learners of English as a Second Language
Jiang	Selective Integration of Linguistic Knowledge in Adult Second Language Learning
Barcroft	Effects of Opportunities for Word Retrieval During Second Language Vocabulary Learning
Mori, Sato, Shimuzi	Japanese Language Students' Perceptions on Kanji Learning and Their Relationship to Novel Kanji Word Learning Ability
Lee	Effects of Textual Enhancement and Topic Familiarity on Korean EFL Students' Reading Comprehension and Learning of Passive Form
Kaushanskaya, Marian	Bilingual Language Processing and Interference in Bilinguals: Evidence From Eye Tracking and Picture Naming
	Language Learning, 2006
Ferré, Sánchez-Casas, Guasch	Can a Horse Be a Donkey? Semantic and Form Interference Effects in Translation Recognition in Early and Late Proficient and Nonproficient Spanish-Catalan Bilinguals
Rubenfeld, Clément, Lussier, Lebrun, Auger	Second Language Learning and Cultural Representations: Beyond Competence and Identity
Abbott	ESL Reading Strategies: Differences in Arabic and Mandarin Speaker Test Performance
McDonough, Mackey	Responses to Recasts: Repetitions, Primed Production, and

Table 18 (cont'd)

	Linguistic Development
Vandergrift, Goh, Mareschal, Tafaghodtari	The Metacognitive Awareness Listening Questionnaire: Development and Validation
de Groot	Effects of Stimulus Characteristics and Background Music on Foreign Language Vocabulary Learning and Forgetting
Marsden	Exploring Input Processing in the Classroom: An Experimental Comparison of Processing Instruction and Enriched Input
Comajoan	The Aspect Hypothesis: Development of Morphology and Appropriateness of Use
Schauer	Pragmatic Awareness in ESL and EFL Contexts: Contrast and Development
Toth	Processing Instruction and a Role for Output in Second Language Acquisition
GeESLin, Guijarro-Fuentes	Second Language Acquisition of Variable Structures in Spanish by Portuguese Speakers
Kondo-Brown	How Do English L1 Learners of Advanced Japanese Infer Unknown Kanji Words in Authentic Texts?
Gullberg	Handling Discourse: Gestures, Reference Tracking, and Communication Strategies in Early L2
	Language Learning, 2005
Gass, Mackey, Ross-Feldman	Task-Based Interactions in Classroom and Laboratory Settings
Csizér, Dörnyei	Language Learners' Motivational Profiles and Their Motivated Learning Behavior
Sueyoshi, Hardison	The Role of Gestures and Facial Cues in Second Language Listening Comprehension
Paribakht	The Influence of First Language Lexicalization on Second Language Lexical Inferencing: A Study of Farsi-Speaking Learners of English as a Foreign Language
De Angelis	Interlanguage Transfer of Function Words
Nicoladis	The Acquisition of Complex Deverbal Words by a French-English Bilingual Child
García Mayo, Lázaro Ibarrola, Liceras	Placeholders in the English Interlanguage of Bilingual (Basque/Spanish) Children
Erdener, Burnham	The Role of Audiovisual Speech and Orthographic Information in Nonnative Speech Production
Rydland, Aukrust	Lexical Repetition in Second Language Learners' Peer Play Interaction
Clachar	Creole English Speakers' Treatment of Tense-Aspect Morphology in English Interlanguage Written Discourse
Lee	Facilitating and Inhibiting Factors in English as a Foreign Language Writing Performance: A Model Testing With Structural Equation Modeling
Carroll	Input and SLA: Adults' Sensitivity to Different Sorts of Cues to

Table 18 (cont'd)

	French Gender
Kempe, Brooks	The Role of Diminutives in the Acquisition of Russian Gender: Can Elements of Child-Directed Speech Aid in Learning Morphology?
Williams, Lovatt	Phonological Memory and Rule Learning
Dekydtspotter, Outcalt	A Syntactic Bias in Scope Ambiguity Resolution in the Processing of English-French Cardinality Interrogatives: Evidence for Informational Encapsulation
Major, Fitzmaurice, Bunta, Balasubramanian	Testing the Effects of Regional, Ethnic, and International Dialects of English on Listening Comprehension
Wang, Koda	Commonalities and Differences in Word Identification Skills Among Learners of English as a Second Language
Kiss, Nikolov	Developing, Piloting, and Validating an Instrument to Measure Young Learners' Aptitude
Sparks, Javorsky, Philips	Comparison of the Performance of College Students Classified as ADHD, LD, and LD/ ADHD in Foreign Language Courses
	Language Learning, 2004
Félix-Brasdefer	Interlanguage Refusals: Linguistic Politeness and Length of Residence in the Target Community
Derwing, Rossiter, Munro, Thomson	Second Language Fluency: Judgments on Different Tasks
Wayland, Guion	Training English and Chinese Listeners to Perceive Thai Tones: A Preliminary Report
Jung	Topic and Subject Prominence in Interlanguage Development
Laufer, Goldstein	Testing vocabulary knowledge: Size, strength, and computer adaptiveness
Zhang	Processing constraints, Categorical Analysis, and the Second Language Acquisition of the Chinese Adjective Suffix -de(ADJ)
Sasaki	A Multiple-Data Analysis of the 3.5-Year Development of EFL Student Writers
Liao, Fukuya	Avoidance of Phrasal Verbs: The Case of Chinese Learners of English
Albert, Kormos	Creativity and Narrative Task Performance: An Exploratory Study
Ishida	Effects of Recasts on the Acquisition of the Aspectual Form -te i-(ru) by Learners of Japanese as a Foreign Language
Gardner, Masgoret, Tennant, Mihic	Integrative Motivation: Changes During a Year-Long Intermediate-Level Language Course
Sanz, Morgan-Short	Positive Evidence Versus Explicit Rule Presentation and Explicit Negative Feedback: A Computer-Assisted Study
Yashima, Yashima, Shimizu	The Influence of Attitudes and Affect on Willingness to Communicate and Second Language Communication
Loewen	Uptake in Incidental Focus on Form in Meaning-Focused ESL Lessons

Table 18 (cont'd)

Language Learning, 2003	
Belz, Kinginger	Discourse Options and the Development of Pragmatic Competence by Classroom Learners of German: The Case of Address Forms
Phakiti	A Closer Look at Gender and Strategy Use in L2 Reading
Geeslin	A Comparison of Copula Choice: Native Spanish Speakers and Advanced Learners
Stevenson, Schoonen, de Glopper	Inhibition or Compensation? A Multidimensional Comparison of Reading Processes in Dutch and English
Jensen, Vinther	Exact Repetition as Input Enhancement in Second Language Acquisition
Hu	Phonological Memory, Phonological Awareness, and Foreign Language Word Learning
Vandergrift	Orchestrating Strategy Use: Toward a Model of the Skilled Second Language Listener
Gass, Svetics, Lemelin	Differential Effects of Attention
Verspoor, Lowie	Making Sense of Polysemous Words
Akamatsu	The Effects of First Language Orthographic Features on Second Language Reading in Text
Pulido	Modeling the Role of Second Language Proficiency and Topic Familiarity in Second Language Incidental Vocabulary Acquisition Through Reading
Izumi	Processing Difficulty in Comprehension and Production of Relative Clauses by Learners of English as a Second Language
Noels, Pelletier, Clément, Vallerand	Why Are You Learning a Second Language? Motivational Orientations and Self-Determination Theory
Baker, MacIntyre	The Role of Gender and Immersion in Communication and Second Language Orientations
Noels	Learning Spanish as a Second Language: Learners' Orientations and Perceptions of Their Teachers' Communication Style
MacIntyre, Baker, Clément, Donovan	Sex and Age Effects on Willingness to Communicate, Anxiety, Perceived Competence, and L2 Motivation Among Junior High School French Immersion Students
Ross	A Diachronic Coherence Model for Language Program Evaluation
Mackey, Oliver, Leeman	Interactional feedback and the incorporation of feedback: An exploration of NS-NNS and NNS-NNS adult and child dyads
Williams, Lovatt	Phonological Memory and Rule Learning
Language Learning, 2002	
Hansen, Umeda, McKinney	Savings in the Relearning of Second Language Vocabulary: The Effects of Time and Proficiency
Snellings, Van Gelderen, De Glopper	Lexical Retrieval: An Aspect of Fluent Second–Language Production That Can Be Enhanced
Qian	Investigating the Relationship Between Vocabulary Knowledge and Academic Reading Performance: An Assessment

Table 18 (cont'd)

	Perspective
Holowka, Brosseau-Lapré, Petitto	Semantic and Conceptual Knowledge Underlying Bilingual Babies' First Signs and Words
Dewaele, Pavlenko,	Emotion Vocabulary in Interlanguage
Barcroft	Semantic and Structural Elaboration in L2 Lexical Acquisition
Carson, Longhini	Focusing on Learning Styles and Strategies: A Diary Study in an Immersion Setting
Braidi	Reexamining the Role of Recasts in Native-Speaker/Nonnative-Speaker Interactions
Collins	The Roles of L1 Influence and Lexical Aspect in the Acquisition of Temporal Morphology
Storch	Patterns of Interaction in ESL Pair Work
	Language Learning, 2001
Barcroft	Acoustic Variation and Lexical Acquisition
Lee	Interlanguage Development by Two Korean Speakers of English With a Focus on Temporality
Matsumura	Learning the Rules for Offering Advice: A Quantitative Approach to Second Language Socialization
Lin	Syllable Simplification Strategies: A Stylistic Perspective
Iwashita, McNamara, Elder	Can We Predict Task Difficulty in an Oral Proficiency Test? Exploring the Potential of an Information-Processing Approach to Task Design
Reynolds	Language in the Balance: Lexical Repetition as a Function of Topic, Cultural Background, and Writing Development
Roca de Larios, Marín, Murphy	A Temporal Analysis of Formulation Processes in L1 and L2 Writing
Hulstijn, Laufer	Some Empirical Evidence for the Involvement Load Hypothesis in Vocabulary Acquisition
Muter, Diethelm	The Contribution of Phonological Skills and Letter Knowledge to Early Reading Development in a Multilingual Population
Tyler	Resource Consumption as a Function of Topic Knowledge in Nonnative and Native Comprehension
Ellis, Basturkmen, Loewen	Learner Uptake in Communicative ESL Lessons
Montrul	Causatives and Transitivity in L2 English
Day, Shapson	Integrating Formal and Functional Approaches to Language Teaching in French Immersion: An Experimental Study
DeKeyser Sokalski	The Differential Role of Comprehension and Production Practice
Leow	Attention, Awareness, and Foreign Language Behavior
Bardovi-Harlig	Another Piece of the Puzzle: The Emergence of the Present Perfect
Lyster	Negotiation of Form, Recasts, and Explicit Correction in Relation to Error Types and Learner Repair in Immersion Classrooms

Table 18 (cont'd)

Williams	Learner-Generated Attention to Form
	Language Learning, 2000
Warden	EFL Business Writing Behaviors in Differing Feedback Environments
Muranoi	Focus on Form Through Interaction Enhancement: Integrating Formal Instruction Into a Communicative Task in EFL Classrooms
Johnson, Prior, Artuso	Field Dependence as a Factor in Second Language Communicative Production
Bardovi-Harlig	Adverbials and the Acquisition of Simple Past Morphology
Bardovi-Harlig	Adverbials and Morphology in Reverse-Order Reports
Wharton	Language Learning Strategy Use of Bilingual Foreign Language Learners in Singapore
Jarvis	Methodological Rigor in the Study of Transfer: Identifying L1 Influence in them Interlanguage Lexicon
Kormos	The Role of Attention in Monitoring Second Language Speech Production
Rodríguez, Sadoski	Effects of Rote, Context, Keyword, and Context/Keyword Methods on Retention of Vocabulary in EFL Classrooms
de Groot, Keijzer	What Is Hard to Learn Is Easy to Forget: The Roles of Word Concreteness, Cognate Status, and Word Frequency in Foreign-Language Vocabulary Learning and Forgetting
Onwuegbuzie, Bailey, Daley	The Validation of Three Scales Measuring Anxiety at Different Stages of the Foreign Language Learning Process: The Input Anxiety Scale, the Processing Anxiety Scale, and the Output Anxiety Scale
Oliver	Age Differences in Negotiation and Feedback in Classroom and Pairwork
Murphy	Compounding and the Representation of L2 Inflectional Morphology
	Language Learning, 1999
Gass, Mackey, Alvarez-Torres, Fernández-García	The Effects of Task Repetition on Linguistic Output
Shehadeh	Non-Native Speakers' Production of Modified Comprehensible Output and Second Language Learning
Mori	Epistemological Beliefs and Language Learning Beliefs: What Do Language Learners Believe About Their Learning?
Cheng, Horwitz, Schallert	Language Anxiety: Differentiating Writing and Speaking Components
Wade-Woolley	First Language Influences on Second Language Word Reading: All Roads Lead to Rome
Abrahamsson	Vowel Epenthesis of /sC(C)/ Onsets in Spanish/Swedish Interphonology: A Longitudinal Case Study
Gholamain, Geva	Orthographic and Cognitive Factors in the Concurrent

Table 18 (cont'd)

	Development of Basic Reading Skills in English and Persian
Taylor, Kirsch, Jamieson, Eigor	Examining the Relationship Between Computer Familiarity and Performance on Computer-Based Language Tasks
Riney, Takagi	Global Foreign Accent and Voice Onset Time Among Japanese EFL Speakers
Wang, Lee	L2 Acquisition of Conflation Classes of Prenominal Adjectival Participles
Skehan, Foster	The Influence of Task Structure and Processing Conditions on Narrative Retellings
Eviatar, Leikin, Ibrahim	Phonological Processing of Second Language Phonemes: A Selective Deficit in a Bilingual Aphasic
Carlisle	The Modification of Onsets in a Markedness Relationship: Testing the Interlanguage Structural Conformity Hypothesis
Cichocki, House, Kinloch, Lister	Cantonese Speakers and the Acquisition of French Consonants
Major	Chronological and Stylistic Aspects of Second Language Acquisition of Consonant Clusters
Major, Kim	The Similarity Differential Rate Hypothesis
Stockman, Pluut	Segment Composition as a Factor in the Syllabification Errors of Second-Language Speakers
Hardison	Bimodal Speech Perception by Native and Nonnative Speakers of English: Factors Influencing the McGurk Effect
Munro, Derwing	Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners
	Language Learning, 1998
Myles, Hooper, Mitchell	Rote or Rule? Exploring the Role of Formulaic Language in Classroom Foreign Language Learning
Laufer, Paribakht	The Relationship Between Passive and Active Vocabularies: Effects of Language Learning Context
Derwing, Munro, Wiebe	Evidence in Favor of a Broad Framework for Pronunciation Instruction
Munro, Derwing	The Effects of Speaking Rate on Listener Evaluations of Native and Foreign-Accented Speech
Belmechri, Hummel	Orientations and Motivation in the Acquisition of English as a Second Language Among High School Students in Quebec City
Shirai, Kurono	The Acquisition of Tense-Aspect Marking in Japanese as a Second Language
Schmitt	Tracking the Incremental Acquisition of Second Language Vocabulary: A Longitudinal Study
Hoover, Dwivedi	Syntactic Processing by Skilled Bilinguals
Lotto, de Groot	Effects of Learning Method and Word Type on Acquiring Vocabulary in an Unfamiliar Language
Schoonen, Hulstijn, Bossers	Metacognitive and Language-Specific Knowledge in Native and Foreign Language Reading Comprehension: An Empirical Study Among Dutch Students in Grades 6, 8 and 10

Table 18 (cont'd)

Juffs	Main Verb Versus Reduced Relative Clause Ambiguity Resolution in L2 Sentence Processing
	Language Learning, 1997
van den Branden	Effects of Negotiation on Language Learners' Output
Inagaki	Japanese and Chinese Learners' Acquisition of the Narrow-Range Rules for the Dative Alternation in English
Buck, Tatsuoka, Kostin	The Subskills of Reading: Rule-space Analysis of a Multiple-choice Test of Second Language Reading Comprehension
van Hell, Mahn	Keyword Mnemonics Versus Rote Rehearsal: Learning Concrete and Abstract Foreign Words by Experienced and Inexperienced Learners
de Groot, Poot	Word Translation at Three Levels of Proficiency in a Second Language: The Ubiquitous Involvement of Conceptual Memory
MacIntyre, Noels, Clément	Biases in Self-Ratings of Second Language Proficiency: The Role of Language Anxiety
Purpura	An Analysis of the Relationships Between Test Takers' Cognitive and Metacognitive Strategy Use and Second Language Test Performance
Carlisle	The Modification of Onsets in a Markedness Relationship: Testing the Interlanguage Structural Conformity Hypothesis
Robinson	Individual Differences and the Fundamental Similarity of Implicit and Explicit Adult Second Language Learning
Polio	Measures of Linguistic Accuracy in Second Language Writing Research
White, Bruhn-Garavito, Kawasaki, Pater, Prévost	The Researcher Gave the Subject a Test about Himself: Problems of Ambiguity and Preference in the Investigation of Reflexive Binding
	Language Learning, 1996
Lin, Hedgcock	Negative Feedback Incorporation Among High-Proficiency and Low-Proficiency Chinese-Speaking Learners of Spanish
Gu, Johnson	Vocabulary Learning Strategies and Language Learning Outcomes
Ying	Multiple constraints on processing ambiguous sentences: Evidence from adult L2 learners
Avila, Sadoski	Exploring new applications of keyword method to acquire English vocabulary
Kobayashi, Rinnert	Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background
Nicoladis, Genesee	A longitudinal study of pragmatic differentiation in young bilingual children
Donato, Antonek, Tucker	Monitoring and assessing a Japanese FLES program: Ambiance and achievement
Akiyama, Williams	Spatial components in the use of count nouns among English speakers and Japanese speakers of English as a second language
Elder	The Effect of Language Background on "Foreign" Language

Table 18 (cont'd)

	Test Performance: The Case of Chinese, Italian, and Modern Greek
Juffs, Harrington	Garden path sentences and error data in second language sentence processing
Carrell, Prince, Astika	Personality types and language learning in an EFL context
Lawson, Hogben	The vocabulary-learning strategies of foreign-language students
Sasaki, Hirose	Explanatory variables for EFL students' expository writing
	Language Learning, 1995
Yuan	Acquisition of base-generated topics by English-speaking learners of Chinese
Lockhart, Ng	Analyzing talk in ESL peer response groups: Stances, functions, and content
Takano, Noda	Interlanguage dissimilarity enhances the decline of thinking ability during foreign language processing
de Groot, Hoeks	The development of bilingual memory: Evidence from Word Translation by trilinguals
Whalen, Menard	L1 and L2 writers' strategic and linguistic knowledge: A model of multiple-level discourse processing
Klein	Second versus third language acquisition: Is there a difference?
de Groot, Comijs	Translation recognition and translation production: Comparing a new and an old tool in the study of bilingualism
Rose, Ono	Eliciting speech act data in Japanese: The effect of questionnaire type
Zhang	Semantic differentiation in the acquisition of English as a second language
Chalhoub-Deville	A contextualized approach to describing oral language proficiency
Young	Conversational styles in language proficiency interviews
Harley, Howard, Hart	Second language processing at different ages: Do younger learners pay more attention to prosodic cues to sentence structure?
Munro, Derwing	Foreign accent, comprehensibility, and intelligibility in the speech of second language learners
Robinson	Task complexity and second language narrative discourse
	Language Learning, 1994
Robinson	Universals of word formation processes: Noun incorporation in the acquisition of Samoan as a second language
Pearson, Fernandez	Patterns of interaction in the lexical growth in two languages of bilingual infants and toddlers
Major	Chronological and stylistic aspects of second language acquisition of consonant clusters
Verhoeven	Transfer in bilingual development: The linguistic interdependence hypothesis revisited
Clément, Dörnyei, Noels	Motivation, self-confidence, and group cohesion in the foreign language classroom

Table 18 (cont'd)

Ellis, Tanaka, Yamazaki	Classroom interaction, comprehension, and the acquisition of L2 word meanings
Yano, Long, Ross	The effects of simplified and elaborated texts on foreign language reading comprehension
Umbel, Oller	Developmental changes in receptive vocabulary in Hispanic bilingual school children
MacIntyre, Gardner	The subtle effects of language anxiety on cognitive processing in the second language
Shimron, Sivan	Reading proficiency and orthography: Evidence from Hebrew and English
Simmons-McDonald	Comparative patterns in the acquisition of English negation by native speakers of French Creole and Creole English
Macdonald, Yule, Powers	Attempts to improve English L2 pronunciation: The variable effects of different types of instruction
Jin	Topic-prominence and subject-prominence in L2 acquisition: Evidence of English-to-Chinese typological transfer
Hamilton	Is implicational generalization unidirectional and maximal? Evidence from relativization instruction in a second language
	Language Learning, 1993
Sasaki	Relationships among second language proficiency, foreign language aptitude, and intelligence: A protocol analysis
Heilenman, McDonald	Processing strategies in L2 learners of French: The role of transfer
Ellis, Beaton	Psycholinguistic determinants of foreign language vocabulary learning
Sasaki	Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach
Horiba, van den Broek, Fletcher	Second language readers' memory for narrative texts: Evidence for structure-preserving top-down processing
Donin, Silva	The relationship between first- and second-language reading comprehension of occupation-specific texts
Zuengler	Encouraging learners' conversational participation: The effect of content knowledge
Gardner, MacIntyre	On the measurement of affective variables in second language learning
Tamamaki	Language dominance in bilinguals' arithmetic operations according to their language use
Luppescu, Day	Reading, dictionaries, and vocabulary learning
Geva, Ryan	Linguistic and cognitive correlates of academic skills in first and second languages
Cochocki, House, Kinlock, Lister	Cantonese speakers and the acquisition of French consonants
Koster, Koet	The evaluation of accent in the English of Dutchmen
Pearson, Fernandez,	Lexical development in bilingual infants and toddlers:

Table 18 (cont'd)

Oller	Comparison to monolingual norms
	Language Learning, 1992
Bacon, Finneman	Sex differences in self-reported beliefs about foreign-language learning and authentic oral and written input
Danan	Reversed subtitling and dual coding theory: New directions for foreign language instruction
Anderson-Hsieh, Johnson, Koehler	The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure
Buck	Listening comprehension: Construct validity and trait characteristics
Wang, Thomas	The effect of imagery-based mnemonics on the long-term retention of Chinese characters
Samimy, Tabuse	Affective variables and a less commonly taught language: A study in beginning Japanese classes
Eisterhold Carson, Kuehn	Evidence of transfer and loss in developing second language writers
Kobayashi, Rinnert	Effects of first language on second language writing: Translation versus direct composition
Johnson	Critical period effects in second language acquisition: The effect of written versus auditory materials on the assessment of grammatical competence
Yule, Powers, MacDonald	The variable effects of some task-based learning procedures on L2 communicative effectiveness
Carrell	Awareness of text structure: Effects on recall
Stockman, Pluut	Segment composition as a factor in the syllabification errors of second-language speakers
	Language Learning, 1991
MacIntyre, Gardner	Language anxiety: Its relationship to other anxieties and to processing in native and second languages
Buczowska, Weist	The effects of formal instruction on the second-language acquisition of temporal location
Ephratt	Piaget's nominal realism from a linguistic point of view
Fotos	The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations?
Hamp-Lyons, Henning	Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts
Elley	Acquiring literacy in a second language: The effect of book-based programs
Bamford, Mizokawa	Additive-bilingual (immersion) education: Cognitive and language development
Thompson	Foreign accents revisited: The English pronunciation of Russian immigrants
Verhoeven	Predicting minority children's bilingual proficiency: Child,

Table 18 (cont'd)

	family, and institutional factors
Rost, Ross	Learner use of strategies in interaction: Typology and teachability
Hadden	Teacher and nonteacher perceptions of second-language communication
Day, Shapson	Integrating formal and functional approaches to language teaching in French immersion: An experimental study
Segalowitz	Does advanced skill in a second language reduce automaticity in the first language
	Language Learning, 1990
Bardovi-Harlig, Hartford	Congruence in native and nonnative conversations: Status balance in the academic advising session
Segalowitz, Hebert	Phonological recoding in the first and second language reading of skilled bilinguals
Yule, Macdonald	Resolving referential conflicts in L2 interaction: The effect of proficiency and interactive role
Cook	Timed comprehension of binding in advanced L2 learners of English
Griffiths	Speech rate and NNS comprehension: A preliminary study in time-benefit analysis
Sasaki	Topic prominence in Japanese EFL students' existential constructions
Register	Influences of typological parameters on L2 learners' judgments of null pronouns in English
Lennon	Investigating fluency in EFL: A quantitative approach
Si-Qing	A study of communication strategies in interlanguage production by Chinese EFL learners
Ramage	Motivational factors and persistence in foreign language study
Nayak, Hansen, Krueger, McLaughlin	Language-learning strategies in monolingual and multilingual adults
Fouly, Bachman, Cziko	The divisibility of language competence: A confirmatory approach
Olshtain, Shohamy, Kemp, Chatow	Factors predicting success in EFL among culturally different learners
Dörnyei	Conceptualizing motivation in foreign-language learning

REFERENCES

REFERENCES

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*, 207-245.

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of organizational research methods trends in design, measurement, and data-analysis topics. *Organizational Research Methods, 12*, 69-112.

Altman, M. (2004). Statistical significance, path dependency, and the culture of journal publication. *The Journal of Socio-Economics, 33*, 651-663.

Anzures-Cabrera, J., & Higgins, J. P. T. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods, 1*, 66-80.

Bangert, A. W., & Baumberger, J. P. (2005). Research and statistical techniques used in the *Journal of Counseling & Development*. *Journal of Counseling & Development, 83*, 480-487.

Brown, J. D. (1997). Skewness and kurtosis. *Shiken: JALT Testing & Evaluation SIG Newsletter, 1*, 20-23.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley.

Brutus, S., Gill, H. & Duniewicz, K. (2010). State-of-science in industrial and organizational psychology: A review of self-reported limitations. *Personnel Psychology*, 63, 907-936.

Cameron, J., & Pierce, W. D. (1996). The debate about rewards and intrinsic motivation: Protests and accusations do not alter the results. *Review of Educational Research, 66*, 39-51.

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods* 2004 7: 151-167.

Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., et al. (1981). A method for assessing the quality of a randomized control trial. *Control Clinical Trials, 2*, 31-49.

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials. *Journal of the American Medical Association, 291*, 2457-2465.

Chaudron, C. (2001). Progress in language classroom research: Evidence from *The Modern Language Journal*, 1916-2000. *The Modern Language Journal, 85,* 57-76.

Cheung, S. F., & Chan, D. K.-S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology, 89*, 780-791.

Cohen J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145-153.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426-443.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 97-1003.

Combs, J. G. (2010). Big samples and small effects: Let's not trade relevance and rigor for power. *Academy of Management Journal, 53*, 9-14.

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in educational research* (pp. 150-178). Washington, DC: Brookings Institute.

Cortina, J. M., & Landis, R. S. (2011). The earth is *not* round ($p = .00$). *Organizational Research Methods, 14*, 332-349.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116-127.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitude and instructional methods: A handbook on research for interactions*. New York: Irvington.

Crookes, G. (1991). Power, effect size, and second language research. Another researcher comments. *TESOL Quarterly, 25*, 762-765.

DeKeyser, R. (2009, October). *Variable interaction in SLA: Much more than a nuisance*. Plenary address given at the Second Language Research Forum, East Lansing, MI.

DeKeyser, R., & Schoonen, R. (2007). Editors' announcement. *Language Learning, 57*, ix-x.

Dinsmore, T. H. (2006). Principles, parameters, and SLA: A retrospective meta-analytic investigation into adult L2 learners' access to Universal Grammar. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 53-90). Philadelphia: John Benjamins

DeVaney, T. A. (2001). Statistical significance, effect size, and replication: What do the journals say? *The Journal of Experimental Education, 69*, 310-320.

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomized and nonrandomized studies of health care interventions. *Journal of Epidemiology & Community Health, 52*, 377-384.

Egbert, J. (2007). *Quality analysis of journals in TESOL and applied linguistics*. *TESOL Quarterly*, *41*, 157–171.

Ellis, N. C. (2000). Editorial statement. *Language Learning, 50*, xi–xiii.

Felser, C. (2005). Experimental psycholinguistic approaches to second language acquisition. *Second Language Research, 21*, 95-97.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23*, 89–105.

Fidler, F., & Cumming, G. (2007). Lessons learned from statistical reform efforts in other disciplines. *Psychology in the schools, 44*, 441-449.

Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory & Psychology, 12*, 825-853.

Fish, L. (1988). Why multivariate methods and usually vital. *Measurement and Evaluation in Counseling and Development, 21*, 130-137.

Flahive, D., & Ehlers-Zavala, F. (2010, March). *Power analysis in applied linguistics research*. Paper presented at the Annual Conference of the American Association for Applied Linguistics, Altanta, GA.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Gass, S. (1993). Editorial: Second language acquisition: Cross-disciplinary perspectives. *Second Language Research, 9*, 95-98.

Gass, S. (2009). A survey of SLA research. In W. Ritchie & T. Bhatia (Eds.) *Handbook of Second language acquisition* (pp. 3-28). Bingley, UK: Emerald.

Gass, S., Fleck, C., Leder, N., & Svetics, I. (1998). Ahistoricity revisited: Does SLA have a history? *Studies in Second Language Acquisition, 20*, 407-421.

Gass, S., Mackey, A., & Ross-Feldman, L. (2005). Task-based interactions in classroom and laboratory settings. *Language Learning, 55*, 575-611.

Gelman, A., Hill, J., & Yajima, M. (2009). Why we (usually) don't have to worry about multiple comparisons. Manuscript in preparation.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex, and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist, 97*, 310-316.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 357-376). New York: Russell Sage Foundation.

Goodwin, L. D., & Goodwin, W. L. (1985a). An analysis of statistical techniques used in the Journal of Educational Psychology, 1979-1983. *Educational Psychologist, 20*, 13-21.

Graham, J. M. (2008). The General Linear Model as structural equation modeling. *Journal of Educational and Behavioral Statistics, 33*, 485-506.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston: Heinle & Heinle.

Hattie, J. (1987). Measuring the effects of schooling. *Australian Journal of Education, 36*, 5-13.

Hauser, E. (2001, October). *The statistical power of second language acquisition research: A review*. Paper presented at the Pacific Second Language Research Forum, University of Hawaii at Manoa.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic.

Henning, G. (1986). Quantitative methods in language acquisition research. *TESOL Quarterly, 20*, 701–708.

Henson, R. K., Hull, D. M., & Williams, C. S. (2009). Methodology in our education research culture: Towards a stronger collective quantitative proficiency. *Educational Researcher, 39*, 229-240.

Humphreys, L. G. (1978). Doing research the hard way: Substituting analysis of variance for a problem in correlational analysis. *Journal of Educational Psychology, 70*, 873-876.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Journal Article Reporting Standards Working Group (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839-851.

Jüni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of randomised controlled trials. In M. Egger, G. D. Smith, & D. G. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd ed., pp. 87-108). London: BMJ.

Keck, C., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Philadelphia: Benjamins.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *The Journal of Experimental Education, 69,* 280-309.

Keselman, H. J., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350-386.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746-759.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Klingner, J. K., Scanlon, D., & Pressley, M. (2005). How to publish in scholarly journals. *Educational Researcher, 34,* 14-20.

Kubanyiova, M. (2008). Rethinking research ethics in contemporary applied linguistics: The tension between macroethical and microethical perspectives in situated research. *The Modern Language Journal, 92,* 503-518.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Lazaraton, A. (1991). Power, effect size, and second language research. A researcher comments. *TESOL Quarterly, 25,* 759-762.

Lazaraton, A. (2000). Current trends in research methodology and statistics in applied linguistics. *TESOL Quarterly, 34,* 175-181.

Lazaraton, A., Riggenbach, H., & Ediger, A. (1987). Forming a discipline: Applied linguists' literacy in research methodology and statistics. *TESOL Quarterly, 21*, 263-277.

Lent, R. H., Aurbach, H. A., & Levin, L. S. (1971). Research design and validity assessment. *Personnel Psychology, 24,* 247-274.

Li, S. (2010). *Corrective feedback in perspective: The interface between feedback type, proficiency, the choice of target structure, and learners' individual differences in working memory and language analytic ability*. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Lightbown, P. M. (2000). Anniversary article: Classroom second language research and second language teaching. *Applied Linguistics, 21,* 431-462.

Lipsey, M. W. (2009). Identifying interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 147-158). New York: Russell Sage Foundation.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Loewen, S. (2005). Incidental focus on form and second language learning. *Studies in Second Language Acquisition, 27*, 361-386.

Loewen, S., & Gass, S. (2009). The use of statistics in L2 acquisition research. *Language Teaching, 42*, 181-196.

Lykken, D. E. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151-159.

Lyster, R., & Izquierdo, J. (2009). Prompts versus recasts in dyadic interaction. *Language Learning, 59*, 453-498.

Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Erlbaum.

Mackey, A., & Gass, S. M. (2006). Pushing the methodological boundaries in interaction research: An introduction to the special issue. *Studies in Second Language Acquisition, 28*, 169-178.

Mackey, A., & Gass, S. (in press). *Research methodologies in second language acquisition*. London: Basil Blackwell.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407-449). Oxford: Oxford University Press.

Mackey, A., & Sachs, R. (in press). Older learners in SLA research: A first look at working memory, feedback, and L2 development. *Language Learning*.

Magnan, S. S. (1994). From the editor: The MLJ tradition and the challenges ahead. *The Modern Language Journal, 78*, 7-9.

Matrixx Initiatives Inc. v. Siracusano. 09-1156. (2010).

Matthews, M. S., Gentry, M., McCoach, D. B., Worrell, F. C., Matthews, D., & Dixon, F. (2008). Evaluating the state of a field: Effect size reporting in gifted education. *The Journal of Experimental Education, 77*, 55-65.

Meara, P. (1995). [Book review of *Statistical techniques for the study of language and language behavior*.] *Language Learning, 45*, 341-343.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37*, 113-115.

Moja, L. O., Telaro, E., D'Amico, R., Moschetti, I., Coe, L., & Liberati, A. (2005). Assessment of methodological quality of primary studies by systematic reviews: Results of the metaquality cross sectional study. *British Medical Journal, 330*, 1053-1057.

Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology, 49*, 103-120.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*, 364-386.

Nassaji, H. (in press). Significance tests and generalizability of research results: A case for replication. In G. Porte (Ed.), *Replication research in applied linguistics*. Cambridge: Cambridge University Press.

Nekrasova, T., & Becker, T. (2009). Effectiveness of practice: A research synthesis and quantitative meta-analysis. Manuscript in preparation.

Nicoladis, E. & Krott, A. (2007). Word family size and French-speaking children's segmentation of existing compounds. *Language Learning, 57*, 201-228.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*, 417–528.

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717–761). Malden, MA: Blackwell.

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 3–50). Philadelphia: Benjamins.

Nunan, D. (1991). Methods in second language classroom-oriented research: A critical review. *Studies in Second Language Acquisition, 13*, 249-274.

Nunan, D. (1996). Issues in second language acquisition research: Examining substance and procedure. In W. C. Ritchie & T. K. Bhatia (Eds.), *The handbook of second language acquisition* (pp. 349-374). San Diego: Academic Press.

Ortega, L. (2005). Methodology, epistemology, and ethics in instructed SLA research: An introduction. *The Modern Language Journal, 89*, 317-327.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics, 30*, 85-110

Pica, T. (1997) Second language teaching and research relationships: A North American view. *Language Teaching Research, 1*, 48-72.

Pigott, T. D. (2009). Handling missing data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 399-416). New York: Russell Sage Foundation.

Plonsky, L. (2009, October). *"Nix the null": Why statistical significance is overrated*. Paper presented at the Second Language Research Forum (SLRF), East Lansing, MI.

Plonsky, L. (2011, February). *Data reporting practices and study quality in L2 research*. Paper presented at the SLS Spring Symposium, East Lansing, MI.

Plonsky, L. (in press-a). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning, 61*.

Plonsky, L. (in press-b). Replication, meta-analysis, and generalizability. In G. Porte (Ed.), *Replication research in applied linguistics*. Cambridge: Cambridge University Press.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning, 61*, 325-366.

Plonsky, L., & Oswald, F. L. (2010, March). *Interpreting mean differences in L2 research: Cohen's benchmarks for* d-*values, revisited*. Paper presented at the conference of the American Association for Applied Linguistics, Atlanta, GA.

Plonsky, L., & Oswald, F. L. (in press). How to do a meta-analysis. In A. Mackey & S. Gass (Eds.), *A guide to research methods in second language acquisition*. London: Basil Blackwell.

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

Polio, C. (in press). Replication in published applied linguistics research: An historical perspective. In G. Porte (Ed.), *Replication research in applied linguistics*. Cambridge: Cambridge University Press.

Polio, C., & Gass, S. (1997). Replication and reporting: A commentary. *Studies in Second Language Acquisition, 19,* 499-508.

Porte, G. (2010) *Appraising research in second language learning: A practical approach to critical analysis of quantitative research (2nd ed.).* Amsterdam: John Benjamins.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160-164.

Publication manual of the American Psychological Association (6th ed.). (2010). Washington, DC: American Psychological Association.

Pulido, D. (2004). The relationship between text comprehension and second language incidental vocabulary acquisition: A matter of topic familiarity? *Language Learning, 54,* 469-523.

Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis*. New York: Routledge.

Read, J. (2007). Towards a new collaboration: Research in SLA and language testing. *New Zealand Studies in Applied Linguistics, 13,* 22-35.

Richard, F. D., Bond, C. F. Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331-363.

Rothstein H. R., Sutton A. J., & Borenstein M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: Wiley.

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133-164). Philadelphia: Benjamins.

Salgado, J. F. (1998). Sample size in validity studies of personnel selection. *Journal of Occupational and Organizational Psychology, 71*, 161-164.

Selinker, L., & Lakshmanan, U. (2001). How do we know what we know?; Why do we believe what we believe? *Second Language Research, 17*, 323-325.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training researchers. *Psychological Methods, 1*, 115-129.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

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

Sheen, Y. (2007). The effects of corrective feedback, language aptitude, and learner attributes on the acquisition of English articles. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 301-322). Oxford: Oxford University Press.

Smith, B., & Lafford, B. A. (2009). The evaluation of scholarly activity in computer-assisted language learning. *Modern Language Journal, 93*, 868-883.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60*, 263-308.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, 102, 989-1004.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 435-452). New York: Russell Sage Foundation.

Teleni, V., & Baldauf, R. B. (1989). Statistical techniques used in three applied linguistics journals: *Language Learning*, *Applied Linguistics*, and *TESOL Quarterly*, 1980-1986: Implications for readers and researchers. Unpublished research report. (ERIC Document Reproduction Service No. ED312905)

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *The Journal of Experimental Education, 70*, 80-93.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423-432.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent *JCD* research articles. *Journal of Counseling and Development, 76*, 436-441.

Tracy, J. L., Robins, R. W., Sherman, J. W. (2009). *Journal of Personality and Social Psychology, 96*, 1206-1225.

Trikalinos, T. A., et al. (2004). Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *Journal of Clinical Epidemiology, 57*, 1124-1130.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. The Journal of Experimental Education, 67, 335-341.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*, 473-481.

Valdman, R. (1998). A note from the Editor: 20th anniversary of SSLA. *Studies in Second Language Acquisition, 20*, 463-470.

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The study design and implementation assessment device (Study DIAD). P*sychological Methods, 13*, 130-149.

VanPatten, B., & Williams, J. (2002). *Research criteria for tenure in second language acquisition: Results from a survey of the field*. Unpublished manuscript, The University of Illinois at Chicago.

Wa-Mbaleka, S. (2006). *A meta-analysis investigating the effects of reading on second language vocabulary learning*. Unpublished doctoral dissertation, Northern Arizona University, Flagstaff, AZ.

Waring, H. Z. *(2009). M*oving out of IRF (initiation-response-feedback): A single case analysis. *Language Learning, 59*, 796-824.

Wells, C. S., & Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychology in the Schools, 44*, 495-502.

Wells, K., & Littell, J. H. (2009). Study quality assessment in systematic reviews of research on intervention effects. *Research on Social Work Practice, 19*, 52–62.

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Willson, V. L. (1980). Research techniques in *AERJ* articles: 1969 to 1978. *Educational Researcher, 9,* 5-10.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6*, 413-442.

Yirmiya, N., Erel, O., Shaked, M. Solomonica-Levi, D. (1998). Meta-Analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin, 124*, 283-307.