CHAPTER 2
Methodological Synthesis in Quantitative L2 Research: A Review of Reviews and a Case Study of Exploratory Factor Analysis

Luke Plonsky\textsuperscript{a} and Talip Gonulal\textsuperscript{b}
\textsuperscript{a}Northern Arizona University and \textsuperscript{b}Michigan State University

Research synthesis and meta-analysis provide a pathway to bring together findings in a given domain with greater systematicity, objectivity, and transparency than traditional reviews. The same techniques and corresponding benefits can be and have been applied to examine methodological practices in second language (L2) research (e.g., Plonsky, 2013). In the first half of this paper, we integrate findings, trends, and critiques from a number of syntheses to both illustrate the potential of this approach and to promote more methodologically informed research practices. Our emphasis here is on study designs and sampling practices. In the second half, we provide an example of a methodological synthesis that reviews the use of one particular statistical technique as applied in L2 research: exploratory factor analysis (EFA). Here and throughout the chapter, we provide specific recommendations for primary research as well as for future efforts to synthesize methodological practices in the field.

Keywords research synthesis; meta-analysis; research methods; factor analysis

Introduction

Research synthesis and meta-analysis comprise techniques that can improve on traditional literature reviews in many ways. Namely, these procedures provide enhanced levels of systematicity, objectivity, and transparency, thus also providing a fuller and more precise description of effects across studies in a particular domain.

The same benefits resulting from the set of procedures that make up the synthetic process (e.g., locating primary studies, coding for features and effects;...
see Plonsky & Oswald, in press) are by no means limited to addressing substantive questions. A synthetic approach can also be employed fruitfully to describe and evaluate methodological phenomena. A small but growing number of such reviews of second language (L2) research have appeared in recent years, both independent of and in conjunction with meta-analyses (i.e., systematic reviews involving the aggregation of effect sizes). This line of research employs synthetic methods to describe and evaluate the presence of research and reporting practices in a given domain, whether broadly or narrowly conceived. In parallel to primary research, synthetic studies of this type treat primary studies as participants that are surveyed to collect methodologically oriented data.

The first half of this chapter presents findings from studies related to methodological practices in quantitative L2 research as observed in syntheses covering a variety of subdomains within the field. Our objective here and throughout is to bring together findings, trends, and critiques from methodological syntheses to both illustrate the potential of this set of techniques and to promote more methodologically informed research practices. In the second half, we provide a detailed example of such a review focusing on the use of EFA in L2 research. This study, unlike most others reviewed in this chapter, is not limited to any particular substantive domain and is thus relevant to a broad range of interests in the field. Here and throughout the chapter, we also maintain a forward-facing outlook by providing recommendations for the practice of primary quantitative research based on our review. We also encourage further consideration of methodological issues in future syntheses of L2 research.

Quantitative Research Practices: A Review of Reviews

Syntheses of quantitative methods in L2 research have focused mainly on three issues: (a) study design (e.g., random assignment), (b) statistical analyses (e.g., which statistics have been used and how often), and (c) reporting practices (e.g., reliability estimates, effect sizes). Although there is inevitably a degree of overlap among these, our chapter will focus primarily on reviews of study design, which is perhaps the most critical among these and which is also relevant to all L2 research. (For an in-depth discussion based on syntheses of reporting practices, see Larson-Hall & Plonsky, 2015.)

The methods used by researchers to carry out review studies in this area have largely (but not exclusively) adhered to well-established research synthetic and meta-analytic procedures (see Plonsky & Oswald, in press). By this we mean that researchers, first, systematically collect primary studies based on a set
of parameters for inclusion/exclusion. Next a coding scheme is developed to extract information from each study related to the methodological features and practices of interest. This phase of the synthesis can be understood in parallel to primary data collection where participants (in this case, studies) are surveyed to better understand the phenomena under investigation. In the case of meta-analysis, this phase also involves recording effect sizes to later be averaged. Finally, once the data are collected, the synthesist analyzes the data pertaining to the methodological features being studied. In many cases, the analyses are rather simple and straightforward (e.g., frequency counts of different research and reporting practices). As described below, however, other analyses are more complex, taking into account the relationships among methodological features and other study characteristics (e.g., year of publication, publication venue, effect sizes).

Our review here examines findings from studies that have examined methodological features of primary studies. Studies of this type present systematic reviews of research and reporting practices (e.g., statistical analyses), whether particular to a given domain as in Plonsky and Gass (2011) or across domains as in Gass (2009). In some cases results from methodological reviews were produced in the context of larger projects related to substantive findings of a given area of L2 research. Norris and Ortega’s (2000) meta-analysis, for example, not only synthesized results pertaining to the effects of L2 instruction in their sample but analyzed and critiqued its methodological practices (e.g., use of pretesting) as well. In other cases, the interest of reviewers has been primarily or exclusively methodological (e.g., Plonsky, 2013).

We have sought to bring together in this paper—a review of reviews—the findings, trends, and critiques regarding methodological practices as found in secondary-level analyses of quantitative L2 studies. Our goal in doing so is to draw attention both to study findings themselves and to the potential of this budding line of research. We also hope that shedding light on methodological practices will lead to improved research practices that will then more accurately and informatively contribute to L2 theory and practice.

**Motivations in Methodological Syntheses**

Before getting into these studies and, in particular, results pertaining to research designs, we want to draw attention to four different but nonexclusive motivations driving the recent line of methodological reviews in L2 research. Several syntheses have been motivated by an interest in describing and defining the field as a kind of window into the methodological culture of L2 research generally and/or of different journals. Gass (2009), for example, chose to survey data
analysis, measures, and statistics used in second language acquisition (SLA) research, rather than “considering details of what is known about SLA” (p. 3). Notably absent from some studies with this goal is any explicit interest in assessing quality or in criticizing the methods observed. That is, their interest is in simply describing (not evaluating) methodological practices.

Other methodological syntheses, however, have been expressly motivated by and concerned with quality. Unlike the previous type of review, such research seeks not only to describe but to evaluate and comment on the field’s practices with the intention to improve future research as well. One inherent challenge in these studies is the subjectivity involved in defining and operationalizing quality. Plonsky (2013), for example, defined it as “the combination of (a) adherence to standards of contextually appropriate, methodological rigor in research practices and (b) transparent and complete reporting of such practices” (p. 657). The instrument employed in his study was informed by several sources including (a) the Publication Manual of the American Psychological Association (APA, 2010) and other APA publications (Journal Article Reporting Standards Working Group, 2008; Wilkinson & Task Force on Statistical Inference, 1999), (b) L2 journal guidelines and editorials (Chapelle & Duff, 2003; Ellis, 2000), and (c) numerous measures of study quality employed in the meta-analytic literature to weight study effects (e.g., Valentine & Cooper, 2008). Nevertheless, there is no widely agreed-upon definition of study quality, even for a particular set of methodologies that might fall under the “quantitative” rubric. Given the relative scarcity and novelty of quality-oriented reviews, we should expect—if not hope—further refinement of operational definitions and instrumentation. A related challenge to this type of review lies in identifying features associated with quality that are relevant to a potentially diverse set of primary studies. Future assessments of quality will hopefully also seek to develop definitions and operationalizations specific to the domains they address.

A third motivation or interest in methodological syntheses has been to examine the relationship between study features and 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-Haase & Thompson, 2004, p. 478). To be clear, this type of analysis does not suggest that methodological quality necessarily bears directly on study outcomes. For example, though Plonsky (2011) found larger effects for studies that randomly assigned participants to experimental conditions and that reported reliability estimates, no one would claim that randomly assigning
participants to experimental conditions or reporting reliability coefficients causes larger or smaller effects. On the other hand, 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. These are treated as empirical matters in methodological syntheses and, therefore, this line of investigation has explored possible relationships between study features and outcomes/results in order to better understand how L2 research has been carried out and to determine how these variables may interact (see Plonsky & Oswald, 2014).

The fourth motivation behind methodological syntheses of quantitative L2 research has been to examine changes and/or improvements taking place over time in the field’s designs, analyses, and so forth. Gass (2009) and Lazaraton (2005), for example, both tracked chronological developments in L2 researchers’ designs and analyses. The goals of these studies generally overlap with those seeking to better understand the culture of L2 research (e.g., motivation #1 above) and/or those seeking to evaluate methodological quality in the field (motivation #2; e.g., Chaudron, 2001; Plonsky & Gass, 2011). In this case, however, the focus is not on overall patterns but on changes and improvements taking place.

Design Issues
A number of reviews have examined design-related features in samples of primary L2 research, relying on categorization schemes similar to the model described above. Our review here considers all (or nearly all) syntheses of this type conducted in the context of L2 research. Fortunately, many of the same issues have been addressed systematically across syntheses, affording greater comparability and generalizability of findings related to methodological practices. In one such study, Gass (2009) tallied articles across four journals as quantitative, qualitative, theoretical, introductory, quantitative+qualitative, or descriptive. Using a somewhat simpler scheme, Lazaraton (2005), classified 524 studies in four journals as quantitative (86%), qualitative (13%), or mixed (1%). Cohen and Macaro’s (2010) count of 419 studies in five L2 journals found the sample to be 18% descriptive, 44% exploration of relationships, and 28% experimental. Similarly and most recently, looking across 606 quantitative studies published in two L2 journals, Plonsky (2013) found that a much greater proportion of L2 research is nonexperimental than experimental in nature, and
that far more studies are carried out in labs than with intact classes. His study also found an interaction between the design orientation (experimental vs. nonexperimental) and the study context: Whereas nonexperimental research was especially common in labs, experimental studies, though less common overall, were almost twice as likely to have been carried out in a classroom environment (see similar findings in Nunan, 1991, and Plonsky & Gass, 2011). This result is perhaps surprising given the experimental control afforded to researchers working in lab contexts; it may also be indicative, however, of the inherent connection between classroom-based research and an interest in testing the effects of instructional treatments.

These and other design-related features have been studied not only in the aggregate but with respect to other study features such as time, quality (rigor), and study outcomes (i.e., effect sizes) as well. In the case of the former, for example, Gass (2009) commented that L2 researchers’ reliance on quantitative data has increased over the years. And Plonsky (2014b) showed the percentage of (quasi-)experimental studies to have increased from the 1990s (22%) to the first decade of the 2000s (33%), a finding he interpreted as a potential indicator of theoretical maturity in the field (see also Oswald & Plonsky, 2010; Plonsky & Oswald, 2014).

Other findings related to design preferences brought to light in methodological reviews have been more concerned with quality and rigor in quantitative L2 research. Much of the attention here has been on design elements associated with (quasi-)experiments, leaving a gap in our understanding of features of associational and descriptive designs. For instance, several meta-analyses have noted a lack of certain features associated with internal and external validity in experimental designs such as random assignment to experimental conditions, inclusion of control/comparison groups, pretesting, and delayed posttesting. Looking across domains, Plonsky (2013) found only about a third (37%) of experimental studies to have assigned participants to conditions randomly. This finding, however, appears to vary across different substantive areas of L2 research. Whereas 31% of the studies in Norris and Ortega’s (2000) synthesis of the effects of instruction employed random assignment, over half of the experimental studies did so in Plonsky and Gass’s (2011) sample of interactionist research.¹

Results for pretesting, another feature associated with experimental rigor and control (see Plonsky & Gurzynski-Weiss, 2014), are more encouraging and more homogeneous across domains. The percentage of experimental studies that included a pretest in their design has been found to be fairly high, ranging from approximately two-thirds to nearly all of the studies being
synthesized (Lee, Jang, & Plonsky, in press; Norris & Ortega, 2000; Plonsky, 2011, 2013; Plonsky & Gass, 2011). The exception in this case may be in computer-assisted language learning (CALL) research. Following their review of 47 CALL studies, Macaro, Handley, and Walter (2012) remarked that researchers “often claimed that progress had been measured even though no pre-testing had apparently been carried out” (p. 26).

The third feature relevant to experimental quality and control is the use of delayed posttests, which has been found to vary somewhat more across domains of L2 research. Whereas only 38% of the experimental studies in Plonsky (2013) included a delayed posttest to measure the durability of an intervention, 47% did so in Norris and Ortega (2000), 79% in Plonsky and Gass (2011), and nearly all the studies in Mackey and Goo’s (2007) meta-analysis of the effects of interaction. The use of delayed posttesting also appears to be increasing over time. As with the shift from non experimental to experimental studies reported above, we view movement toward addressing the longevity of experimental effects as a marker of the field’s theoretical and methodological maturity. Likewise, pretesting has also increased in recent decades across domains of L2 research (Plonsky, 2014b; although not in the interactionist domain according to Plonsky & Gass, 2011). Surprisingly, the use of both comparison groups and random assignment to experimental conditions has dropped from the 1990s (91% and 40%) to the 2000s (84% and 34%, respectively; Plonsky, 2014b). The drop in the latter feature may be attributed in part to the increase in pretesting, because both techniques can be used to ensure comparability between groups prior to an intervention. The increase in classroom-based research, where random assignment is more difficult or perhaps unethical in some circumstances, may also have come into play here.

In addition to changes occurring over time, some of the features discussed here have been found to be associated with study outcomes (i.e., effect sizes) with the focus, again, on features particular to (quasi-)experiments. As we might expect, several studies have found larger effects in lab- compared to classroom-based studies. Lee et al.’s (in press) meta-analysis of the effects of pronunciation instruction found a mean effect for between-groups contrasts in classroom- and lab-based studies of $d = 0.65$ and 0.95, respectively. Similar patterns have also been observed in a variety of other domains such as corrective feedback (Li, 2010; $d = 0.50$ and 1.08, respectively), and strategy instruction (Plonsky, 2011; $d = 0.43$ and 0.79, respectively). Likewise, Plonsky and Gass (2011) found much smaller effects for interactional studies that did not include vs. included a delayed posttest in their design ($d = 0.20$ and 0.90, respectively).
Each of the findings here related to interactions between methods and results have the potential to contribute to theoretical, practical, and methodological discussions and decisions in their respective domains. And although space limitations prevent us from examining additional analyses of this type, we hope to make clear two points. First, the relationships between methods and outcomes (effect sizes) are indeed present in different bodies of L2 research. Future meta-analyses and methodological reviews would do well to continue to investigate them as a means to inform subsequent research. And second, although some interactions between methods and outcomes can be explained naturally and intuitively, others—likely most—require expertise in both methodological and substantive issues relevant to the domain in question. In other words, it is not always immediately apparent why certain methodological practices would be associated with larger or smaller effects, thus necessitating researcher interpretation.

**Sampling Issues**

Another critical feature of any study, sampling, has also been analyzed across primary studies. Sample-related discussions generally center around two main issues, which we now briefly address in turn: quantity (i.e., sample size) and quality (i.e., sample representativeness and generalizability).

The first of these issues, sample size, is closely related to the likelihood that a study will detect a statistically significant effect or relationship (i.e., statistical power; Cohen, 1988). When relying on the flawed but commonly adhered-to notion of statistical significance and the associated practice of null hypothesis significance testing (NHST; see Norris, 2015; Plonsky, in press), a lack of power can lead to Type II error (false negative), thus introducing a threat to internal—and by extension, external—validity. With NHST, obtaining a sample size large enough to achieve statistical significance is therefore critical but often overlooked.

Levels of statistical power available to L2 researchers have been examined from various angles at the meta-analytic level. One approach has been to calculate post hoc power—the likelihood of observing a statistically significant effect or relationship if one truly exists in the population. Like statistical significance, effect size and sample size conjointly and inversely impact power; the larger the effect, the smaller the sample that is needed to detect a statistical relationship and vice versa. Based on their observed/meta-analytic $d$ value and mean $n$, Plonsky and Gass (2011) calculated post hoc power in interactionist L2 research to be $.56$. Similarly, Plonsky (2013) observed post hoc power of $.57$ across various domains of quantitative L2 research. More recently,
Plonsky (2014a) found post hoc power across 61 meta-analytic effects and sample sizes to range widely, from .06 to .99. These findings indicate that considerable amounts of L2 research may frequently lack the power to yield a statistically significant indication of intended effects and relationships.

Several additional factors also contribute to and ripple outward from what Plonsky (2013) referred to as the “power problem” (p. 678) in L2 research including: (a) the rarity of power analyses (Chaudron, 2001; Plonsky & Gass, 2011); (b) overuse of statistical tests in primary studies (the median number of inferential tests per study was found to be 18 in Plonsky, 2013); (c) frequently violated statistical assumptions, non-normal data, and outliers (e.g., Plonsky, Egbert, & LaFlair, in press); and (d) omission of nonstatistical results, which create an upwardly biased set of effects at the meta-analytic level and which fail to appropriately inform future theory, research, and practice.

One obvious, if perhaps oversimplified, solution to the power problem is to increase sample sizes. Other suggestions include bootstrapping the available data, conducting fewer subgroup comparisons, and employing more multivariate tests as a means to preserve experiment-wise power (e.g., Cohen, 1968; Plonsky et al., in press; Raykov & Marcoulides, 2008, chap. 1). Under the NHST approach, these are certainly useful techniques. We would argue, though, that even if the field were to adhere strictly to these recommendations, our focus would still be misguided. In other words, even if we could fix the power problem, our results would still be based on the flawed notion of statistical significance. A much more fruitful approach, as advocated by numerous methodologists across the social sciences including our own (e.g., Larson-Hall, 2010; Norris, 2015; Plonsky & Oswald, 2014), would involve focusing instead on point estimates (effect sizes) and estimation (confidence intervals) along with a synthetic mindset and the understanding that no single study can provide a conclusive answer to any question worth asking (Norris & Ortega, 2006; Plonsky, 2012, in press; Porte, 2012).

The other issue related to sampling is not concerned with the threat to internal validity posed by low statistical power resulting from small samples but, rather, with external validity or generalizability. Several scholars have argued that L2 research has not been conducted across many of the contexts and learner demographics that it seeks to generalize to. As DeKeyser, Alfi-Shabtay, and Ravid (2010) put it, “almost every sample has been one of convenience” (p. 416). If verified empirically, the natural consequence of such limited sampling is a severe lack of generalizability (Macaro et al., 2012; Tarone, 2013); for pedagogy the consequence is that practice cannot be accurately informed and thus these populations of learners cannot be best served (Ortega, 2005).
Despite these concerns, to our knowledge demographics sampled in L2 research have only been empirically examined twice, in Norris and Ortega (2000) and Plonsky (2014a). Despite a difference in the scope of these two studies, their findings are much the same. Among other results, Norris and Ortega’s and Plonsky’s reviews showed that the vast majority of the field’s samples (a) are younger adults, (b) are recruited from among university students, (c) live in the United States or in one of a small set of other countries in western Europe or east Asia, and (d) speak English as a first or second language. In other words, the findings confirm and expand on Ortega’s (2009) assertion that “most of what we know . . . pertains mostly to formal learning by (highly literate) adolescents and adults in schools and universities” (p. 145; see also Spinner, 2011). Despite what many of us may perceive to be a field of scholars with international and diverse backgrounds and interests, the demographics sampled in quantitative L2 research—and, hence, the generalizability of many of our results—are perhaps no better than those of our sister discipline, psychology, which is often criticized for its overreliance on samples drawn from participant pools of U.S. college students (Shen et al., 2011; Wintre, North, & Sugar, 2001).

In the first half of this chapter, we have reviewed findings of several methodological syntheses with a particular focus on quantitative study designs and sampling practices. We have also described the motivations and processes employed by methodological syntheses. In the remainder of the paper, we provide an example of this type of research with a focus on one particular statistical technique: Exploratory Factor Analysis (EFA).

**Use of Exploratory Factor Analysis (EFA) in L2 Research**

SLA is a relatively new and developing field. As such, researchers in this area have borrowed methodological practices from other disciplines (Selinker & Lakshmanan, 2001). In particular, the statistical techniques used in SLA “have not developed from within the field” (Loewen & Gass, 2009, p. 181) but are mostly drawn from sister disciplines such as psychology and education. Over the years, SLA researchers have used not only common inferential statistics such as t tests, chi-square, and analyses of variance but also more advanced statistical techniques, like structural equation modeling and factor analysis (Gass, 2009; Lazaraton, 2000, 2005; Loewen et al., 2014). However, perhaps due in part to its noninnovative nature in terms of statistical methods, SLA lacks field-specific standards for employing and reporting on such techniques. This lack of standards is compounded by the lack of statistical literacy among L2 researchers (Loewen et al., 2014), both of which merit urgent attention.
Although some L2 researchers have called for more rigorous L2 methods, including sound research designs and clear reporting practices (Brown, 2004; Loewen & Gass, 2009; Norris & Ortega, 2000; Plonsky, 2011, 2013; Plonsky & Gass, 2011), very little research has investigated how well L2 researchers employ statistical methods to analyze their data.

Taking Gass (2009) and Lazaraton (2005) as a starting point, we set out with a narrower focus than previous methodological reviews. That is, we sought to explore the extent to which L2 researchers used a specific family of statistical methods in accordance with standards of methodological rigor. Such a review is helpful for revealing whether L2 researchers check, for instance, the assumptions of the tests they use, the reliability of measures employed, and the soundness of their reporting practices. This section of the chapter reviews the use of a particular statistical tool, exploratory factor analysis—one of the more problematic statistical methods (Bandalos & Boehm-Kaufman, 2009)—in L2 research, and discusses how such focused methodological syntheses might contribute to the field.

Factor analysis (FA) is a series of complex structure-analyzing procedures. There are two general types of factor analysis: confirmatory factor analysis (CFA) and EFA. CFA is generally preferred when researchers have a theory that clearly specifies a precise number of factors, whereas EFA is preferred when researchers do not have any particular expectations about the number of factors involved (Fabrigar & Wegener, 2012). These factor analytic methods have been used extensively in L2 research for data reduction, hypothesis testing, instrument development, or summarizing patterns of interrelationships (e.g., Brown, 2010b; Henson, Capraro, & Capraro, 2004; Pett, Lackey, & Sullivan, 2003; Thompson, 2004). In fact, the use of FA in L2 research goes back to the 1940s, when Wittenborn and Larsen (1944) were the first L2 researchers to apply EFA to investigate the differences between high- and low-achieving German L2 students (cited in Loewen & Gass, 2009). Since then, a substantial number of L2 scholars have applied EFA in their research (Loewen & Gonulal, in press).

Frequency does not, however, imply proper use. In fact, with the exception of structural equation modeling, EFA is perhaps the most problematic inferential analysis in terms of the types and numbers of questionable practices employed by quantitative social science researchers in general (Bandalos & Boehm-Kaufman, 2009). EFA is an inherently subjective process requiring a series of researcher judgments (Henson et al., 2004), each of which may “render distinct results when certain conditions are satisfied” (Keiffer, 1999, p. 77). With these issues in mind, and given the popularity of EFA in L2 research, our synthesis
seeks to review and critically evaluate EFA practices in L2 research and to shed light on the major issues in the application of EFA. In doing so, we build on previous reviews of EFA in other fields (e.g., Conway & Huffcut, 2003; Costello & Osborne, 2005; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Henson & Roberts, 2006; Preacher & MacCallum, 2003). Note that, due to the very limited number of CFA applications in the L2 literature and the different assumptions required for CFA, the main focus of this review is on EFA.²

Major Issues in EFA

Previous reviews of EFA practices in non-L2 research have found the state of the art to be “routinely quite poor” (Fabrigar et al., 1999), pointing out a number of methodological issues EFA researchers need to consider. These issues include decisions on (a) the appropriateness of EFA, (b) the factor extraction model used, (c) the factor retention criteria, (d) the factor rotation methods, and (e) the interpretation of the factor solution. Because each of the many decisions can lead to different factor analytic results, EFA researchers should clearly express and justify their decisions (Henson et al., 2004). Following is an overview of these issues.

Factorability of the Data

The first decision that researchers confront is to determine whether the data are amenable to factor analysis. Because EFA is based on a correlation matrix, the variables should be on an interval or continuous scale and normally distributed (Field, 2009). In addition, sample size needs to be taken into consideration due to the sensitivity of correlations to outliers and non-normal distributions. Unfortunately, in the EFA literature there is no consensus on the sample size necessary to employ EFA. Tabachnick and Fidell (2013), for example, suggest that there should be at least 300 cases for EFA, whereas according to Hair, Anderson, Tathan, and Black (1995) a sample size of 100 is enough. Alternatively, recommendations regarding sample size relate to the specific number of subjects or items per variable. Although the exact number required is disputed, with estimates ranging from 3 to 20 subjects or items per variable (Gorsuch, 1983, 1990, 2003; Pett et al., 2003; Tabachnick & Fidell, 2013; Thompson, 2004), 10 to 15 is the most common suggestion (Field, 2009). Due to myriad rules of thumb, it is more sensible to consider each data set separately and then to decide whether that particular sample size is enough for the questions posed and the variables investigated. One additional and perhaps easier way to check whether the sample is big enough is to look at the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy results (available in SPSS). KMO values range
from 0 to 1, with higher values representing better sampling adequacy, though again, hard and fast criteria are not available.

Factor Extraction Model

The two major factor models pertinent to EFA include the component model PCA and the common factor model (i.e., EFA). In conceptual terms, the difference between PCA and EFA lies in how the models treat variance; PCA analyzes variance whereas EFA analyzes covariance (Tabachnick & Fidell, 2013). In other words, PCA does not differentiate between variance that is shared versus unique among variables, but EFA does. Even though PCA results may be very similar to EFA results in many cases, there are other cases in which meaningful and noteworthy differences occur. Conway and Huffcutt (2003) note that:

If a researcher’s purpose is to understand the [underlying] structure of a set of variables (which will usually be the case), then use of a common factor model [EFA] such as principal axis or maximum likelihood factoring represents a high-quality decision. If a researcher’s purpose is pure reduction of variables . . . then use of PCA represents a high-quality decision. (p. 150–151)

Previous reviews of EFA in non-L2 research have found that primary researchers often fail to report whether an EFA or PCA was used and/or that they applied an EFA that was incompatible with the goals of the study (e.g., Fabrigar et al., 1999; Conway & Huffcut, 2003). For instance, though differences in actual findings may be subtle, PCA was often used for situations in which an EFA (e.g., principal axis factoring or maximum likelihood) would have been more appropriate (Fabrigar et al., 1999).

Factor Retention Criteria

A third important decision is to determine the number of factors to retain. Options available in most statistical packages include the Kaiser-1 rule (i.e., “eigenvalues greater than one” rule), Joliffe’s criterion (i.e., eigenvalues greater than 0.7), visual inspection of a scree plot, cumulative percentage of variance for extracted factors, and parallel analysis. Different techniques sometimes render different numbers of factors to retain (Fabrigar et al., 1999; Gorsuch, 1990, 2003). The Kaiser-1 rule, also known as EV > 1, is the most frequently used criterion, probably because it is the default option in most statistical packages (e.g., SPSS). However, it tends to overestimate or underestimate the number of factors to retain, depending on the communalities (Comrey & Lee, 1992;
It is therefore advisable to use multiple factor retention criteria to extract the correct number of factors (assuming an EFA approach).

**Factor Rotation Method**

The main purpose of factor rotation is to get more interpretable solutions compared to unrotated solutions. There are two main types of rotations: *orthogonal rotations*, in which factors are assumed to be uncorrelated or independent, and *oblique rotations*, which allow for correlated factors. Because most factors related to human cognition and language learning can be assumed to be related in some way, the most appropriate choice in L2 research is generally an oblique rotation. (For further details on rotation options, see Fabrigar & Wegener, 2012; Field, 2009; and Loewen & Gonulal, in press).

**Interpretation and Reporting on EFA**

Interpretation includes examining which items or variables load on which factors, and giving each factor a name or theme based on their substantive content. This process is challenging because it is a subjective, theoretical, and inductive task (Pett et al., 2003). Henson and Roberts (2006) highlight that the meaningfulness of factors depends primarily on the researcher’s interpretation. Thus, a detailed investigation of the content of each factor is important (Comrey & Lee, 1992).

Given the many decisions and options in each step of carrying out an EFA, it is crucial for readers to be able to assess researchers’ EFA practices and results. Conway and Huffcutt (2003) and Pett et al. (2003) note that EFA researchers should report all the necessary decisions, including theoretical rationale for the use of (exploratory) factor analysis, factor extraction, retention and rotation models used, and justifications for the choice of these methods. Further, descriptive statistics, eigenvalues, communalities, percentage of variance accounted for by the extracted factors, and the full factor loading matrix (whenever possible) should be reported as well (Conway & Huffcutt, 2003; Fabrigar et al., 1999; Henson et al., 2004).

**The Present Review**

As can be seen, carrying out an EFA necessarily involves numerous decision points on the part of the researcher (Henson et al., 2004), each of which may alter the results of the analysis (Pett et al., 2003). The complex nature of EFA taken together with the current state of statistical literacy among L2 researchers (Loewen et al., 2014) motivated this synthesis, which sought to review and critically evaluate the EFA decisions and reporting practices in L2 research. Specifically, our synthesis sought to answer the following research questions:
1. To what extent have L2 researchers used FA in accordance with common standards of methodological rigor?
2. What types of statistical information have been reported with these FA studies?

Method
Selection of EFA Studies: In order to investigate the use of factor analytic applications in L2 research, journals that frequently publish EFA studies were initially identified from keyword searches (e.g., FA, PCA, common factors, and so on) in four academic databases, namely ERIC, LLBA, PsycInfo, and Google Scholar. A number of L2 journals were found to publish EFA studies, but the following five journals were selected for this review because of the higher frequency with which they publish EFA studies: Applied Linguistics, Language Learning, Modern Language Journal, TESOL Quarterly, and System. Second Language Research and Studies in Second Language Acquisition were not included in the list because these journals have published very few EFA studies. The articles in these journals were manually reviewed over a publication period of 14 years (2000 to 2013). In total, 51 articles were identified that used one or more EFAs as either a primary or secondary statistical method. It is worth noting that there are some substantive areas in the field that employ EFA more often than other areas. These include research related to motivation \( (k = 13) \), learner beliefs \( (k = 9) \), and willingness to communicate and language anxiety \( (k = 9) \).

Coding Procedure
Once collected, a coding scheme was created based on previous reviews of EFA from other fields (Conway & Huffcutt, 2003; Fabrigar et al., 1999; Henson et al., 2004; Henson & Roberts, 2006) and each study \( (K = 51) \) was coded by the second author for features and reporting practices such as factor extraction model, factor retention criteria, and factor rotation method. In addition, these studies were coded for sample/variable ratio, number of variables, number of factors retained, percentage of variance accounted for by retained factors, and so on. An additional trained rater coded a random sample of 10 studies (approximately 20% of the total). The intercoder agreement was .97 for numerical variables such as sample size, number of factors, and percentage of variance. For categorical variables such as type of rotation and factor retention criteria, both simple percentage agreement and Cohen’s Kappa were calculated. Overall intercoder agreement was found to be 89% with a kappa of .67 (SE = .08, 95% confidence intervals = .60–.75, \( p < .001 \)), which is acceptable (Landis &
Table 1 Descriptive results of exploratory factor analysis applications

<table>
<thead>
<tr>
<th>Variable</th>
<th>$k$</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Median</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>51</td>
<td>25</td>
<td>2278</td>
<td>381.8</td>
<td>252</td>
<td>457.5</td>
</tr>
<tr>
<td>Number of variables factored</td>
<td>50</td>
<td>4</td>
<td>78</td>
<td>24.9</td>
<td>24</td>
<td>14.3</td>
</tr>
<tr>
<td>Ratio of sample size to variables</td>
<td>50</td>
<td>3</td>
<td>76</td>
<td>17.7</td>
<td>12</td>
<td>16.7</td>
</tr>
<tr>
<td>Number of factors extracted</td>
<td>50</td>
<td>1</td>
<td>10</td>
<td>4.5</td>
<td>4</td>
<td>2.2</td>
</tr>
<tr>
<td>Total variance explained</td>
<td>39</td>
<td>20%</td>
<td>81%</td>
<td>60%</td>
<td>58%</td>
<td>13%</td>
</tr>
</tbody>
</table>

Note: Sample size ($k$) indicates the number of studies reporting corresponding information.

Results

Because EFA is based on correlations, sample size plays a vital role in the soundness of an EFA application. In fact, one might expect that the extremely low sample sizes used in L2 research (median = 19; see Plonsky, 2013) raises concerns regarding the possibility of employing EFA in L2 research. Nonetheless, as can be seen in Table 1, the sample size (median = 252) in this methodological synthesis is well above 19, and would be considered sufficient for EFA based on the guidelines suggested by Comrey and Lee (1992). However, as stated before, there are several rules of thumb for determining sample size discussed in the EFA literature. Proposals for minimum sample sizes range from 100 cases on the low end (e.g., Hair et al., 1995) to 500 cases on the high end (e.g., Tabachnick & Fidell, 2013). Another approach to evaluating whether a sample size is large enough is to calculate the ratio of participants to variables (Stevens, 1996). In the present sample, the median number of participants per variable is 12, which satisfies Field’s (2009) suggestion that at least 10–15 participants per variable should be included. Although the overall sample size and item-to-participant ratio meet the generally accepted criterion for adequacy in EFA, in a substantial subset of studies ($k = 10$) researchers employed factor analysis with less than 100 participants. Still more concerning, in nearly half of the studies ($k = 22, 43\%$) the ratio of participants to variables was smaller than 10.

Another important finding relates to the total percentage of variance explained by extracted factors. Field (2009) suggested that the cumulative percentage of total modeled variance should be minimally around 55–65%. In the present data, the extracted factors explained, on average, 59.8% of the total variance.
variance, which meets Field’s recommendation. However, employing EFA did not account for much variance in a number of studies ($k = 7$ explained less than 50% of the total variance) in this sample. At the low end, Asenci´on-Delaney and Collentine’s (2011) five-factor solution explained only 20% of the shared variance. However, although this cumulative percentage of total variance is only about a third of what Field suggests, it might still be considered useful information, owing to its origin in such a large corpus-based study. This finding raises the question of whether Field’s 55–65% criterion is a reasonable cut-off level in the context of distinct types of L2 research. In some cases, it might thus be more reasonable to reconsider such benchmarks, depending on the nature of the phenomena under investigation and the overall sample size being factor analyzed.

Turning now to the quality of EFA decisions in L2 research, the results reveal several patterns in the omission of information. Table 2 presents the frequencies and percentages of EFA reporting practices in L2 research. For instance, the first section of Table 2 indicates that L2 researchers have a tendency to not report checking to see whether their data were suitable for factor analysis. Although it is speculative, such a tendency may also explain the low cumulative percentage of variance found in several studies ($k = 7$, with less than 50% of the total variance). The same omission of information continues at other decision points. For instance, almost one-third ($k = 17$, 33.3%) of the studies failed to specify what factor extraction model they used. Given the fact that methodologists still debate whether PCA or EFA is preferable (Bandalos & Boehm-Kaufman, 2009), it is perhaps not surprising that L2 researchers did not explicitly state whether they used PCA or EFA. Among those reporting the factor model used ($k = 34$), slightly more than half (55.8%) used principal components analysis (PCA). This finding is consistent with previous reviews from other fields in that PCA was mostly preferred over EFA, probably because PCA is the default option in many statistical software packages (Conway & Huffcutt, 2003; Henson & Roberts, 2006; Henson et al., 2004). However, in cases in which the studies reported using a PCA, the majority used PCA to explore underlying relationships rather than reduce or consolidate variables. Such improper use of PCA may be due to confusion among researchers regarding differences between PCA and EFA, prompting many researchers to use them interchangeably. As mentioned before, though PCA and EFA are conceptually different analyses, they often produce very similar results.

Regarding decisions made on the number of factors or components to retain, $EV > 1$ was the most common method (31.4%), which is, again, the default option in most statistical packages. However, $EV > 1$ has been shown
### Table 2  Frequencies and percentages of exploratory factor analysis reporting practices

<table>
<thead>
<tr>
<th>Feature</th>
<th>$k$</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factorability of correlations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaiser-Meyer-Olkin (KMO)</td>
<td>12</td>
<td>23.5</td>
</tr>
<tr>
<td>Bartlett’s test of sphericity</td>
<td>11</td>
<td>21.6</td>
</tr>
<tr>
<td><strong>Loading magnitude</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.30-.39</td>
<td>24</td>
<td>47.1</td>
</tr>
<tr>
<td>.40-.49</td>
<td>3</td>
<td>5.9</td>
</tr>
<tr>
<td>.50 or higher</td>
<td>2</td>
<td>3.9</td>
</tr>
<tr>
<td>Not reported</td>
<td>22</td>
<td>43.1</td>
</tr>
<tr>
<td><strong>Factor extraction method</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Principal Component Analysis (PCA)</td>
<td>19</td>
<td>37.3</td>
</tr>
<tr>
<td>Principal Axis Factoring (PAF)</td>
<td>6</td>
<td>11.8</td>
</tr>
<tr>
<td>Maximum Likelihood (MLH)</td>
<td>9</td>
<td>17.6</td>
</tr>
<tr>
<td>Not reported</td>
<td>17</td>
<td>33.3</td>
</tr>
<tr>
<td><strong>Factor retention criterion</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EV &gt; 1</td>
<td>16</td>
<td>31.4</td>
</tr>
<tr>
<td>Scree plot</td>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>Multiple criteria</td>
<td>13</td>
<td>25.5</td>
</tr>
<tr>
<td>Percentage of variables explained</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>Not reported</td>
<td>19</td>
<td>37.3</td>
</tr>
<tr>
<td><strong>General rotation strategy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orthogonal</td>
<td>20</td>
<td>49.4</td>
</tr>
<tr>
<td>Oblique</td>
<td>21</td>
<td>41.0</td>
</tr>
<tr>
<td>Not reported</td>
<td>10</td>
<td>19.6</td>
</tr>
<tr>
<td><strong>Rotation type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct oblimin</td>
<td>17</td>
<td>33.3</td>
</tr>
<tr>
<td>Varimax</td>
<td>19</td>
<td>37.3</td>
</tr>
<tr>
<td>Promax</td>
<td>3</td>
<td>5.9</td>
</tr>
<tr>
<td>Mixed</td>
<td>2</td>
<td>3.9</td>
</tr>
<tr>
<td>Not reported</td>
<td>10</td>
<td>19.6</td>
</tr>
<tr>
<td><strong>Reported correlation matrix</strong></td>
<td>12</td>
<td>23.5</td>
</tr>
<tr>
<td><strong>Reported factor loading matrix</strong></td>
<td>37</td>
<td>72.5</td>
</tr>
<tr>
<td><strong>Reported EV for factors retained</strong></td>
<td>15</td>
<td>29.4</td>
</tr>
<tr>
<td><strong>Reported percentage of variance for factors retained</strong></td>
<td>25</td>
<td>49.0</td>
</tr>
<tr>
<td><strong>Reported communalities</strong></td>
<td>8</td>
<td>15.7</td>
</tr>
<tr>
<td><strong>Complex variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addressed</td>
<td>7</td>
<td>13.7</td>
</tr>
<tr>
<td>Not addressed</td>
<td>19</td>
<td>37.3</td>
</tr>
<tr>
<td>Not applicable</td>
<td>25</td>
<td>49.0</td>
</tr>
<tr>
<td>EFA only method</td>
<td>10</td>
<td>19.6</td>
</tr>
</tbody>
</table>
to be less reliable than other methods (Velicer, Eato, & Fava, 2000; Cortina, 2002) due to its arbitrary nature (e.g., an eigenvalue of 1.05 is significant but a value of .95 is not). Interestingly, the scree test and the percentage of variance explained by extracted factors were rarely used as the single criterion, but these methods were frequently combined with the EV>1 method (25.5%). Considering the complex nature of EFA, using multiple factor retention criteria in combination can help researchers extract the most appropriate number of factors. Unfortunately, in a substantial number of studies ($k = 19, 37.3\%$), researchers did not report how they determined the number of factors to retain.

A large majority (approximately 81\%) of the studies reported their specific factor rotation method. Contrary to previous reviews of EFA (Conway & Huffcutt, 2003; Henson & Roberts, 2006) in which orthogonal rotation was found to be the method of choice, the present review found an almost even split between oblique and orthogonal rotation (41\% and 39.4\%, respectively). The most popular oblique method was direct oblimin (33.3\%) whereas the most common orthogonal method was varimax (37.3\%). Only two of the 51 studies used both rotation methods in combination. Considering the highly correlated nature of many variables investigated in L2 research, we might have expected higher use of oblique rotation. Furthermore, few authors justified their choice of orthogonal rotation.

With respect to the presentation of factor analytic results, reporting practices were far from complete, perhaps due in part to limited journal space. For example, very few researchers (23.5\%) provided a correlation matrix. Another aspect of data-reporting problems is that many of these studies (70.6\%) typically omit the eigenvalues for each factor. However, Henson and Roberts (2006) suggested that the eigenvalue for at least the first factor not extracted should be reported for external evaluation. This eigenvalue becomes particularly important when the EV>1 rule is used as the only retention criterion. Another important finding relates to complex variables that load on two or more factors. Complex variables can be problematic if multiple-loading items are used as major descriptors of a factor (Pett et al., 2003). In fact, complex variables are to be expected in many EFA studies. However, only a small number of researchers (37.3\%) report whether they addressed complex variables and, if so, how. Perhaps these complex variables even went unnoticed.

**Discussion**

When viewed in its entirety, this review reveals that EFA is often applied in a relatively unsophisticated way in L2 research, parallel to previous reviews
from other fields (Conway & Huffcutt, 2003; Henson & Roberts, 2006). In a fair number of studies, L2 researchers made less than optimal decisions regarding exploratory factor analytic procedures and/or showed poor reporting practices. Almost one-third of the studies failed to specify at least one of the three methods: factor extraction method, factor retention method, and factor rotation method.

This methodological synthesis raises the question of why EFA is not more accurately applied in L2 research. There are several potential explanations. First, although the researchers who conduct EFA may be more statistically literate than the average L2 researcher, using an advanced statistical method apparently requires further statistical knowledge. Indeed, many researchers in applied linguistics do not consider themselves very competent in more advanced statistical procedures such as factor analysis (see Loewen et al., 2014). Second, closely related to the first observation is that there is very little in the way of field-specific guidance on EFA. In fact, to our knowledge, there are only a few treatments of this topic in the L2 domain. The first to our knowledge is Hatch and Lazaraton (1991). This source, like others which have appeared occasionally and which provide generally light treatments on the topic (e.g., Bachman, 2004; Richards, Ross, & Seedhouse, 2012), however, provides no guidance for using statistical packages, an issue addressed in Loewen and Gonulal (in press), a guide to PCA and EFA using SPSS. Brown’s (2009a, 2009b, 2009c, 2010a, 2010b) series of JALT publications on the use of FA in L2 research is also particularly accessible, though seldom referenced in the domains of L2 research beyond language testing. Third, the APA publication manual provides limited information regarding EFA reporting practices, hence, researchers may have little recourse to guidance when it comes to publishing and reporting their factor analytic studies. Fourth, though speculative, these conditions may prompt L2 researchers to rely on previously published studies when conducting EFA, which in turn replicates less than optimal methodological and reporting practices.

Rather than focusing solely on the deficiencies in this area, we prefer to look ahead and to consider how EFA practices might be improved. With that goal in mind, we have provided an overview of one particular use of EFA in L2 research which we hope might serve as a kind of model for future studies (see sample study). We also hope that our brief review of EFA and the issues raised here will encourage L2 researchers, reviewers, and editors to take more responsibility for improving the use of EFA in L2 research. We recommend that such standards as well as future EFA users consult Bandalos and Boehm-Kaufman (2009),
and Loewen and Gonulal (in press) for more detailed information regarding EFA practices. We would also hope to see greater attention given to EFA in the methodological training provided in graduate programs in applied linguistics.

A Sample Study

Background
Loewen et al. (2009) investigated L2 learners’ perspectives on the role of grammar instruction and error correction in the L2 classroom. Although previous studies have taken into consideration both teachers’ and students’ beliefs on this issue, learner beliefs have received less attention than teacher beliefs, even though such beliefs may influence the effectiveness of classroom instruction. More specifically, Loewen et al. attempted to answer the following research questions:

- What underlying constructs are present in L2 learners’ responses to a questionnaire regarding their beliefs about grammar instruction and error correction?
- To what extent can the underlying constructs of learners’ beliefs distinguish L2 learners studying different target languages?

Method
A questionnaire consisting of 37 Likert-scale questions (13 of them used as distractors) regarding beliefs about L2 grammar instruction and error correction was used. A total of 754 L2 learners completed the questionnaire.

Statistical Tools
An EFA was chosen because the researchers had no a priori expectations regarding the number and nature of underlying factors. The factorability of the data was checked through the KMO test (.89) and Barlett’s test of sphericity ($p < .001$). These tests and the variable-to-participant ratio (31) showed that the data set was appropriate for factor analysis. PCA was selected for the factor extraction method. The EV > 1 criterion was chosen as the basis for extracting the appropriate numbers of factors. The initial factor solution was rotated using direct oblimin. Factor loadings of .30 or above were considered significant. Further, the factor scores calculated from the EFA were used in a subsequent discriminant function analysis to determine if students studying different L2s varied in their responses to the factors.
Data Reporting Practices

The EFA produced six factors with eigenvalues greater than 1. These factors accounted for 55% of the total variance. The researchers provided detailed descriptions and interpretations of the factors. Further, the researchers reported the factor loading matrix, the percentage of variance for factors, and the communalities. However, the possibility of complex variables was not addressed. Overall, the researchers made generally optimal decisions regarding exploratory factor analytic procedures and reported their decisions and findings appropriately.

Conclusion

Synthetic research carries considerable potential not only as a means to summarize substantive findings in a given domain. As we have shown in this chapter, the same set of procedures can also be applied to describe and evaluate methodological phenomena within or across such domains. Indeed, syntheses of quantitative methods in L2 research are a new yet promising avenue by which we can detect and address problems in study design, instrumentation, statistical analyses, and so forth. By doing so, we might also be able to make pointed suggestions for improving methodological practices in the field. With this potential in mind, we suggest that all research syntheses and meta-analyses also review methodological features along with substantive features and study outcomes. In addition, future research syntheses and meta-analyses should consider possible relationships between methodological practices and outcomes, when appropriate.

In addition to describing findings from several methodological reviews, we have also provided an example of this kind of review, focusing on the use of EFA in L2 research. Finally, in order to build on and expand this line of research and, consequently improve future primary research, we would encourage researchers to conduct similar, focused reviews of other popular yet perhaps error-prone statistical techniques (e.g., multiple and logistic regression) as a means for determining the status quo and charting a course for improved practice in the future.

Final revised version accepted 12 December 2014

Notes

1 We have to be careful not to confuse random assignment of experimental conditions with random sampling of participants, which is very rare in L2 research. In fact, to our knowledge, Norris and Ortega (2000) is the only synthesis where the latter
practice was coded for, and their study observed the practice of random sampling from a known population in only one primary report.

2 EFA is used as a generic term to include both EFA and principal components analysis (PCA).

3 Other common factor models include principal axis factoring, maximum likelihood, unweighted least squares, generalized least squares, alpha factoring, and image factoring.

References


Plonsky, L. (2014a, February). Sampling, power, and generalizability in L2 research (Or, why we might as well be flipping coins). Keynote presentation at the Second Language Studies Symposium, East Lansing, MI.


