METHODOLOGICAL REFORM IN QUANTITATIVE SECOND LANGUAGE RESEARCH: EFFECT SIZES, BAYESIAN HYPOTHESIS TESTING, AND BAYESIAN ESTIMATION OF EFFECT SIZES

A Dissertation

by

REZA NOROUZIAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Michael A. De Miranda
Committee Members,	Victor L. Willson
	Mary Margaret Capraro
	Hector H. Rivera
Head of Department,	Michael A. De Miranda

August 2018

Major Subject: Curriculum and Instruction

ABSTRACT

This dissertation consists of three manuscripts. The manuscripts contribute to a budding "methodological reform" currently taking place in quantitative second-language (L2) research.

In the first manuscript, the researcher describes an empirical investigation on the application of two well-known effect size estimators, eta-squared (η^2) and partial eta-squared (η_p^2), from the previously published literature (2005 - 2015) in four premier L2 journals. These two effect size estimators express the amount of variance accounted for by one or more independent variables. However, despite their widespread reporting, often in conjunction with ANOVAs, these estimators are rarely accompanied by much in the way of interpretation. The study shows that η_p^2 values are frequently being misreported as representing η^2. The researcher interprets and discusses potential consequences related to the long-standing confusion surrounding these related but distinct estimators.

In the second manuscript, the researcher discusses a Bayesian alternative to *p*-values in t-test designs known a "*Bayes Factor*". This approach responds to pointed calls questioning why null hypothesis testing is still the go-to analytic approach in L2 research. Adopting an open-science framework, the researcher (a) re-analyzes the empirical findings of 418 L2 t-tests using the *Bayesian hypothesis testing*, and (b) compares the Bayesian results with their conventional, *null hypothesis testing* counterparts. The results show considerable differences arising in the rejections of the null hypothesis in certain cases of previously published literature. The study provides field-wide recommendations for improved use of null hypothesis testing, and introduces a free, online software package developed to promote Bayesian hypothesis testing in the field.

In the third manuscript, the researcher provides an applied, non-technical rationale for using Bayesian estimation in L2 research. Specifically, the researcher takes three steps to achieve my goal. First, the researcher compares the conceptual underpinning of the Bayesian and the Frequentist methods. Second, using real as well as carefully simulated data, the researcher introduces and applies a Bayesian method to the estimation of standardized mean difference effect size (i.e., Cohen's *d*) from t-test designs. Third, to promote the use of Bayesian estimation of Cohen's *d* effect size in L2 research, the researcher introduces a free, web-accessed, point-and-click software package as well as a suite of highly flexible R functions.

DEDICATION

To my beloved family, adoring wife, and our beautiful daughter Ryka

ACKNOWLEDGEMENTS

I am deeply grateful to the members of my committee, Drs. de Miranda, Willson, Rivera, and Capraro, for their guidance and support throughout the course of this dissertation. I especially owe a debt of thanks to my advisor, Dr. de Miranda, for all his time, support, and mentorship. Dr. de Miranda's ideas were integral to every phase of this dissertation and to the entire process leading up to it.

Many thanks are also due to Drs. Bruce Thompson, Oi-man Kwok, and William (Bill) Jefferys who invested their time and intellect in me each in a unique way. Dr. Thompson taught me that methodology is about thinking, Dr. Kwok made that multivariate, and Bill added a Bayesian element to all methods. I also thank Dr. Luke Plonsky for his steadfast support and invigorating encouragements. Luke's abundant insights were instrumental in my understanding of the current methodological needs in L2 research.

I am also thankful to Drs. Rand Wilcox, Jeff Cumming, Mike Smithson, Ken Kelley, Richard Morey, E-J. Wagenmakers, John Kruschke, and Amir Nikooienejad for their invaluable insights.

Additionally, I am indebted to my many unknown colleagues in the StackExchange community, especially Stack Overflow, and Cross Validated communities. The power of these communities to disseminate detailed knowledge on advanced programming, mathematical statistics, machine learning, and data science topics is beyond description.

Finally, I would like to extend my heartfelt gratitude to my exceptional parents, and my adoring wife who patiently accompanied me and provided me unconditional support throughout this long journey.

CONTRIBUTORS AND FUNDING SOURCES

This work was supervised by a dissertation committee consisting of Dr. de Miranda of the Department of Teaching, Learning & Culture, Dr. Capraro of the Department of Teaching, Learning & Culture, Dr. Willson of the Department of Educational Psychology, and Dr. Rivera of the Department of Educational Psychology.

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Overview

Since its inception in the latter half of 1960s, second language acquisition (SLA) has taken several turns establishing itself as a subfield of applied linguistics (see Ortega, 2013; Selinker & Lakshmanan, 2001). As with any growing field, most of these movements have been substantive in nature. For example, in early 1990s, admonitions regarding the fact that second language (L2) acquisition is deeply rooted in learner-embedded activities (see Norouzian & Eslami, 2016) laid the groundwork for the *social turn* of SLA (Block, 2003; Firth & Wagner, 1997; Gass, Lee, & Roots, 2007) opening up opportunities to probe into learners' role relationships and social identities (Lantolf, 1996). Indubitably, such reform movements have elevated the status of SLA to the point that SLA is now expected to "be of use outside the confines of the field and contribute to overall knowledge about the human capacity for language" (Ortega, 2013, p. 1).

But attribution of such characteristics as *transdisciplinarity* (i.e., the ability of SLA to go beyond its own perimeter to inform other language sciences) to SLA is revelatory of a yet broader fact. Precisely, the fact that we are a science, and as such, we follow scientific methods to find the answer to an inquiry. Our scientific practices manifest themselves in the scholarly research that we conduct and the subsequent conclusions that we draw from it. Thus, it requires no lengthy argument that research methods are integral to our identity as a science. Moreover, when there is a consensus among practicing scientists that a field reached a level of theoretical

maturity (see Ortega, 2005, 2013), it seems befitting for them to naturally engage in a serious discourse surrounding its methodological development (see Byrnes, 2013; Gass, in press). This is partly due to the fact that theories are not directly testable, no matter how elegant they might be. Through research methods, theories are always turned into testable models. In reality, then, models act as proxies for theories. A researcher tests a model, and then attributes the results to the theory on which the model is premised. In this sense, every researcher is a modeler even without being consciously aware of it. Consequently, research methods play a crucial role in the decisions made by researchers as regards designing, executing, and reporting empirical pieces of research and hence have no "ancillary status in our work" (Byrnes, 2013, p. 825).

As noted earlier, the field of SLA has traditionally focused almost exclusively on theoretical and practical issues (Selinker & Lakshmanan, 2001). In the last decade or so, however, researchers have begun to reflect on—and even examine empirically—the field's methods. In the next section, I will provide the directions of this budding *methodological reform* currently taking place in SLA research setting the stage for the three studies presented in the following chapters.

Methodological Reform in SLA Research

Recent methodological reform efforts in SLA research span a wide spectrum of topics guided by the assumption that "[p]rogress in any of the social sciences depends on sound research methods, principled data analysis, and transparent reporting practices; the field of second language acquisition (SLA) is no exception" (Plonsky & Gass, 2011, pp. 325-326). Recently, Byrnes (2013), the editor of the *Modern Language Journal*, made reference to this body of work as a "*methodological turn*" (p. 825) taking place in L2 research (also see the most recent commentray by the editor of *Studies in Second Language Acquistion*; Gass, in press).

Indeed, efforts to improve data analysis practices constitute the backbone of what has been dubbed the "*methodological turn*" in L2 research. Among other issues, L2 researchers have taken up (a) the relative value of statistical vs. practical significance (Plonsky & Oswald, 2014), (b) reporting practices and data transparency (Plonsky, 2013), (c) novel analytical approaches such as bootstrapping (LaFlair, Egbert, & Plonsky, 2015), (d) statistical literacy among researchers (Loewen et al., 2014), robust statistics (Larson-Hall, 2012b), and data visualization (Hudson, 2015).

A closer perusal of this body of methodological L2 research, however, reveals that two themes are more prominently emphasized than others. First, the importance of reporting estimates of effect size. Second, discouraging the common use of null hypothesis significance tests. These two key issues have emerged in frequent published studies reviewing the *quantity* of use of effect sizes (Plonsky, 2013), providing field-specific guidelines for interpreting effect sizes using published literature (Plonsky & Oswald, 2014), and critical reviews targeting common misuses of null hypothesis significance tests (Nassaji, 2012; Norris, 2015). Each of these two issues merits further consideration as is discussed next.

Effect Sizes

Effect sizes are certainly not new to L2 researchers. Clarion calls for the use of effect sizes in L2 research have been sounded for over a quarter of a century now (Crookes, 1991; Hatch & Lazaraton, 1991; Lazaraton, 1991). However, this *bottom-up* approach to require effect size reporting by individual L2 researchers has not been the only venue to incite change. Soon, reporting of effect sizes was required using a *top-down* approach by a group of premier L2 journals. At the time of this writing, at least eight L2 journals require effect sizes to be included in reports of quantitative L2 research: *Foreign Language Annals*, *Language Learning*, *Language*

Learning & Technology, *Language Testing*, *Modern Language Journal*, *Second Language Research*, *Studies in Second Language Acquisition*, and *TESOL Quarterly*.

Definitions of an Effect Size

It is very probable that if an L2 researcher intends to understand what an effect size is s/he will encounter one or more of the following definitions: (a) "an effect size is a statistic quantifying the extent to which sample statistics diverge from the null hypothesis" (Thompson, 2006, p. 187), (b) "an effect size measures the degree to which such a null hypothesis is wrong" (Grissom & Kim, 2012, p. 5), (c) "effect size [is] a quantitative reflection of the magnitude of some phenomenon . . . of interest" (Kelley & Preacher, 2012, p. 140), or that it is best (d) "to use the phrase 'effect size' to mean the degree to which the phenomenon is present in the population" (Cohen, 1988, pp. 9-10).

The occasional difficulty that may be faced when trying to better understand the definitions of effect size presented above is mainly methodological, and largely a function of how familiar L2 researchers are with the fact that theories are never directly tested. Indeed, these well-known definitions of effect size are immensely informed by the modeling frameworks that they are founded upon. Specifically, the definitions represent a statistical view of the world in which a single study is always assumed to work with a sample of participants randomly drawn from one or more target populations (depending on the study design). This being one of our modeling assumption, the make-up of the participants in one's study is assumed to have been determined by randomness as are the results from such a study. However, as a study garners a larger pool of participants, the make-up of the participants in the study as well as the results of it

do a better job of being reflective of the population(s) of interest. Note that because of the omnipresent element of randomness, no measure or index could ensure that we make correct inferences about the population unless we have at our disposal the data from a very large number of participants. Rather, the idea is that, randomness aside, what measure or index could show us the magnitude of the effect that might arise from the introduction of a treatment from the study at hand? Most profitably, then, effect size is a quantitative index that could be used to measure the outcome of one study or provide the basis for comparing the outcomes of a series of studies (Olejnik & Algina, 2003).

Why Effect Sizes in L2 Research

The emphasis on effect sizes in SLA research, in line with American Psychological Association's (2009) guidelines, has been mainly motivated by (a) the pointed calls to supplement information from null hypothesis significance test (NHST) results, and (b) the fact that effect sizes provide the basis for cumulative knowledge. Below, I delve deeper into both these frequently discussed issues.

Supplementing Null Hypothesis Significance Tests (NHST)

In SLA research, a number of methodological reviews criticize the fact that null hypothesis testing is "the go-to analytic approach" in the field (Norris, 2015, p. 97). Vocal advocates for effect size reporting in SLA (Larson-Hall, 2016; Norris, Ross, & Schoonen, 2015; Plonsky & Oswald, 2014) often argue that to measure the result of a study, practical significance of the results (i.e., effect sizes) should take precedence over the statistical significance of the results (i.e., *p*-values). The main argument against reliance on the result of a null hypothesis test

is that a "*p*[-value] is jointly affected by sample size and the magnitude of the relationship in question" (Plonsky & Oswald, 2014, p. 879). Using modern technology, it is easy to find out how exactly *p*-values are affected by sample size. Thanks to the recent L2 methodological research, information about both the group sample sizes (Plonsky, 2013) and the size of effects (Plonsky & Oswald, 2014) that are commonly found in L2 research is currently available. This information can help us more realistically examine the problems associated with the use of *p*-values specifically in L2 research. For this purpose, suppose that a researcher is interested in assessing the effect of Synchronous Computer-Mediated Communication (SCMC) on improving English as a Foreign Language (EFL) learners' oral proficiency (e.g., Norouzian & Eslami, 2013). Let us consider two groups of learners (i.e., control and experimental groups) for this study. First, we consider each group sample size is 20, then we increase that to 50, and finally 100. Groups of size 20 or so are believed to be the average in major domains of L2 research (Plonsky, 2013) as well as in the interactionist tradition of SLA (Plonsky & Gass, 2011), but 50 and 100 are selected above the average so that the effect of sample size on the *p*-value from such a study could be better understood. Figure 1 (to explore Figure 1 see https://github.com/izeh/j/blob/master/1.r) shows the effect of increasing the group sample size on the *p*-value and Cohen's *d* effect size when the underlying Cohen's *d* effect size for direct feedback is assumed to be .1, a relatively small underlying effect.

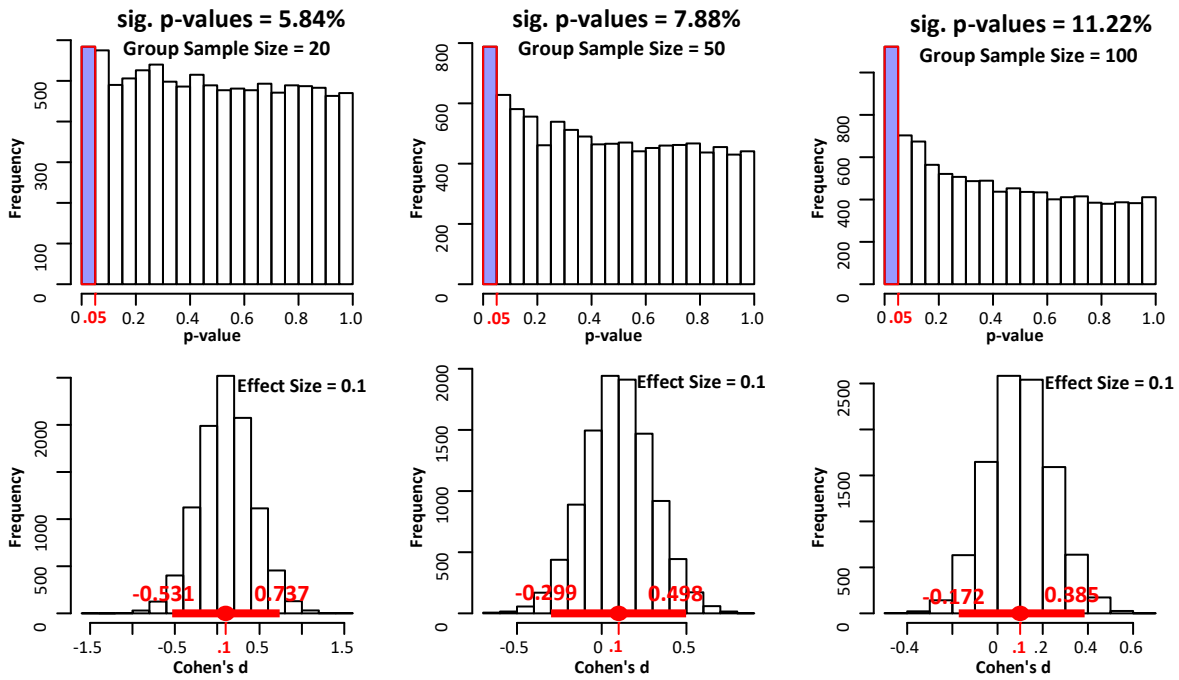

Figure 1. Results of 10,000 replications of an experimental study with two groups.

It is helpful to think of each column in Figure 1 as displaying the results (i.e., *p*-values and effect sizes) of our SCMC study conducted by 10,000 independent L2 researchers (unaware of the fact that the underlying effect of SCMC is quantified by a Cohen's *d* of .1). This way, based on the leftmost column, about 6% of these researchers who chose group sample sizes of 20 participants for their study called their study finding statistically significant ($p < .05$). However, as groups sample sizes increased to 60 (middle column), and 100 (rightmost column), the proportion of the researchers who declared their results to be statistically significant increased to about 8% and 11%, respectively. Thus, as group sample sizes increase by a factor of 5 (from 20 to 100), the likelihood of finding a significant effect almost doubles (from ~6% to ~11%). Therefore, in terms of sample size, increase of group sample sizes does lead to higher chances of

rejection when the underlying effect size is relatively small (here Cohen's *d* is .1). Noteworthy is also the fact that our 10,000 researchers have also obtained different effect sizes (the bottom row). Our researchers, however, are more likely to obtain an estimate of effect size that is closer to .1, the assumed underlying effect of SCMC, as they employ larger groups of participants in their study. To sum up, with increasing the group sample sizes, we are more likely to call any finding significant ($p < 0$). Conversely, increasing the group sample sizes leads to higher precision in our estimation of the size of the underlying effect of a phenomenon in question (here SCMC). Therefore, reliance on effect sizes is more consistent with the view that large-sized studies provide a more accurate picture regarding the validity of an L2 theory or hypothesis that is of interest to a researcher.

Basis for Cumulative Knowledge

Closely related to the discussion in the previous section is the fact that effect sizes provide a meaningful index measuring the outcome of an empirical study. Unlike *p*-values, effect sizes represent a reflection of their unknown population parameter when obtained from replication attempts. As shown in the bottom row of Figure 1, replication attempts converge to the true underlying effect of a treatment when averaged (the filled circles in Figure 1). Thus, if we intend to systematically synthesize the results of a line of research in a specific domain of inquiry, it is best to use effect sizes for this purpose. Precisely for this reason, effect sizes form the backbone of *meta-analyses*. To date, over 90+ meta-analyses (see Plonsky & Oswald, 2014) in L2 research aiming at synthesizing the results of a vast body of literature have been conducted (see Norris & Ortega, 2006). Without the use of effect sizes, meta-analysis would have been possible albeit with potential difficulty in interpreting the obtained results. But as is discussed

next in the *statement of the problem* section, the importance of effect sizes goes beyond the conventional approaches to data analysis to what are known as Bayesian methods.

Statement of The Problem

Despite the commendable efforts made to promote effect size reporting both to alleviate the problems associated with the null hypothesis testing approach, and facilitate cumulative science in SLA, a host of critical issues remain unexplored. First, I concur with Lazaraton (2009) in the belief that most methodological works in SLA have taken an "uncritical stance towards the use of statistics in SLA" and that they only "provide evidence of the increased quantity . . . of statistical SLA research" (p. 415). Specifically, the correct understanding and application of effect sizes affect the quality of SLA research and are not ensured solely by requiring its reporting (see Norouzian & Plonsky, in press). This being understood, what seems to be needed for such methodological work on effect sizes is to provide "evidence that SLA researchers are, in fact, using statistical procedures more APPROPRIATELY" (Lazaraton, 2009, p. 415, emphasis in original). Indeed, in an era of point-and-click analyses (see Mizumoto & Plonsky, 2015), choices regarding effect sizes and other statistical results may be made based on program defaults rather than on an accurate understanding of the data. This is particularly likely to occur in the case of effect sizes, which, despite their presence in published L2 research, are not generally accompanied by much in the way of interpretation. Thus, an empirical investigation focusing on widely reported effect size estimates especially in extended designs (i.e., designs with more than two groups) should respond to the concerns raised over the lack of studies examining the accuracy of application of statistical concepts (here effect size estimators) in published reports of SLA research (Lazaraton, 2005, 2009).

Second, the role of hypothesis testing, as a prevailing inference- and decision-making approach (Kruschke & Liddell, 2017), has solely been considered (e.g., Larson-Hall, 2016; Norris, 2015) through the lens of null hypothesis significance testing (NHST). Recent advances in applied statistics (Johnson, 2016; Morey, Romeijn, & Rouder, 2016) as well as new recommendations offered by learned societies (e.g., American Statistical Association, 2016), however, emphasize the use of Bayesian model selection (i.e., Bayesian hypothesis testing) culminating in a new evidence-quantifying index called "*Bayes Factor*" replacing the widely-criticized *p*-values. To the best of my knowledge, no previous methodological L2 research has either examined the use of or applied this Bayesian alternative to study the possible differences in the inferential conclusions between the conventional null hypothesis significance test results and those of the Bayesian hypothesis testing approach. Such a more recent approach to inference and decision-making is additionally supported by the most recent inferential framework known as "*new statistics*" (Kruschke & Liddell, 2017).

Third, Loewen et al. (2014) recently conducted a survey measuring the *statistical literacy* of a sizable number of practicing applied linguists (n = 331) in the field. In their questionnaire (see Appendix B in Loewen et al., 2014), the statistical terms "Bayes" or "Bayesian" was not even included. This is indicative of the fact that statistical knowledge regarding Bayesian thinking in the field is considerably low and its novelty is not yet fully appreciated. Furthermore, as it stands, the burgeoning yet multidisciplinary literature on Bayesian methods contributed to by mathematical psychology (Kruschke, 2015), cognitive science (Etz & Vandekerckhove, in press), and mathematical statistics (Gönen, Johnson, Lu, & Westfall, 2005) is highly impenetrable making it unfit for use by practicing L2 researchers. This is while the Bayesian estimation methods especially those using effect sizes could be highly beneficial to a field like

L2 research which is known to suffer from "impoverished samples sizes" (Norris et al., 2015, p. 1). Additionally, to promote the use of novel statistical methods such as Bayesian methods, L2 researchers often allude to the lack of easy to use and access statistical tools (see Mizumoto & Plonsky, 2015). Thus, to raise the statistical literacy of L2 researchers (Gonulal, Loewen, & Plonsky, 2017), a non-technical and applied resource providing a rationale for the use of Bayesian estimation and to promote the use of the method user-friendly software packages that enable the use of Bayesian estimation are critically warranted.

In response to these three gaps, this dissertation consisting of three manuscripts, whose overview and extended summaries appear in the following chapters, is summarized in Figure 2.

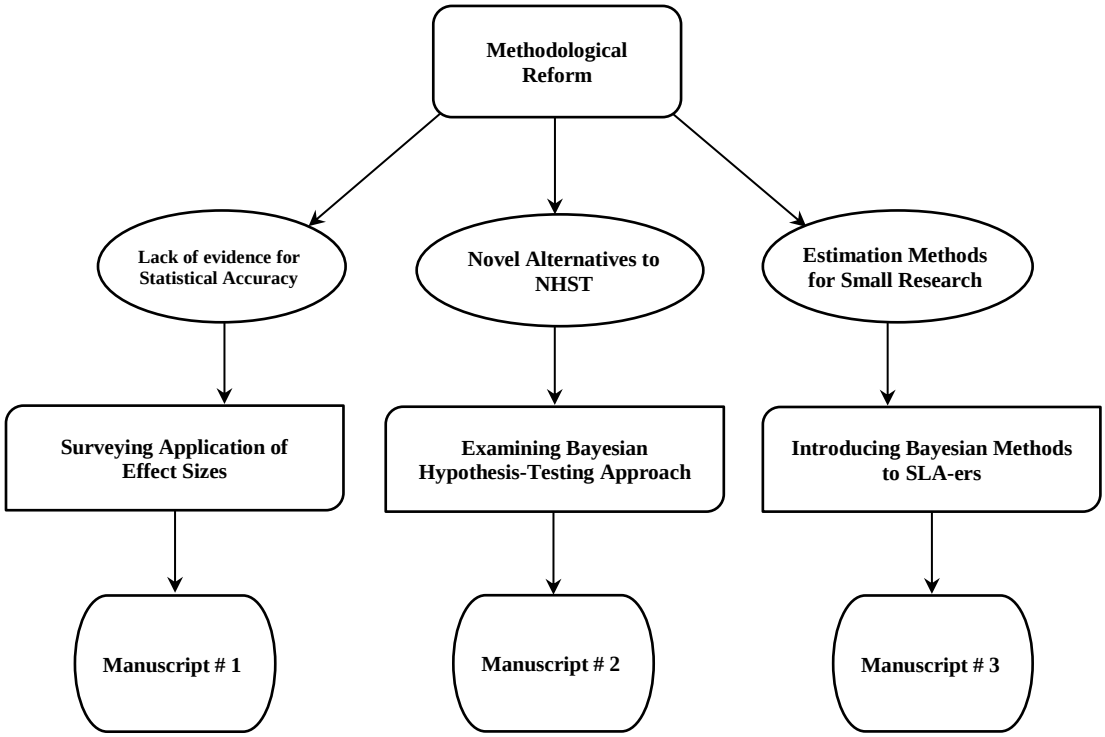

Figure 2. Schematic framework for the dissertation studies.

CHAPTER II

AN EMPIRICAL INVESTIGATION OF TWO EFFECT SIZE MEASURES IN L2 RESEARCH: RESOLVING A LONG-STANDING CONFUSION*

Overview

Eta-squared (η^2) and partial eta-squared (η_p^2) are effect sizes that express the amount of variance accounted for by one or more independent variables. These indices are generally used in conjunction with ANOVA, the most commonly used statistical test in second language (L2) research (Plonsky, 2013). Consequently, it is critical that these effect sizes are applied and interpreted appropriately. The present study will examine the use of these two effect sizes in L2 research. We begin by outlining the statistical and conceptual foundation of and distinction between η^2 and η_p^2. We then review the use of these indices in a sample of published L2 research (N = 156). The study will empirically show the possible instances of η_p^2 values being misreported as representing η2 in four well known L2 Journals. Additionally, the study will interpret and discuss potential causes and consequences related to the long-standing confusion surrounding these related but distinct estimators. Within the context of reform efforts in quantitative L2 research, the current study seeks to respond to the recent, pointed calls for improving study quality (Plonsky, 2014) and statistical literacy (Loewen et al., 2014) in the field.

Introduction

It has been almost three decades since Cohen (1988) wisely noted that "a moment's thought suggests that it [effect size] is, after all, what science is all about" (p. 532). With this

* Part of this chapter has appeared as "Eta-and partial eta-squared in L2 Research: A cautionary review and guide to more appropriate usage" by Norouzian, R. & Plonsky, L. 2018. *Second Language Research,* 34(2), 257 – 271, Copyright 2018 by Authors.

position in mind, some have gone as far as to argue that failing to appropriately report estimates of effect sizes amounts to "a kind of withholding of evidence" (Grissom & Kim, 2012, p. 9). In the case of L2 research, however, effect sizes are still a relatively novel concept. Historically, the field has relied very heavily on statistical significance and null hypothesis significance testing (*p* values) (see Norris, 2015; Plonsky 2015). It is only in the last decade or so that we have seen a shift in favor of effect sizes and practical significance, which can be attributed both to influential advocates (e.g., Norris & Ortega, 2000; Plonsky & Oswald, 2014) and journal editors. We know of at least eight L2 journals that now require effect sizes to be included in reports of quantitative research: *Foreign Language Annals*, *Language Learning*, *Language Learning & Technology*, *Language Testing*, *Modern Language Journal*, *Second Language Research*, *Studies in Second Language Acquisition*, and *TESOL Quarterly*.

Two of the most commonly employed effect sizes are eta-squared (η^2), and partial eta-squared (η_p^2), which are used in conjunction with ANOVA and its variants. We have chosen, therefore, to examine these two effect sizes in terms of how they are reported and interpreted in L2 research. We are concerned that, in an era of point-and-click analyses (see discussion in Mizumoto & Plonsky, 2016), choices regarding effect sizes and other statistical results may be made based on program defaults rather than on an accurate understanding of the data. This is particularly likely to occur in the case of effect sizes, which, despite their presence in published L2 research, are not generally well understood. The result of an overreliance on statistical packages together with the relative lack of detailed knowledge about effect sizes carries the risk of erroneous reporting, mislabeling, and faulty interpretations.

The present study builds on the momentum surrounding methodological reform in applied linguistics, including concerns expressed in recent years over, for example, study quality

(Plonsky, 2013, 2014) and statistical literacy (Loewen et al., 2014). Such discourse responds to repeated calls for examining how "APPROPRIATELY" (e.g., Lazaraton, 2009, p. 415, emphasis in original) different statistical concepts are being employed.

Eta Squared and Partial Eta Squared In ANOVA Models

Although L2 researchers often report effect sizes such as eta-squared (η^2), such values are rarely accompanied by much in the way of interpretation (Plonsky & Oswald, 2014). One reason for this is that there appears to be a good deal of confusion surrounding the terminology of "*proportion of variance*" (Grissom & Kim, 2012, p. 181) effect sizes. Therefore, for the purposes of clarity, some conceptual explanation of what these indices express is warranted.

Consider, for example, a study wherein the researcher was interested in analyzing the effect of an experimental treatment across conditions (e.g., condition 1, condition 2, control). Conceptually, the focus of analysis is on group differences as regards the dependent variable (e.g., usually a measure of L2 knowledge or learning). Statistically speaking, scores on the dependent variable scores contain the amount and source of variance caused by treatment effects (Thompson, 2006).

Proportion of variance effect sizes in the η^2 family partition the amount of total variation in the dependent variable (e.g., knowledge as measured) to determine how much of the variation is separately accounted for or explained by each independent variable (i.e., explained sum of squares or SOS). Also taken into account by the η^2 family is how much of the DV variation is left unexplained (i.e., unexplained or error SOS). Thus, total variation in the DV can be described in terms of explained and unexplained variance.

We suspect that, despite the now-widespread reporting of eta-squared (η^2), many L2 researchers may not be aware of the differences among its variants, most notably *partial* eta-

squared (η_p^2). Further complicating this matter is the mislabeling of η^2 and η_p^2 by certain early versions of SPSS, the most frequently used statistical software package in L2 research (Loewen et al., 2014). The likelihood of this error in the context of L2 research is supported by evidence presented in other fields. Levine and Hullett (2002) and Pierce, Block, and Aguinis (2004) found widespread misreporting and misinterpretation of η^2 and η_p^2 in published studies in communication and psychology, respectively. Both studies also cite the mislabeling of η_p^2 values as η^2 in early versions of SPSS.

In the two sections that follow, we provide a brief overview of these two important effect size indices. We also illustrate the points being made with heuristic examples.

Classical Eta-Squared

Imagine an intervention study in which four treatment conditions are compared on a single dependent variable. To examine the relationship of interest here, we would likely use a one-way ANOVA. The effect size in this case, η^2 (also called the squared correlation ratio), is computed using Kerlinger's (1964) classical formula (p. 203) as:

$$\eta^2 = \frac{SOS_A}{SOS_{TOTAL}}. \quad (1)$$

Note that in one-way designs, there is only one independent source (SOS_A; treatment) of variance to explain some portion of the total variation in the dependent variable (SOS_{TOTAL}; L2 knowledge). The numerator of the effect size estimator then represents variability that is attributable to the only independent variable we have (e.g., treatment condition). Therefore, an η^2 of, say, 0.35 (or 35%), indicates that we can account for 35% of the total variation in L2 knowledge as measured. The rest of the total SOS remains unexplained (i.e., SOS_{Error} = 65%), and may be due to individual differences, measurement error, or any number of other factors.

Partial Eta-Squared

One-way designs can certainly be found in L2 research. However, designs with multiple independent variables are likely much more common due to the multivariate nature of L2 learning, knowledge, use, and so forth (Brown, 2015). In such cases, the conceptual approach embodied by η^2 can be extended to apply to multi-way or factorial ANOVA. However, we now may have multiple sources of independent effects leading to a distinction that must be drawn between the *classical* η^2 and *partial* η^2 (Bakeman, 2005; Richardson, 2011).

Building on the example from above, imagine a 3×4 design in which *proficiency level* (with 3 levels) and treatment condition (with 4 levels) are jointly examined to explain the variation in learners' (N = 120) scores on a subsequent grammar test (i.e., dependent variable). Table 1 presents the hypothetical results for this two-way design.

Table 1
Hypothetical results of a fixed-effects 3×4 ANOVA (N = 120)

Source	*SOS*	*df*	*MS*	$F_{obtained}$	$p_{obtained}$	η^2	η_p^2	Inflation%
Treatment	80	3	26.67	57.60	2.57E-22	**0.39**	**0.62**	57.69%
Proficiency	70	2	35.00	75.60	2.94E-21	**0.34**	**0.58**	70.83%
Treat.×Prof.	5	6	0.83	1.80	0.106	**0.02**	**0.09**	272.73%
Error	50	108	0.46					
Total	205	119	1.72					

Note. Eta-squared values and their corresponding partial eta-squared values appear in bold.
Treat. = Treatment; Prof. = Proficiency.
Inflation% = $(\eta_p^2 - \eta^2) / \eta^2 \times 100$, this shows how different η_p^2 and η^2 can be in this two-way design.

If we want to quantify any of the independent variables' contributions to the variation observed in post-test scores, we can do so by invoking the *classical* η^2 in each case. But a different form of η^2 may be computed as well. Cohen (1965) implicitly introduced a new variant of η^2 (now often denoted by η_p^2) in multi-way designs which was similar to the classical η^2 formula with "other

nonerror sources of variance being partialled out [from the denominator]" (p. 105). Later, Cohen (1973) emphasized that this new variant is distinct from the classical η^2 and may be called "*partial* η^2" (p. 108 italics in original). Thus, in multi-way designs, the term *partial* refers to removing all other possible sources of effect in the design except the one of interest in the denominator of equation (1) and the error/unexplained variance.

In our two-way design, which includes two main effects and one interaction effect, partial eta-squared (η_p^2) for treatment condition (A) can be computed as:

$$\eta_p^2 = \frac{SOS_A}{SOS_A + SOS_{Error}}. \quad (2)$$

Thus, $\eta_p^2 = 80 / (80 + 50) = .62$ [90% CI: .494, .659]. Likewise, η_p^2 for the effect of proficiency level (SOS_B) can be computed in a similar fashion with other independent sources (i.e., treatment, and the treatment $\times$proficiency interaction) removed from the denominator:

$$\eta_p^2 = \frac{SOS_B}{SOS_B + SOS_{Error}}. \quad (3)$$

Therefore, for *proficiency*, $\eta_p^2 = 70 / (70 + 50) = .58$ [90% CI: .458, .632]. And for the interaction effect (SOS_{A*B}), we will have:

$$\eta_p^2 = \frac{SOS_{A*B}}{SOS_{A*B} + SOS_{Error}}. \quad (4)$$

Thus, regarding the interaction effect, $\eta_p^2 = 5 / (5 + 50) = .09$ [90% CI: .000, .133]. Note that because in one-way designs there is only one source of effect, no difference in the denominator of the *classical* and *partial* eta-squared formulas exists. In other words, because there are no other effects to be partialled out, eta-squared and partial eta squared are identical in one-way designs. However, as shown in Table 1, for our two-way design, η_p^2 values are invariably

larger—often much larger—than their η^2 counterparts. This occurs because the *partial* η^2 formula is partialling out the other nonerror terms (i.e., proficiency: SOS_B and proficiency×treatment: SOS_{A*B}) from the denominator for each effect, thus augmenting the outcome (see Grissom & Kim, 2012; Pedhazur, 1997). It is therefore critical that care be taken to report and interpret these indices appropriately.

Assumptions and Rationale of the Study

Having laid out the conceptual and statistical reasoning behind η^2 and η_p^2, in the present study, we seek to examine the use and interpretation of these two indices. The study is motivated by several factors that, in coordination, may create conditions that are counterproductive for the field's progress. First, although effect sizes are regularly reported, they are not often interpreted and even less often are they interpreted meaningfully (Plonsky & Oswald, 2014). Second, ANOVA designs are exceedingly common and therefore highly influential in L2 research. The family of effect sizes for this set of techniques is particularly prone to error, however, due to very similar and often ambiguous or even misleading labels, as described in the previous section. This problem, observed in other social sciences, is only compounded by a lack of general statistical literacy in the field (Loewen et al., 2014). With these issues in mind, we anticipate that erroneous reporting of these frequently used effect sizes is likely to occur in L2 research. Therefore, in this study we examine the use of η^2 and η_p^2 as a means to improve future research practices in the field. With these concerns in mind, the present study sought to answer the following question: To what extent does published L2 research demonstrate erroneous reporting of η_p^2 as representing η^2?

Method

In this section, methods used to select L2 journals, criteria for choosing individual L2 studies are discussed. Also explained are the procedures and analyses followed.

Journal Selection and Search Criteria

In order to collect a representative sample of L2 research, we first consulted previous surveys of L2 research practices (e.g., Egbert, 2007; Gass, 2009; Lazaraton, 2005; Plonsky, 2013) as well as L2 research methods textbooks providing various L2 journals' descriptions (Perry, 2011) and other documents discussing L2 journals (VanPatten & Williams, 2002). There is, of course, no consensus as to which journals are most prominent or influential in the field. In the end, we decided to survey the following five journals: *Applied Linguistics, Language Learning, Language Teaching Research, Modern Language Journal,* and *System.* This sample is by no means exhaustive, but we would argue that it does provide generally representative view of quantitative L2 research.

In order to gain a current view of this domain, we limited our search to studies published from 2005 to 2015. In line with previous reviews (e.g., Gass, 2009), we excluded from consideration forums, short reports, commentaries, review articles, and book reviews. We then examined all studies that included variants of multi-way ANOVA (repeated measures, factorial, ANCOVA- henceforth, multi-way ANOVA studies). The total sample consisted of 156 studies. Our goal to include multi-way designs was because, as discussed in the previous section, in these studies η^2 and η_p^2 lead to different results. Thus in these designs, mistakenly reporting η_p^2 as η^2 presents a distorted view of the results. Figure 3 shows the distribution of the sampled studies across the period 2005 through 2015.

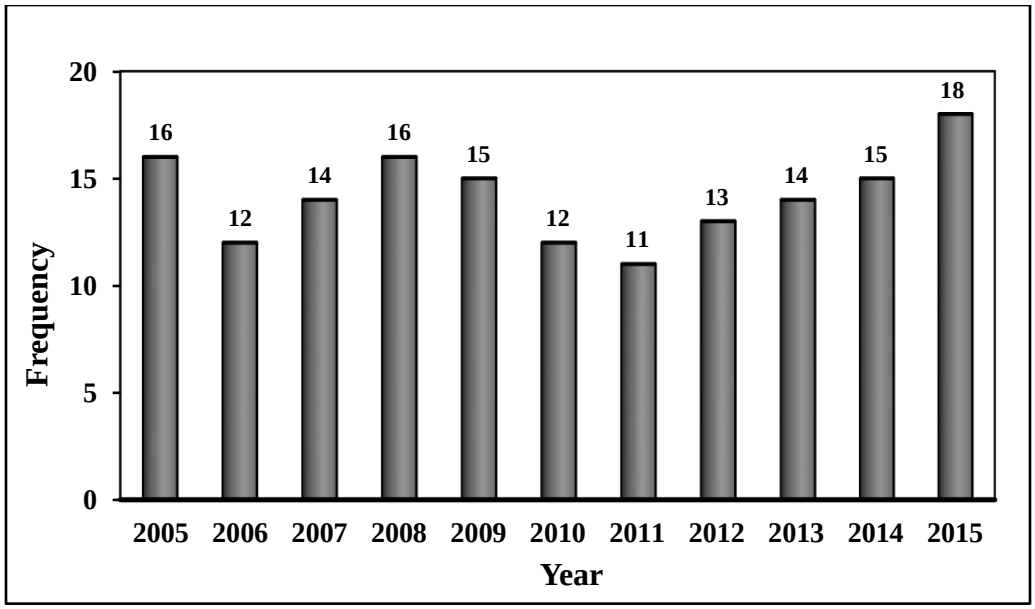

Figure 3. Distribution of multi-way ANOVA studies over time.

Procedures and Analyses

In order to address our research question, following best practices in synthetic research (Plonsky & Oswald, 2015), each study in the sample was systematically coded for the design type (repeated measures, factorial, ANCOVA), model (fixed-, random-, mixed-effects), and sampling unit distribution (balanced, unbalanced) applied. We also extracted from each study *F* values, degrees of freedom, and descriptive statistics (Mean, SD). We then conducted secondary analyses by using any or a combination of the following three methods, as appropriate.

First, in line with previous studies that have examined the reporting and interpreting of η^2 effect sizes (Levine & Hullett, 2002; Pierce et al., 2004), we computed the sum of the η^2 values for every multi-way design in papers that reported them (i.e., $\sum\eta^2_{max}$ limit check). When the sum for a multi-way design exceeded 1 (or equivalently 100%), the values were assumed to be representing instances of η_p^2 labeled erroneously as η^2. This method was applied to all 156 multi-

way studies we collected. Using this technique, we found 17 studies with this type of erroneous reporting.

Second, we applied Cohen's (1973) *partial* eta-squared meta-analytic equation which is computed as:

$$\eta_p^2 = \frac{df_A(F_A)}{df_A(F_A) + df_{Error}}. \quad (5)$$

Equation 5 was used to evaluate if the values reported and labeled as η^2 in reality were η_p^2. Being "purely algebraic [i.e., insensitive to the design type and model]" (Cohen, 1973: 107), this equation was applied to all designs types (e.g., repeated measures, ANCOVA) and models of multi-way analysis (i.e., fixed-, mixed-, and random-effects). If the answer from the manual calculations matched within rounding error those in the primary published report, we concluded that η_p^2 values were mistakenly presented as η^2. Also, when possible (i.e., when the design was fixed-effects with all relevant error and effect terms reported), we used Haase's (1983) meta-analytic equation which for a two-way design is computed as:

$$\eta^2 = \frac{df_A(F_A)}{df_A(F_A) + df_B(F_B) + df_{A*B}(F_{A*B}) + df_{Error}}. \quad (6)$$

Equation 6 was used to correctly compute η^2 values in fixed-effects multi-way designs. For the second method, when no match was found between our calculation of η^2 or η_p^2 and those reported in the original paper, the analysis in question was excluded from our study. The second method was also uniformly applied to all 156 multi-way studies and resulted in the identification of an additional 12 studies with erroneous reporting of η^2, which also confirmed and extended the results of method 1.

A third method for identifying erroneous reporting was applied when full summary tables (i.e., with all sum of squares, *df*s, *F* values made available) were reported. Using these data, we separately computed the η^2 and η_p^2 values to compare them with the values appearing in the original published studies. The third method led to the identification of additional 5 studies which had inaccurately presented η_p^2 effect size as representing η^2.

Results

The present study was intended to determine the extent of erroneous reporting of η_p^2 as representing η^2 in quantitative L2 research published between 2005 and 2015.

Previous studies in other fields (Levine & Hullett, 2002; Pierce et al., 2004) were only able to show that η_p^2 values were mistakenly reported as η^2 if the sum of η^2 values in a multi-way design exceeded 1 or 100% (method 1). As noted in the Method, we have sought to gain a more comprehensive view of this practice by employing additional equations (method 2) and in some cases directly computing η_p^2 and η^2 effect sizes from summary tables (method 3).

Table 2 presents the sum of η^2 values (i.e., $\sum\eta^2$; method 1) for the studies in our sample along with relevant data retrieved from these studies, and other features specific to the methods used for our secondary calculations. All 34 studies in Table 2, which we have anonymized, have incorrectly reported η_p^2 as representing η^2 values.

Table 2

Summary of 34 studies erroneously reporting η_p^2 as representing η^2

No.	Study ID (Year)	Design	Statistics Reported	Software used	Main effect η^2 values reported	Interaction effect η^2 values reported	$\Sigma\eta^2$ values Computed %
1-	Study 87 (2011) *Main Analysis*	2*2 ANOVA	*F, Dfs, SD, Mean*	n.r.	n.r.	.12	n.a. [1]
2-	Study 101 (2012) *Main Analysis*	2*3 ANCOVA	*Full Summary Table*	SPSS version 18	.452 .17	n.r.	n.a.
3-	Study 147 (2015) *Main Analysis*	4*4 RM [2] ANOVA	*F, Dfs, SD, Mean*	n.r.	.93	.90	183%
4-	Study 39 (2007) *Analysis 1*	2*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.754	.193	n.a.
5-	Study 43 (2007) *Analysis 2*	2*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.820 .123	n.r.	n.a.
6-	Study 41 (2007) *Analysis 3*	2*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.831 .654	n.r.	148.5%
7-	Study 127 (2014) *Analysis 1*	2*3 RM ANOVA	*F, Dfs, SD, Mean*	SPSS version 16.0	.73 .40	.50	163%
8-	Study 128 (2014) *Analysis 2*	2*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.72 .29	.54	155%
9-	Study 14 (2005) *Analysis 1*	2*3 RM ANOVA	*Full Summary Table*	n.r.	.29 .66	.59	154%
10-	Study 16 (2005) *Analysis 2*	2*3 RM ANOVA	*Full Summary Table*	n.r.	.30 .46	.56	132%
11-	Study 11 (2009) *Main Analysis*	2*2 ANOVA	*Full Summary Table*	SPSS	.06 .24	.00	n.a.
12-	Study 9 (2005) *Main Analysis*	2*6 ANOVA	*Full Summary Table*	n.r.	.013 .105	.009	n.a.
13-	Study 31 (2007) *Main Analysis*	One-Way [3] ANCOVA	*F, Dfs, SD, Mean*	n.r.	.69 .52	.12	133%

Table 2 (continued)

14-	Study 114 (2013) *Analysis 1*	2*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.63 .54	.62	179%
15-	Study 117 (2013) *Analysis 2*	2*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.67 .55	.52	174%
16-	Study 126 (2014) *Experiment 3*	2*2 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.19 .37	n.r.	n.a.
17-	Study 41 (2007) *Main Analysis*	2*3 ANOVA	*F, Dfs, SD, Mean*	n.r.	.28 .10	.06	n.a.
18-	Study 30 (2007) *Main Analysis*	2*3 ANOVA	*F, Dfs, SD, Mean*	n.r.	.048 .024	n.r.	n.a.
19-	Study 21 (2006) *Analysis 1*	3*2 RM ANOVA	*F, Dfs*	n.r.	.99	.76	175%
20-	Study 18 (2006) *Analysis 2*	2*2 RM ANOVA	*F, Dfs*	n.r.	.68 .43	.40	151%
21-	Study 23 (2006) *Analysis 3*	2*3 RM ANOVA	*F, Dfs*	n.r.	.97	.38	135%
22-	Study 12 (2005) *Experiment 1*	3*4 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	n.r.	.373	n.a.
23-	Study 15 (2005) *Main Analysis*	4*5 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	n.r.	.396	n.a.
24-	Study 47 (2008) *Main Analysis*	2*2 ANOVA	*F, Dfs, SD, Mean*	n.r.	.620 .473	n.r.	109.3%
25-	Study 51 (2008) *Main Analysis*	2*2*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.332 .803	n.r.	113.5%
26-	Study 56 (2008) *Main Analysis*	2*2 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.448 .798	n.r.	124.6%
27-	Study 44 (2008) *Main Analysis*	2*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.31 .21	n.r.	n.a.
28-	Study 41 (2008) *Main Analysis*	2*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.25 .05	.13	n.a.
29-	Study 39 (2008) *Main Analysis*	2*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.13	.08	n.a.
30-	Study 76 (2010) *Main Analysis*	2*2 ANCOVA	*F, Dfs, SD, Mean*	SPSS GLM	.30 .44 .34	.01	109%

Table 2 (continued)

31-	Study 84 (2010) *Main Analysis*	2*4 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.62	n.r	n.a.
32-	Study 137 (2014) *Main Analysis*	3*3 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.346	.244	n.a.
33-	Study 152 (2015) *Main Analysis*	2*2*2 RM ANOVA	*F, Dfs, SD, Mean*	n.r.	.15 .13 .01	.01 .13 .03	n.a.
34-	Study 143 (2015) *Main Analysis*	4*2 ANOVA	*F, Dfs, SD, Mean*	n.r.	.97	.83	180%

Note. First seven studies had a balanced design. "n.r." = not reported. "n.a." = not applicable.
[1] Not applicable: either summary table was presented (method 3) or equation's 5 outcome matched that in the original report (method 2).
[2] RM = Repeated measures.
[3] One-way ANCOVA's summary table terms are algebraically similar to those of two-way ANOVA.

As can be seen in Table 2, mistakenly reporting partial eta squared as representing eta squared is not uncommon in published quantitative L2 research. More precisely, this error occurred in 34 of the 156 studies in our sample, or 22%. Figure 4 shows the breakdown of the misreported studies in the 156 multi-way ANOVA studies published between 2005 and 2015. In Figure 4 the proportion of studies which misreported eta squared to the total multi-way ANOVAs in each year is represented by the cross-hatched columns. One important observation is that inaccurately presenting partial eta squared as representing eta squared is still present in recent L2 research. This might be due to that fact that multi-way ANOVA and its variants are frequently and increasingly employed to answer different substantive questions in L2 research (Plonsky, 2014).

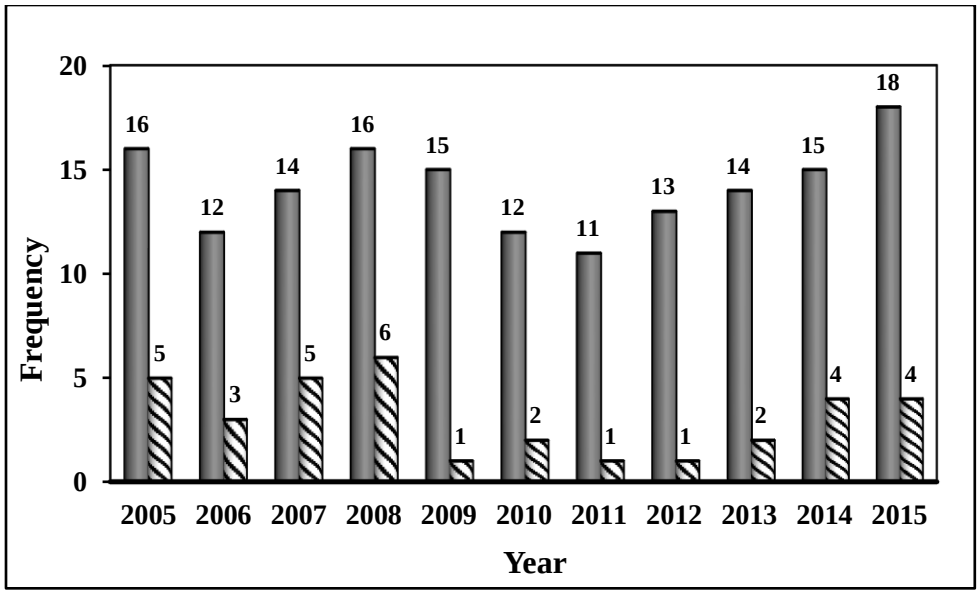

Figure 4. Multi-way ANOVA studies that presented partial eta squared as representing eta squared (crosshatched bars).

At this point, assuming that L2 researchers did not largely learn how to use these two variants of effect size from each other (i.e., independence of observation), the question of interest is: what is the actual proportion of this erroneous reporting and how prevalent it is across all L2 studies that use these two measures of effect size to report their findings? Since this was the first survey of this type in L2 research, no specific prior knowledge in L2 research is available to refer to as a knowledge base. Thus, with a very broad prior (i.e., Beta[1.2, 1.2]), the Bayesian estimation of the actual proportion of this erroneous reporting of the two effect size estimators can be shown to be around 15.90% - 28.74%. The result of this Bayesian estimation is also shown graphically in Figure 5.

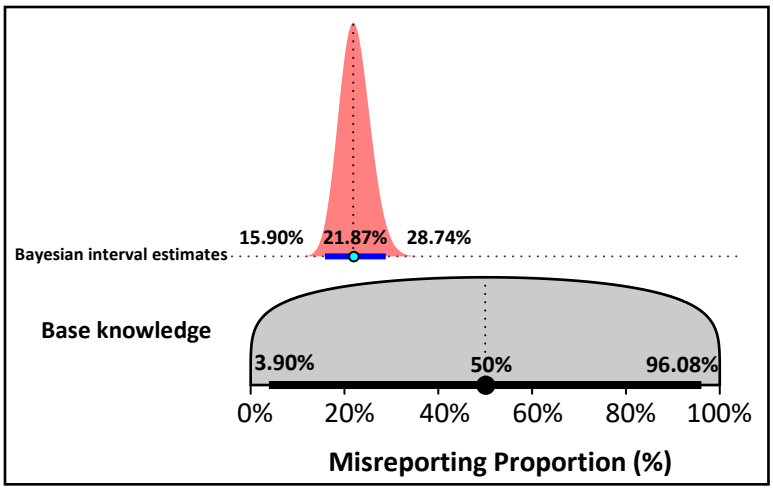

Figure 5. Bayesian estimation of the proportion of misreported L2 studies.

Discussion

The confusion between η^2 and η_p^2, found in the present study to be widespread in quantitative L2 research, can lead to a series of, at least, four problems which affect interpretations of findings to varying degrees. Some actual examples here may convey the general tenor of these problems.

First, while reporting η_p^2 values in place of η^2 values does not change the rank ordering of effects within a single study (see Table 1 for example), η_p^2 and η^2 values use different denominators in their formulas. Put succinctly, the base of each η_p^2 value differs in nature from another one in the *same* design because η_p^2 values do not share a common base (i.e., denominator). Therefore, cross-effect comparison of η_p^2 values is not meaningful (Olejnik & Algina, 2000; Pedhazur, 1997). Interpretations are especially problematic, however, when η_p^2 values (often expressed in percentages), either correctly labeled as η_p^2 or erroneously presented

as η^2, are used to indicate that they have explained a certain amount of *total* variation in the dependent variable as exemplified in Table 3.

Table 3
Misinterpreting η_p^2 as proportion of total variance

Study ID	η_p^2 mislabeled as η^2	Author(s) Interpretation
Study 21 (2006)	7% 69% 19%	In all cases, η_p^2 values taken to account for total variation
Study 47 (2008)	.177	The effect of interaction explained 17.7% of variance in the dependent variable
Study 15 (2005)	.105	about 11% of variability explained in the dependent variable

Second, η^2 values are often upwardly biased (an issue not discussed here, but see Grissom & Kim 2012) and particularly so when the effects are based on small samples, which is often the case in L2 research (Plonsky, 2013). Therefore, erroneously reporting η_p^2 as η^2 in a multi-way study can inflate an already-biased η^2 effect size even further. Pedhazur (1997) warned that "[b]ecause partial η^2 tends to be larger than η^2, I am afraid that novices will be [more] inclined to use it" (p. 509). Thus, it is critical not to look at the effect sizes in a single study "from a bigger-is-better standpoint" (Bakeman, 2005, p. 380). We were able to estimate the inflation due to mistaken reporting in some of the studies by using summary table information or by applying equation 5 for η_p^2 and equation 6, when applicable, for η^2. The inflation percentage may vary from one study to the next. Table 4 presents some examples along with the amount of inflation observed.

Table 4
Inflation resulting from mistakenly presenting η_p^2 for η^2

Study ID	Effect	η_p^2 mislabeled as η^2	η^2	Inflation% [a]
Study 101 (2012)	main	.452 [b]	.127	72%
Study 13 (2005)	interaction	.59	.32	46%
Study 7 (2005)	main	.46	.27	41%
Study 64 (2009)	main	.06	.04	33%

[a] Inflation% = $(\eta_p^2 - \eta^2) / \eta_p^2 \times 100$.
[b] As reported by the original authors to 3 decimal places.

Third, Cohen's (1988) benchmarks for interpreting effect sizes are arbitrary and should not be applied in L2 research or elsewhere (Cohen, 1988). Even so, the frequently used Cohen's (1988, p. 283) proportion of variance effect size cutoff points (i.e., small = .0099; medium = .0588; large = .1379) may only relate to "partial eta squared" values and not to those of "eta squared" in multi-way designs (Richardson, 2011). Thus, employing Cohen's benchmarks (error 1) and erroneously applying them to eta squared (error 2) creates a "double-error" situation. For example, if η^2 is erroneously chosen to be benchmarked against Cohen's (1988) cut-offs, one may interpret the magnitude of a given effect as "small". However, for the same effect, if "η_p^2" is compared against Cohen's benchmarks it may be interpreted as "large".

It is useful at this point to recall that effect sizes are descriptive statistics that leave the decision about the importance of an observed effect to the community of researchers in any specific domain given (a) their understanding of the phenomenon they study, (b) prior studies in the same domain, (c) the predictions of theory, (d) practical implications, (e) the design and

instrumentation from which the effect was derived, and so forth. Looking ahead, we recommend that Cohen's conventions be dropped in favor of researchers' direct and explicit comparison of the effects in related literature as well as these and other considerations (see Thompson, 2006; Plonsky & Oswald, 2014).

The Practical Roots

The findings of this study, which reveal somewhat widespread misuse of a common statistic, prompt us to consider why this problem exists (and persists). One explanation might be the lack of appropriate reference material. In examining 14 texts on L2 research methods at our immediate disposal, the materials available to L2 researchers do not appear to address adequately the distinction between η^2 and η_p^2. For example, Larson-Hall (2012a), a brief and generally quite useful chapter-length overview of statistics used in L2 research, briefly commented that "Effect sizes for ANOVA results are also of the same type as the correlation but use the Greek letter eta (η) and are called *eta-squared* or *partial eta-squared*" (p. 249, italics added). However, no clear distinction is made between η^2 and η_p^2. In Phakiti (2014), another generally strong reference, no clear distinction between the use of eta- and partial eta-squared is made (see p. 205 and pp. 283-300).

Other L2 research methods textbooks we reviewed likewise lacked sufficient discussion of the difference between the η^2 and η_p^2 effect size measures. Dörnyei (2007), for example, first provided a brief account of eta squared followed by presenting the formula for computing η^2. However, the only passing reference to η_p^2 was made later in the context of ANCOVA: "The good news about SPSS output is that next to the significance value we find the '*partial eta squared*' index, *which is an effect size*" (p. 223, italics added). A discussion on η^2 and η_p^2,

however, did appear in Larson-Hall (2016) where a number of the same considerations addressed here as regards these two variants were usefully and clearly explained (see p. 149).

We suggest that future texts that discuss ANOVA explain all the terms that appear in a full table of summary results (Thompson, 2006). It would be particularly useful if such a rubric would detail how all terms in the ANOVA results, including but not limited to η^2 and η_p^2, are (a) computed and (b) related to each other. When a reader is able to ascertain the relationship between all the terms in an ANOVA summary table, the distinction between η^2 and η_p^2 becomes more meaningful. As a final note, we would add that in many studies using ANOVA and its variants, researchers will want to go beyond the initial analysis to often perform pair-wise comparisons of groups' mean scores. In such cases, it is not sufficient to report only the effect size for the ANOVA main and interaction effects; rather, an eta-squared (equivalent to a point-biserial correlation; r_{pb}) or a standardized mean difference effect size such as Cohen's *d* for the comparison of interest should be reported and interpreted as well.

Conclusion

"[A]ny effect size that is chosen from possible alternatives should be technically [and nominally] appropriate" (Grissom & Kim 2012, p. 9). Evidence we provide in this paper contains numerous and recent examples of erroneous reporting of often large η_p^2 effect sizes in multi-way designs misinterpreted and mislabeled as η^2.

The distinction we draw in this paper between η^2 and η_p^2 is in no way semantic or statistical nit-picking. These effect sizes are increasingly reported throughout quantitative L2 research. They also have immense potential to inform our understanding of L2 learning and use. Clarity about these indices, their reporting, and interpretation is, in fact, critical to arrive at

appropriate conclusions regarding L2 theory and practice and, at the same time, to preventing misinterpretations that compromise work in this field.

We remind readers that, like many measures of effect size and in contrast to the dichotomous result embodied by *p* values, an η^2 value provides a continuously-expressed result within a *single* multi-way ANOVA. Thus, reporting η_p^2 values *alone* which (a) lack comparability advantages within a study and (b) are often larger than η^2 values (see Discussion section) may lead to erroneous interpretation. Added to the above problems is that η_p^2 values depend on the model of analysis (i.e., fixed, random, and mixed). That is, for a given study, if we run the data analysis under fixed-, random-, or mixed-effects models, values of η_p^2 for some treatment effects can change. Presenting the reasoning behind this dynamic requires knowledge of "Mean Squares Expectation Rubric" which falls outside the scope of the present paper (but see Thompson, 2006). Based on these considerations, we encourage L2 researchers to compute, report and interpret, by default, (classical) η^2 for all ANOVA-based analyses. This approach will provide an estimate of variance accounted for that is more stable as well as comparable *within* a single multi-way study. However, it is also useful for researchers to report η_p^2 along with η^2 to avoid the possibility of erroneous reporting and interpretation. In addition, in multi-way designs, reporting η_p^2 facilitates the calculation of power for an effect and thus using the size of that effect as the basis for planning the sample size for a relevant, future research. Thus, eta- and partial eta-squared serve different purposes which legitimizes the presence of both estimates in published multi-way ANOVA studies (see Notes). Finally, it is critical to note that reporting confidence intervals for η_p^2 values is both highly recommended and possible via several statistical packages. For example, one can use the function "`peta.ci`" in the first author's R package (Norouzian, de Miranda, & Plonsky, under review) available at (https://github.com/rnorouzian/i/blob/master/i.r).

Unfortunately, confidence intervals for eta-squared are more complex (often roughly approximated) than those for partial eta-squared, and not currently widely available.

In closing, the results of this study do not present an ideal state of statistical proficiency in L2 research. Nevertheless, we are hopeful that the field's momentum toward methodological reform—a movement to which the present study seeks to contribute—will continue to improve L2 research and reporting practices thereby leading to a clearer understanding of language learning and use.

Notes

A more detailed discussion on the application of eta-squared (η^2) and partial eta-squared (η_p^2) in L2 research depending on the substantive nature of the independent variables is currently available at Open Science Framework (https://osf.io/aymqd/) as supplementary to the present study.

CHAPTER III

A BAYESIAN APPROACH TO MEASURING EVIDENCE IN L2 RESEARCH: AN EMPIRICAL INVESTIGATION

Overview

Null hypothesis testing has long-since been "the go-to analytic approach" in quantitative second-language (L2) research (Norris, 2015, p. 97). To many, however, years of reliance on this approach has resulted in a crisis of inference across the social and behavioral sciences (e.g., Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016). As an alternative to the null hypothesis testing approach, many such experts recommend the *Bayesian hypothesis testing* approach. Adopting an open-science framework, the present study (a) re-evaluates the empirical findings of 418 t-tests from published L2 research using the Bayesian hypothesis testing, and (b) compares the Bayesian results with their conventional, null hypothesis testing counterparts as observed in the original reports. The results show that the Bayesian and the null hypothesis testing approaches generally arrive at similar inferential conclusions. However, considerable differences arise in the rejections of the null hypothesis. Notably, 64.06% of cases when *p*-values fell between .01 and .05 (i.e., evidence to reject the null), the Bayesian analysis found the evidence in the primary studies to be only at an "anecdotal" level (i.e., insufficient evidence to reject the null). Practical implications, field-wide recommendations, and an introduction to free online software (rnorouzian.shinyapps.io/bayesian-t-tests/) for Bayesian hypothesis testing are discussed.

Introduction

Recent advances in the social science research methods have been embraced by a wide array of social and behavioral sciences. Similarly, in second language (L2) research, several influential works (e.g., Larson-Hall, 2016; Norouzian & Plonsky, in press; Norris, 2015; Plonsky & Oswald, 2014) and special issues (e.g., Language Learning 65, Suppl. 1) have been devoted to a budding "methodological reform" currently taking place. These reform efforts are made under the assumption that "[p]rogress in any of the social sciences depends on sound research methods, principled data analysis, and transparent reporting practices; the field of second language acquisition (SLA) is no exception" (Plonsky & Gass, 2011, pp. 325-326).

Methodologically speaking, one of the most challenging tasks facing L2 researchers is making reasonable inferences when extending their study findings to the larger populations of interest. Indeed, a good share of recent methodological works within L2 research consists of criticisms against the common practice of null hypothesis testing to make such inferences (e.g., Norris, 2015; Norris et al., 2015; Plonsky & Oswald, 2014). These criticisms are mainly motivated by the fact that, despite their widespread use, *p*-values resulting from the formal testing of a null hypothesis (H_0) provide misleading measures as to whether an empirically obtained effect from a sample of participants generalizes to the larger population of interest or not (Francis, 2016; Ioannidis, 2005; Thompson, 2006). Pointed calls discouraging the common use of *p*-values in applied research have now been made at the societal level as well. Most notably, the American Statistical Association (ASA) recently released an unequivocal statement on the matter, arguing that "a *p*-value does not provide a good measure of evidence regarding a model or hypothesis" (American Statistical Association, 2016).

Consequently, to some experts, years of reliance on *p*-values has contributed to what might be termed as the 'Crisis of Inference' across social and behavioral sciences (Dienes & Mclatchie, 2017; Kruschke & Liddell, 2017; Pashler & Wagenmakers, 2012; Rouder et al., 2016). At its core, such a crisis stems from the lack of confidence in the inferential conclusions about the real-world effects of various research phenomena based on *p*-values from the individual studies targeting those phenomena.

While these criticisms are important in raising our collective awareness about the problems associated with *p*-values and the null hypothesis testing approach, we argue that what is critically needed is knowing about the alternatives to *p*-values, and how such alternatives compare with—and, in many cases, improve on—*p*-values.

Null Hypothesis Testing

The conventional paradigm to make a formal inference from which *p*-values result is known as null hypothesis testing. The idea is that when a researcher finds an effect from a single study with a specified sample of participants, s/he must first assume that there is no effect (i.e., effect size is zero) from his/her study in the actual population of participants (i.e., H_0; Null Hypothesis position). In order for this approach to be applied appropriately, the researcher must then theoretically think of infinitely repeating the exact same study with different samples of participants the same size of his/her own study from the population. Because the make-up of participants in each of these repetitions of the study could be different, the found study effects in these repetitions might differ from each other forming a theoretical mass of obtained effects. Some of resultant effects in this mass may have occurred more frequently, and some less making

some areas of the mass to be higher and other areas lower in terms of frequency (sometimes forming a bell-curve of some kind).

With this theoretical mass of effects at hand, a (two-tailed) *p*-value is simply obtained by examining the mass to find out the probability of the effect actually found by a researcher or more extreme (i.e., larger in absolute value) than that. For example, for a simple L2 pre-post study with 30 participants which has found a Cohen's *d* effect size of .3, the resulting *p*-value is graphically shown as the two red-shaded areas of the grey-colored mass of the study's theoretical effect values in Figure 6 (to explore Figure 6 see https://github.com/izeh/l/blob/master/1.r). In this case, these two red-shaded areas in the tails constitute 11.11% of the entire theoretical mass of effect size values for this example. Thus, the probability known as the *p*-value for this example is .1111.

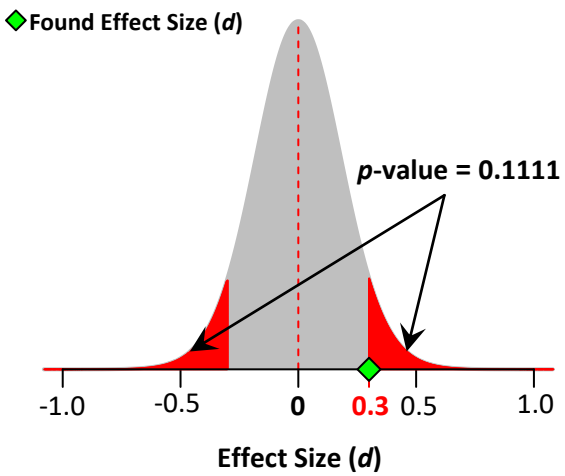

Figure 6. The process of obtaining a p-value from a pre-post design.

As an evidence-measuring index, a *p*-value has a fine-grained classification that indicates the strength of the evidence a *p*-value provides *against* the null hypothesis position (H_0). Table 5 shows this fine-grained classification for *p*-values (see Wasserman, 2004, p. 157). However, the

common practice is that when a *p*-value is smaller than .05 (or 5% of the theoretical mass of effects), a researcher can conclude s/he has evidence against the null hypothesis position, as the effect s/he has found from her/his study compares with 5% or less of the theoretical mass of effects. Because such areas are distant from 0 (our null hypothesis and the center of the mass in Figure 6), the conventional conclusion is that the null hypothesis is unlikely to be true. Thus, we should reject the null hypothesis.

Table 5
P-values Classificatory Scale (Wasserman, 2004)

p-value	Strength of Evidence
< .001	Decisive evidence against H_0
.001 - .01	Substantive evidence against H_0
.01 - .05	Positive evidence against H_0
> .05	No evidence against H_0

Bayesian Hypothesis Testing

Bayesian hypothesis testing takes a completely different approach to the hypothesis testing process. Specifically, in the first step the obtained effect from a study (i.e., an obtained effect size) is tested against a range of innumerable hypotheses. The result of this first step is referred to as a *Likelihood Function.* In a likelihood function, the obtained data gets a chance to be benchmarked against all possible hypotheses. Thus, in addition to the theoretical mass of effects based around a null (i.e., 0), Bayesian hypothesis testing allows for any other alternative value to be the basis for the theoretical mass of effect values. Such alternative hypotheses are innumerable and thus require a reasonable specification (see principle 1 in the next section). For

example, if for illustration purposes, we take only four possible alternative effect values (i.e., H_1: "**.5**", H_2: "**1**", H_3: "**1.5**", H_4: "**2**") in addition to H_0 (i.e., **0**) for our example of a simple pre-post study with 30 participants, then we can show the null as well as four alternative theoretical masses of effects side by side in Figure 7 (to explore Figure 7 see https://github.com/izeh/l/blob/master/2.r).

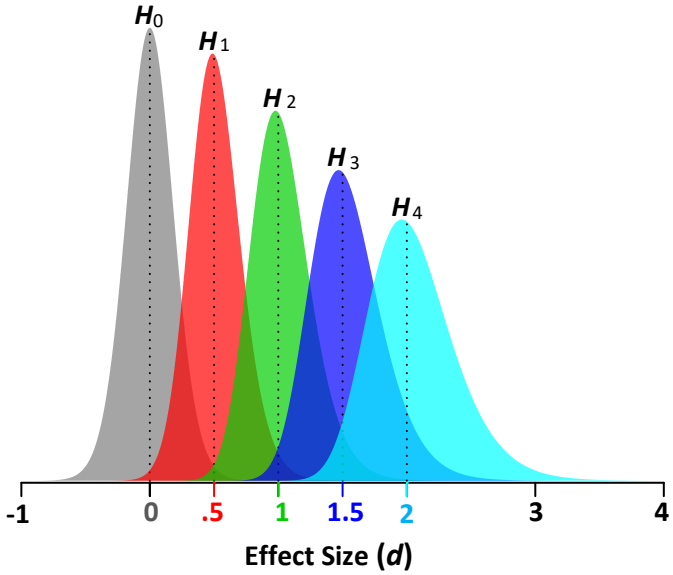

Figure 7. Four theoretical masses of effects based on four alternatively hypothesized effect size values.

The Bayesian hypothesis testing then benchmarks the obtained effect size (in our example Cohen's *d* of .3) against the theoretical mass based around the null hypothesis (see Figure 7), as well as benchmarking the same obtained effect size against the theoretical masses based around all alternative hypotheses. The value obtained from each set of benchmarking provides the probability of the observed effect based on its respective hypothesis. The idea is then to simply compare (by division) these two sets of probabilities (i.e., from alternatives and null) to arrive at what is called a "*Bayes Factor*". Thus, a Bayes factor is a statistic that expresses

the comparative evidence for one hypothesis (e.g., alternative hypothesis) over another hypothesis (e.g., null hypothesis). A Bayes factor provides a naturally comparative metric replacing a *p*-value as a widely-criticized evidence-measuring index (see Etz & Vandekerckhove, in press; Kruschke & Liddell, 2017; Morey, Wagenmakers, & Rouder, 2016; Rouder et al., 2016; Rouder, Speckman, Sun, Morey, & Iverson, 2009). To better understand a Bayes factor, let us apply the concept to a meaningful L2 quantitative research example. We will then formalize the steps involved in its computation.

Suppose a researcher is studying the effect of an L2 treatment on the development of explicit knowledge (DeKeyser, 2015; Ellis, 2009; Lyster & Sato, 2013) of 60 high-intermediate English as a Foreign Language (EFL) learners with respect to the English indefinite article "*a/an*". Following the treatment at the post-testing stage, the goal is to evaluate the difference in the level of the explicit knowledge of the members of the treatment group (n = 30) and the control group (n = 30) with respect to the target linguistic form. In both groups, the development of explicit knowledge is measured using a *grammaticality judgement test* (GJT) that includes 20 target errors (see Ellis, 2009). A common scoring method (see Mackey & Gass, 2016; Shintani & Ellis, 2013) for GJTs is a dichotomous scheme (i.e., 0 for not identifying an erroneous form, and 1 for successfully identifying an error). Therefore, the minimum score that a participant can obtain on such a GJT is 0 and the maximum score is 20 with all other possible scores (i.e., 1, 2, 3, …, 19) lying in between. Let us simulate such a study and then conduct both the null as well as the Bayesian hypothesis testing to evaluate the results. The simulated study's design and raw scores for all 60 EFL participants in the two groups are graphically shown in Figure 8 (to explore Figure 8 see https://github.com/izeh/l/blob/master/3.r).

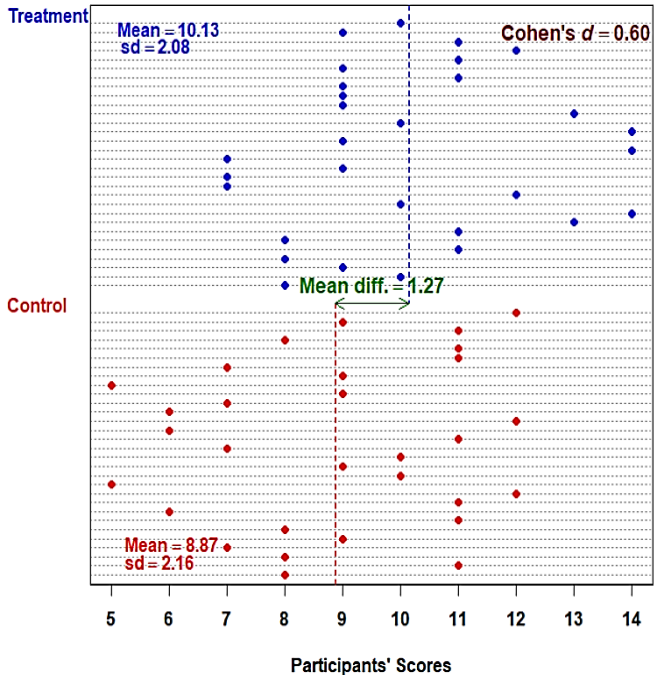

Figure 8. Participants' scores on GJT. Each grey, horizontal, dotted line represents a participant (1, 2, …, 60). The vertical dashed lines denote the mean of each group. Mean diff. = difference between the means of the two groups.

As indicated in Table 6, based on the result of this simulated study we can conclude that the L2 treatment has been effective in expanding the explicit knowledge of the treatment group with respect to the target linguistic form (t (58) = **2.31**, $\boldsymbol{p}$ **= .02**, $\boldsymbol{d}$ **= .60**, 95% $CI_{(d)}$ [.08, 1.12]). Furthermore, since we followed a null-hypothesis testing approach, we reject the null hypothesis position (H_0) that the effect of this treatment is zero in the actual population of high-intermediate EFL learners. Given this positive evidence against the H_0 (see Table 5), we can claim that the significantly positive effect of the treatment extends beyond the 60 high-intermediate EFL participants in the present study and to the far larger population of high-intermediate EFL learners.

Table 6
Post-Test Results for EFL Learners in a Simulated Study (N = 60)

	Descriptive				Inferential	
Groups	*n*	*M* (*SD*)	*ES* (*d*)	95% $CI_{(d)}$	*t* (*df*)	*p*
Treatment	30	10.13 (2.08)	.60	[0.08, 1.11]	2.31(58)	0.024
Control	30	8.87 (2.16)				

Note. ES = Effect size; *d* = Cohen's *d*.

Is the claim we have just made reliable? Can this significantly positive result *really* generalize to the actual population of high-intermediate EFL learners? No.

Indeed, the data and simulation we have just presented is based on a population in which there was absolutely zero effect for the L2 treatment in the population. Nevertheless, the observed effect including the statistically significant *p*-value and the corresponding *d* value well above zero are both entirely possible. The null-hypothesis testing approach, in this case, led us to an erroneous conclusion. Consequently, any practical implications from such a study are completely invalid as well.

As noted earlier, Bayesian philosophy's approach to our running research problem is different from that of the null hypothesis testing. Systematically speaking, the Bayesian approach lays out the following three principles:

1- *Specify the alternative hypotheses*. In addition to the null hypothesis which describes only one possibility for the effect (i.e., effect is "0") of the study in the actual population, specify a set of reasonable alternative effects informed by previous findings or the general sizes of effects in your field. The researcher describes all different possibilities for the effect size of the study in the actual population.

2- *Obtain a comparative measure*. Divide the likelihood of your observed effect under the alternative hypothesis by that under your null hypothesis to obtain a "comparative measure" indicating the extent to which a hypothesis (e.g., alternative hypothesis or null hypothesis) your study data supports. For example, if you obtain 2, this means your alternative hypothesis is two times more strongly supported by your data. Call this comparative measure a "Bayes Factor".

3- *Interpret the comparative measure*. Instead of rejecting/not rejecting a hypothesis, interpret your obtained "Bayes Factor" on a classificatory scale. The scale provides a useful guide, but is not meant to provide a rigid set of benchmarks. Researchers may evaluate their obtained Bayes factor (Bayesian counterpart of *p*-values) against Table 7 to evaluate the extent to which a hypothesis (i.e., Alternative or Null) their study data supports (Jeffreys, 1961, p. 432).

Let us apply these three Bayesian principles to our running example of the efficacy of an L2 treatment in developing the explicit knowledge of 60 high-intermediate EFL learners.

Specifying the alternative. As shown previously (see Figure 7), Bayesian hypothesis testing requires specifying a set of alternative hypotheses in addition to the null hypothesis. One of the most commonly employed, scale-free metrics used to specify alternatives is an effect size. An effect size such as Cohen's *d* directly measures the effect of a treatment and is commonly used in L2 research (Larson-Hall, 2016; Plonsky & Oswald, 2014).

Table 7
Bayes Factor Classificatory Scale

Bayes Factor $\left(\frac{Alternative}{Null}\right)$	Strength of Evidence
> 100	Decisive evidence for *Alternative*
10 - 30	Very strong evidence for *Alternative*
3 - 10	Substantial evidence for *Alternative*
1 - 3	Anecdotal evidence for *Alternative*
1	Hypothesis Insensitive Evidence (No evidence for either hypotheses)
1/3 - 1	Anecdotal evidence for *Null*
1/10 - 1/3	Substantial evidence for *Null*
1/30 - 1/10	Strong evidence for *Null*
1/100 - 1/30	Very strong evidence for *Null*
< 1/100	Decisive evidence for *Null*

Unlike the null hypothesis, which is represented by a single statement that the size of effect in the actual population of interest (here high-intermediate EFL learners) is "0", alternative hypotheses on the size of effect in a population of interest almost always consist of innumerable values. That is, when we think about sizes of effect for our study in the actual population of EFL learners, a range of possible values could be considered. One useful way to specify a reasonable range for alternative sizes of effect in the actual population is to consider the sizes of effects in the general domain L2 research. Fortunately, a resource for doing so in the context of L2 research is available. Specifically, Plonsky and Oswald (2014) studied the magnitude of Cohen's *d* effect size in 346 primary L2 studies and 91 meta-analyses of L2 research. As for the magnitude of Cohen's *d* in L2 research, Plonsky and Oswald (2014) found that the effect size

could often be as large as +1. Even so, direction in Cohen's *d* effect size is arbitrary. Thus, a researcher specifying the alternative sizes of effect, and not certain about the direction of an effect in the population might want to take a neutral position and consider that effects can go both directions creating a two-sided (i.e., two-tailed) alternative. Therefore, when specifying the alternative, it is possible to consider that the alternative sizes of effects could be as large as reported by Plonsky and Oswald (2014) in either direction. That is, the most frequently expected sizes of an effect in the general domain of L2 research could often range between -1 and +1. Conversely, effect sizes outside -1 and +1, though possible, are much less frequently expected in L2 research. At this point, we should use a weighting scheme such that our highly-expected effect sizes (ranging from -1 to +1) are upwardly weighted and effect sizes outside this range receive successively less and less weight. Here we use the extensive research conducted in psychology which has led to the specification of a default form of alternatively hypothesized effect sizes (e.g., Ly, Verhagen, & Wagenmakers, 2016; Morey, Romeijn, et al., 2016). The technical specifics of this particular distribution of effect sizes are discussed in various sources (Rouder et al., 2009). But it is important to note that this weighting scheme for effect sizes is widely known as a *Cauchy* (after Augustin-Louis Cauchy) distribution with a scale (similar to standard deviation) of ".707", and is centered at "0". Figure 9 (to explore Figure 9 see https://github.com/izeh/l/blob/master/4.r) shows this default alternative distribution of effect sizes in which the part between -1 and +1 shows a much higher weight, indicating a higher likelihood of being observed. Values for effect size outside this range (i.e., –1 and +1), by contrast, are less likely to be found in L2 research, and thus the distribution assigns much smaller weight to such alternative effect size values. For example, we can all agree that it is very unlikely that a treatment in L2 research finds an effect of + 6. Thus, such an effect size value (and its

negative counterpart – 6) is located in the tail area to denote that such a value may not be a good alternative value from the perspective of a researcher.

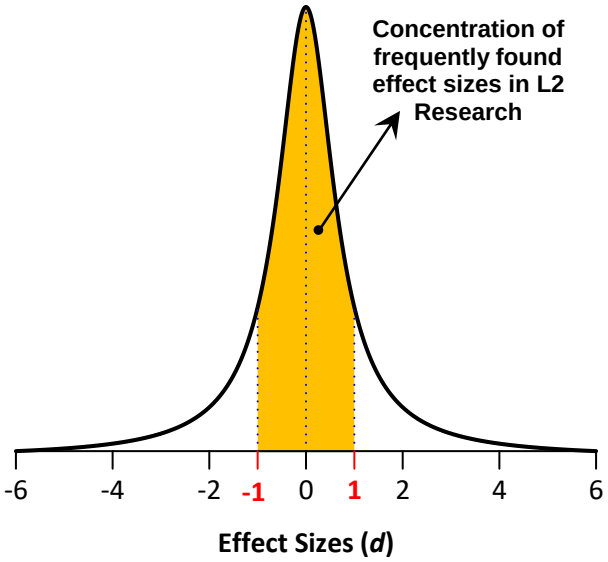

Figure 9. A default distribution of alternatively hypothesized effect sizes in L2 research.

Obtain a comparative measure. As with other, more familiar statistics, a Bayes factor, as a comparative measure, can be obtained using a software package. Here we use a free, web-accessed, and interactive software package developed by the first author of the present study available at (rnorouzian.shinyapps.io/bayesian-t-tests/) to obtain the Bayes factor for our example of the effect of an L2 treatment on developing the explicit knowledge of 60 high-intermediate EFL learners. As noted previously, the process of obtaining a Bayes factor begins by: (a) placing a researcher's obtained effect value under all *alternatively hypothesized* sizes of effect in the mass of theoretical values specified under principle 1 (see Figure 9), (b) placing the same researcher's obtained effect value under the *null hypothesized* size of effect in the population of high-intermediate EFL learners (i.e., "zero"), and finally (c) dividing (a) by (b) to

obtain a comparative measure (the Bayes factor) indicating the extent to which either hypothesis is better supported by the researcher's obtained results. Figure 10 provides the main panel of the software. The required settings for our example are indicated by the red arrows.

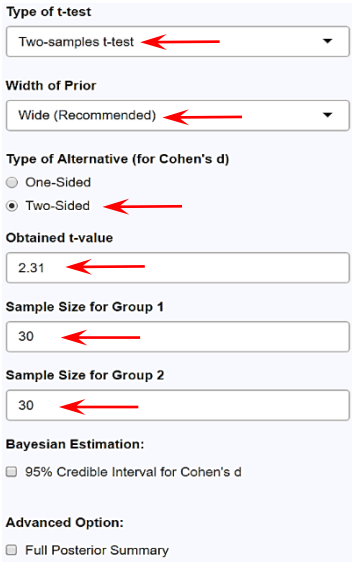

Figure 10. A screenshot of the Bayesian for t-tests software. The red arrows indicate the settings used for the example in the text.

The software has additional Bayesian capabilities that allows conducting Bayesian estimation, and replacing *confidence intervals* with a Bayesian alternative known as a "*Credible Interval*" for Cohen's *d* effect size (Norouzian, De Miranda, & Plonsky, under review). As the output indicates, the Bayes factor comparing the probability of obtaining the effect size of .6 (our simulated study Cohen's *d* effect size) given the alternative hypothesis to that given the null hypothesis is **2.34** $\left(\text{i.e., } \frac{Alternative}{Null}\right)$. This Bayes factor value results from the three steps described earlier in this section. The numerator results from the application of Bayesian framework which integrates all the alternative hypotheses resulting in .0695. The denominator is simply the height

(i.e., density) of the observed effect (i.e., Cohen's *d* of .6) under the mass of effects based around the null hypothesis which in this case is .0296. The Bayes factor (i.e., 2.34) is simply the result of the division of the two values $\left(\text{i.e., } \frac{.0695}{.0296} = 2.34\right)$. Given that the result of this division is larger than 1, we have found some evidence for the alternative hypothesis. Specifically, our simulated study's results are 2.34 times more strongly supported by the alternative hypothesis than by the null hypothesis (H_0).

Interpret the comparative measure. Recall that our simulated L2 study's result (i.e., $t\,(58) = 2.31$, $p = .02$, $d = .60$, 95% $CI_{(d)}$ [.08, 1.12]) led us to erroneously reject the null hypothesis, and claim that the L2 treatment can produce a significantly positive effect in the actual population of high-intermediate EFL learners. However, the Bayes factor comparing such an alternative hypothesis to the null hypothesis is only 2.34 in favor of the alternative hypothesis. Based on Table 7, we only have *anecdotal evidence* (i.e., very weak) for accepting the statement that our treatment can have any positive effect in the actual population of high-intermediate EFL learners. In other words, the obtained Bayes factor for the exact same study leads us to conclude that the obtained effect size, Cohen's *d* of .6, could have been a random finding applicable only to the particular 60 high-intermediate EFL participants that we studied and not to the actual population of high-intermediate EFL learners.

The Study

The *t*-test design in our example, one of the most common designs in L2 research (Larson-Hall, 2016), reveals the potential difficulty faced by L2 researchers when making an inference about the efficacy of their studies in larger populations of L2 learners. Specifically, the *p*-value of .02, provided *positive evidence* (see Table 5) against the null hypothesis yet the

obtained Bayes factor, for the same study provided *anecdotal evidence* (see Table 7) against the null hypothesis. To practically understand the extent of the disagreements in the conclusions reached using two evidence-measuring indices and make field-wide recommendations, this study seeks to re-analyze the empirical findings from 418 previously published t-tests from four well-known L2 journals using *p*-values, and Bayes factors. The detailed comparison between these two approaches were motivated by the following research question: Do Bayes factors (*Bayesian hypothesis testing*) and *p*-values (*null hypothesis testing*) differ in agreement over the strength of empirical findings from a representative sample of *t*-tests from published L2 research?

Method

In the following section, methods used to select L2 journals, criteria for choosing individual L2 studies are explained. Also, procedures and analyses followed are detailed. Additionally, all the codes, software and data are made available online and linked within the discussions when appropriate.

Journal Selection and Inclusion Criteria

To select the L2 journals for the present study, we consulted (a) previous surveys of L2 research practices (e.g., Egbert, 2007; Gass, 2009; Lazaraton, 2005; Plonsky, 2013), (b) Journal Citation Reports (JCR, with no impact factor size considerations), (c) L2 method textbooks providing various L2 journals' descriptions (e.g., Perry, 2011), and (d) miscellaneous documents surveying L2 journals' perceived quality (VanPatten & Williams, 2002).

There is, of course, no consensus as to which journals are most prominent or influential in the field. In the end, we selected the following four journals: *Language Learning, Modern Language Journal, Studies in Second Language Acquisition,* and *System.* It is important to note that as long as the sampled primary L2 studies from these journals employ a significance level

that is common in L2 research (e.g., .05), and include sample sizes that are commonly found in the general domain of L2 research, the results of the present study can provide a reasonable basis for offering field-wide recommendations as to how the conclusion of L2 t-test studies are changed if the conventional null hypothesis testing approach is replaced with the Bayesian hypothesis testing approach.

Given the frequency with which t-tests are employed in L2 research, in each of the four selected L2 journals, we limited our search to studies published in the 2014 and 2015 volumes. In line with previous reviews (e.g., Gass, 2009), we excluded from consideration forums, short reports, commentaries, review articles, and book reviews. Initially, data from 712 t-tests from 119 studies that employed t-tests were extracted. However, our technical inclusion criteria of collecting t-tests that (a) were not used in pair-wise comparisons following (post-hoc) a larger analysis (often ANOVAs), (b) were not used for planned comparisons, (c) were not from relational analyses that use t-tests (e.g., regression) and (d) if from independent-samples, were calculated under the assumption of equality of group variances, decreased the final number of the sampled t-tests to 418. This sample included 172 independent-samples t-tests, and 243 paired-samples t-tests, and 3 one-sample t-tests. The complete raw dataset for the present study is publicly available at (https://raw.githubusercontent.com/izeh/l/master/l.csv).

Procedures and Analyses

In order to address our research question, we followed a standard data collection procedure as recommended in synthetic research (see Plonsky & Oswald, 2015). First, each study in the sample was systematically coded for its type of t-test (i.e., independent-samples, paired-samples, one-sample). Then, we extracted from each study its corresponding *t*-value(s),

samples sizes (i.e., n_1, and if independent-samples t-test n_1 and n_2), degrees of freedom, and the *p*-value(s). The analyses were conducted using a computer program developed by the first author in the R language for statistical computing (R Core Team, 2017). The complete R program developed to analyze the data for the present study is publicly available at (https://github.com/izeh/l/blob/master/d.r). Essentially, the program first distinguished between the type of the t-tests in the sampled primary studies, and accordingly performed a secondary null hypothesis test for each t-test in the primary studies to obtain the exact *p*-values (up to 9 digits). Then, the program separately performed a Bayesian hypothesis test to obtain an exact Bayes factor (up to 9 digits) for each t-test. The computation of Bayes factors in each case followed the Bayesian principles described in the previous section. Next, the null hypothesis testing results, and the Bayesian hypothesis testing results were categorized according to their respective classificatory scales presented earlier in Table 5 and Table 7. Finally, the R program reorganized the categorized results of the comparisons between the two methods of inference (i.e., the null hypothesis approach vs. the Bayesian hypothesis testing approach) to create a contingency table. Importantly, the contingency table enables comparing the conclusions of the t-tests in the sampled primary L2 studies according to the null hypothesis approach with those according to the Bayesian hypothesis testing approach.

Results

The present study was intended to examine the extent to which the inferential conclusions of a representative sample of L2 t-test studies published between 2014 – 2015 might differ according to the method of inference employed. As noted in the Method, the result of the comparisons between the conclusions of these primary studies using the Bayesian and null hypothesis testing approaches could be best made using a contingency table. The contingency

table allows for the results from the two methods of inference to be benchmarked against their respective classificatory scales presented in Tables 5 and 3. Table 8 provides a 7 (rows) × 4 (columns) contingency table that contains the frequency outcomes of the comparisons made between the two methods. The last column and the last row titled *marginal* provide the sum of each column and each row, respectively. Additionally, the sum of the entire columns marginals and row marginals equal the entire set of the t-tests collected from the primary studies.

Table 8
Comparison of Bayesian and Null Hypothesis Testing Results for 418 T-Tests

	p-value				
Bayes Factor	Decisive (0 – .001)	Substantial (.001 – .01)	Positive (.01 – .05)	None (.05 – 1)	Marginal
Decisive (> 100)	88	0	0	0	*88*
Very Strong (30 – 100)	19	2	0	0	*21*
Strong (10 – 30)	1	36	0	0	*37*
Substantial (3 – 10)	0	28	20	0	*48*
Anecdotal H_1 (1 – 3)	0	0	41	15	*56*
Anecdotal H_0 (1/3 – 1)	0	0	3	87	*90*
Substantial H_0 (1/10 – 1/3)	0	0	0	78	*78*
Marginal	*108*	*66*	*64*	*180*	418

To better establish the relationship between the inferential conclusions of the two methods of inference and evaluate their comparative distribution, the results of the comparisons are also graphically shown in Figure 11 with the marginals in percentages (to explore Figure 11 and Table 8 see https://github.com/izeh/l/blob/master/5.r).

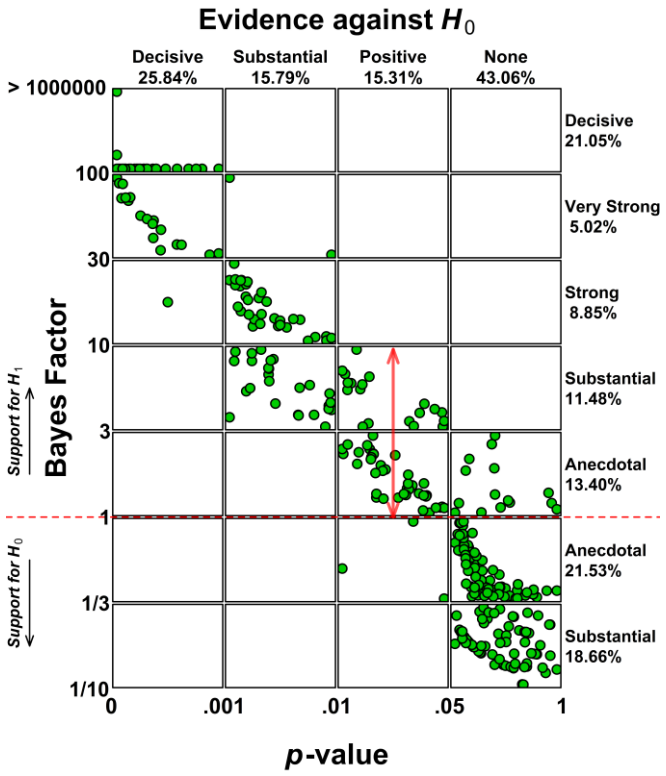

Figure 11. Relationship between the conclusions of the Bayesian and null hypothesis testing approaches. Side percentages indicate the proportion of conclusions in each category out of the total sample size (N = 418). Bayes factors above the red, dashed line support alternative (H_1) and those below the line support H_0. The red, double-headed arrow indicates the range of Bayes factors for p-values falling between .01 and .05.

As shown numerically in Table 8 and graphically in Figure 11, a clear relationship between the results of the Bayesian hypothesis testing and those of the null hypothesis testing is observed. Specifically, small *p*-values which provide "Decisive" (0 – .001) evidence against the null hypothesis correspond to very large Bayes factors that also provide "Decisive" (> 100) or "Very Strong" (30 – 100) evidence against the null hypothesis. The pattern of agreements between the two methods of inference is also seen for *p*-values that provide "Substantial" (.001 - .01) evidence against the null hypothesis. These *p*-values correspond to Bayes factors that likewise provide either "Strong" (10 – 30) or "Substantial" (3 – 10) evidence against the null hypothesis. However, a

critical disagreement seems to arise between the two methods of inference over what the conventional *p*-value approach labels as "Positive" (.01 – .05) evidence against the null hypothesis. For 64.06% of such *p*-values, the corresponding Bayes factors provide only "anecdotal" evidence for the alternative hypothesis. In other words, the amount of evidence that under the null hypothesis testing approach leads to the rejection of a null hypothesis (*p*-values between .01 and .05), from the Bayesian hypothesis perspective is "not worth more than a bare mention" (Jeffreys, 1961, p. 432). In the final category where *p*-values find "no" (> .05) evidence against the null hypothesis, the corresponding Bayes factors also mainly provide no evidence against the null hypothesis. Thus, no decision-changing disagreements between the conclusions of the two methods of inference for this category exist.

It is wise to change the prior on effect size and re-analyze the data to inspect stability of the results obtained above. For this purpose, we use one other prior specification. This specification is a *Cauchy*(0, 1) which is wider than the prior used in the previous section. Under this specification, still 64.06% (41 out of 64) of the *p*-values falling between .01 and .05 have corresponding Bayes factors that only provide "anecdotal" evidence for the alternative hypothesis. The results are graphically shown in Figure 12 numerically in Table 9.

Table 9

Comparison of Bayesian and Null Hypothesis Testing Results for 418 T-Tests (wider prior)

Bayes Factor	*p*-value				
	Decisive (0 – .001)	Substantial (.001 – .01)	Positive (.01 – .05)	None (.05 – 1)	Marginal
Decisive (> 100)	87	0	0	0	*87*
Very Strong (30 – 100)	18	1	0	0	*19*
Strong (10 – 30)	1	35	0	0	*38*
Substantial (3 – 10)	0	30	14	0	*44*
Anecdotal H_1 (1 – 3)	0	0	41	15	*48*
Anecdotal H_0 (1/3 – 1)	0	0	3	62	*70*
Substantial H_0 (1/10 – 1/3)	0	0	0	111	*112*
Marginal	*108*	*66*	*64*	*180*	418

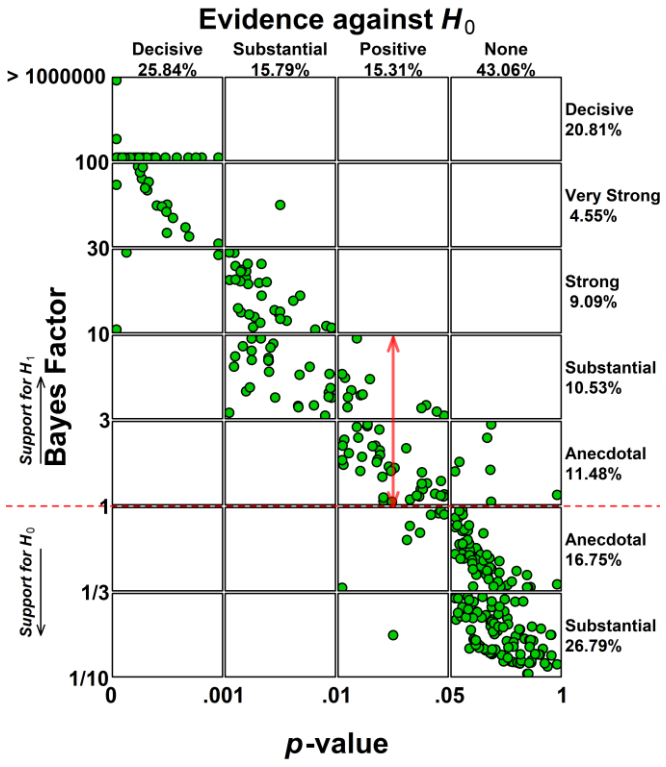

Figure 12. Relationship between the conclusions of the Bayesian and null hypothesis testing approaches with wider prior specification (i.e., Cauchy[0, 1]).

Discussion

Several seminal, theoretical works (e.g., Benjamin et al., in press; Dienes & Mclatchie, 2017; Johnson, 2016; Kruschke & Liddell, 2017; Rouder et al., 2016) along with the American Statistical Association (2016) have called for "alternative measures of evidence such as likelihood ratios or Bayes factors" (American Statistical Association, 2016, p. 2) in place of the current null hypothesis testing-based measures (i.e., *p*-values). We empirically implemented these recommendations in a representative sample of quantitative L2 studies that had used t-tests designs. For these studies, we compared the inferential conclusions of the Bayesian hypothesis testing approach with their conventional, null hypothesis testing counterparts. Here we provide two distinct implications arising from our study.

A New Threshold for Statistical Significance: Replication and Estimation

Our empirical findings raise a critical concern that the commonly adopted thresholds (ranging from .01 to .05) for declaring a statistically significant finding in quantitative L2 research could allow for a potentially high false discovery rates (Benjamin et al., in press; Ioannidis, 2005). In brief, false discovery rate refers to "the proportion of true null effects among the total number of statistically significant findings" (Benjamin et al., in press, p. 8). Specifically, the disagreement between the Bayesian and null hypothesis testing approaches over the sufficiency of the evidence against the null hypothesis when *p*-values fall between .01 – .05 suggests that adoption of more stringent (i.e., lower) thresholds for researchers to declare a statistically significant finding could reconcile the two methods of inference. How low? Johnson (2013) and Benjamin et al. (in press) both reason that such a threshold must be .005. Notice that even in our empirical results in Figure 12, t-tests from studies whose *p*-values are between .001 – .01 correspond to Bayes factors that either provide "Substantial" (3 – 10) or "Strong" (10 – 30) evidence against the null (and thus for the alternative hypothesis). Therefore, it can be understood that lowering the threshold for statistical significance to .005 finds support from the Bayesian stand point. For the primary studies in our sample, the entire distribution of *p*-values is shown in Figure 13 (to explore Figure 13 see https://github.com/izeh/l/blob/master/6.r).

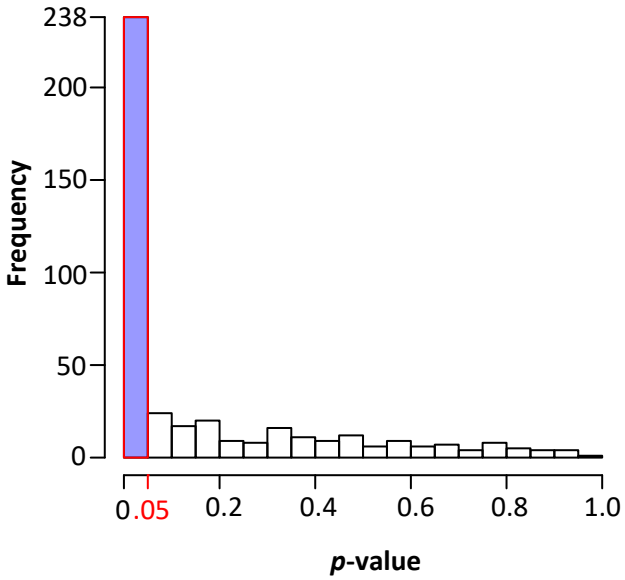

Figure 13. Distribution of p-values in the primary studies

As expected, a majority of the published results (238 t-tests) in the primary studies had found statistical significance under the traditional statistical threshold of .05. It would be interesting to see what proportion of these significant findings will remain significant if we employ the recommended threshold of ".005". As depicted in Figure 14 (to explore Figure 14 see https://github.com/izeh/l/blob/master/7.r), 154 tests in the primary studies remain statistically significant; a 35.29% $\left(\text{i.e.,}\frac{238 - 154}{238} \times 100\right)$ reduction in declarations of significant findings in an overall sample of 418 tests.

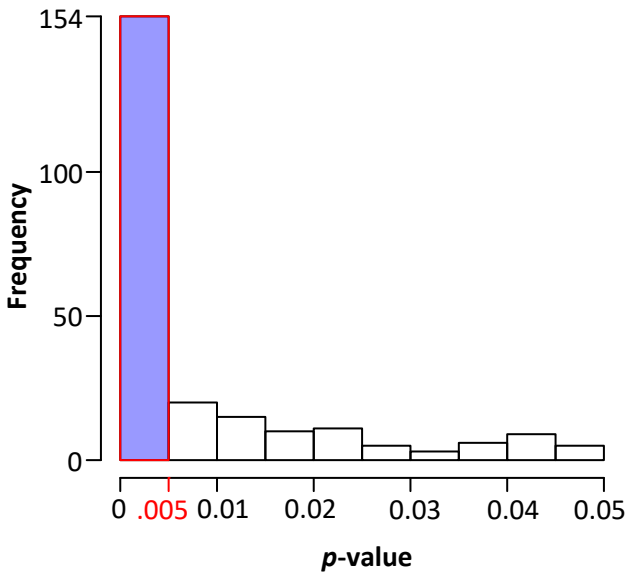

Figure 14. Distribution of significant p-values in the primary studies.

A critical question is then what we gain in return for reducing the number of findings considered to be statistically significant?

We believe, two importantly related gains accrue from adopting this new threshold for the field. First, gains in replicability and reproducibility rates. Although it garnered little attention until recently (but see Porte, 2012), replication has long been known to be the "*sine qua non* of research" (Thompson, 2004, p. 150, italics in original). Recent large-scale, international replication efforts in other fields such as psychology (Open Science Collaboration, 2015) have provided empirical evidence that original studies with *p*-values smaller than ".005" are nearly twice as much likely to be replicated and verified by an independent replication attempt under the original study's stated conditions compared to that for the original studies with larger *p*-values up to ".05" (see Benjamin et al., in press; Johnson, 2013). In other words, if we believe that "[t]he essence of the scientific method involves observations that can be repeated and verified by others" (American Psychological Association, 2009, p. 12), then the higher

reproducibility rates in the field seems to provide one reasonable rationale for embracing a stricter threshold for statistical significance.

Second, adopting the new convention for statistical significance will also assist us in paying attention to findings that provide more accurate information about the size (i.e., magnitude) and the direction (i.e., sign) of underlying (population) effects. Gelman and Carlin (2014) convincingly argue that "when researchers use small samples and noisy measurements to study small [underlying] effects . . . a significant result is often surprisingly likely to be in the wrong direction and to greatly overestimate an [underlying] effect" (p. 1). In fact, the issues of "impoverished sample sizes" (Norris et al., 2015, p. 1) and the modest size of the underlying effects in L2 research (Plonsky & Oswald, 2014) are both well documented. In essence, the crux of the argument made by Gelman and Carlin (2014) is that if we use the ".05" as the threshold for detecting statistically significant results, while using small sample sizes to estimate small underlying (population) effects, then the likelihood that our obtained, statistically significant results are overestimates of their underlying population effects (i.e., exaggeration rate) or have additionally the wrong sign (i.e., misdirection rate) could be considerably high. A full demonstration of the points raised by Gelman and Carlin (2014) and software to implement them is provided in the Supplementary Documents for the present study found at (https://github.com/izeh/l/blob/master/i.pdf). Together with increasing the sample size for a research study, lowering the threshold for statistical significance might be a reasonable way to prevent the statistically significant findings that could potentially obscure our views regarding the size and the sign of the underlying population effects in the field.

Bayesian Thinking: Researcher Involvement

Another distinct implication from the present study relates to the advantages gained from applying "Bayesian thinking" to the inference process as we described it throughout this study. Specifically, we believe two advantages accrue from employing Bayesian thinking. First, researchers are not passive to the processes that determine the generalizability of their findings. In pursuit of objectivity, substantive researchers have traditionally been advised to determine their study generalizability through inferences that only consider a null hypothesis position. As a new possibility, Bayesian hypothesis testing asks that researchers use their substantive knowledge, practical experience, and prior research to specify the alternative hypotheses that could compete with the null hypothesis position. When no such knowledge is believed to be sufficiently available, or there are doubts in how best the alternative hypothesis distribution could be specified, we recommend the default alternative specification that we described in the present study (see principle 1). It is also possible that a researcher uses a number of reasonable alternative hypothesis distributions, and then check the stability of her/his Bayesian results. Thus, Bayesian hypothesis testing requires researchers' involvement and transparency at every step of the inference-making process.

Second, Bayesian thinking offers a philosophically sounder approach than null hypothesis testing regarding the inference process. Specifically, the null hypothesis testing process to reject a null hypothesis can be summarized by the following sentential logic:

Premise 1: If H_0 is true, then observation *D* is *unlikely* to happen
Premise 2: Observation *D* happened

Conclusion: Null Hypothesis is *probably* not true (i.e., $p < .05$, decision: reject H_0)

Vulnerability of null hypothesis testing logic can be shown using simple examples that following this logic can lead to erroneous rejection of a valid null hypothesis (see Cohen, 1994; Pollard & Richardson, 1987). Consider the following example:

Premise 1: If Bob is an American (H_0), then it is unlikely that Bob is a U.S. Senator (*Expectation under* H_0)

Premise 2: Bob is a U.S. Senator (*Observation*)

Conclusion: Bob is probably not an American! (i.e., $p < .05$, reject H_0)

The null hypothesis is a reasonable hypothesis; not many Americans are U.S. Senators. However, following the null hypothesis testing logic, we must reject the "Bob is an American" hypothesis! This is because the observation "Bob is a U.S. Senator" was not expected under the null hypothesis, and there is no other competing hypothesis under which to evaluate the same observation. Specifically, if instead of following the null hypothesis testing logic, one could specify an alternative under which the observation could be evaluated, the erroneous conclusion would not be made. For example, under the *alternative* hypothesis that "Bob is not an American", the observation that "Bob is a U.S. Senator" has a "zero" probability:

$$P(\text{Bob is a U.S. Senator } \textbf{given that} \text{ Bob is not an American}) = .00000000 \quad (1)$$

However, out of roughly 300,000,000 Americans, only 100 are U.S. Senators. Thus, under the null hypothesis that "Bob is an American" the observation that "Bob is a U.S. Senator" results in the following probability:

$$P(\text{Bob is a U.S. Senator } \textbf{given that} \text{ Bob is an American}) = .00000033 \quad (2)$$

Now, forming a simple ratio (just like Bayes factors) comparing the observation under the null hypothesis to that under the alternative makes clear that "Bob is an American" (the null hypothesis) is infinitely more likely than "Bob is not American" (i.e., .00000033 / 0 = + Infinity). Note that the *p*-value associated with same problem erroneously led to the rejection of the null hypothesis that "Bob is an American".

Thus, specification of alternatives and comparing the observed result under the alternative with that under the null is the philosophical advantage of the Bayesian thinking that, when applied to the inference process, helps avoiding incorrect rejection of null hypotheses.

Conclusion

Testing of hypotheses is often performed to distinguish the random findings (noise) from replicable ones (signal). The common use of *p*-values does not allow for the reliable detection of the true signals in the field. We proposed an empirically-informed modification for better use of *p*-values. We also introduced a new way for the reliable detection of the true signals in the field. This new way is based on Bayesian hypothesis testing and, instead of a *p*-value, results in a Bayes factor. We hope that the present study has offered solutions as to how improve our ability to detect true signals. Because only after the existence of a true signal is ensured, can one proceed to measuring the size of the obtained effect and thereby meaningfully contribute to L2 theory and practice.

Notes

A supplementary document for the present study is found at: https://github.com/izeh/l/blob/master/i.pdf where further tools and L2 research examples are discussed.

CHAPTER IV

THE BAYESIAN REVOLUTION IN L2 RESEARCH: AN APPLIED APPROACH

Overview

Frequentist methods have long-since dominated in quantitative L2 research (Norris, 2015). Recently, however, a number of fields have begun to embrace an alternative known as the Bayesian method (see e.g., Kruschke, Aguinis, & Joo, 2012). Using an open-source approach, this article provides an applied, non-technical rationale for Bayesian methods in L2 research. Specifically, we take three steps to achieve our goal. First, we compare the conceptual underpinning of Bayesian and Frequentist methods. Second, using real as well as carefully simulated examples, we introduce and apply Bayesian methods to estimate effect sizes from t-test designs. Third, to promote the use of Bayesian methods in L2 research, we introduce a free, web-accessed, point-and-click software package (rnorouzian.shinyapps.io/bayesian-t-tests/) as well as a suite of flexible R functions (https://github.com/rnorouzian/i/blob/master/i.r). Additionally, we demonstrate Bayesian methods for secondary analysis. Practical and theoretical dimensions of a "Bayesian revolution" for L2 research are discussed.

Introduction

Recent years have seen repeated calls to reform the conventional data analysis practices in the social and behavioral sciences (e.g., Dienes & Mclatchie, 2017; Etz & Vandekerckhove, in press; Kruschke & Liddell, 2017; Morey, Romeijn, et al., 2016). Most prominent among these calls, however, has been one to shift emphasis away from *Frequentist* methods to *Bayesian* methods. Three critical ingredients are required for such a shift to take place in L2 research. First, in order to embrace the Bayesian method, we would need to address the difference between

the conventional, Frequentist method and the Bayesian method. Second, Bayesian methods are to be adapted to be used with a commonly employed set of designs in L2 research (e.g., t-test designs). Third, and as a very practical matter, software packages that handle Bayesian analyses must be available to a wide audience of users. It is the goal of this article to address these three key issues and, in doing so, to encourage and enable the use of Bayesian methods in L2 research. All the discussions are accompanied by informationally-rich visuals, and various demonstrations to establish the critical links needed to understand the basics of Bayesian methods with minimal use of technical terms or mathematical expressions. Additionally, following an open-science approach, all the tools, data, and scripts to reproduce the visuals and replicate the analyses are made publicly available to the reader.

Frequentist and Bayesian Methods: An Introduction

To appreciate the difference between the Frequentist and the Bayesian methods, it is best to apply these methods to a simple research problem. Suppose a researcher administers a single-item survey to determine the *real proportion* of language minority families in a state with a large population of English Language Learners (ELLs) that prefer *bilingual* education (*B*) over *monolingual* education (*M*) for their children (e.g., Bedore, Peña, Joyner, & Macken, 2011; Farruggio, 2010; Ramos, 2007). In this case, parents' preference for the "bilingual" or the "monolingual" (i.e., English-only) education indicates the binary nature of the data that is sought. Given the available resources, the responses from 100 randomly selected parents are collected, 55 of whom prefer "*B*". Thus, the *obtained proportion* of the parents that prefer the bilingual education in this sample is 55%. By contrast, 45% of the parents prefer "*M*" for their children. Also, the 95% confidence interval values for the obtained proportion (i.e., 55%) of parents preferring the "*B*" are: [44.72%, 64.96%]. Figure 15 shows the proportion of preferences

for the bilingual education (*B*) as each parent in the sample (n = 100) responds (i.e., "*B*" or "*M*") to the survey question (to explore Figure 15 see https://github.com/izeh/i/blob/master/1.r).

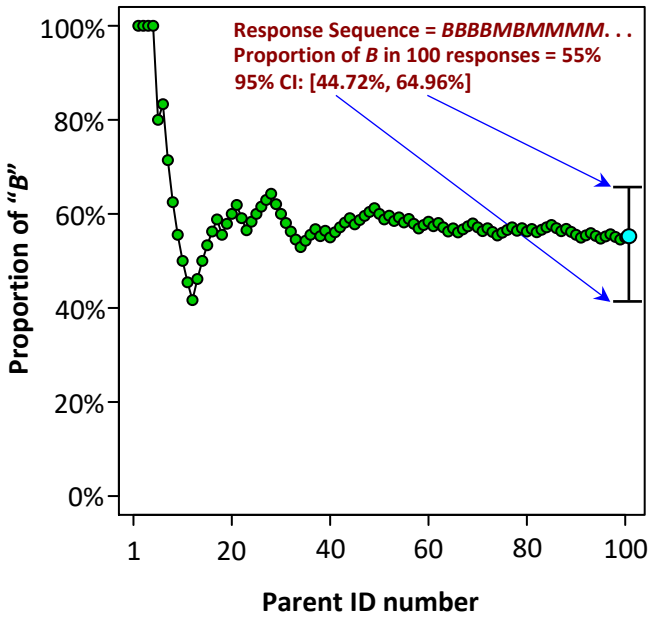

Figure 15. Proportion of preferences for bilingual education. "B" denotes preference for bilingual education and "M" denotes preference for monolingual education.

At this point, the critical question is: Given that we have data from only 100 parents in the state, can we discover the *real* proportion of preferences for bilingual education *in the entire state*? This question has a Frequentist as well as a Bayesian answer.

From the Frequentist perspective, the answer to this question relies on the Frequentist theory. According to this theory, there is surely one objective answer to the question above. However, there will always be uncertainty in any one answer (i.e., point estimate; here 55%) obtained from any one study with a limited sample size (e.g., 100 parents). To incorporate this uncertainty in any obtained answer, Frequentists use a confidence interval (CI) whose interpretation requires close attention. For example, the 95% Frequentist confidence interval of

[44.72%, 64.96%] obtained from our above survey (see Figure 15) "would indicate that over long-run frequencies [i.e., infinitely many repetitions of the survey], 95% of the confidence intervals constructed in this manner (e.g., with the same sample size, etc.) would contain the true population value" (Depaoli & van de Schoot, 2017, p. 257). To better understand the nature of this interpretation, Figure 16 shows a possible set of results from only 20 such repetitions (to explore Figure 16 see https://github.com/izeh/i/blob/master/2.r). The filled circles represent the observed proportion of the parents that prefer bilingual education in each of these 20 repetitions of the survey. The solid horizontal lines passing through the filled circles are the 95% confidence intervals for the obtained proportion of preferences for "*B*" in each of these 20 repetitions of the bilingual education survey.

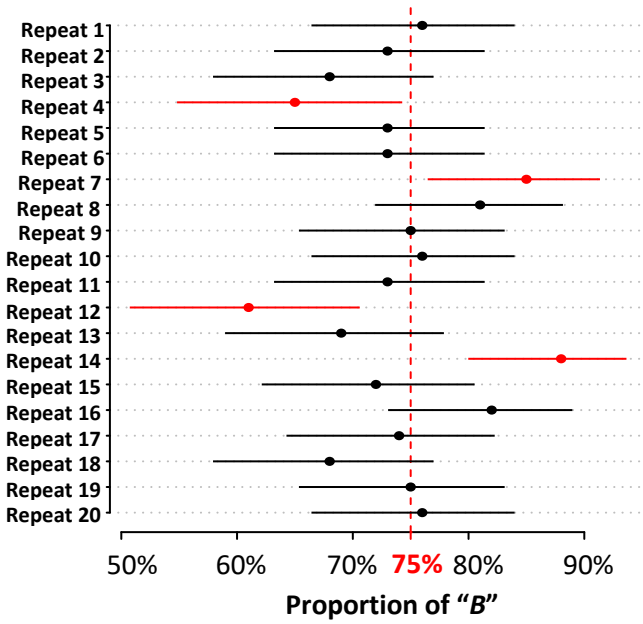

Figure 16. Twenty repetitions of the same bilingual education survey. The vertical red line represents the real (i.e., state-wide) proportion of preferences for bilingual education.

Let us assume for the sake of this demonstration that the real proportion of preferences for bilingual education in the population of parents is 75% (as shown by the vertical red line in

Figure 16). In this case, some of the obtained proportions (filled circles) in these 20 repetitions have either egregiously underestimated or overestimated the real proportion of preferences for bilingual education. These observed proportions are indicated in red as are their associated 95% confidence intervals, which do not contain the real proportion of preferences for bilingual education (i.e., 75%). Of course, 20 repetitions are not infinitely many repetitions. In theory, if repeated infinitely many times, 95% of the obtained confidence intervals will contain the real proportion of preferences for "*B*" that our researcher is interested in. Based on this perspective, the Frequentist answer to the *critical question* relies on a procedure that in the long run can be correct with a specified correction rate (e.g., 95%) and a specified error rate (e.g., 5%). Consequently, the so-called 95% confidence level often attached to an obtained CI in reality applies to a Frequentist, long-run procedure in which infinitely many intervals are assumed; it does not denote that a single interval obtained from a single study has captured the population value with 95% certainty (see Thompson, 2006, p. 204).

From the Bayesian perspective, however, this long-run procedure and the subsequent interpretation is considered unnecessarily complex. That is, such a Frequentist interpretation not only is not desired, but it also could be a source of confusion for a researcher wanting to interpret her/his single study's obtained results. Surely, what one seeks to have is X% certainty that a single obtained interval from her/his study has captured the population value.

The Bayesian method does not require thinking in Frequentist terms. Rather, it starts from the position that when a parameter is unknown (e.g., proportion of parents preferring "*B*"), then it is wiser to think of it as a variable (rather than having a single answer as in the Frequentist method) with a full range of possibilities governing its magnitude. As one of the ways to apply this view to our bilingual survey from above, the Bayesian method might begin by asking our

researcher to use the prior empirical findings relevant to the phenomenon under study, and/or the theoretically defensible expectations for the phenomenon under study to define an expected range for the *real* proportion of the preferences for "*B*" prior to conducting the survey. Given such knowledge, some of the values in this expected range may be more strongly expected and some less. The resultant expected range along with the weights given to the individual values in it lead to the formation of a "prior" distribution. For example, a review of past literature might reveal that (a) the proportion of language minority parents that prefer bilingual education has been varying between 60% and 80% in the state of interest, (b) higher literacy rates, and socio-economic status of the parents in language minority families have been reported, and (c) the long-term efforts and investments in promoting bilingual education in that state have been constantly increasing. Based on this knowledge, the values of proportion found to be smaller than 60% or larger than 80%, although possible, are logically less likely to represent the real proportion of preferences for "*B*" in the population of parents. Figure 17 shows a possible prior distribution (see next section for prior appropriateness) that would match the researcher's expectations described above (to explore Figure 17 see https://github.com/izeh/i/blob/master/3.r). Displayed for better visualization, the upward-pointing arrows in the middle denote the higher weights given to individual values between 60% and 80%. By contrast, the downward-pointing arrows denote the successively lower weights assigned to individual values outside 60% and 80%. Such a weighting scheme often results in prior distributions that resemble a bell-curve of some kind peaked over the expected range (here 60% - 80%).

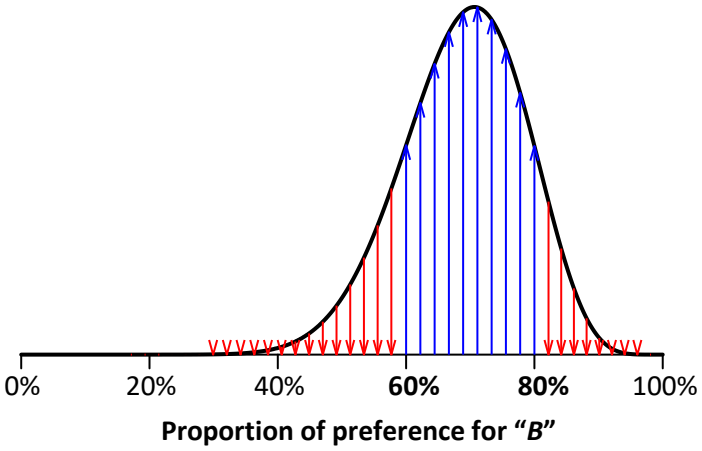

Figure 17. Prior distribution for the proportion of preference for bilingual education.

Now that the prior distribution is at hand, the next step is to obtain the likelihood function for the obtained proportion of preferences for "*B*". The likelihood function is easy to obtain because, depending on the nature of the study data, the likelihood functions are either well known or easy to construct. In our case, because the nature of the survey data is binary (i.e., "*B*" or "*M*"), the likelihood function is a "Binomial" one. All we need to do is to input the number of parents who preferred "*B*" (i.e., 55), and the total number of parents surveyed (i.e., 100) to a Binomial formula, and indicate the place for the unknown proportion of preferences for "*B*" in the formula by an "x" perhaps using a software package (see https://github.com/izeh/i/blob/master/4.r for an R implementation). Figure 18 shows the likelihood function for our example (to explore Figure 18 see https://github.com/izeh/i/blob/master/5.r).

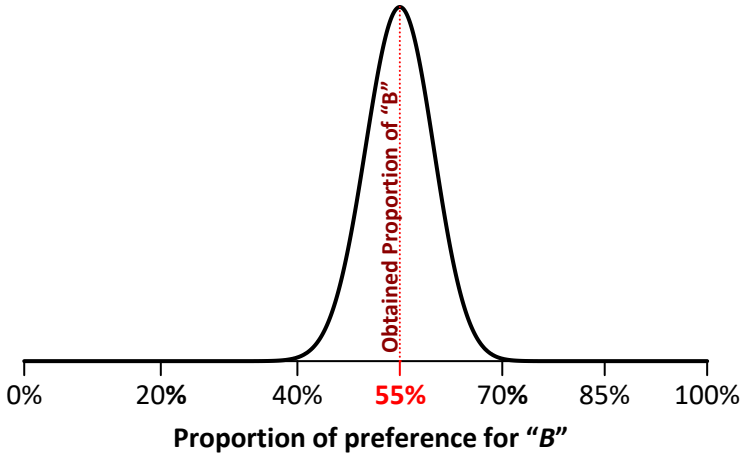

Figure 18. Likelihood function for the proportion of preference for bilingual education.

In terms of weighting, the likelihood function automatically assigns the highest weight to the obtained proportion of "*B*" (i.e., 55%). This is almost always the case because, as implied earlier, likelihood functions are simply fixed, well-known formulas that operate solely on the basis of the obtained data. Thus, they recognize the obtained proportion of preferences for "*B*" as the most likely estimate of the real proportion of preferences for "*B*" in the population of parents and all other possible estimates further away from this estimate as successively less and less likely.

Now that we have the two essential ingredients of a Bayesian method (i.e., prior and likelihood), it is time for the Bayesian mantra:

$$\textit{Prior} \times \textit{Likelihood} \propto \textit{Bayesian Result} \qquad (1)$$

where "$\propto$" (is proportional to) denotes the fact that a Bayesian result from this equation remains proportional to its proper form until scaled by a normalizing constant (see Gelman, Carlin, Stern, & Rubin, 2014). For simplicity's sake, the reader may take "$\propto$" as "=". Equation 1 is the only equation in Bayesian methods applied to ANY research problem. And the Bayesian result

obtained is the only result that an expert researcher will need to describe and interpret. At no point will one need to refer to the infinitely many repetitions [i.e., long-run frequencies] of the exact same survey necessary under the Frequentist paradigm. Essential to know is that the Bayesian result is better known as the "*Posterior*". Per our Bayesian mantra, the posterior is obtained by multiplying the prior distribution by the likelihood function. Figure 19 illustrates this multiplication to obtain the posterior for our example (to explore Figure 19 see https://github.com/izeh/i/blob/master/6.r).

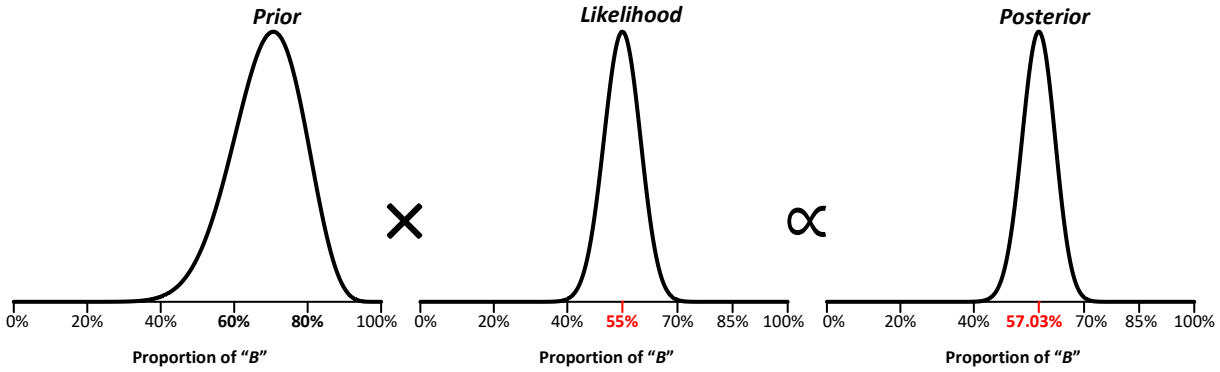

Figure 19. Steps to obtaining the Bayesian result (i.e., posterior) for estimating the proportion of preferences for bilingual education.

At this point, we can more precisely concentrate on our obtained posterior. Figure 20 shows the posterior for our example with more details added to it to help the accurate interpretation of our Bayesian results (to explore Figure 20 see https://github.com/izeh/i/blob/master/7.r). As is discussed next, these details provide direct insights into finding out what the real proportion of parents' preferences for "*B*" in the entire state (population) could be.

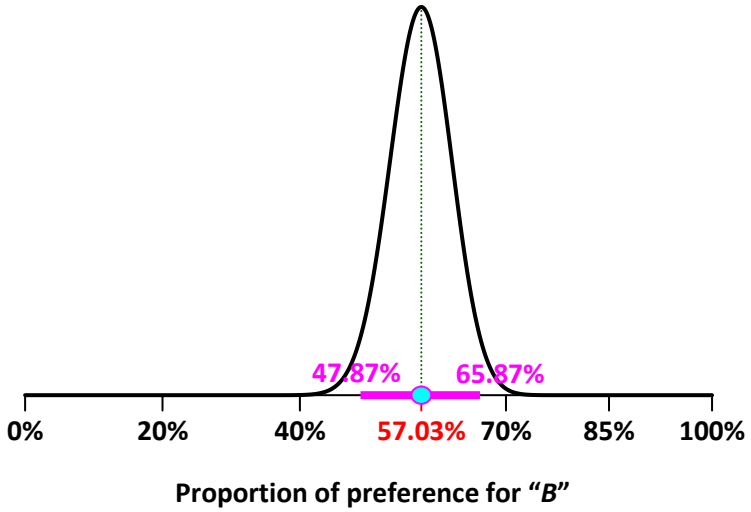

Figure 20. Posterior distribution for the proportion of preference for bilingual education.

The confidence interval-like horizontal line segment at the bottom of Figure 20 covers 95% of the highly weighted areas of the posterior. Values of proportion inside this 95% range are more credibly likely to represent the real proportion of preferences for "*B*" than others in the posterior. As such, this confidence interval-like range is often referred to as a "Credible Interval". Such a credible interval is quite helpful in describing and interpreting a posterior.

With the help of this credible interval, our researcher is now able to state that the *real* proportion of preferences for "*B*" in the population of parents could credibly range between 47.87% and 65.87%. Notice the brevity and the directness with which a single obtained Bayesian credible interval describes the candidate values representing the real proportion of preferences for "*B*" in the population of parents. Also, note that the values of proportion closer to the center (filled circle) of this credible interval are still more likely to represent the real proportion of preferences for "*B*" in the population of parents than others. As we discussed above, none of

these informative properties could be interpreted from a single obtained Frequentist confidence interval.

Putting Priors to the Test

In the previous section, we discussed that a Bayesian method starts by choosing a prior. Often, however, the prior distribution picked for estimating a parameter must pass a test for it to prove plausible. There are several ways of evaluating the plausibility of a prior depending on the nature of the parameter at hand, as well as the type of prior selected. In the case of estimating the proportion of parents supporting bilingual education described in the previous section, we used a type of prior that belonged to the "Beta" family. Beta priors are naturally bounded between 0 (or 0%) and 1 (100%). Thus, they could be one possible prior type for estimating a parameter (e.g., proportion, eta squared effect size; see Norouzian & Plonsky, in press) that ranges between 0 (or 0%) and 1 (or 100%). Albert (2009) suggests that a beta prior distribution may be specified "through statements about the percentiles of the distribution" (p. 23). In non-technical terms, even if past research shows that the proportion of language minority parents that prefer "*B*" for their children varies between 60% and 80%, we might not exactly know how well such findings do at representing the true proportions of preference for "*B*" across the state. That said, it would be perhaps unrealistic to think that the degree of representativeness for previous findings could be fairly high or fairly low. If one chooses to express this degree of representativeness in percentages (i.e., from not representative; 0% to completely representative; 100%), then conservatism dictates that a reasonable range for this representativeness could start from mid-low (e.g., 40%) to mid-high (e.g., 60%). This means we can specify different priors that separately take this range for representativeness (i.e., 40%, . . ., 50%, . . ., 60%) for the past research findings into consideration and then obtain the corresponding posteriors under all such priors. To

do this, we suggest using our suite of R functions accessible by running the following in R or RStudio®:

```
source("https://raw.githubusercontent.com/rnorouzian/i/master/i.r")
```

The first step would be to obtain a set of priors (e.g., 10) that incorporate the range of 40% to 60% for the representativeness of past survey findings (i.e., 60% - 80%). The R function "`beta.id`" is designed for this purpose:

```
I = beta.id(Low = "60%", High = "80%", Cover = seq(.4, .6, l = 10))
```

Now, we have 10 different prior specifications each of which incorporating in it a different level of representativeness (i.e., 40% (or .4) - 60% (or .6)) for the past research findings (i.e., 60% - 80%), all stored in "`I`". Each of these 10 prior distributions can be individually inspected using the R function "`prop.priors`". For example, to see the last (i.e., 10th) prior which was also used in the previous section (see Figure 17) we can use:

```
prop.priors(a = I$a[10], b = I$b[10], dist.name = "dbeta", show.prior = TRUE)
```

Or to see the first prior displayed below in Figure 21 we can use:

```
prop.priors(a = I$a[1], b = I$b[1], dist.name = "dbeta", show.prior = TRUE)
```

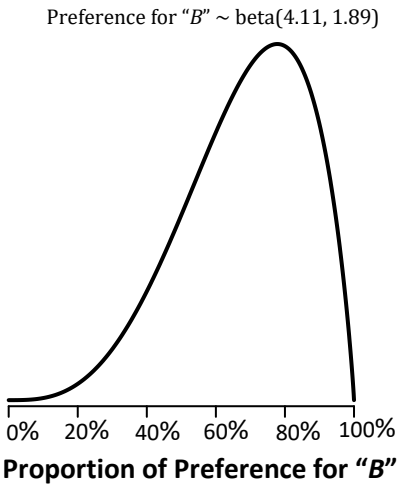

Figure 21. The first prior distribution for the preference for "B".

The next step is to obtain the Bayesian result (i.e., posterior) using all these different priors one at a time and compare their resultant 95% credible intervals. Egregious differences among the 95% credible intervals would indicate that our results are sensitive to uncertainty about the representativeness of past research findings. When such notable differences occur, we have failed the test of robustness under our choices of prior. To perform these analyses and compare their 95% credible intervals, we can once again use the function "`prop.priors`":

```
prop.priors(n = 100, yes = 55, a = I$a, b = I$b, dist.name = "dbeta", scale =
.1, top = 1.055)
```

The Bayesian posteriors along with their 95% credible intervals are provided in Figure 22.

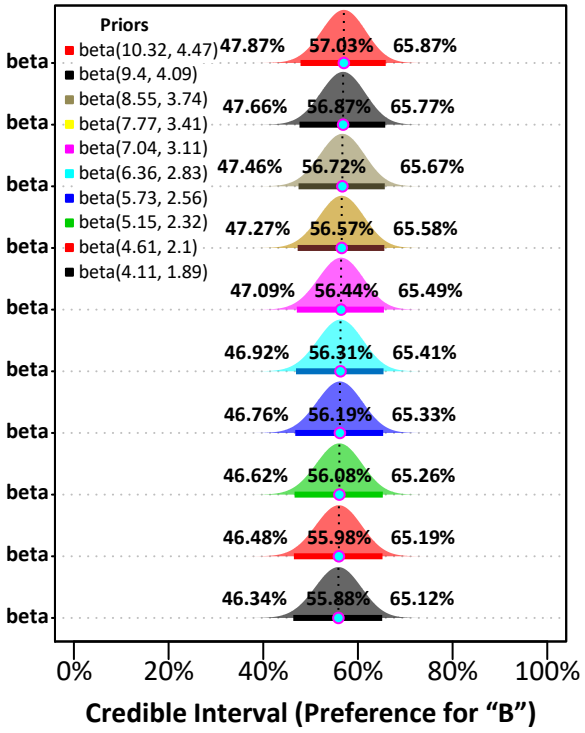

Figure 22. Bayesian posterior credible intervals under various Beta priors.

As can be seen, although the priors are different, the posteriors are fairly aligned with each other with no egregious differences among their 95% credible intervals. After taking a reasonable set of candidate priors, the visual inspection of the credible intervals is critical in demonstrating the robustness of results under the choice of priors. In the following sections, we will see that in various situations, the nature of the parameter at hand and the type of common priors employed to describe it allow us to conduct other forms of robustness analyses.

Skepticism and Lack Of Prior Knowledge

In some cases, prior knowledge is absent, diminished, or its credibility might be under question. In such situations, priors that concentrate their weight on (i.e., are peaked over) a certain range for a parameter may be easily prone to biasing a Bayesian result (i.e., posterior).

Defining a prior distribution that expresses the state of neutrality or a lack of knowledge is one way to avoid such potential biases. Several seminal works have looked at this issue from perspectives that require both space and technical background knowledge (e.g., *information theory*; Jaynes, 2003; *invariance to transformation*; Jeffreys, 1961; *contribution of prior measured in datapoints*; Liang, Paulo, Molina, Clyde, & Berger, 2008). In this introductory discussion, however, we tend to simply refer to priors that express a lack of or minimal prior knowledge as "broad" or "minimally informative". As we shall see, reasonableness must always play a role in defining such priors depending on the nature of the parameter at hand and the type of prior meant to be used with it.

Let us use an actual example in which lack of prior knowledge is best evident. In chapter II, we surveyed the application of two effect size variants, eta-squared (η^2) and partial eta-squared (η_p^2), in a sample of 156 uses of these two effect sizes from various L2 journals published between 2005 and 2015. Surprisingly, we found that in 34 cases, the primary authors of the published L2 research had erroneously reported and interpreted partial eta-squared effect size in place of eta-squared effect size (for consequences of this misreporting see Ch. 2). This indicated that 21.79% (i.e., $\frac{34}{156} \times 100$) of the collected sample of L2 studies had misreported these two effect size variants. But assuming that L2 researchers did not largely learn how to use or distinguish these two variants of effect size from each other (i.e., independence of observation), the question of interest is: What is the actual proportion of this erroneous reporting and how prevalent it is across all L2 studies that report these two measures of effect size? Since this was the first survey of this type in L2 research, no specific prior knowledge in L2 research is available to refer to as a knowledge base. Also, similar studies in sister fields such as psychology (Pierce et al., 2004) and communication (Levine & Hullett, 2002) tend to only narratively

describe the existence of a confusion in using the two variants of effect size among researchers in their respective fields without offering much quantifiable evidence. With such highly restricted knowledge base, defining an informative prior distribution may not be possible. What is needed, however, is a "broad" or "minimally informative" prior distribution that assigns almost equal weights to most possible values (i.e., 0% - 100%) representing the misreporting rate of eta- and partial eta-squared as two measures of effect size in L2 research. Many Bayesian analysts have argued that it is always wise to exclude extremely unrealistic values that may not represent the possible magnitude of the parameter (here the misreporting rate of the two effect sizes) under estimation (e.g., Gelman et al., 2014; Kruschke, 2015; McElreath, 2016). In our case, to assume that the misreporting rate of the two measures of effect sizes, eta- and partial eta-squared, in L2 research could be close to ~0% or ~100% is unequivocally unrealistic. A broad prior then could be one (a) whose effective weight concentration spans over most possible values for the misreporting rate excluding the unrealistic ones (e.g., ~0% and ~100%) and thus (b) which is not skewed toward a particular side in the parameter range (i.e., is symmetric slightly pivoting on 50%). One such broad prior is shown in Figure 23. Figure 23 can be easily replicated using our R function "`prop.update`":

```
prop.update(a = 1.2, b = 1.2, show.prior = TRUE, prior.scale = .5, top = 1.6)
```

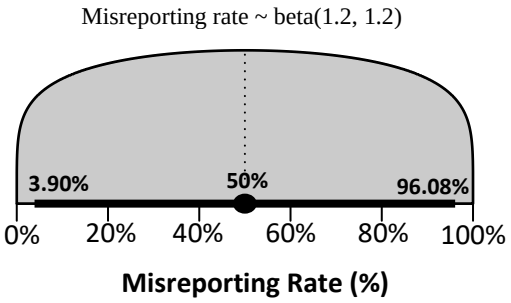

Figure 23. A broad prior expressing lack of knowledge for misreporting rate.

With this broad prior at hand, we can proceed with estimating the proportion of misreporting eta-squared (η^2) and partial eta-squared (η_p^2), as two measures of effect size, in published L2 research. The function "`prop.update`" can be called again to see how our broad prior knowledge is changed in light of the 34 cases of effect size misreporting out of 156 applications of these two effect size variants found in Chapter II :

```
prop.update(yes = 34, n = 156, a = 1.2, b = 1.2, scale = .2, top = 5, prior.scale =
1.3)
```

The result of our analysis is displayed in Figure 24.

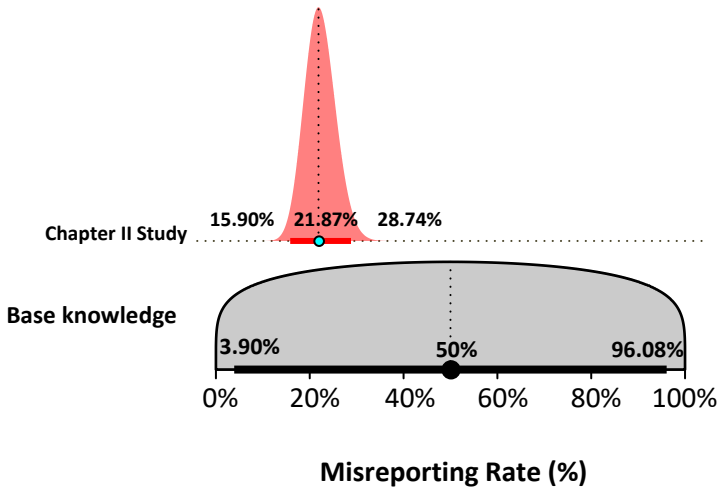

Figure 24. Updating a broad knowledge base in light of misreporting cases found in Chapter II.

Our analysis returns a misreporting rate of 15.90% - 28.74%. But is the Bayesian result obtained robust to the choice of prior? We can again put our choice of prior to the test and visually examine the robustness of our Bayesian results. Since this time (as opposed to the case of the bilingual survey in the previous section) no specific source of prior knowledge is available, and we have used a "broad" prior, we can select from a variety of different families of priors in

addition to “Beta”. These other families are first positioned such that they, just like our Beta prior, cover the entire range of 0% to 100% for misreporting rate (i.e., our parameter) of the two effect sizes but then cut for any additional coverage for values that do not fall within 0% to 100%. For example, the familiar Normal distribution which is naturally boundless (i.e., goes from –infinity to +infinity) is first positioned so that, like our Beta prior, it reflects neutrality and symmetry (e.g., pivoting around .5 or 50%) but then cut everywhere except for areas falling between 0 and 1. Here we use two other families of distributions in addition to “Beta”, namely “Normal” and “Cauchy” (see next section on effect size). Both of these distributions are naturally symmetric, but we can position them between 0 (or 0%) and 1 (or 100%) while pivoting them around .5 (or 50%). Note that in R and some other software packages (e.g., JAGS, WinBUGS), distribution names start with a “`d`” (standing for *density*). Examples include “`dnorm`” for Normal, “`dcauchy`” for Cauchy, and “`dbeta`” for Beta distribution. We can use the R function “`prop.priors`” to test these three prior families all at once:

```
prop.priors(a = c(1.2, .5, .5), b = c(1.2, 1, 1), dist.name = c("dbeta", "dnorm",
"dcauchy"), scale = .075, top = 1.4, yes = 34, n = 156)
```

The resultant posteriors along with their 95% credible intervals are shown in Figure 25.

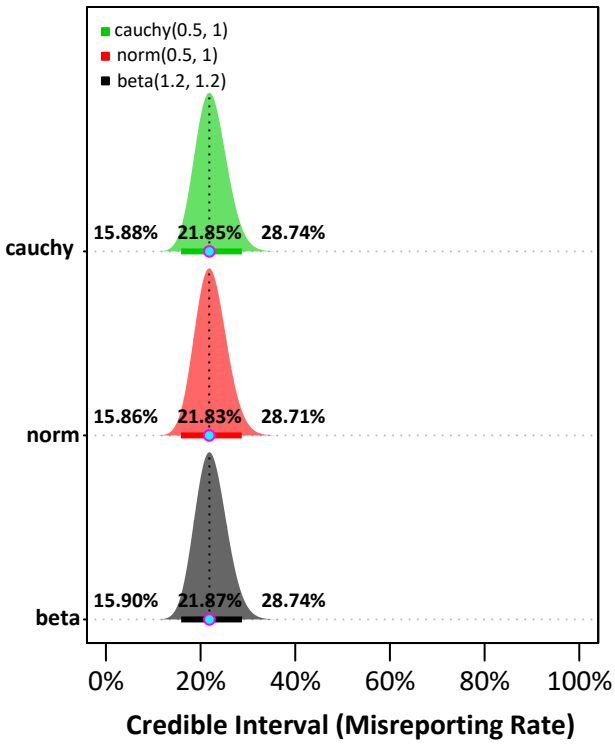

Figure 25. Posterior results under different families of priors.

As shown in Figure 25, the results under these three families of priors barely change. Indeed, even if the width (spreadoutness) of Normal and the Cauchy priors are increased by a factor of 10 (i.e., from 1 to 10) no major change in the posteriors occurs:

```
prop.priors(yes = 34, n = 156, a = c(1.2, .5, .5), b = c(1.2, 10, 10),
dist.name = c("dbeta", "dnorm", "dcauchy"), scale = .075, top = 1.4)
```

Figure 26 shows the result of the ten-fold increase in the width of the Normal and Cauchy priors. Because of these fairly stable results, it is safe to believe that the misreporting rate of the two measures of effect size, eta- and partial eta-squared, in L2 quantitative research ranges between ~15.9% and ~28.7% as indicated by our 95% high-density credible intervals.

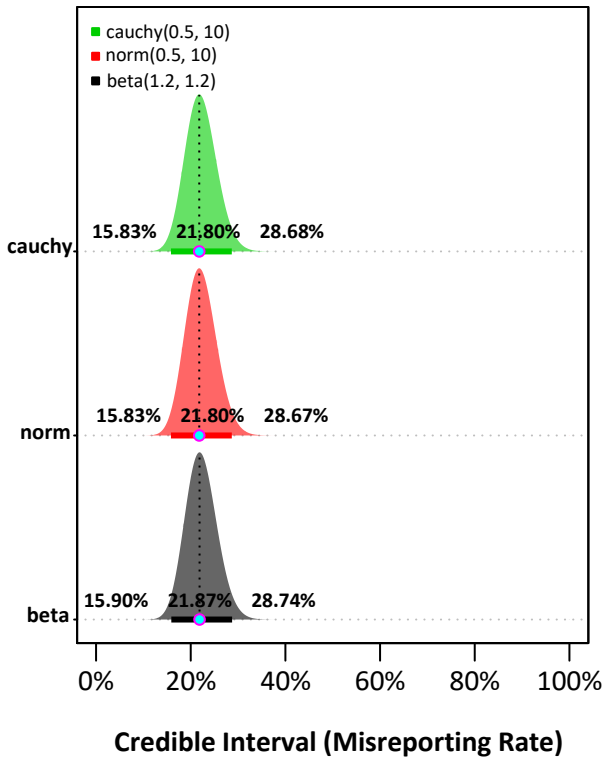

Figure 26. The result of a ten-fold increase in the width of the Normal and Cauchy priors.

Letting Priors Arise

Many of us as applied linguists would agree that the knowledge generated from our studies must play a role in informing future replication efforts (see Marsden, Morgan-Short, Thompson, & Abugaber, in press; Porte, 2012 for a fuller discussion of replication in L2 research). Bayesian methods are uniquely designed so that each future replication could build on the knowledge generated by any number of replication works conducted before it (see Note). This feature of Bayesian methods is so boundless that it is often said that *yesterday's posterior is today's prior* (see Lindley, 2000). To better see this in action, suppose that two other surveys at two different points in time had targeted the preference of language minority parents for bilingual education before our current survey discussed in the previous section. A Bayesian

framework allows us to cumulatively incorporate these two other surveys' results into our current survey in a step-wise fashion. That is, one can (a) start with a broad knowledge base, (b) use that broad knowledge base as a prior for the first available survey to obtain the posterior, (c) use that posterior as prior for the second survey to obtain a second posterior, and finally (d) use the posterior of the second survey as prior for the current survey, obtain the final posterior, and describe it using 95% credible intervals as the most current result. This step-wise Bayesian updating process is implemented in our R function "`prop.update`". To use the function, suppose the first and oldest survey came from 70 parents, 27 (39%) of whom preferred bilingual education, and the second survey was based on 84 parents, 31 (37%) of whom favored bilingual education for their ELLs (English Language Learners). Recall that our current survey (see Figure 15) showed that 55 out of the 100 parents support bilingual education. Now, a call to function "`prop.update`" can be made to incorporate both of the previous surveys' results in our current replication survey using a broad prior base:

```
prop.update(n = c(70, 84, 100), yes = c(27, 31, 55), a = 1.2, b = 1.2, dist.name =
"dbeta", scale = .086, top = 1.6)
```

The result of this step-wise Bayesian updating is shown in Figure 27. As can be seen, we started from a very broad knowledge base that allowed us to believe almost any proportion (0% - 100%) could be a candidate value for representing the proportion of parents that prefer bilingual education. But then this broad knowledge base was updated by the first survey conducted on the matter. Still, the second survey built on both the initial knowledge base as well as the result of the first survey and this updating went on until the most recent survey was carried out. Other than *letting the priors arise* in the process rather than specifying them in advance, the end result of such updating processes is one final posterior that, founded upon previous replication

attempts, will concentrate narrowly on the proportion values (or any other parameter of interest) that represent the parents' view regarding bilingual education. In the later sections, we will return to this updating process to generate a prior based on the findings of previous replication research and extend it to situations where our parameter of interest is a standardized mean difference effect size (i.e., Cohen's *d*).

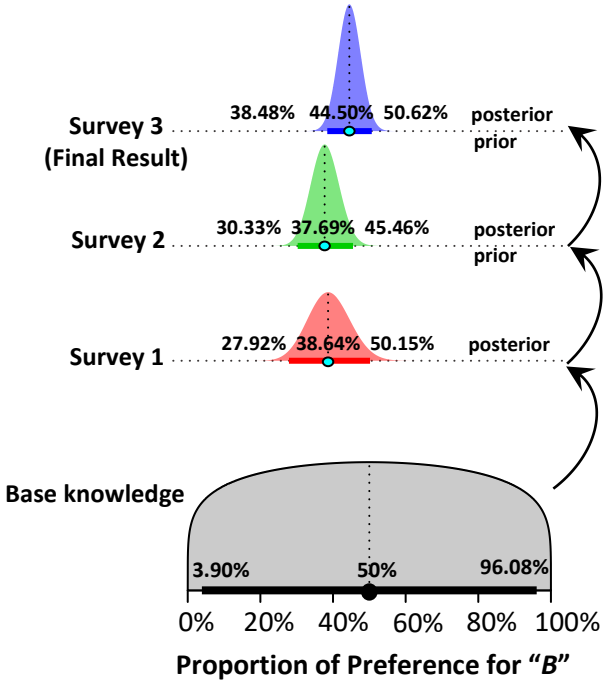

Figure 27. Step-wise updating of three bilingual education surveys using a broad prior.

In the next section, we present an application of Bayesian methods for one of the most commonly employed statistical analyses in L2 research, the *t-test*, (Larson-Hall, 2016). Through the Bayesian method, we add a new application to t-tests so that in addition to being used for testing the validity of a null hypothesis, t-tests become vehicles for estimation of the real effect (i.e., effect size) of a treatment. In addition to a repository of highly flexible R functions, we also

introduce a free, online, point-and-click software package (rnorouzian.shinyapps.io/bayesian-t-tests/) that painlessly automates some of the steps involved. As will be shown, this Bayesian application of t-tests can also be used for the Bayesian estimation of the real effect of a treatment (i.e., effect size) from a previously published study using only the basic information available in that study.

Bayesian Methods as Applied in t-test Designs in L2 Research

The Bayesian method discussed in the previous section also applies to designs that use *t-tests,* which are ubiquitous in L2 research (Larson-Hall, 2016; Linck & Cunnings, 2015). And the approach that we take to run Bayesian t-tests, an "effect size" approach, concurs in the belief that "the primary product of a research inquiry is one or more measures of effect size, not *p*-values" (Cohen, 1990, p. 1310). To be clear, t-tests are analytic tests that are used to evaluate *if there is an effect* (i.e., null hypothesis testing; *p*-value) for a treatment in pre-post designs (paired-samples t-test), experimental designs with two groups (independent samples t-test), and one-sample designs (one-sample t-test), the last of which is less commonly found in L2 research (see Larson-Hall, 2016, p. 270). The marriage of the Bayesian methods and the effect sizes from such designs allows for estimating the real size of an effect for a treatment from the above-mentioned designs. In our view, this significantly adds to the applicability and utility of t-tests in L2 research.

Let us then apply the Bayesian t-test method to a meaningful L2 research example as we did when discussing survey data in the previous sections. Suppose a researcher is interested in finding out the real effect of an L2 treatment on improving the explicit knowledge (DeKeyser, 2015; Lyster & Sato, 2013) of 60 high-intermediate English as a Foreign Language (EFL)

learners with respect to Type III conditionals (e.g., *If I had arrived earlier, I could have caught the bus*). The schematic design of this study is shown in Figure 28.

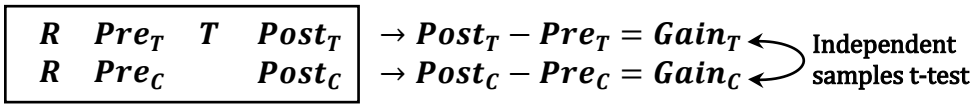

Figure 28. Pre-post-control design layout. R = Random assignment; T = Treatment; C = Control; Pre = Pre-test; Post = Post-test.

Based on this *Pre-Post-Control* design, the participants are randomly assigned to either the treatment group (n = 30) or the control group (n = 30) to protect the study outcome from some of the design's internal validity threats, e.g., regression to the mean (see Campbell & Stanley, 1963). Then, following the pre-test and treatment, the researcher administers a posttest to measure the difference in the level of the explicit knowledge of Type III conditionals gained by the two groups. To measure explicit knowledge (see Ellis, 2009), both groups are to complete an untimed error correction test (ECT) consisting of 15 sentences 10 of which contain different number grammatical errors in the use of Type III conditionals. The scoring scheme used for Type III conditionals often involves awarding a combination of half-points and whole-points depending on what feature (e.g., correcting the past modal: *"would / could / . . ."* 1 point, correcting the past participle form: *"caught"* .5 point) in the *conditional* or the *main* clause is appropriately corrected by a participant (see Izumi, Bigelow, Fujiwara, & Fearnow, 1999). In total, 25 points are allowed on the entire error correction test. Recent research (e.g., Shintani, Ellis, & Suzuki, 2014) suggests that it is reasonable to believe that this scoring scheme would result in scores complying with the assumption that such scores belong to normally shaped populations. Finally, as Campbell and Stanley (1963) indicate, for a *Pre-Post-Control* design, as

in our case, it is wise "to compute for each group pretest-posttest gain scores and to compute a *t* [i.e., independent samples *t*] between experimental and control groups on these gain scores" (p. 23). With these details in mind, let us simulate such a study, and then employ a Bayesian independent-samples t-test to estimate its possible effect. Figure 29 graphically shows the design, raw gain scores, and the immediate results of this simulated study (to explore Figure 29 see https://github.com/izeh/i/blob/master/11.r).

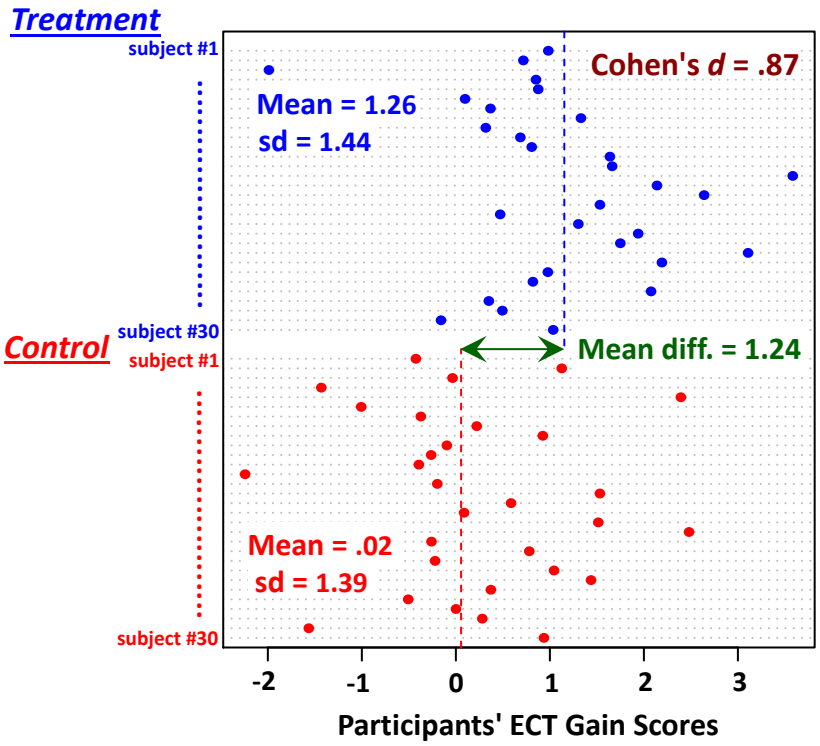

Figure 29. The design and raw gain scores (posttest – pre-test) of the participants in the simulated study. ECT = Error Correction Test. Each grey, horizontal, dotted line denotes a participant. The vertical dashed lines denote the mean of each group's gain scores. Mean diff. = difference between the means of groups' gain scores.

Table 10 presents the full descriptive and the conventional Frequentist results (e.g., confidence interval) of this study.

Table 10

Frequentist Study Results for EFL Learners in the Simulated Study (*N* = 60)

	Descriptive (Gain Scores)				Inferential	
Group	*n*	*M* (*SD*)	*ES* (*d*)	95% $CI_{(d)}$	*t* (*df*)	*p*-value
Treatment	30	1.26 (1.44)	**.87**	[.33, 1.40]	3.35 (58)	.001
Control	30	0.02 (1.39)				

Note. *M* = Mean; ES = Effect size; *d* = Cohen's *d*.

The effect size (i.e., *d* = .87) along with its 95% confidence interval (i.e., 95% $CI_{(d)}$ [.33, 1.40]) obtained from our simulated study (Table 10) are both subject to Frequentist interpretations. Recall from our discussion in the previous sections that from the Frequentist perspective these results can be theoretically seen as just one set of possible results from among many more in the long chain of repetitions of the exact same study on Type III conditionals. For example, let us assume that in reality our L2 treatment is able to produce an effect quantified by a Cohen's *d* effect size of .5, then a possible set of results from only 20 repetitions of our exact same study on Type III conditionals is presented in Figure 30 (to explore Figure 30 see https://github.com/izeh/i/blob/master/9.r).

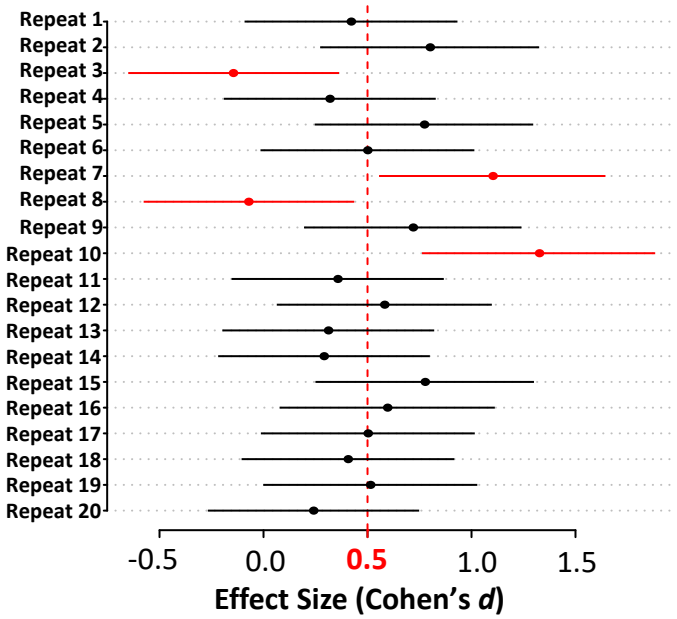

Figure 30. Twenty repetitions of the same study on Type III conditionals. The vertical red line represents the real (i.e., population) size of effect for the L2 treatment.

As with the survey example, here again some of the obtained effect sizes along with their 95% confidence intervals from these 20 repetitions fail to capture the real effect of treatment (i.e., .5), as indicated in red. And our obtained results (i.e., d = .87; 95% CI [.33, 1.40]) could be "red" results, as is the case in four of the 20 repetitions here. Again, while in the long-run, the Frequentist procedure is correct (i.e., contains the true effect) in 95% of infinitely many repetitions of such a study, this assurance does not mean that our single obtained CI from our single study on Type III conditionals contains the true effect of the treatment with 95% certainty. Now imagine what a formidable challenge as well as confusion this might create for a researcher wanting to interpret the single obtained effect size, and its 95% confidence interval; certainty in a long-run procedure rather than in the single obtained result. Certainly, here the only *critical*

question of interest is: What is the *real* effect of the L2 treatment on improving high-intermediate EFL learners' explicit knowledge of Type III conditionals?

Once again, the Bayesian method begins by asking our researcher about her/his expectation regarding the range of effect sizes reported in the previous research or the general domain of L2 research. We take a general approach here which appeals to a broader domain of L2 research. This makes such a Bayesian approach broadly accessible and provides a default prior distribution for Cohen's *d* effect size applicable to a wide range of domains in L2 research. To do so, we first draw on the results of Plonsky and Oswald (2014) who studied the magnitude of Cohen's *d* effect size in 346 primary L2 studies and 91 meta-analyses of L2. The researchers found that *d* values in L2 research could often be as large as +1. Even so, conservativism dictates that one takes a neutral position and consider that Cohen's *d* effect size theoretically can be positive and negative. As such, it is safer to consider that the expected sizes of effect could be as large as reported by Plonsky and Oswald (2014) in either a positive or negative direction (i.e., –1 and +1). Now that the range of likely effect sizes are at hand, it is time to assign higher weights to our expected range and successively lower weights to other effect size values outside this range. We use a "*Cauchy*" (named so in Augustin-Louis Cauchy's honor) weighting scheme to achieve this. A Cauchy weighting scheme, to be shown shortly, puts higher weights on the values of effect size between –1 and +1 than does the more familiar standard normal weighting scheme (see Rouder et al., 2009). The technical specifics of the resultant prior distribution of effect sizes following the Cauchy weighting scheme are well documented (Ly et al., 2016; Rouder et al., 2016). It is, however, worth noting that our prior distribution of effect sizes has a width (akin to standard deviation) of ".707", and is centered at "0" (i.e., our neutral position between positive

and negative effect size values). Figure 31 shows this recommended prior distribution of effect sizes in which the area between –1 and +1 has received the highest weights (to explore Figure 31 see https://github.com/izeh/i/blob/master/d.r). Note that in theory, Cohen's *d* effect size has no bound. That is, it can be infinitely large in either direction. However, we can all agree that effect sizes beyond ± 6 are very unlikely. Thus, the largest values of effect size displayed in Figure 31 are ± 6 with two ± ∞ signs indicating the theoretical bounds of Cohen's *d* effect size.

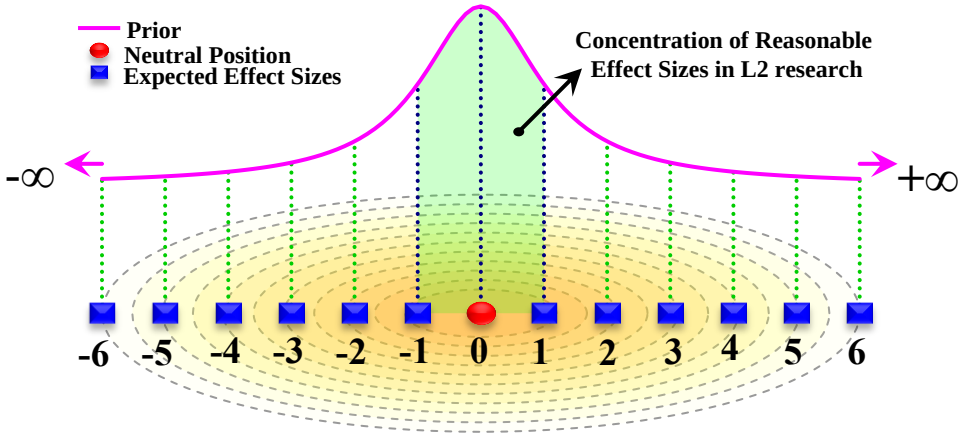

Figure 31. Recommended prior distribution for Cohen's d effect size in L2 research informed by Plonsky and Oswald (2014).

In Figure 31, the dashed oval lines represent the main weighting domain of the prior. That is, the domain within which the possible effect size values in L2 research (e.g., −6 to +6) could receive various amounts of weight. The yellow color that spreads out from within the center of the dashed oval lines fades away as we move toward the large values of effect in the tails. This is to emphasize the fact that as we move from our neutral position (i.e., "0") toward the tails, the weights assigned to the individual effect size values successively decrease.

With the prior specified, the next steps involve determining the likelihood, applying the Bayesian mantra (Equation 1) to arrive at the posterior (i.e., the Bayesian result), and then obtaining a credible interval for the effect size to help the final interpretation. However, given the wide application of t-tests in real L2 research and the challenges inherent in learning to use new statistical applications that permit Bayesian analyses, here we introduce a free, point-and-click, web-accessed software package developed by the first author of the present study to automate these processes. This software package is found at rnorouzian.shinyapps.io/bayesian-t-tests/. The software will painlessly provide the posterior and the credible interval for effect sizes for the three, common t-test designs described above. For wider flexibility in terms of using a variety of different priors and robustness checks, we also provide easy to use R functions. The software has additional Bayesian capabilities that enable performing Bayesian hypothesis testing, and replacing *p*-values with a Bayesian alternative known as a *Bayes Factor*. The issue of Bayes Factors/Bayesian Hypothesis Testing/Model Selection, however, falls outside the scope of the present study (for details see Ch. II). Figure 32 provides a snapshot of the main panel of the software.

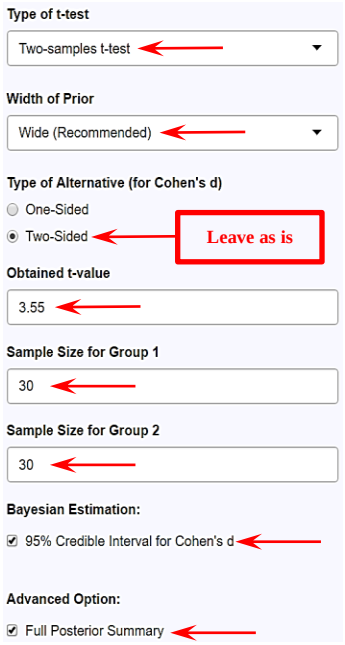

*Figure 32***.** A snapshot of the "Bayesian for t-tests" software. The red arrows indicate the settings used for the example in the text.

To use the software in our example, we do not need to provide the raw data shown in Figure 29. Rather, only the following information is required: (1) the type of t-test, (2) the width of the prior, (3) the obtained t-value, (4) the groups' sample sizes. These four pieces of information for our example study on Type III conditionals are indicated by *red arrows* in Figure 32. Figure 33 (explore the software output) summarizes the software's Bayesian result (i.e., posterior) for our running example.

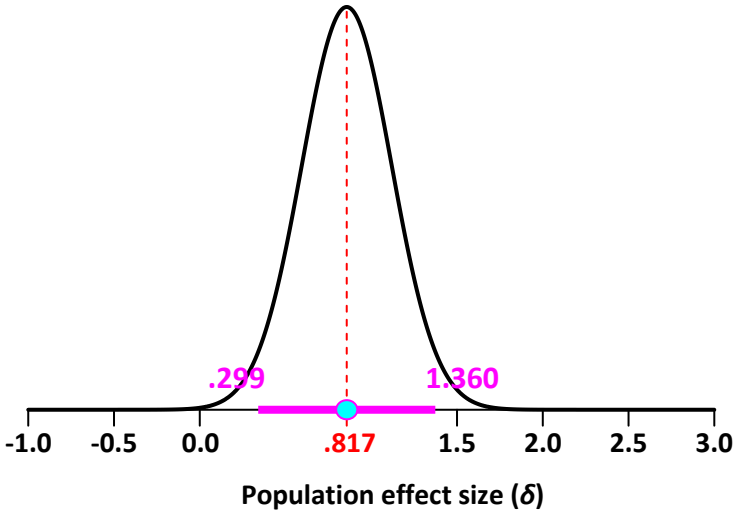

Figure 33. Posterior results for the effect of an L2 treatment on improving 60 high-intermediate EFL learners' explicit knowledge of Type III conditionals.

Now our 95% Bayesian credible interval can help us think about the *real* effect of our treatment, as measured in terms of Cohen's *d* effect size, on improving high-intermediate EFL learners' explicit knowledge of Type III conditionals. Here, we can directly state that the *real* effect size for our treatment could range between .299 to 1.360. One of the appealing features of the software is that it automatically provides the corresponding Frequentist results along with the Bayesian results. For our example, the Frequentist 95% confidence interval limits for effect size are: [.380, 1.445]. Again, this confidence interval is theoretically only one of the infinitely many possible confidence intervals that can result from repetitions of our study, and thus we cannot take its 95% confidence level as 95% certainty that this single obtained confidence interval contains the true effect of our treatment on improving explicit knowledge of Type III conditionals. Research shows that the temptation to erroneously interpret a Frequentist confidence interval as if it is a Bayesian credible interval is considerably high despite the fact

that such an interpretation is not permissible under the Frequentist framework (Albert, 2009; Gelman et al., 2014; Kruschke, 2015; McElreath, 2016).

Putting Priors on Cohen's *d* Effect Size to the Test

As noted earlier, it is always recommended and useful to test the robustness of the Bayesian result (i.e., posterior) for any research parameter against the choice of prior, and effect size is no exception. Here again the nature of effect size and type of priors commonly used with it should govern how one might want to go about choosing priors for such sensitivity analyses. Specially, the intrinsic meaning of effect size as a research result should guide us in determining (a) how wide priors on an effect size could be, and (b) where to center the priors as a pivot point. Given these two considerations, one possible way to start the robustness analysis is to use different families of priors that cover a realistic range for effect size (e.g., −6 to +6) while they might differ in distributing their weight over this realistic range. Note that too wide or too narrow specifications of prior in the case of effect size could easily lead to the assignment of undue weights to values for effect size that might not realistically need such amounts of weight. For example, prior specifications for effect size that are too narrow may unrealistically ignore effect sizes that are slightly larger than |1|, and too wide of a specification may give fairly large effect size (e.g., > |3|) more weight than required. Let us use two families of priors, namely Normal, and Cauchy. These two prior families for effect size (Cohen's *d*) could be used when they are each pivoted at "0" (a neutral position) and their width set to "1" and "1.25" (two reasonably wider settings compared to .707 used in the previous section). This plan leads to four different prior specifications: *Cauchy*(0, 1), *Normal*(0, 1), *Cauchy*(0, 1.25), *Normal*(0, 1.25).

As in the case of proportions in the previous section, the goal is to evaluate the robustness of the Bayesian result obtained in Figure 33 under four different prior specifications. To do this, we can

use function "`d.priors`" which uses the t-value (`t`), group samples sizes (`n1` or/and `n2`), pivot point for priors (`m`), and the width of prior (`s`):

```
d.priors(t = 3.55, n1 = 30, n2 = 30, m = 0, s = rep(c(1, 1.25), 2),
dist.name = c(rep("dcauchy", 2), rep("dnorm", 2)), scale = .6, top = .9)
```

The result of our analyses is illustrated in Figure 34.

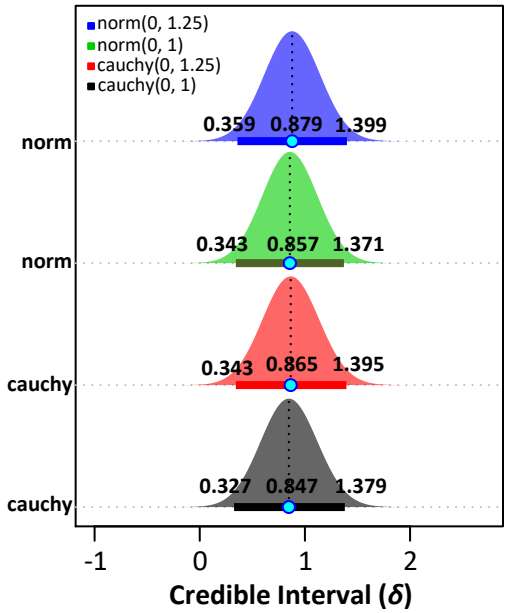

Figure 34. The credible intervals under different families and specifications of prior.

As can be seen, the 95% credible intervals under these different priors still range from ~.3 to ~1.4. Thus, it is safe to believe that under such reasonably different prior specifications (i.e., wider and of different families), our Bayesian result for our study on type III conditional is reasonably stable. The interested reader may use other families of priors such as a standard *t* distribution (`dist.name = "dt", s = 0`) with a few degrees of freedom (e.g., `m = 5`) to see that the credible intervals are still robust to this other reasonable expression of prior knowledge on the effect size in the example of Type III conditionals.

Doing Bayesian Estimation on the Published Literature: An Actual Study Example

The discussion in the previous section should imply the ease with which full Bayesian estimation of effect sizes can be performed even on previously published studies. As an even more concrete example, consider Gurzynski-Weiss and Baralt (2014). Using a pre-post design, one of the questions that the authors investigated was the effect of the interaction mode (i.e., computer-mediated communication [CMC] vs. face-to-face [FTF]) when providing 24 intermediate-level learners of Spanish as a foreign language (SFL) with opportunities to modify their output during interactional feedback episodes with their teacher. After eliciting their data via stimulated recall protocols (see Mackey & Gass, 2016), the authors conducted a paired-samples t-test to answer their research question, finding t (23) = 5.03, with descriptive results favoring the FTF environment. This is enough information for us to perform a secondary Bayesian estimation of the effect size on this study using the default prior proposed in the previous section. Changing the software settings to a paired-samples t-test, and inputting the sample size of 24, and the obtained t-value of 5.03 will provide us with the result in Figure 35 (explore the software output).

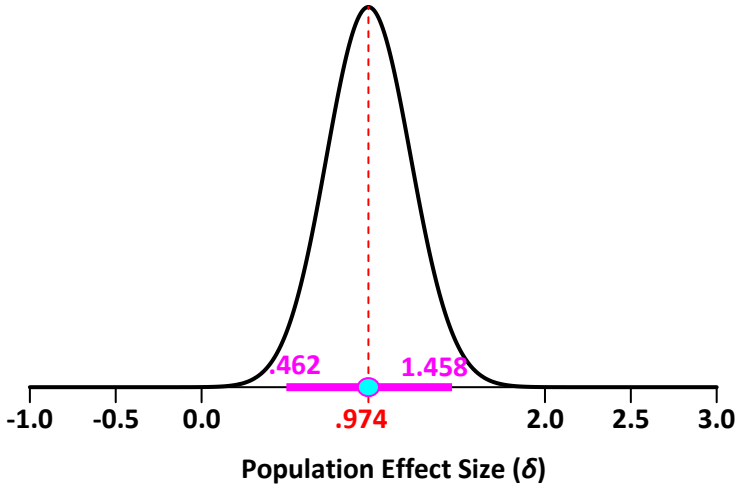

Figure 35. Posterior distribution for the effect size found in Gurzynski-Weiss and Baralt (2014) for the superiority of FTF environment over CMC environment in affording more opportunities for modified output.

Succinctly put, if Gurzynski-Weiss and Baralt (2014) had conducted a Bayesian estimation for their study, they could have interpreted their results as directly and concisely as follows: the *real* superiority of the FTF over CMC in providing more opportunities for intermediate SFL learners to modify their output in interactional feedback episodes is quantified by Cohen's *d* estimates ranging between .462 and 1.458. Although not reported in Gurzynski-Weiss and Baralt (2014), using the software, the corresponding 95% confidence interval for their effect size, which is subject to a Frequentist interpretation, is: [.522, 1.516]. We encourage the informed reader to perform various robustness analyses on these results following our demonstration in the previous section.

Performing a secondary Bayesian analysis on one or more previously published studies is not only advantageous in providing a Bayesian interpretation of the previous research findings but also in effectively informing (as a cumulative prior) a future replication study. Recall from our previous discussions that *yesterday's posterior is today's prior* (see Lindley, 2000). For

example, suppose previous research has shown that the advantage of FTF environments over CMC environments has been found to be fluctuating in three previous studies. More specifically, in the first study with n = 44, the result has indicated a smaller advantage for FTF over CMC (t (43) = 2.36), for the second replication study with n = 36 the result shows a moderate advantage (t (35) = 3.39), and the third replication study with n = 52 found a small advantage for FTF over CMC (t (51) = 1.59). We can use these studies' results together as prior for Gurzynski-Weiss and Baralt (2014). To do so, we can use as knowledge base a *Cauchy*(0, 1) as a reasonably informative prior for effect size using R function "`d.update`" from our repository:

```
d.update(t = c(2.36, 3.39, 1.59, 5.03), n1 = c(44, 36, 52, 24), scale = .21, top =
1.7, m = 0, s = 1, dist.name = "dcauchy", prior.scale = 2, margin = 1.5)
```

The result of this updating is shown in Figure 36.

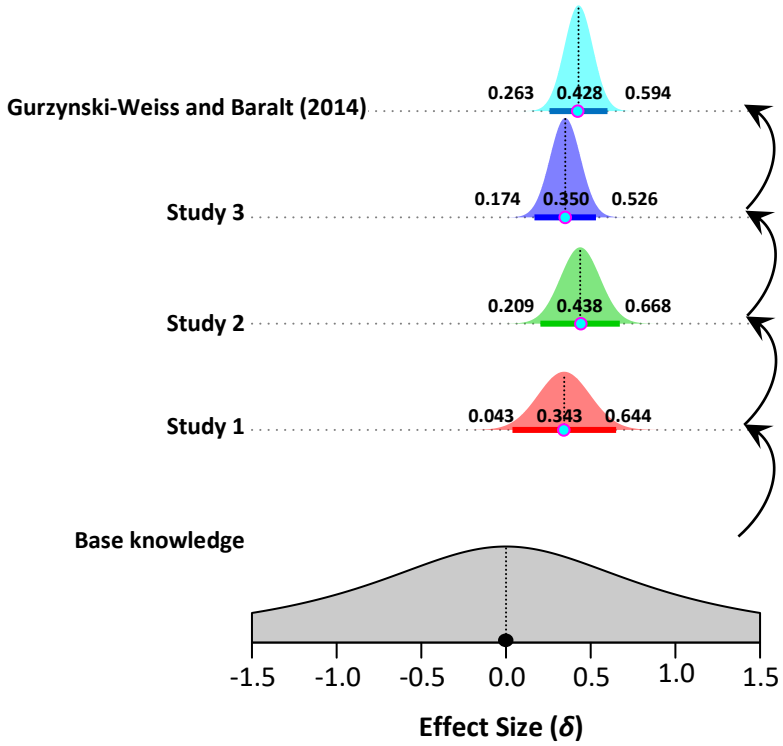

Figure 36. Step-wise Bayesian updating of three replication attempts to use them as prior for Gurzynski-Weiss and Baralt (2014).

When a researcher in a given subdomain of L2 research intends to adopt a Bayesian approach for her/his replication study, s/he can (a) perform secondary Bayesian analyses on previous studies, regardless of whether the initial studies conducted a Bayesian estimation; (b) obtain the full posterior from those previous research works; and then (c) use the final posterior obtained in that step-wise updating process as the prior for her/his intended replication study. Such a practice is very consistent with the spirit of Bayesian methods which heavily rely on past research to inform a current replication study (see Note).

Conclusion

There is a statistical view of the world that has long permeated the scientific literature. By the basic rules of this world, there are good reasons to believe what we report as "findings" from our studies might not represent the reality we are attempting to capture. To learn about that reality, however, two solutions exist.

The first solution relies on a procedure that assumes repeating one's exact same study ad infinitum, providing a specified certainty (e.g., 95%) in capturing the true effect in question from this long-run procedure (Frequentism). Under this approach, the interpretation of a single observed interval estimate (i.e., confidence interval) must be made in the context of the Frequentist procedure i.e., over long-run frequencies, 95% of the confidence intervals theoretically constructed in the process (see Figures 2 and 16) would contain the true population value and not in terms of the single interval estimate obtained (see Depaoli & van de Schoot, 2017; Thompson, 2006). This Frequentist interpretation likely escapes the awareness of many applied researchers.

The second solution, which we advocated in the present paper, translates the theoretical repetitions assumed in the Frequentist paradigm into a prior distribution. That is, a prior is a practical way for expressing defensible expectations for reality rather than thinking about reality in Frequentist terms.

By nature, reasonableness and conservatism must always govern the use of Bayesian statistics. Choices of priors must be transparent as they are an orderly form of knowledge presentation (Edwards, Lindman, & Savage, 1963). Decisions made at every step of the analyses must be defensible. And researchers must routinely evaluate the robustness of the obtained results and report them to their audience (for a complete checklist of points to consider when conducting a Bayesian analysis see Depaoli & van de Schoot, 2017). However, we argue that with Bayesian methods taking a central stage in L2 research, we will enter a new era marked by (a) constructive criticisms and academic debates over key issues in the assessment and development of L2 theory, (b) more precise attention to past research findings to come up with defensible priors, and (c) a focus on meaningful research parameters worthy of being estimated (e.g., effect sizes).

These three advantages from Bayesian methods, we believe, best characterize the need for a "Bayesian revolution" in L2 research. Thus, we hope the applied, and non-technical approach that we adopted in this paper could be a first step for the field in that direction.

Notes

In practice, exact replication of previous research (i.e., repeating an original study while keeping all experimental conditions the same as the original study), as we discuss here, is rarely encountered in L2 research (e.g., Norouzian, 2015; Norouzian & Eslami, 2016; Norouzian & Farahani, 2012; Norouzian & Plonsky, in press). The broader framework for synthesizing outcomes of multiple studies when differences between studies (due to differences among

sampled participants in the studies and differences in treatments, settings etc.) also exist is the form of random-effects meta-analysis (Cooper, Hedges, & Valentine, 2009). Bayesian methods are capable of seamlessly handling random-effects meta-analysis even in the face of a limited number of primary studies available, a problem often restricting the use of random-effects meta-analysis under the Frequentist framework. The topic of Bayesian meta-analysis falls outside the scope of the present treatment. The interested reader is referred to Berry, Carlin, Lee, and Müller (2011, Sec. 2.4), Smith, Spiegelhalter, and Thomas (1995), Spiegelhalter, Abrams, and Myles (2004, Ch. 8), Stangl and Berry (2000), and Sutton and Abrams (2001) for a foundational introduction. The R packages "`bayesmeta`" (Röver, 2017) and "`bmeta`" (Ding & Baio, 2016) both provide efficient implementation of Bayesian random-effects meta-analyses for a variety of study outcome metrics (e.g., standardized mean difference effect size).

CHAPTER V

CONCLUSIONS

In each of the three studies described in the preceding chapters, we presented study-specific discussions, field-wide recommendations, and technical conclusions. As such, here we take a general approach to conclude this dissertation by first enumerating the goals that we aimed to achieve, and then summarizing the outcomes of the three studies.

The present dissertation was intended to attain several goals within the context of L2 methodological reform. Specifically, we sought to:

1. *respond* to the repeated calls for examining how "APPROPRIATELY" (e.g., Lazaraton, 2009, p. 415, emphasis in original) some statistical concepts are being employed in L2 published research (also see Lazaraton, 2000, 2005) as opposed to *how often* certain methods are used.
2. *contribute* to the current state of *statistical literacy* among L2 researchers (Lazaraton, Riggenbach, & Ediger, 1987; Loewen et al., 2014).
3. *offer* a free software package as well a more flexible R package to promote the actual use of Bayesian methods in the field of L2 research and make such methods fully available to L2 researchers (Mizumoto & Plonsky, 2015).
4. *provide* field-wide recommendations to improve reporting practices, and thereby study quality (Norris, Plonsky, Ross, & Schoonen, 2015).
5. *introduce* the field of L2 research to modern statistical methods that may provide a more valid basis for generalizing results to wider populations (Larson-Hall, 2012b).

6. *help* prevent misinterpretations that compromise L2 theory development and research in various ways (Norris, 2015).
7. *promote* "quantitative reasoning" (Norris, Ross, & Schoonen, 2015, p. 2) in the field of L2 research.

In the first manuscript, we sought to respond to the pointed calls for investigating appropriateness of statistical concepts in quantitative L2 research (Lazaraton, 2005, 2009). Specifically, two variants of effect size, eta-squared (η^2) and partial eta-squared (η_p^2), generally used in conjunction with AN(C)OVAs, showed to require more close attention when applied in L2 research. Evidence obtained from the first manuscript alluded to a long-standing confusion among L2 researchers with regards to the correct application and interpretation of these two effect size estimators. Findings from chapter II not only should alert L2 researchers to paying more careful attention to using eta-squared (η^2) and partial eta-squared (η_p^2), but they also highlight the consequences of misapplying these effect size measures as regards L2 theory development future study planning. Although previous research has emphasized the importance statistical literacy among L2 researchers (e.g., Gonulal et al., 2017; Loewen et al., 2014), there seems to be a need for taking more concrete steps in promoting quantitative reasoning and discussing why such a line of reasoning should be viewed as an integral part of an L2 research work.

In the second manuscript, we extended the concept of hypothesis testing to the Bayesian framework. Specifically, we empirically applied the Bayesian hypothesis testing methods to a representative sample of published L2 research and contrasted the result with the traditional null hypothesis significance testing (NHST) approach. The results revealed that the two methods of inference disagree over a critical area of decision making. For about 65% of results that a

researcher under the NHST declares a significant result and thus rejects the null hypothesis (i.e., that there is no effect for the treatment), Bayesian hypothesis testing found the strength of the evidence to be only at an "anecdotal" (insufficient to reject) level. Interestingly, when we used a different prior specification, the same results were obtained. The empirical results showed that the methods of inference could be reconciled if threshold for declaring a significant result is decreased to about .005. This result is in line with theoretical findings of Benjamin et al. (in press) and Johnson (2013) as well as recommendations of the American Statistical Association (2016) emphasizing that the current practice of NHST leads to high false discovery rates. We also provided free software to facilitate the use of Bayesian model selection methods. These empirical as well as practical steps need to be still extended to more complex designs and research problems both to better understand the divergence between the Bayesian and traditional methods and to enable the use of Bayesian hypothesis testing methods in complex designs.

In the third manuscript, we sought to extend the Bayesian methods to the issue of effect sizes; the primary product of a research inquiry (see Cohen, 1990). we presented a comprehensive primer on Bayesian methods, provided a comprehensive R package to conduct Bayesian estimation, and perform diagnostic test to examine its sensitivity (i.e., stability) under a wide variety of priors. For each unique research situation presented, proper decision-making strategies and line of quantitative reasoning were expounded. It is critical for L2 researcher is be informed of the basis and applications of Bayesian methods in L2 research. This need is especially motivated mainly by the fact that quantitative L2 research is often based on small groups of participants (Larson-Hall & Herrington, 2010) and that a body of knowledge regarding the size of various treatment effects in L2 research is available (Plonsky & Oswald, 2014).

Future research must expand on the ideas and the tool developed in Chapter IV, as complex research designs require complex analytic solutions to answer complex research questions. It is imperative to note that the studies provided in this dissertation cumulatively help in building a case for the importance of research methods and methodological expertise in L2 research and perhaps other branches of social and behavioral sciences.

The importance of methodological knowledge comes into view when we review the process by which research findings are published and thereby made available to theoreticians, other fellow researchers. A lack of methodological expertise in the field not only restricts research-as-produced but it also restricts the potential of peer review to lead to higher quality journal articles. Finally, once a study is published, methodological acumen is again required in order for consumers to be able to adequately and critically interpret the findings and the process by which they were derived.

In closing, I should stress the importance of two specific areas where Bayesian thinking can be specifically of significance. First, given Bayesian methods' use of the knowledge generated by prior studies, these methods can greatly enhance our understanding of systematic reviews of literature. Specifically, Bayesian methods are capable of providing more realistic estimates of how effective an L2 treatment of interest in light of several pieces of research using that treatment. Finally, the area of measurement can benefit from Bayesian methods. Psychometric issues such as reliability of scores produced by a researcher-developed instrument can be improved by obtaining Bayesian reliability estimates. The obtained Bayesian estimates can serve to inform current substantive interpretation and form the basis for future adjustments to measurement instruments.

REFERENCES

Albert, J. (2009). *Bayesian computation with R* (2nd ed.). New York: Springer.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

American Statistical Association. (2016). *American statistical association releases statement on statistical significance and p-values: Provides principles to improve the conduct and interpretation of quantitative science*. Retrieved from http://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior research methods, 37*(3), 379-384. doi:10.3758/bf03192707

Bedore, L. M., Peña, E. D., Joyner, D., & Macken, C. (2011). Parent and teacher rating of bilingual language proficiency and language development concerns. *International Journal of Bilingual Education and Bilingualism, 14*(5), 489-511.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., . . . Johnson, V. E. (in press). Redefine Statistical Significance. *Nature Human Behavior*.

Berry, S. M., Carlin, B. P., Lee, J. J., & Müller, P. (2011). *Bayesian adaptive methods for clinical trials*. NY, New York: CRC press.

Block, D. (2003). *The social turn in second language acquisition*. Washington, DC: Georgetown University Press.

Brown, J. D. (2015). Why bother learning advanced quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 9-20). New York, NY: Routledge.

Byrnes, H. (2013). Notes from the Editor. *Modern Language Journal, 97*(4), 825-827.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Cohen, J. (1965). Some statistical issues in psychological research. In B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York, NY: McGraw-Hill.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33*(1), 107-112.

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

Cohen, J. (1990). Things I have learned (so far). *American psychologist, 45*(12), 1304 -1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American psychologist, 49*(12), 997-1003.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.

Crookes, G. (1991). Power, effect size, and second language research: Another researcher comments. *TESOL Quarterly, 25*(4), 762-765.

DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 94-112). London: Routledge.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological methods, 22*(2), 240.

Dienes, Z., & Mclatchie, N. (2017). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic bulletin & review*, 1-12. doi:10.3758/s13423-017-1266-z

Ding, T., & Baio, G. (2016). bmeta: Bayesian Meta-Analysis and Meta-Regression. R package version 0.1.2. available at: https://CRAN.R-project.org/package=bmeta.

Dörnyei, Z. (2007). Research methods in applied linguistics.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological review, 70*(3), 193-242.

Egbert, J. (2007). Quality analysis of journals in TESOL and applied linguistics. *TESOL Quarterly, 41*(1), 157-171.

Ellis, R. (2009). Measuring implicit and explicit knowledge of a second language. In R. Ellis, S. Loewen, C. Elder, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 31–64). Bristol: Multilingual Matters.

Etz, A., & Vandekerckhove, J. (in press). Introduction to Bayesian Inference for Psychology. *Psychonomic Bulletin and Review*.

Farruggio, P. (2010). Latino immigrant parents' views of bilingual education as a vehicle for heritage preservation. *Journal of Latinos and Education, 9*(1), 3-21.

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *Modern Language Journal, 81*(3), 285-300.

Francis, G. (2016). Equivalent statistics and data interpretation. *Behavior research methods, 49*, 1524-1538. doi:10.3758/s13428-016-0812-3

Gass, S. (2009). A historical survey of SLA research. In T. K. Bhatia & W. C. Ritchie (Eds.), *The new handbook of second language acquisition* (pp. 3-28). Bingley, UK: Emerald.

Gass, S. (in press). Commentary 1: SLA and study abroad: A focus on methodology. *System*.

Gass, S. M., Lee, J., & Roots, R. (2007). First and Wagner (1997): New ideas or a new articulation? *Modern Language Journal, 91*, 788-799.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641-651.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (2nd Ed.): Chapman & Hall/CRC Boca Raton, FL, USA.

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-sample t test. *The American Statistician, 59*(3), 252-257.

Gonulal, T., Loewen, S., & Plonsky, L. (2017). The development of statistical literacy in applied linguistics graduate students. *International Journal of Applied Linguistics, 168*(1), 5-33. doi: 10.1075/itl.168.1.01gon

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). NY, New York: Routledge.

Gurzynski-Weiss, L., & Baralt, M. (2014). Exploring learner perception and use of task-based interactional feedback in FTF and CMC modes. *Studies in Second Language Acquisition, 36*(1), 1-37.

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*: New York, NY: Newbury House Publishers.

Hudson, T. (2015). Presenting quantitative data visually. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 78-105). NY, New York: Routledge.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med, 2*(8), e124.

Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the output hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition, 21*(3), 421-452.

Jaynes, E. T. (2003). *Probability theory: The logic of science*: Cambridge university press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). New York: Oxford University Press.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences, 110*(48), 19313-19317. doi:10.1073/pnas.1313476110

Johnson, V. E. (2016). Bayes' Factors. *eLS*, 1-6. doi:10.1002/9780470015902.a0005851.pub2

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological methods, 17*(2), 137.

Kerlinger, F. (1964). *Foundations of behavioral research*. New York, NY: Holt, Rinehart and Winston.

Kruschke, J. K. (2015). *Doing Bayesian data analysis* (2nd ed.). Boston: Academic Press.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*(4), 722-752.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, 1-29. doi:10.3758/s13423-016-1221-4

LaFlair, G., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 46-77). NY, New York: Routledge.

Lantolf, J. P. (1996). SLA theory building: "Letting All the Flowers Bloom!". *Language Learning, 46*(4), 713-749. doi:10.1111/j.1467-1770.1996.tb01357.x

Larson-Hall, J. (2012a). How to run statistical analyses. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 245-274). Oxford, UK: Blackwell.

Larson-Hall, J. (2012b). Our statistical intuitions may be misleading us: Why we need robust statistics. *Language Teaching, 45*(4), 460-474.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York: Routledge.

Larson-Hall, J., & Herrington, R. (2010). Improving Data Analysis in Second Language Acquisition by Utilizing Modern Developments in Applied Statistics. *Applied Linguistics, 31*(3), 368-390. doi:10.1093/applin/amp038

Lazaraton, A. (1991). Power, effect size, and second language research: A researcher comments. *TESOL Quarterly, 25*(4), 759-762.

Lazaraton, A. (2000). Current trends in research methodology and statistics in applied linguistics. *TESOL Quarterly, 34*(1), 175-181.

Lazaraton, A. (2005). Quantitative research methods. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 209-224). Mahwah, NJ: Erlbaum.

Lazaraton, A. (2009). The use of statistics in SLA: A response to Loewen & Gass (2009). *Language Teaching, 42*(3), 415.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*(4), 612-625.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*(481), 410-423.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185-207. doi:10.1111/lang.12117

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician, 49*(3), 293-337.

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly*, *48*(2), 360-388.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19-32. doi:10.1016/j.jmp.2015.06.004

Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In P. García Mayo, M. Gutierrez-Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 71-92). Amsterdam: Benjamins.

Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design* (2nd ed.). New York: Routledge.

Marsden, E. J., Morgan-Short, K., Thompson, S., & Abugaber, D. (in press). Replication in second language research: Narrative and systematic reviews, and recommendations for the field. *Language Learning*.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan* New York: CRC Press.

Mizumoto, A., & Plonsky, L. (2015). R as a lingua franca: Advantages of using R for quantitative research in applied linguistics. *Applied Linguistics*, *37*(2), 284-291.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6-18.

Morey, R. D., Wagenmakers, E.-J., & Rouder, J. N. (2016). Calibrated Bayes factors should not be used: A reply to Hoijtink, van Kooten, and Hulsker. *Multivariate behavioral research, 51*(1), 11-19.

Nassaji, H. (2012). Significance tests and generalizability of research results: A case for replication. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 92-115). Cambridge: Cambridge University Press.

Norouzian, R. (2015). Does teaching experience affect type, amount, and precision of the written corrective feedback?. *Journal of Advances in English Language Teaching, 3,* 93-105.

Norouzian, R., De Miranda, M., & Plonsky, L. (under review). The Bayesian revolution in L2 research: An applied approach.

Norouzian, R., de Miranda, M. A., & Plonsky, L. (under review). A Bayesian approach to measuring evidence in L2 research: An empirical investigation.

Norouzian, R., & Eslami, Z. (2013). Does synchronous computer-mediated communication improve EFL learners' oral proficiency? *Modern English Teacher, 22*(3), 40-42.

Norouzian, R., & Eslami, Z. (2016). Critical perspectives on interlanguage pragmatic development: An agenda for research. *Issues in Applied Linguistics, 20*(1), 25-50.

Norouzian, R., & Farahani, A. A. K. (2012). Written Error Feedback from Perception to Practice: A Feedback on Feedback. *Journal of Language Teaching and Research, 3*(1), 11-22.

Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research, 34*, 257-271.

Norouzian, R., & Plonsky, L. (in press). Correlation and Simple Linear Regression in Applied Linguistics. In A. Phakiti, P. I. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology*. New York, NY: Palgrave.

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning, 65*(S1), 97-126.

Norris, J. M., & Ortega, L. (2006). *Synthesizing research on language learning and teaching*. Philadelphia: John Benjamins.

Norris, J. M., Ross, S. J., & Schoonen, R. (2015). Improving second language quantitative research. *Language Learning, 65*(S1), 1-8.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary educational psychology, 25*(3), 241-286.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological methods, 8*(4), 434-447.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943-951.

Ortega, L. (2005). Methodology, epistemology, and ethics in instructed SLA research: An introduction. *Modern Language Journal, 89*(3), 317-327.

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning, 63*, 1-24. doi:10.1111/j.1467-9922.2012.00735.x

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science a crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528-530.

Pedhazur, E. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd Ed.). Fort Worth, TX: Harcourt Brace.

Perry, F. L. (2011). *Research in applied linguistics: Becoming a discerning consumer* (2nd ed.). New York, NY: Routledge.

Phakiti, A. (2014). *Research methods in linguistics: Experimental research methods in language learning*. London, UK: Bloomsbury Academic.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary Note on Reporting Eta-Squared Values from Multifactor ANOVA Designs. *Educational and Psychological Measurement, 64*(6), 916-924. doi:10.1177/0013164404264848

Plonsky, L. (2013). Study quality in SLA. *Studies in Second Language Acquisition, 35*(04), 655-687.

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal, 98*(1), 450-470.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning, 61*(2), 325-366.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912.

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 106-128). New York, NY: Routledge.

Pollard, P., & Richardson, J. (1987). On the probability of making Type I errors. *Psychological bulletin, 102*(1), 159.

Porte, G. (2012). *Replication research in applied linguistics*. New York: Cambridge University Press.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ramos, F. (2007). What do parents think of two-way bilingual education? An analysis of responses. *Journal of Latinos and Education, 6*(2), 139-150.

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*(2), 135-147.

Rouder, J., Morey, R., Verhagen, J., Province, J., & Wagenmakers, E. J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8*(3), 520-547. doi:10.1111/tops.12214

Rouder, J. N., Speckman, P., Sun, D., Morey, R., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review, 16*(2), 225-237.

Röver, C. (2017). Bayesian random-effects meta-analysis using the bayesmeta R package. *arXiv preprint arXiv:1711.08683*.

Selinker, L., & Lakshmanan, U. (2001). How do we know what we know?; Why do we believe what we believe? *Second Language Research, 17*(4), 323-325. doi:10.1177/026765830101700401

Shintani, N., & Ellis, R. (2013). The comparative effect of direct written corrective feedback and metalinguistic explanation on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing, 22*(3), 286-306.

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language Learning, 64*(1), 103-131.

Smith, T. C., Spiegelhalter, D. J., & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in medicine, 14*(24), 2685-2699.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Sussex, England: Wiley.

Stangl, D., & Berry, D. A. (2000). *Meta-analysis in medicine and health policy*. New York: CRC Press.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical methods in medical research, 10*(4), 277-303.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. NY, New York: Guilford Press.

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.

Wasserman, L. (2004). *All of statistics: a concise course in statistical inference*. New York, NY: Springer Science & Business Media.