New Directions in Quantitative Hispanic Sociolinguistics

by

Paul Adrian Kidhardt

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved May 2015 by the
Graduate Supervisory Committee:

Álvaro Cerrón-Palomino, Chair
Verónica González-López
Barbara Lafford

ARIZONA STATE UNIVERSITY

August 2015
ABSTRACT

The present thesis explores how statistical methods are conceptualized, used, and interpreted in quantitative Hispanic sociolinguistics in light of the group of statistical methods espoused by Kline (2013) and named by Cumming (2012) as the “new statistics.” The new statistics, as a conceptual framework, repudiates null hypothesis statistical testing (NHST) and replaces it with the ESCI method, or Effect Sizes and Confidence Intervals, as well as meta-analytic thinking. In this thesis, a descriptive review of 44 studies found in three academic journals over the last decade (2005 – 2015), NHST was found to have a tight grip on most researchers. NHST, much discredited outside of linguistics, confused authors who conflated the theories of Fisher and Neyman-Pearson, who themselves battled acrimoniously until the end of their publishing lives. Within the studies reviewed, with exceptions, dichotomous thinking ruled the quantitative approach, and binary reporting ruled the results and discussions. In addition, this thesis revealed that sociolinguistics, at least within the studies reviewed, is not exactly a “statistical monoculture” as suspected by Gorman and Johnson (2013), rather ANOVAs have joined Goldvarb’s logistic regression in its dominance. As described insightfully by Plonsky (2015), these two methods are exposed as extensions of the dichotomous thinking that attaches itself to NHST. Further, little evidence was found that the methods of the new statistics were being implemented in a coordinated fashion, including far too few meta-analyses. As such, quantitative Hispanic sociolinguistics, and linguistics in general, were shown to be vulnerable to problems with reliable quantitative theory building.
DEDICATION

I dedicate this thesis to my loving mother who is gifted with both mathematical and creative geniuses that far exceed my own, if I have any, and whose intellectual honesty I model.
ACKNOWLEDGMENTS

I acknowledge the years of perseverance and hard work of Dr. Barbara Lafford who introduced the world of quantitative analysis to me, who accompanied me down my winding academic path toward the subjects herein that have ignited my interest, and for her exacting standards in academic production that I will attempt to model, to best of my abilities, over the coming years. I will always be thankful to have worked under Dr. Lafford, one of the towering figures of Hispanic Linguistics.

I also acknowledge the true academic leadership and companionship of Dr. Verónica González who makes me happy on sight because I know she will always have a question to stump me. Dr. González’s extraordinary versatility across both empirical and theoretical domains of Hispanic Linguistics raise the bar for me as a thinker and academic. More than once, Dr. González has challenge me to think deeply, carefully, and broadly before committing my thoughts to paper, a characteristic of maturity that I will cultivate in all that comes next.

I acknowledge my thesis chair, Dr. Álvaro Cerrón-Palomino, who has single-handedly guided me over the years from being a devout prescriptivist to a respectful and understanding descriptivist. This change has been more than academic; it has reflected across all domains of daily life, changing how I approach each new person, that is, openly, observantly, and empathically – the traits of a good sociolinguistic fieldworker, and human being.
TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................... vi
LIST OF TABLES ............................................................................................................ vii

CHAPTER

INTRODUCTION .........................................................................................................1
  Overview ................................................................................................................. 1
  Statement of the Problem ....................................................................................... 2
  Difficulties with Statistics ..................................................................................... 3

THEORETICAL BACKGROUND.....................................................................................6
  The New Statistics ................................................................................................. 6
  Null Hypothesis Statistical Testing ......................................................................... 7
  Effect Sizes ........................................................................................................... 20
  Confidence Intervals ............................................................................................. 21
  Meta-analysis ........................................................................................................ 25

REVIEW OF THE LITERATURE .............................................................................27
  Literature ............................................................................................................... 27
  Research Questions ............................................................................................... 33

METHOD ....................................................................................................................35
  Journal Selection ................................................................................................... 35
  Study Selection ..................................................................................................... 37
  Data Selection and Coding .................................................................................... 39
<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESULTS</td>
<td>50</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>63</td>
</tr>
<tr>
<td>Null Hypothesis Statistical Testing</td>
<td>63</td>
</tr>
<tr>
<td>Confidence Intervals</td>
<td>83</td>
</tr>
<tr>
<td>Effect Sizes</td>
<td>85</td>
</tr>
<tr>
<td>Meta-Analysis</td>
<td>86</td>
</tr>
<tr>
<td>Sample Sizes</td>
<td>88</td>
</tr>
<tr>
<td>Statistical Monoculture?</td>
<td>92</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>97</td>
</tr>
<tr>
<td>Scientific Sociolinguistics?</td>
<td>97</td>
</tr>
<tr>
<td>Research Integrity</td>
<td>100</td>
</tr>
<tr>
<td>Null Hypothesis Statistical Testing</td>
<td>100</td>
</tr>
<tr>
<td>Toward a Cumulative Quantitative Hispanic Sociolinguistics</td>
<td>104</td>
</tr>
<tr>
<td>New Directions</td>
<td>105</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>108</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>119</td>
</tr>
<tr>
<td>A   METHODS USED</td>
<td>119</td>
</tr>
<tr>
<td>B   SELECTED STUDIES</td>
<td>121</td>
</tr>
<tr>
<td>Figure</td>
<td>Title</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>The Dance of the Confidence Intervals (CIs) and $p$ value Volatility</td>
</tr>
<tr>
<td>2</td>
<td>Inverse Probability</td>
</tr>
<tr>
<td>3</td>
<td>Validity</td>
</tr>
<tr>
<td>4</td>
<td>Replicability</td>
</tr>
<tr>
<td>5</td>
<td>Magnitude</td>
</tr>
<tr>
<td>6</td>
<td>Meaningfulness</td>
</tr>
<tr>
<td>7</td>
<td>Equivalence</td>
</tr>
<tr>
<td>8</td>
<td>Quality and Success</td>
</tr>
<tr>
<td>9</td>
<td>Failure</td>
</tr>
<tr>
<td>10</td>
<td>Reification</td>
</tr>
<tr>
<td>11</td>
<td>Theory Conflation</td>
</tr>
<tr>
<td>12</td>
<td>Effect Sizes Reported</td>
</tr>
<tr>
<td>13</td>
<td>Confidence Intervals Appear</td>
</tr>
<tr>
<td>14</td>
<td>Meta-Analysis Possible</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Theoretical Differences Between Fisher and Neyman-Pearson</td>
<td>18</td>
</tr>
<tr>
<td>2: Final Study Choices by Journal</td>
<td>38</td>
</tr>
<tr>
<td>3: Coding Scheme Categories, Variables, Values, and Definitions</td>
<td>44</td>
</tr>
<tr>
<td>4: Conflation of NHST Constructs</td>
<td>47</td>
</tr>
<tr>
<td>5: Top Two Methods Represented Per Journal</td>
<td>53</td>
</tr>
</tbody>
</table>
Chapter 1

INTRODUCTION

Overview

The present thesis focuses on quantitative methodology in Hispanic sociolinguistic research and its aim is to examine critically the way sociolinguistic evidence is analyzed. As Coupland (2007) points out:

Variationist sociolinguistics is self-consciously bullish about its empirical discovery procedures. It is clearly an empiricist research tradition. Its epistemology – its research philosophy – is grounded in neutral observation, minimizing observer-effects, and objective analysis of data through precise and replicable quantitative procedures. (p. 24)

Yet quantitative practices in Hispanic sociolinguistics might be falling short of best practices in the use and interpretation of statistics. As suggested by Bailey and Tillery (2004), “over the last twenty years…the concern for methodological rigor has lessened considerably” (p. 11). They underscore that “the results in quantitative sociolinguistic research are sometimes as much a consequence of the methodology used as the behavior of the informants” (p. 27). Some sociolinguists (Bayley, Cameron, & Lucas, 2013; Gorman & Johnson, 2013; Roy, 2011), perhaps in a fiat of methodological conscience, are beginning to question some of their long-standing quantitative rituals, although not addressing quantitative issues as often nor as thoroughly as notable Second Language Acquisition (SLA) researchers (Larson-Hall, 2010; Plonsky, 2011a, 2011b, 2013, 2014; Plonsky, Egbert, & Laflair, 2014; Plonsky & Gass, 2011).
Statement of the Problem

Much of the quantitative research in sociolinguistics in general, and in Hispanic variationist sociolinguistics in particular, is based on statistical practices that have been rejected in the hard sciences and increasingly disfavored in the social and behavioral sciences. The problem is summarized well by Geeslin (2014) in her tour de force on how sociolinguistics and SLA go hand in hand: “it is likely that we will see a shift toward...more sophisticated statistical models” (p. 149). Geeslin predicates her argument on the fact that “the most frequently employed practices in linguistics would be frowned upon by those in the field of statistics” (Geeslin, 2014, p. 148).

While Hispanic sociolinguistics continues its reliance on increasingly outmoded statistical practices such as statistical testing, the social and behavioral sciences, and other areas of linguistics, are adopting the new statistics. The new statistics include a group of statistical methods that, to a good degree, address the shortcomings of relying on traditional statistical testing, also known as Null Hypothesis Statistical Testing, or NHST. NHST is a mashup of theoretic approaches that has been “vigorously excoriated for its inappropriateness” (Rozeboom, 1960, p. 428) for many years, and it appears to be in the crosshairs of some of the most statistically inclined textbook writers in linguistics (Gries, 2013b; Larson-Hall, 2010; Turner, 2014). The present thesis examines the methodological trends in the use and interpretation of NHST and the methods of the new statistics as they appear in Hispanic sociolinguistic research published in peer-reviewed journals over the last decade (2005 – 2015). The most common methods are evaluated in
light of current best statistical practices in the social and behavioral sciences, with special attention being given to NHST and new statistics reporting practices.

My intention here is not to criticize investigators for the decisions they have made, but rather to describe a compelling new statistics framework in which future projects can be planned and executed as effectively as possible. This framework, emphasizing best practices and openness to new ideas, is becoming increasingly more understood, used, and valued by journal editors and reviews and researchers in other disciplines of the social and behavioral sciences, such as education, psychology, sociology, political science, public policy, and law (Barnett & Feeley, 2011; Cumming, 2012, 2014; Kline, 2013; Schmidt & Hunter, 2014).

**Difficulties with Statistics**

Statistics remain a bugbear for students, professors, and applied researchers in many fields. In fact, the systematic study of how statistics are understood and communicated is its own subfield within education and psychology. Beyth-Marom, Fidler, and Cumming (2008) study statistical cognition, which the authors consider important for evidence-based practice (EBP) in statistics and statistics education. EBP often applies to laboratory and medical settings, but statistical cognition applies far more broadly as the “processes, representations, and activities involved in acquiring and using statistical knowledge” (p 22). Prior to Beyth-Marom et al. (2008), statistical reasoning was a term already widely used (Joan Garfield, 2002; Joan Garfield & Gal, 1999), as was statistical literacy (Ben-Zvi & Garfield, 2004; Gal, 2002; Wallman, 1993). These areas of inquiry remain active because understanding, interpreting, and communicating statistics
continue to challenge students, professors, researchers, and journal editors (J. Garfield & Ben-Zvi, 2010; Iten, 2015).

In linguistics, the understanding and use of statistics remains equally challenging, even for specialists. Writing on empirical methods in cognitive linguistics, J. Newman (2010) admits:

I have not found this particularly easy, in part because even those who are more statistically sophisticated than I am seem to have more questions than answers when it comes to deciding what ‘best practice’ is supposed to be. (p 92)

Spanish author Cantos Gómez (2013) prefaces his book on statistical methods in language and linguistic research by lamenting that statistical methods are “mostly ignored or avoided because of lack of training, and also fear and dislike” (p. xi). Plonsky (2014), writing for applied linguists, comments on the need for better statistical education in linguistics in his list of proposed quantitative research reforms. Within sociolinguistics, the comparatively small number of quantitative methodologists who write on the topic of statistical reform focus on the need for more sophisticated modeling techniques (Bayley et al., 2013; Gries, 2013a; Roy, 2011), and say relatively little about statistical education in sociolinguistics.

When statistical testing is introduced as a part of a more comprehensive class on research methodology in linguistics, linguistics students may perform like regular students of college statistics classes. Although there do not appear to be studies on the statistical cognition of linguistics students specifically, there are studies on the performance of students who take introductory statistics classes. These studies show, for
example, that statistics students do not understand the “big picture” of NHST (Iten, 2015). Although they can perform procedures mechanistically, they do not have a strong understanding of the concepts, logic, and uses of statistical hypothesis testing. While focusing on NHST in their analysis of quantitative and qualitative data, Smith (2008) found that procedural knowledge is not a predictor of overall understanding. They reveal that students fail to understand the role of indirect reasoning and inference in generating and interpreting NHST results. NHST presents an ongoing challenge, not just for students, but also for current day researchers in several disciplines including linguistics. As Larson-Hall (2010) writes in her well-researched book on the use of statistical methods and SPSS in second language research:

As I have gotten immersed in statistics literature while writing this book, I have discovered that many people who are knowledgeable about statistics are dismayed by the almost fetishistic and uncomprehending use of significance tests by most applied researchers. (p. 96)
Chapter 2

THEORETICAL BACKGROUND

The New Statistics

Gigerenzer (2004), one of the most authoritative scholars on statistical testing, states: “Statistical rituals largely eliminate statistical thinking in the social sciences” (p. 587). As shown in his long publishing record, Gigerenzer’s indignation is focused on the rituals associated with null hypothesis statistical testing (NHST). Many other authors, especially more recently, are joining Gigerenzer to rail against NHST. Cumming (2012) is among them, uniting with Gigerenzer and others to assert that NHST promotes a set of “mindless” procedures that lead to confusion, error, and ultimately countervail efforts to build cumulative disciplines dedicated to high-integrity, reliable theory building (Engman, 2013; Gigerenzer, 2004; Gross, 2014; Makel, 2014). In his book Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis, Cumming (2012) proposes a set of statistical methods he calls the “new statistics.” Under the new statistics framework, NHST is repudiated and driven to the margins as misleading and dangerous “old statistics,” while more reliable, valid, and informative methods take its place. Cumming’s new statistics is based on the use of effect sizes, confidence intervals, and meta-analyses. Many of these procedures have existed for more than a century, and these methods have been grouped before (Morgan, Gliner, & Harmon, 2006), but only recently have they been proposed and named “the new statistics” as an articulated solution to the intractable problems associated with NHST.
The *Publication Manual of the American Psychological Association* (American Psychological Association, 2010) unambiguously calls for the use of the new statistics. The new statistics’ aim is “to switch emphasis…from dichotomous thinking to estimation thinking and meta-analytic thinking” (Cumming, 2012, p. 20). For language researchers there are many positive reasons to adopt the new statistics framework, and many important reasons to deemphasize NHST.

**Null Hypothesis Statistical Testing**

The phrase ‘statistically significant,’ generally referring to $p$ values that are small, such as less than 0.05, has done disservice to the understanding of truth, proof, and uncertainty. (Blackstone, 2015, p. 11)

Across the social and behavioral sciences, null hypothesis significance testing (NHST) is commonly considered to be the most objective and scientific way to organize, interpret, and learn from data. Indeed it is the most heavily relied upon epistemological doctrine in quantitative sociolinguistics. Sociolinguists, and many other applied researchers (Schmidt & Hunter, 2014), often equate NHST to scientific hypothesis testing in general, treating it as if it were central to science. This belief, however, is demonstrably false. It “is tantamount to stating that physics, chemistry, and the other physical sciences are not legitimate sciences because they do not test their hypotheses using statistical significance testing” (Schmidt & Hunter, 2014). Close examination reveals that NHST by itself does not elevate social and behavioral science hypothesis testing to the level of “hard” or objective and rigorous science; in fact, it appears to be intrinsically flawed, overused, misapplied, and misinterpreted. Authors who promote the
new statistics underscore the problems attached to NHST, and they urge applied
researchers to discontinue statistical testing as the *sine qua non* statistical approach to
scientific inquiry and replace it with a set of statistical methods that address and answer
research questions with transparency and integrity.

Null hypothesis statistical testing is far easier to understand when it is introduced
in terms of its simple execution and the limited information conveyed by the *p* values it
generates. Essentially, all statistical test procedures are the same. The difference between
a sample result and the value of its corresponding parameter is summarized as a test
statistic specified in terms of the ratio of a null hypothesis divided by the sampling error.
The ratio is then converted into a conditional probability, or *p* value, based on the
theoretical sampling distribution associated to the appropriate test statistic, given the
assumptions. The result is statistical output containing one or many *p* values. These *p*
values receive enormous attention in linguistic and sociolinguistic research when they are
simply a “statistical finding” (Gigerenzer & Marewski, 2015), that is, a *p* value is a
*soupçon* of inconclusive discovery information (Glaser, 1999; Martin & Bridgmon, 2012;
Schneider, 2015). Immediate attention needs to be paid to what a *p* value communicates
in plain language. In its most simple form:

\[ p \text{ is a measure of the plausibility that the observed or more extreme data would } \]
\[ \text{coexist in their current configuration, as collected by the researcher, in that } \]
\[ \text{quantity, if they were generated by a completely random process.} \]

This definition embodies the intention for *p* of the original theorist, Fisher (1973),
and it clarifies its limits. “Plausibility” (or we might use the word “chance”) is
conceptually clearer than nesting the word “probability” within the larger conditional probability of the null being true given the data plus assumptions. Detecting the underlying “process” actually addresses the researcher’s goal, and “completely random” is the null hypothesis, which is set to the value of completely true.

In the definition given above, $p$ is influenced by the variability expressed as randomness, or lack of randomness, that includes both random effects as naturally present in the world as it naturally occurs, and sampling error, that is, unrepresentativeness of the sample due to researcher error, such as measurement error. Therefore, it is hard to know from a $p$ value alone exactly how much can be attributed to sampling error, or to random or non-random effects from the real world. As follows from this argument, and as discussed below, $p$ is not a measure of sampling error.

The concept of randomness is crucial to understanding why statistical testing was created in the first place. Randomness is often misunderstood (Gigerenzer, 1993). Randomness has to do with the process, not the outcome of the process, that is, the variables themselves. Thus, when we think of randomness, we should conceptualize and focus on how the underlying process is proceeding, here called the stochastic process, which generates the random variables, or stochastic variables. The concept of stochasticity is fundamental to probability theory and extends to assumptions in many statistical tests, for example, the random selection of cases or participants. When we think of random as associated to $p$ values, we should think about the process that generated the variables, not the variables themselves.
If Fisherian $p$ were appropriately labeled it would be called something akin to *Random Process Coefficient* (RPC). This RPC would be limited to data discovery and essentially serve as a “statistical finding” for the researcher as they plan their next steps, as argued by Gigerenzer and Marewski (2015). The statistical cognition surrounding NHST, or perhaps *Random Process Detection Procedure*, that generates the RPC would lead researchers to focus on what comes after the procedure; that is, further data collection, data assumption testing, cleaning and imputation, exploring the research question, estimating magnitude of effects, and reporting results using confidence intervals.

Yet how can a Random Process Coefficient be understood in real terms? Set together with $p$ values, for a RPC/$p$ value of .01, for example, the $p$ value is saying that, given the amount of data collected, there is a 1% chance that the observed data would coexist in their current configuration if they were generated by a completely random process. Another way to look at, one out of a 100 times the data that the researcher collected together might be due to an underlying process that is completely random. Another example, with a RPC/$p$ value of .051, $p$ is saying that, given the amount of data collected, there is a 5.1% chance that the observed data would coexist in their current configuration if they were generated by a completely random process; Or 5.1 out of 100 times the underlying process might be completely random. Finding a 1% as compared to a 5.1% chance of a completely random process should make no difference to the researcher in pursuing their research, interrogating their research hypotheses, and making
the trip through effect sizes and confidence intervals when appropriate. Stopping at a Random Process Coefficient or $p$ value is like ending a road trip at the first gas station.

The RPC/$p$ dialogue:

Researcher: “$P$-value, what is the plausibility/chance that the data I pulled together was generated by some sort of process that was completely random?”

$P$-value: “Researcher, there is a 6% chance ($p = .06$) that a completely random process generated your data, that is, six out of 100 times.”

Researcher: “Thank you. What about the goodness of my research hypothesis, replicability, magnitude of effects?”

$P$-value: “Researcher, there is a 6% chance that…”

The upshot: our $p$ values simply do not say much!

Once a researcher and reader understand the limits of NHST, all of the NHST fallacies begin to make sense. Before exploring these fallacies, it is reasonable to question, Should linguists care?

Linguists in the areas of psycholinguists, corpus linguists, and SLA are increasingly recognizing the fall of the $p$-paradigm. In his article entitled *Quantitative considerations for improving replicability in CALL and applied linguistics*, Plonsky (2015) is now placing NHST and its fallacies at the center of the discussion on quality in quantitative SLA: “The larger issue at hand here, though, is that of our field’s extremely heavy reliance on null hypothesis significance testing” (p. 235). While sociolinguists are all but silent on the issue, SLA’s quantitative methodologists are on the heels of the statisticians of the APA. As for the hard sciences, there is “no there, there” because
NHST is not considered scientific hypothesis testing, and it is not relied upon (Schmidt & Hunter, 2014). There is no reason why Hispanic sociolinguistics should follow far behind SLA, psycholinguistics, corpus linguistics, and many other fields in the social and behavioral sciences in terms of implementing best statistical practices.

The fallacies detailed by Kline (2013) and outlined below have accumulated over the years, burying the simple definition of \( p \) under misconceptions borne of empiricist enthusiasm to make \( p \) mean far more than it does. As Cohen (1994) put it: “NHST does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!” (p. 997).

**Odds Against Chance Fallacy.** The Odds Against Chance Fallacy (Carver, 1978) is, according to Kline (2013), the most pervasive of the bunch. This fallacy concerns the false belief that \( p \) indicates the probability that our results are from sampling error. When committing this fallacy, the researcher concludes that when \( p < .05 \) it means that there is less than a 5% likelihood that the finding is due to chance. This misconception extends to the notion that \( p \) values sort results into those of “chance” when \( H_0 \) not rejected, and those due to “real” effects when the \( H_0 \) is rejected (Kline, 2013). This is erroneous. When \( p \) is calculated, the \( H_0 \) is set to be true, that is there is an underlying process that is 100% random, as I defined above: “generated by a completely random process.” In probability (always between 0 and 1), an underlying process that appears totally randomizing, aggravated by sampling error (measurement error, etc.), is the only explanation, and it is already taken to be 1.00. Therefore, it is illogical to view \( p \) as a measure of the likelihood of sampling error (Schneider, 2015). There is no statistical technique that determines the
probability that various causal factors, such as sampling error, acted on any result in particular (Kline, 2013). A \( p \) value does not provide causal inference as suggested by the Odds Against Chance Fallacy.

**Causality Fallacy.** This is the false belief that statistical significance means that an underlying causal mechanism has been identified.

**Local Type I Error Fallacy.** This fallacy occurs when \( p < .05 \), given \( \alpha = .05 \), is understood to mean that there was a 5% likelihood that the decision that was just taken to reject \( H_0 \) was a Type I error (Kline, 2013; Pollard, 1993). This fallacy confuses the conditional probability of a Type I error of \( \alpha = p(\text{Reject } H_0 | H_0 \text{ true}) \) with the conditional posterior probability of a Type I error given that \( H_0 \) has been rejected, or \( p(H_0 \text{ true } | \text{Reject } H_0) \). Yet \( p \) values are conditional probabilities of the data they are tested against, so they do not apply to specific decisions to reject the null hypothesis, \( H_0 \). This is because any specific decision to reject the null is either correct or incorrect, so no probability can be associated with it other than 0 or 1. Only with sufficient replication can we determine whether a decision to reject \( H_0 \) in a specific study was correct or incorrect.

**Inverse Probability Fallacy.** The Inverse Probability Fallacy, named by Kline (2013), is also known as the fallacy of the transposed conditional (Ziliak & McCloskey, 2008) and the permanent illusion (Gigerenzer, 1993). This fallacy receives much attention because it is one of the most persistent and ubiquitous across academic and professional disciplines (Engman, 2013). This is the false belief that \( p \) measures the likelihood that \( H_0 \) is true given the data, or \( p(H_0 | D^+) \), which is exactly the opposite of the proper conditional probability. Our \( p \) values are conditional probabilities of the data...
given that the null hypothesis is true, or \( p(D^+ \mid H_0) \). \( D^+ \) refers to the data plus all of the statistical testing assumptions that accompany it.

**Validity Fallacy.** This fallacy, also known as the valid research hypothesis fallacy by Carver (1978), refers to the erroneous belief that the probability that \( H_1 \) is true is > .95 when \( p < .05 \).

**Replicability Fallacy.** This is the mistaken belief that the complement of \( p \), \( (1 - p) \), indicates the probability of finding a statistically significant result in a replication study (Carver, 1978) when it is actually the probability of getting a result even less extreme under \( H_0 \) than the one found. When operating under this fallacy, a researcher would infer that the probability of replication is > .95 if \( p < .05 \).

**Magnitude Fallacy.** This fallacy is the mistaken belief that low \( p \) values mean large effects. Similar to this fallacy, the *slippery slope of significance* described by Cumming (2012) occurs when a researcher describes a result for which \( p < \alpha \) as “significant” without specifying it as “statistically significant,” and then discusses it as if it were practically important or quantifiably large in terms of effects. Significance test probabilities do not say much about effect size because \( p \) values of effect size and sample size are “confounded measures” (Kline, 2013, p. 100). Thus there exists circular logic behind \( p \) values: If the sample size is large, low \( p \) values simply confirm a large sample. As a result, even trivial effects can be statistically significant in studies that use large enough sample sizes, as often occurs in corpus linguistics. The inverse is also true: Large, often substantive, effects may not be statistically significant in small samples.
Meaningfulness Fallacy. This fallacy is the erroneous belief that rejection of $H_0$ confirms $H_1$. According to Kline (2013), this myth reflects two cognitive errors. First, the decision to reject $H_0$ in a single study does not imply that $H_1$ is proven. Second, even if $H_1$ is statistically favored, it does not mean that the substantive hypothesis behind $H_1$ is also correct. Statistical significance does not prove our substantive hypothesis, and there are times when the same $p$ value, or Random Process Coefficient, can align with more than one substantive research hypothesis.

Zero Fallacy. The Zero Fallacy, called the slippery slope of nonsignificance by Cumming (2012) involves what critics call the nil hypothesis. The nil (null) hypothesis is essentially a straw man that researchers knock down because they already know that it is implausible to begin with, especially in non-experimental studies (Perezgonzalez, 2015). For example, a sociolinguist observes a phenomenon, then tests for it. They know that the correlation between variables, or difference between groups (see Equivalence Fallacy below), is not absolutely zero because they have witnessed the phenomenon. As Plonsky (2015) states: “The absence of evidence is not evidence for absence” (p. 235). The Zero Fallacy, then, is the erroneous belief that the population effect size is zero when we fail to reject a nil (null) hypothesis. Statistical testing was originally envisaged for situations in which the researcher had no idea whether a phenomenon was occurring or not occurring (Gigerenzer, 1993). Statistical testing was not meant to be used to mechanistically knock down a series of straw men to fill out methods and results sections in applied research.

Equivalence fallacy. This fallacy occurs when the failure to reject $H_0: \mu_1 = \mu_2$ is understood to mean that the populations are equivalent (Kline, 2013). This is incorrect
because even if the two parameters are equal, $\mu_1 = \mu_2$, variability and the distribution shape can differ.

Quality Fallacy and Success Fallacy. These fallacies are the beliefs that statistically significant results speak to the quality of the study’s design and indicate that a particular study has succeeded. In reality, the way our study is designed and how we gather data (sampling error) can lead to falsely rejecting $H_0$, that is, Type I errors.

Failure Fallacy. This is the mistaken belief that a lack of statistical significance means the study has failed. The old chant of hopeful doctoral students was: “a critical ratio of three, or no PhD” (Gigerenzer & Marewski, 2015, p. 6); that is, to receive their doctorate, students needed to come up with the critical ratio (a sample statistic over its standard error) that today would render something akin to $p < .01$ in our $z$ or $t$ tests (Kline, 2013). If you showed significance, then not only was your study a success, but so were you! If not, your study failed, and perhaps you, too.

Reification Fallacy. This is the incorrect belief that failure to replicate a result is due to a failure to make the same decision about $H_0$ between studies (Kline, 2013). When committing this fallacy, a result is not seen as having been replicated if $H_0$ is rejected in one study but not in another. “This sophism ignores sample size, effect size, and power across different studies” (Kline, 2013, p. 102). Quite simply, and quite often, one of the null hypotheses may not be rejected because the researcher was working with a smaller sample size.

Objectivity Fallacy. This is the false belief that significance testing is an objective method of hypothesis testing while (all) other inference models are seen as
subjective (Gorard, 2006; Kline, 2013). Significance testing has the illusion of being objective when in reality there are other ways to test research hypotheses (Schmidt & Hunter, 2014).

**Sanctification Fallacy.** This fallacy occurs when continuous $p$ values are viewed dichotomously. For example, if $\alpha = .05$, the practical difference between $p = .049$ and $p = .051$ is nugatory. Yet researchers who commit this fallacy consider the first to be important and remarkable, while the second is unimportant and unremarkable.

**Robustness Fallacy** When sample sizes are small or otherwise statistically unrepresentative, classical parametric statistical tests are not robust against outliers or violations of distributional assumptions. When researchers believe that their small sample sizes are indeed robust against violations and assumptions, especially when they do not test for them, they commit the Robustness Fallacy. This fallacy institutionalizes itself when volumes of research is produced and nobody appears to test distributional or other assumptions. Researchers begin to believe that robustness is simply not an issue (Schmidt & Hunter, 2014). This belief is false. Data integrity needs to be checked and reported. This includes testing for missing data and distributional abnormalities, and reporting what steps were taken to remedy them, such as data transformations.

**Theoretical mashup of unknown authorship.** Who can be blamed for the confusion that led to our long list of NHST fallacies? Nobody knows (Gigerenzer, 1993; Schneider, 2015). This makeshift integrated model as the hybrid logic of scientific inference was simply introduced in textbooks without author attribution, not even attribution to the original theorists. Dixon and O'Reilly (1999) called it the *Intro Stats*
method. Many authors have noted that this hybrid model would have been rejected by Fisher, Neyman, and Pearson, although for different reasons, and its composite nature continues to confuse students, professors, and researchers (Kline, 2013; Perezgonzalez, 2015). The differences that lead to most of the confusion are found in Table 1, adapted from Schneider (2015, p. 415).

<table>
<thead>
<tr>
<th>“Significance tests” (R. A. Fisher)</th>
<th>“Hypothesis tests” (Neyman and Pearson)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p values are measures of evidence against H₀</td>
<td>α and β levels provide rules to limit the proportion of decision errors</td>
</tr>
<tr>
<td>Calculated a posteriori from the observed data (random variable)</td>
<td>Fixed values, determined a priori at some specified level</td>
</tr>
<tr>
<td>Applies to any single experiment (short run)</td>
<td>Applies only to ongoing, identical repetitions of an experiment, not to any single experiment (long-run)</td>
</tr>
<tr>
<td>Roots in inductive philosophy: from particular to general</td>
<td>Roots in deductive philosophy: from general to particular</td>
</tr>
<tr>
<td>“Inductive inference”: guidelines for interpreting strength of evidence in data (subjective decisions)</td>
<td>“Inductive behavior”: guidelines for making decisions based on data (objective behavior)</td>
</tr>
<tr>
<td>Based on the concept of a hypothetical infinite population</td>
<td>Based on a clearly defined population</td>
</tr>
<tr>
<td>Evidential, that is, based on the evidence observed</td>
<td>Non-evidential, that is, based on a rule of behavior</td>
</tr>
</tbody>
</table>

Fisher performed “significance tests” while “Neyman and Pearson” performed “hypothesis tests.” For Fisher, a p value was a measure of the evidence against the null hypothesis. For Neyman and Pearson, p values were used to set α and β levels that provided rules to limit the proportion of decision errors. It is worth noting here that Neyman and Pearson were interested in factory efficiency, fewer defects or errors, in the production of goods. Another difference between the theorists was that Fisher’s significance tests were calculated a posteriori from the observed data (short run), while
Neyman and Pearson’s hypothesis test only applied to ongoing, identical repetitions of an experiment (visualize a factory production line), not any single experiment (long run). Fisher’s philosophy was inductive, while Neyman and Pearson’s philosophy was deductive. Fisher made “inductive references” to interpret the strength of the evidence in the data (subjective decisions), while Neyman and Pearson used what they called “inductive behavior” for making decisions based on the data (objective behavior) (Engman, 2013; Gigerenzer, 1993). Fisher rejected Neyman-Pearson’s arguments for "inductive behavior" as "childish" (Fisher, 1955, p. 75) deeming Neyman and Pearson as "mathematicians without personal contact with the Natural Sciences" (p. 69). Another difference between the theorists was how populations were envisaged. Fisher imagined “hypothetical infinite populations,” while Neyman and Pearson imagined a clearly defined population. Also, for fisher, conclusions were based on the evidence observed, that is, it was evidential. Neyman and Pearson felt conclusions should be non-evidential, or based on a rule of behavior (the cut off value of $\alpha$, and binary decision making).

Despite the contradictory elements in the two approaches and the irreconcilable views between the theorists, from the 1940s onwards, applied textbooks in the social and behavioral sciences began to blend Fisher’s and Neyman–Pearson’s approaches into the mashup we now call NHST (Gigerenzer & Marewski, 2015). This hybrid model would most certainly have been rejected by Fisher, and Neyman and Pearson (Perezgonzalez, 2015).
Effect Sizes

Defined broadly, an effect size is the amount of anything of research interest. If we think of effect size as defined by Kelley and Preacher (2012), it is the quantitative reflection of the magnitude of some phenomenon used to address a specific research question. When we set out to quantify a phenomenon of interest, an effect size takes the form of a statistic in our sample and a parameter in our population. In nonexperimental studies, as are most sociolinguistic studies, effect sizes can be described in terms of degrees of covariation between variables of interest.

Most sociolinguists, and probably all SLA researchers, are familiar with effect sizes, even if they do not refer to them in this way. Effect sizes appear as a proportion or percentage, the difference between two means, a correlation, an odds ratio, or a regression slope, among others. For example, when we see Cohen's $d$ and $\eta^2$ (eta-squared), they are simply effect sizes. Effect sizes are estimated based on the data. We take an effect we find in our sample as a point estimate and then we attempt to extrapolate the effect size that exists in the population, as we do when we use a sample mean, $\bar{x}$, to find the corresponding population mean, $\mu$.

Effect sizes can be quantified in original units, like milliseconds in a recording, or they can be transformed and standardized. The most unique characteristic of some effect sizes is that they are unit-free, like Pearson product moment correlation, little $r$, and proportions. Others are metric squares like big $R^2$, an effect size that tells the researcher how much variance is accounted for by their statistical model.
Most people are interested in knowing not only if an effect exists, but how large it is, and how substantively important it is. Cohen (1962) was not the first to introduce effect sizes – that was Neyman and Pearson (1928) – but he did popularized qualitative effect size magnitudes that addressed the questions of size differences and substantive importance. He suggested that his Cohen’s $d = .25$ indicated a small effect size, $d = .50$ corresponded to a medium effect size, and $d = .80$ signified a large effect size (Cohen, 1969). Cohen (1988) provided similar benchmarks for correlation effect sizes. Never did Cohen intend, however, for his descriptors to be conceptualized like T-shirt effect sizes (small, medium, large) and applied to every research area and every research question identically. How can a medium effect size for a drug given to chickens be exactly the same as the medium effect size for a specific instructional technique on the expression of well-formed Spanish subjunctives? Cohen provided us with rough ideas on how to interpret effect sizes, but research communities across the disciplines took his cutoff suggestions literally, applying them blindly without regard to their own empirical literature (Ellis, 2010). The best way for us to judge the relative magnitude for effects of interest within linguistics is to consult relevant meta-analytic studies (Plonsky, 2014). Meta-analytic thinking and meta-analysis as a goal are important parts of the new statistics.

**Confidence Intervals**

The relevance of such confidence intervals must not be underestimated: without a confidence interval it is unclear how well you can generalize from a sample to a population. (Gries, 2013, p. 132)
A confidence interval (CI) can be defined as a point estimate of a parameter or an effect size plus or minus a margin of error (Ellis, 2010). It is a range of plausible values for the index or parameter being estimated, thus it quantifies and visually represents the accuracy of a point estimate. Confidence intervals are relevant whenever an inference is made from a sample to a wider population (Hatcher, 2013). Interval estimates should be given for any effect sizes that involve principal outcomes. It is important to provide intervals for correlations and other coefficients of association or variation whenever it is possible (Cumming, 2014). The *Publication Manual* (APA, 2010) also recommended CIs: “Whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size” (p. 34).

Confidence intervals are typically shown in graphics as error bars represented as lines that extend above and below, or to the left and right around a point that corresponds to a statistic. The APA (2010) specifies the best format for reporting CIs, and they include CI examples for a wide range of measures. Typically, if the assumption of random sampling from a normally distributed population is met, and we are focusing on means (\( \bar{x} \)), a 95% confidence interval is constructed so that the interval will include \( \mu \), the population parameter (the mean we are attempting to estimate accurately), for 95% of an infinite sequence of replications. Any single confidence interval calculated from the data may or may not include \( \mu \), the population parameter, but the researcher never knows which one it is. As such, we are not guaranteed that the population parameter, \( \mu \) in this case, will be included in the confidence interval. We might construct a 95% confidence interval based on the mean in a new sample, but we will see that the center and endpoints
of our new interval will be different from the center and endpoints constructed for our old sample. This is no surprise since confidence intervals are subject to sampling error and assume that other sources of error are nil. Theoretically, in endless repetitions of 95% CIs, roughly 95% of the CIs will include the population parameter, and about 5% of them will miss it. That is, we will miss the true population parameter one (1) out of every twenty (20) attempts.

Figure 1 from Cumming (2014, p. 12) provides an excellent visual overview of confidence intervals as well as the volatile $p$ values that accompany them. From Cumming (2014), described periphrastically here, Figure 1 shows simulated results of 25 replications of an experiment numbered on the left. Each experiment comprised two independent samples with an $n$ of 32. The samples were from normally distributed populations with $\sigma = 20$, and with means that differed by $\mu = 10$. For each experiment, the difference between the sample means (circle) and the 95% confidence interval (CI) for this difference are shown. The $p$ values in the list on the left are two-tailed, for a null hypothesis of zero difference, $\mu_0 = 0$, with $\sigma$ assumed to be unknown (*$.01 < p < .05$, **$.001 < p < .01$, ***$p < .001$; a question mark indicates $.05 < p < .10$). The population effect size is 10, or Cohen’s $\delta = 0.5$, which is conventionally considered a medium-sized effect. Mean differences whose CI does not capture $\mu$ are shown in red. The curve is the sampling distribution of the difference between the sample means; the heavy line spans 95% of the area under the curve.
It is important to recognize that confidence intervals are not just NHST in disguise (Kline, 2013; Thompson, 2006). Null hypotheses are required for NHST, but they are not required for confidence intervals.

Confidence intervals are superior to $p$ values because confidence intervals constructed around a statistic contain all the information that $p$ values contain, but the $p$ values do not contain all the information provided by confidence intervals. Although we should avoid being transfixed by statistical significance, inquiring minds might want to
know that an estimate is significant at the $p < .05$ level if we construct a 95% CI and it does not include zero (usually). Similarly if we construct a 99% CI and we find that it includes zero, or another value from our null hypothesis, we know that the estimate is not significant at the $p < .01$ level. However, CIs are not to be conceptualized or used as a proxy for NHST thinking because “if we mindlessly interpret a confidence interval with reference to whether the interval subsumes zero, we are doing little more than nil hypothesis statistical testing” (Thompson, 1998, pp. 799-800).

Another preferable characteristic of a confidence interval is that degrees of uncertainty in statistical estimates become much more clear. As discussed, statistical testing encourages dichotomous thinking and binary decisions. Dichotomous thinking misleads us to believe that we are making our decisions with much more certainty than what is realistic given all of the statistical assumptions and statistical uncertainty we handle. The ESCI approach, or Effect Size + Confidence Interval approach, often makes uncertainty more clear (and honest) in a visual format. This approach also tells us more about study replicability whereas $p$ values do not (Schmidt & Hunter, 2014).

**Meta-analysis**

One of the basic presumption of science is that of replication. Usually meta-analyses estimate average weighted effect sizes from primary studies, and then they evaluate whether various study factors, like participants or treatments characteristics, explain variation in the effect sizes targeted for analysis. A good meta-analysis summarizes the status of the literature and it suggest new directions. Plonsky (2014)
found this to be one of its benefits in SLA research. Decades ago, Wilkinson and Inference (1999) outlined what Cumming (2012) calls “meta-analytic thinking”:

Reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. ... Reporting effect sizes also informs...meta-analyses needed in future research. ... Collecting intervals across studies also helps in constructing plausible regions for population parameters. (Wilkinson & Inference, 1999, p. 599)

...

Do not interpret a single study’s results as having importance independent of the effects reported elsewhere in the relevant literature. ... The results in a single study are important primarily as one contribution to a mosaic of study effects. (Wilkinson & The APA Task Force on Statistical Inference, 1999, p. 602)
Chapter 3

REVIEW OF THE LITERATURE

My review of the literature includes contributions from the social and behavioral sciences, linguistics, and sociolinguistics. As Gries (2015) states, there is: “a need for linguists to recognize the multitude of techniques that are already regularly applied in fields that struggle with data that pose the exact same distributional challenges that linguists face every day” (p. 6).

Literature

Engman (2013) reported that, as occurs in linguistics, an overwhelming majority of quantitative work in sociology reports levels of statistical significance. He found that significance was reported with little or no discussion of what it actually entails philosophically. Thus Engman adds his name to a long list of authors who aver that this lack of discussion and understanding causes serious problems when analyses are interpreted. He contends that too often significance is understood to represent the probability of the null hypothesis (usually understood as a lack of relationship between two or more variables). This understanding, he states, is “simply erroneous” (p. 257). The first section of Engman’s paper dealt with this common misunderstanding. The second section gave a history of significance testing in the social sciences, with reference to the historical foundations of many common misinterpretations of significance testing. The third section was devoted to a discussion of the consequences of misinterpreting statistical significance for sociology. Engman argued that reporting statistical significance
provides researchers with very little value, and that the consequences of misinterpreting
significance values outweighs the benefits of their use.

Gries (2015) surveyed a variety of methodological problems in current
quantitative corpus linguistics. Some problems discussed were from corpus linguistics in
general, such as the impact that dispersion, type frequencies/entropies, and directionality
(should) have on the computation of association measures as well as the impact that
neglecting the sampling structure of a corpus can have on statistical analyses. Others
involved more specialized areas in which corpus-linguistic work is currently booming,
such as historical linguistics and learner corpus research. For each of these problems, first
ideas as to how these problems can be resolved are provided and exemplified in some
detail. Sociolinguists, many using corpora, would do well to borrow methodological ideas
from corpus linguistics. Corpus linguistics still has work to do, but they have made
“enormous headway” (p. 113), although progress still needs to be made in terms of their
fundamental understanding of distributional hypotheses, quantification of co-occurrence,
handling of underdispersion, and handling of the violation of the fundamental assumption
of data point independence found in chi-squared tests, simple correlations, and binary
logistic regression.

Johnson (2009) advocated for a change of variable rule (Varbrul) software
applications, from Goldvarb to Rbrul. Johnson suggested that Varbrul has been employed
successfully for over three decades to quantitatively assess the influence of multiple
factors on linguistic variables. Having made that claim, however, he added that Goldvarb
lacks flexibility, and it also isolates its users from the wider community of quantitative
linguists. A new version of the variable rule program, Rbrul, he believes, will resolve these concerns; and given that mixed-effects modeling is possible, the program addresses a more serious problem whereby Goldvarb overestimates the significance of effects. He indicated that Rbrul provides superior performance on both simulated and real data sets.

It is true that Goldvarb cannot fit mixed models, and the ability to do so easily is the most toothsome advantage of Rbrul. Most commercial statistical packages also support mixed-model analysis, including the popular and free statistical analysis software R for which Rbrul provides an interface.

Plonsky (2011b) documented the need for meta-analysis in Second Language Acquisition (SLA), and he offered an interesting example for synthesis. The central problem addressed in his meta-analysis was the “extensive yet inconclusive” (p. 993) research on the effects of second language strategy instruction (SI). He performed his meta-analysis aimed at providing a reliable, quantitative measure for the effect of SI, and to describe the relationship between SI and the variables that moderate its effectiveness, such as different learning contexts, treatments, and outcome variables. He conducted a comprehensive search to collect a population of SI studies. He then calculated effect sizes for 61 primary studies that provided a total of 95 unique samples, which he then coded for potential moderators. Plonsky found a small to moderate overall effect of SI ($d = 0.49$). He also found that that the variables that moderate its effectiveness include type and number of strategies, learning context (second language versus foreign language), and the length of intervention. Following a contextualized interpretation of the results, Plonsky’s article concluded with a discussion of the theoretical, practical, and
methodological implications. By using meta-analysis – which is almost entirely ignored in variationist sociolinguistics – Plonsky successfully produced evidence to support claims of a relationship between specific methodological characteristics of primary studies, such as pretesting, random group assignment, reporting of reliability, and the effects of SI they produce.

Plonsky (2013) assessed research and reporting practices in quantitative second language (L2) studies. Plonsky built a sample of 606 primary studies, published from 1990 to 2010, in Language Learning and Studies in Second Language Acquisition. He collected and coded for designs, statistical analyses, reporting practices, and effect sizes. The data collected for his study provided information on a number of characteristics found in L2 research that was previously unknown. These characteristics included the frequency of null hypothesis statistical tests (NHST), the extent to which L2 research was lab versus classroom based, and the inclusion of delayed posttests in experimental research. Plonsky observed many methodological weaknesses in his sample of primary studies, and he suggested that the concerns raised by these weaknesses merit serious attention. According to Plonsky, for example, the Discussion section needs field-wide reform in terms of data analyses and data-reporting practices, as well as a far greater consideration of statistical power and issues related to statistical significance. He hopes to stimulate SLA researchers to reflect on and investigate the methodological ways in which L2 research is carried out and reported. Plonsky recommended that concrete action should be taken by researchers, and related institutions should enact reform. Reform should be pursued to better control experimental designs, address incomplete and
inconsistent reporting practices, and surmount problems with statistical power (the primary indicator of replicability in significance testing). Very clearly, Plonsky’s suggestions are applicable to other fields of linguistic inquiry, including sociolinguistics.

Plonsky (2014) continued to explore methodological practices in quantitative second language (L2) research. Here he employed techniques to examine changes over time in SLA research and reporting practices. Plonsky used 606 primary reports of quantitative L2 research from the journals *Language Learning* and *Studies in Second Language Acquisition*. He surveyed design features, statistical analyses, and data reporting practices. In an effort to reveal changes taking place in the field, Plonsky calculated frequencies and percentages of each feature and compared them across the 1990s and the first decade of the 2000s. The results revealed a large number of changes including an increase in sample sizes and delays in post-testing. Critical data also began to emerge, including effect sizes, reliability estimates, and standard deviations began to accompany means. The range of statistical procedures had not changed, however, and “the field continues its unfortunate reliance on statistical significance” (p. 450). Plonsky’s findings were grouped into three themes, which he discussed in light of previous reviews in SLA and other fields: (a) means-based analyses, (b) missing data, null hypothesis significance testing, and the problem with statistical power, and (c) design preferences. Plonsky concludes with a clarion call for reform targeting field-specific methodological standards and the need for improvements in graduate curricula and training. Plonsky’s call for reform is equally applicable to linguistics in general, and Hispanic variationist sociolinguists in particular.
Referenced by Tagliamonte (2011), Roy (submitted) uses stative possessives to illustrate the statistical techniques used in variable rule (Varbrul) analysis. Roy presented the extended benefits of using logistic regression directly rather than using a variationist-specific statistical technology. Alternative methods of assessing statistical significance were used and Equivalence Constraints were introduced as a way to integrate confidence intervals produced in logistic regression into the analytic methodology of Varbrul analysis.

Roy suggested that alternatives to the model-based assessment of statistical significance for factor weights as a way to assess the statistical consistency of factor groups. He also explored alternative methods for integrating social factors into the statistical model. He then asked: Do the alternatives justify moving from Goldvarb to a full logistic regression model? Roy confirmed and argued for this change. He emphasized that the field of statistics has increased the number of models available for analyzing data. In the face of these advancements, he stated that he did not believe that the small pedagogical and methodological benefit for variationists justify maintaining a non-standard statistical program. Roy averred that keeping a non-standard statistical program will hinder accessibility of successive generations of sociolinguists to advances in regression based approaches that have been made mainstream in statistics, not to mention alternatives such as Bayesian and classification based models that remain unused in contemporary sociolinguistic methodology. Roy suggested that updating the statistical methodology will allow variationist sociolinguists to expand the class of research questions that they can reliably answer. Sociolinguists, he suggested, will be able to
perform fine-grained cross-cohort comparisons once the inherent variability of factor weights are accounted for, and methods will improve for accessing a factor group’s statistical significance. This can occur, he added, while maintaining the analytic methodology used in previous studies conducted with Goldvarb. Roy indicated that adopting a statistical methodology that can be readily extended to other statistical innovations, including mixed-effects models, Bayesian approaches, or even differently distributed response variables, is a benefit that cannot be ignored. According to Roy, “if we want to communicate our results to domains outside of variationist sociolinguistics, we should use a standard statistical approach that can be more readily communicated with other quantitative social (and non-social) scientists as well other quantitative sub-disciplines of linguistics” (p. 14) Roy calls for a change in the “academic homeostasis” (p. 14) that has gripped sociolinguistics for over forty years, all while statistics has gone on to develop new techniques, algorithms and procedures. Roy’s concerns of “academic homeostasis” echo the concerns of (Gorman & Johnson, 2013) of a “statistical monoculture” (p. 236) in sociolinguistics.

**Research Questions**

The following research questions were addressed within the descriptive limits of the present thesis:

RQ1: To what extent can Hispanic sociolinguistics be considered a “statistical monoculture”?

RQ2: To what extent does Hispanic sociolinguistic research reflect an accurate understanding of statistical testing?
RQ3: To what extent does Hispanic sociolinguistic research reflect statistical best practices as defined by the new statistics, namely, the proper use and interpretation of effect sizes and confidence intervals?

RQ4: To what extent does Hispanic sociolinguistic research lend itself to meta-analysis?
Chapter 4

METHOD

To answer the four research questions introduced in Chapter 3, a representative group of studies were identified in journals known to publish Hispanic sociolinguistic research. Although the word “representative” may imply that inferences about the larger population of Hispanic sociolinguistic studies might be made from the selected studies, my objective here was to provide insight into trends in the use and interpretation of statistics in specific journals. Thus, the studies selected, limited to the last decade and to specific academic journals, are the population under study. Providing inferential measures when the actual population is known and studied is not meaningful since there is nothing to infer (Vogt, 2014). As such, this work is descriptive, circumscribing the journals and studies that were selected.

Journal Selection

I restricted the present study to peer-reviewed, academic journals in linguistics, making the assumption that, as occurs in second language acquisition (Plonsky, 2011), journals are the primary means by which Hispanic sociolinguistic research is disseminated. Books, book chapters, and conference materials were excluded.

To identify relevant journals, an initial search for Hispanic sociolinguistic studies was conducted using the databases available through ProQuest, Cambridge Journals Online, De Gruyter Online, ScienceDirect, Springer Link, Taylor & Francis Online, and the Wiley Online Library. The search was restricted to studies published in academic journals between April, 2005 and April, 2015. Eleven journals were identified:
I was going to base my final selection of academic journals on the highest Journal Impact coefficients found in the *Journal Citation Reports* (Thomson Reuters, 2013); however, little to no relation between high impact journals and statistical standards have been found in other areas of social and behavioral research (Tressoldi, Giofré, Sella, & Cumming, 2013). A review of the Hispanic sociolinguistic study search results among the initial twelve journals selected – although more impressionistic than quantified – showed that statistical practices were limited to a handful of methods in each journal depending on the research domain within linguistics, and the types of variables they tended to handle. For example, SLA dedicated much effort to processing ANOVA’s in SPSS using a variety of metric variables, and sociolinguists often used logistic regression with fixed
effects in Varbrul using discrete variables. This study would have to be replicated among all of the journals to verify its results, but some subjective, non-statistical inference to the larger population of Hispanic sociolinguistic studies across the journals would be somewhat reasonable given my initial analysis of statistical methods deployed among all twelve journals during the selection process, but inferences would need to be proven.

In deciding on the final journals, I gave special emphasis to journals that contained a greater number of variationist studies, and more variability in terms of its representativeness of different university locations and university types within the Carnegie Classification of Institutions of Higher Education (Carnegie Foundation, 2010). Journals also had to have at least 10 studies in Hispanic sociolinguistics in the last decade to be considered, and their content had to be accessible. For example, The Modern Language Journal was excluded due to a lack of at least 10 Hispanic sociolinguistic studies according to the criteria given below, and Spanish in Context was excluded due to access restrictions. The final academic journal choices were: Hispania, the Journal of Sociolinguistics, and Language Variation and Change.

**Study Selection**

Hispanic variationist studies were the target. Studies written in English and Spanish were included. Studies that focused mainly on applied linguistics in language learning and second language acquisition, or on corpus linguistics, or psycholinguistics were excluded, although any study within the selected journals that deployed variationist topics and methods throughout were included.
Only quantitative Hispanic variationist studies were included. A study was considered Hispanic if the researcher investigated the use and understanding of the Spanish language in any regional, cultural, or historical context. A study was considered quantitative if it included frequencies and percentages, at minimum in the form of a summary table. Studies that were qualitative were excluded, even if frequencies and percentages were mentioned in the prose. The threshold between qualitative and quantitative studies was whether or not the author chose to create summary tables to make sense of their data. If the author did so, their study was included. All studies that used summary statistics or any statistical method to learn of the data were included. Beyond these selection criteria, I did not attempt to select subsets of studies from the different journals that used similar designs or measures. Designs, measures, and other aspects of quantification in sociolinguistic studies can vary, and they may influence the choice of statistical techniques.

Table 2: Final Study Choices by Journal

<table>
<thead>
<tr>
<th>Journal</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispania</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>Journal of Sociolinguistics</td>
<td>14</td>
<td>31</td>
</tr>
<tr>
<td>Language Variation and Change</td>
<td>18</td>
<td>42</td>
</tr>
<tr>
<td>All</td>
<td>44</td>
<td>100</td>
</tr>
</tbody>
</table>

After examining the data output, I decided to exclude three studies from Language Variation and Change and one from Hispania. These four (4) studies used calculations (frequencies and percentages) and tables that could be considered summary statistics, but they were ineligible for new statistics processing because the authors were not making statistical inferences. Instead, they were simply describing the population
under study. After carefully reviewing articles from among the original twelve academic journals, 44 articles were selected from three academic journals, totaling 12 from *Hispania*, 14 from the *Journal of Sociolinguistics*, and 18 from *Language Variation and Change*, as seen in Table 2. Appendix B lists the final sections, including years, author names, and article titles.

**Data Selection and Coding**

The method, results, and discussion sections of the selected studies were analyzed. I first determined each study’s design. I coded designs as observational, experimental, descriptive, and corpus. These variables were not considered mutually exclusive, but rather each study could be more than one. A study was considered observational if the researcher, or research assistants, personally witnessed linguistic phenomena by dint of their own fieldwork, whether it was directly or through their own recordings, but only statistical correlation, not causation, was possible. An experiment, or pseudo-experiment in strict terms, required some manipulation of variables by the researcher in such a way that some amount of causality could be reasoned or hoped for through qualitative or statistical reasoning. A descriptive study is correlational and if the researcher or their assistants did not observe the linguistic phenomena themselves, using the data they personally collected, I considered it descriptive. The variable *corpus* was important here because "given the sizes of many currently used corpora, even the smallest results will often be significant" (Gilquin & Gries, 2009, p. 17), which is relevant to understanding the (mis)use of statistical testing. Corpus studies were usually
observational and descriptive, especially when the researcher compared the corpus they built personally with corpora built by others.

Next, I focused on the statistical programs used and the statistical methods deployed. The statistical programs are relevant because each statistical package has its own set of options and default results. This can limit or empower the analyst. The statistical methods that researchers selected were important because they helped answer the first research question: To what extent can Hispanic sociolinguistics be considered a “statistical monoculture”? Also, each statistical method has its own, sometimes shared, assumptions to test for, and post hoc and model testing options.

The data reporting variables that I coded follow and expand on Plonsky (2011). Did the author report sample sizes, percentages, frequencies, correlations, means, standard deviations, or perhaps means without standard deviations? This reporting is important for analysts who might want to include studies in meta-analyses, besides meeting minimal reporting standards in many cases.

Since NHST is the firebrand of the new statistics, I focused first on how authors used and interpreted statistical testing. I took special note of their \( p \) value reporting and interpretation of \( p \) values. I sought out a priori \( \alpha \), the symbolic use of more than (\( > \)) and less than (\( < \)), and whether the author stopped to explain what their \( p \) value was actually measuring within the context of their study. Since power testing is desideratum in Neyman-Pearson analyses, I looked for power testing, both a priori and a posteriori. Neyman-Pearson’s dichotomous thinking in absence of power testing is common in the social and behavioral sciences (Murphy, Myors, & Wolach, 2014). This type of
selectiveness within the theories, often due to convention or being unaware of the full procedures, can lead to NHST fallacies.

Although the misapprehensions borne of statistical testing are often called “fallacies,” my purpose was not to expose authors to criticism, but rather to uncover trends among the authors, although this study would be incomplete without providing specific examples in the results. Of high importance, I wanted to know if the researchers discontinued their statistical testing at the \( p \) value, simply reporting \( p \) values as if their conclusions had been made, or as if low \( p \) values indicated large effects (Magnitude), rejection of the null hypothesis confirmed the alternative hypothesis (Meaningfulness), statistical significance meant that an underlying causal mechanism has been identified (Causality), low \( p \) values meant that studies were successful or high \( p \) values meant that studies were unsuccessful (Success/Failure), or arbitrary values, such as \( p = .049 \) versus \( p = .051 \), were considered completely different (Sanctification).

After examining the use and interpretation of statistical testing, I was interested in the use, reporting, and interpretation of the statistical methods grouped together and called the new statistics: effect sizes, confidence intervals, and meta-analysis. For effect sizes I took note of sociolinguists who (against better advice):

1. Applied T-shirt definitions for effect sizes.
2. Estimated effect sizes only for results that were statistically significant.
3. Reported values of effect sizes only as point estimates \( \textit{sans} \) CI.
4. Simply listed effect sizes, or whether they were interpreted in the context of their studies.
After reviewing the use and reporting of effect sizes, I looked for the presence of confidence intervals. I also examined how they were displayed and interpreted in prose. Like effect sizes, confidence intervals are often underreported and misinterpreted in the social and behavioral sciences (Cumming, 2012; Kline, 2013; Schmidt & Hunter, 2014).

Finally, I examined overall statistical reporting to determine if studies could be used in meta-analysis in the future. Meta-analyses can be very sophisticated, and far more than computational output is important, thus my inclusion criteria were very generous. First, I looked for a standardized index that could be an effect size, such as proportions from central tendencies, standardized mean differences (d measures, for example), such as group contrasts of continuous measures, a correlation coefficient (r measures, for example) in linear associations, and odds-ratio for group contrasts of dichotomous measures. I also looked for information on data cleaning and ranges, especially in relation to assumptions about the distribution requirements since estimates can be unstable and affected by outliers. I determined if “not significant” or “n.s.” or “p < 0.05” stood by themselves without exact p values because these notations make those measures hopeless for meta-analysis.

Those who perform meta-analyses also consider a number of biases. They look for preference for statistically significant studies (publication bias), quickly published studies (time lag bias), language bias in favor of English, repeated publishing of the same study, and citation bias. These biases were not considered in the present study. Without considering the biases that meta-analysts will seek, my inclusion were even more liberal. Being aware of biases, however, underscored the illogic of attempting to perform
inference in the present study from the studies found in the three journals selected. Each journal has its own editorial discretion in publishing, which may be seen as preferences by the editors. However, this discretion is likely to be seen as bias by a statistician. This bias, if confirmed, would violate the fundamental assumption of independence required in many, if not most, statistical procedures. It would also mean that the articles considered in the present study were not independent, or stochastic, even if a subset were taken. Inference might be possible for each individual journal from a subset in an effort to infer future editorial biases, but no inference could be made to all journals that publish Hispanic sociolinguistic research based on my biased sample from three journals.

Table 3 details the coding scheme, categories, variables, values, definitions and operationalization. I follow Plonsky (2011) in listing some of the analyses and data reporting variables, but I added variables to Identification, Design, Statistical Tools, Analyses, Data Reporting, and I included the NHST category with its Fallacy variables, and the new statistics categories that included confidence intervals, effect sizes, and meta-analysis. Values are yes (y), no (n), inferred (i), open (any text string), year (number), and “canary” (c). The curious value, “canary,” originates with Kline (2013) who strongly rejects statistical testing that occurs “under the hood” of larger operations when they are used to evaluate assumptions of other statistical tests. He calls them “canary in the coal mine tests…[because] they often depend on other assumptions, such as normality, that may be indefensible” (Kline, 2013, p. 107).
Table 3: Coding Scheme Categories, Variables, Values, and Definitions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values</th>
<th>Definition/operationalization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author(s)</td>
<td>open</td>
<td>Author or authors of each study report</td>
</tr>
<tr>
<td>Institution</td>
<td>open</td>
<td>University, college, or other institution represented by main author</td>
</tr>
<tr>
<td>Location</td>
<td>Open</td>
<td>Country of institution that main author represents</td>
</tr>
<tr>
<td>Journal</td>
<td>Hisp, JS, LVC</td>
<td>Journal in which the study was published, <em>Hispania</em>, <em>Journal of Sociolinguistics</em>, or <em>Language Variation and Change</em></td>
</tr>
<tr>
<td>Year</td>
<td>2005-2015</td>
<td>Year in which the article was published</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observational</td>
<td>y, n</td>
<td>The study did not test a causal relationship, but the author(s) made direct observations of the phenomena.</td>
</tr>
<tr>
<td>Experimental</td>
<td>y, n</td>
<td>Did the study test a causal relationship?</td>
</tr>
<tr>
<td>Descriptive</td>
<td>y, n</td>
<td>The authors did not personally observe the phenomena under study, or personally collect the data.</td>
</tr>
<tr>
<td>Corpus</td>
<td>y, n</td>
<td>The authors used their own or somebody else’s corpus.</td>
</tr>
<tr>
<td><strong>Statistical Tool</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software</td>
<td>open</td>
<td>Was SPSS, SAS, JMP, Stata, Minitab, Goldvarb, Excel, other, or none reported?</td>
</tr>
<tr>
<td><strong>Analyses</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td>y, n, c</td>
<td>Was any type of correlation carried out?</td>
</tr>
<tr>
<td>Chi-square</td>
<td>y, n, c</td>
<td>Were frequency data analyzed using a chi-square test?</td>
</tr>
<tr>
<td>z test</td>
<td>y, n</td>
<td>Was a z test (or a nonparametric equivalent) used?</td>
</tr>
<tr>
<td>t test</td>
<td>y, n</td>
<td>Was a t test (or a nonparametric equivalent) used?</td>
</tr>
<tr>
<td>ANOVA</td>
<td>y, n</td>
<td>Was an ANOVA (or a nonparametric equivalent) used?</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>y, n</td>
<td>Was an ANCOVA (or a nonparametric equivalent) used?</td>
</tr>
<tr>
<td>MANOVA</td>
<td>y, n</td>
<td>Was an MANOVA (or a nonparametric equivalent) used?</td>
</tr>
<tr>
<td>MANCOVA</td>
<td>y, n</td>
<td>Was an ANCOVA (or a nonparametric equivalent) used?</td>
</tr>
<tr>
<td>Reg Regression</td>
<td>y, n</td>
<td>Was regular, Ordinary Least Squares regression carried out?</td>
</tr>
<tr>
<td>Mixed Effects</td>
<td>y, n</td>
<td>Was mixed effects, hierarchical regression carried out?</td>
</tr>
<tr>
<td>SEM</td>
<td>y, n</td>
<td>Was structural equation modeling used?</td>
</tr>
<tr>
<td>DFA</td>
<td>y, n</td>
<td>Was a discriminant function analysis carried out?</td>
</tr>
<tr>
<td>PCA</td>
<td>y, n</td>
<td>Was principle component analysis carried out?</td>
</tr>
<tr>
<td>Factor analysis</td>
<td>y, n</td>
<td>Was a factor analysis carried out?</td>
</tr>
<tr>
<td>Cluster analysis</td>
<td>y, n</td>
<td>Was cluster analysis carried out?</td>
</tr>
<tr>
<td>LR Fixed Effects</td>
<td>y, n</td>
<td>Was logistic regression with only fixed effects used?</td>
</tr>
<tr>
<td></td>
<td>y, n</td>
<td>Assumption</td>
</tr>
<tr>
<td>---------------------</td>
<td>------</td>
<td>--------------</td>
</tr>
<tr>
<td>LR Mixed Effects</td>
<td></td>
<td>y, n</td>
</tr>
<tr>
<td>Other</td>
<td>open</td>
<td></td>
</tr>
<tr>
<td>Assumptions</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Model testing</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Data Reporting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample size</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Percentage</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Mean without SD</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>NHST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p values</td>
<td>open</td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>p &lt; or &gt;</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>A priori alpha</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>p explained</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theory Conflation</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Power a priori</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Power a posteriori</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Fallacy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odds Against Chance</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Local Type I</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Inverse Probability</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Validity</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Replicability</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Magnitude</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Meaningfulness</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Causality</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Zero</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Equivalence</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Quality</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Success</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Failure</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Reification</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Objectivity</td>
<td>y, n</td>
<td></td>
</tr>
<tr>
<td>Sanctification</td>
<td>y, n</td>
<td></td>
</tr>
</tbody>
</table>
In Table 3 above I extend the denomination “canary” to tests that require assumptions to be met, but are not typically tested, such as the chi-square operations that occur in logistic regression and often get reported in Varbrul analyses. I coded for them because authors often interpret them (without assumption tests), but I did not intend to bring them to discussion unless I detected something unusual (like consistently testing assumptions, unfortunately).

For the variable Theory Conflation under the category NHST in Table 3 above, I examined whether authors conflated or confused the theories of Fisher and Neyman-Pearson on any one of the concepts as listed in Table 4. This table was adapted from Perezgonzalez (2015, pp. 9-10).
Table 4: Conflation of NHST Constructs

<table>
<thead>
<tr>
<th>Concept</th>
<th>Fisher</th>
<th>Neyman-Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach</td>
<td>A posteriori.</td>
<td>A priori (in part).</td>
</tr>
<tr>
<td>Research goal</td>
<td>Statistical significance of research results.</td>
<td>Deciding between competing hypotheses.</td>
</tr>
<tr>
<td>Hs under test</td>
<td>Null hypothesis (H₀) to be nullified with evidence.</td>
<td>Main hypothesis (Hₘ) to be favored against research hypothesis (Hₐ).</td>
</tr>
<tr>
<td>Probability distribution of test</td>
<td>As appropriate for null hypothesis (H₀).</td>
<td>As appropriate for main hypothesis (Hₘ).</td>
</tr>
<tr>
<td>Cut-off point</td>
<td>Significance identifies noteworthy results that can be gradated and corrected a posteriori.</td>
<td>Common to critical value of the test (CVₜₑₛₜ) α, β, and minimum effect size (MES); cannot gradate nor corrected a posteriori (in part).</td>
</tr>
<tr>
<td>Sample size calculator</td>
<td>None.</td>
<td>Based on test statistic, effect size, α, and power (1 − β).</td>
</tr>
<tr>
<td>Statistic of interest</td>
<td>p value, as evidence against the null (H₀).</td>
<td>Critical value of the test (CVₜₑₛₜ); p values have no inherent meaning but can be used as a proxy.</td>
</tr>
<tr>
<td>Error prob.</td>
<td>α possible, but not relevant to single studies.</td>
<td>α = Type I error prob. β = Type II error prob. (partly).</td>
</tr>
<tr>
<td>Result falls outside critical region</td>
<td>Ignore result as not significant.</td>
<td>Accept main hypothesis (Hₘ) if power good; in not, conclude nothing.</td>
</tr>
<tr>
<td>Result falls in critical region</td>
<td>Reject null hypothesis (H₀).</td>
<td>Reject main hypothesis (Hₘ) and accept research hypothesis (Hₐ).</td>
</tr>
<tr>
<td>Interpretation of results in critical region</td>
<td>Either a rare event occurred or the null hypothesis (H₀) does not explain the research data.</td>
<td>Hₐ explains research data better than (Hₘ) does (given α).</td>
</tr>
</tbody>
</table>
Perezgonzalez (2015) calls attention to interesting differences between the theories of Fisher and Neyman-Pearson that are often left unclear or unaddressed in linguistics books on quantitative methodology. Most relevant here is the distinction he makes between the null hypothesis of Fisher, $H_0$, and what he calls the “main hypothesis” (p. 4) of Neyman-Pearson, or $H_M$. Our quantitative methodology references usually treat the null hypothesis (Fisher) and main hypothesis (Neyman-Pearson) as if they were identical, or $H_0 = H_M$. This is not surprising since Neyman-Pearson called their $H_M$ a “null hypothesis” just as Fisher did. As Perezgonzalez clarifies, this similarity is “merely superficial” (p. 5). $H_M$ is to be considered at the design stage while $H_0$ is only rarely made explicit; and $H_M$:

\[ \ldots \text{is implicitly designed to incorporate any value below the MES [Minimum Effect Size] – i.e., the a priori power analysis of a test aims to capture such minimum difference} \] (effect sizes are not part of Fisher’s approach); and it is but one of two competing explanations for the research hypothesis ($H_0$ is the only hypothesis to be nullified with evidence). (Perezgonzalez, 2015, p. 5)

In the present thesis this difference is brought to bear on studies that appear to favor the Neyman-Pearson approach, but do so incompletely. The $H_M$, or main

---

Table 4: Conflation of NHST Constructs

<table>
<thead>
<tr>
<th>Concept</th>
<th>Fisher</th>
<th>Neyman-Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Next steps</td>
<td>Rejecting the null hypothesis ($H_0$) does not automatically justify “not $H_0$.” Replication is needed, meta-analysis is useful.</td>
<td>Impossible to know if $\alpha$ error made. Repeated sampling of same population needed, Markov Chain Monte Carlo (resampling) useful.</td>
</tr>
</tbody>
</table>
hypothesis (used for exposition but hereafter I do not distinguish as such), assumes that research hypotheses are presented, a priori statistical power analyses are performed, and error Types I and II are considered as well.
Chapter 5

RESULTS

Results for Research Question 1: To what extent can Hispanic sociolinguistics be considered a “statistical monoculture”?

The first research question was motivated by comments from Gorman and Johnson (2011) in a book chapter section called “Against a Statistical Monoculture” in which they lament that the “collaboration between statisticians and sociolinguists in the 1970s was a fruitful one, but advances in statistics since then have been slow to diffuse into sociolinguistic practice” (p. 236). Bayley (2013) joins Gorman and Johnson (2011) when he comments on the need for more powerful statistical applications and methods in variationist sociolinguistics because, as he optimistically points out, “new creative possibilities for quantitative analysis will doubtless open up” (p. 103). We can include Roy (Submitted) in this sentiment given his concerns of “academic homeostasis” (p. 14) in sociolinguistics. I set out to evaluate whether these concerns were valid for the studies I reviewed.

It is worth noting that some statistical methods use correlation and chi-square in their processing. For example, logistic regression will use both as “canary in the coal mine tests” as Kline (2013, p. 107) calls them. Ostensibly, logistic regression is a nonparametric technique, meaning that the researcher does not need to worry about distributional assumptions like normality or homoscedasticity. However, a test statistic like chi-square does make assumptions about normality, the distribution of cell counts, for example, and it is parametric. If chi-square, a parametric procedure, is nested in
logistic regression, a nonparametric procedure, then logistic regression cannot be considered entirely nonparametric. When a statistical test takes place under the hood of a larger statistical operation, it was coded as “c” for “canary.” While analyzing the output, “c” did not tell a story, except that it was a nested procedure. My focus was on whether the analyst used specific tests deliberately. Deliberate use of statistical tests introduces clearer responsibility for assumption checking, and different data analytic choices change data reporting conventions, interpretations of data output, and quite possibly, justification for and interpretation of statistical testing within the context of each study.

Additionally, the studies reviewed among all three journals often contained more than one statistical method. For example, any single study could use both ANOVA and logistic regression with fixed effects, or a subroutine within a routine (a “canary”) could co-occur. With that in mind, what follows are the numbers and percentages of methods that were present, even if they were present in more than one study.

In aggregate, logistic regression with fixed effects was used 22 times, that is, found in 50% of all studies. Logistic regression with fixed effects were used entirely in Varbrul using Goldvarb X (some authors simply wrote “Varbrul,” so an earlier version could have been used). ANOVAs were used 14 times, appearing in 32% of all studies. Correlation was used appears nine (9) times, or 20%, but 50% of the time they were “canary” variables, meaning they were represented in 10% of all studies deliberately. The studies revealed that chi-square was targeted six (6) times, 14%, but 50% canary variables, as such represented deliberately in 7% of all studies. T-tests were done four (4) times, therefore they were detected in 9% of all studies. Logistic regression with mixed
effects was used four (4) times, that is, detected 9% of the time. Mixed effects OLS regression was used three (3) times, being found 7% of the time. The z-test was used two (2) times, which turns out to be 5% of the time. Regular OLS regression, principal component analysis, and cluster analysis were each used one (1) time, which works out to just under 2% each for a study that uses 44 cases (studies). See Appendix A for a breakdown of the frequencies and percentages of statistical methods detected.

Since the distribution of studies was uneven, it was useful to examine if journals contributed different numbers of statistical methods. My objective was to examine the use of more complex methods: ANOVA, regression, and logistic regression. Focusing on these methods, 57% of the ANOVAs were found in *Hispania*, 36% in JS, and 7% in *Language Variation and Change*. Regular OLS regression was not found in *Hispania* or in the *Journal of Sociolinguistics*, rather *Language Variation and Change* contributed all of them. For mixed effects OLS regression, *Hispania* contributed no studies, *Journal of Sociolinguistics*, 67%, and *Language Variation and Change*, 33%. *Hispania* and *Journal of Sociolinguistics* each contributed 14% of the logistic regressions with fixed effects, and *Language Variation and Change* contributed 73%. Finally, 50% of the logistic regressions with mixed effects came from the *Journal of Sociolinguistics* and *Language Variation and Change* evenly, each contributing two (2) studies each. Totals for the top two methods, logistic regression with fixed effects and ANOVA, are represented in Table 5 below.
Table 5: Top Two Methods Represented Per Journal

<table>
<thead>
<tr>
<th>Journal</th>
<th>ANOVA</th>
<th>JS</th>
<th>LVC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>18% (8)</td>
<td>11% (5)</td>
<td>2% (1)</td>
<td>32%</td>
</tr>
<tr>
<td>Varbrul</td>
<td>[14%]</td>
<td>[14%]</td>
<td>[73%]</td>
<td>Total</td>
</tr>
<tr>
<td>y</td>
<td>7% (3)</td>
<td>7% (3)</td>
<td>36% (16)</td>
<td>50%</td>
</tr>
<tr>
<td>All</td>
<td>27% (12)</td>
<td>32% (14)</td>
<td>41% (18)</td>
<td>100%</td>
</tr>
</tbody>
</table>

Note: In red are the total percentages contributed by each journal to the method indicated to the left. In green, the total percentage of all methods for that particular method. The “All” row shows the percentage that the corresponding journal contributed to the total number of articles (44).

Results for Research Question 2: To what extent does Hispanic sociolinguistic research reflect an accurate understanding of statistical testing?

Using the NHST fallacies as labeled by Kline (2013), I totaled the number of fallacies across all studies. Although any single author could commit fallacies at different rates, each study could only contribute one (1) time to each fallacy total, regardless of the internal same-fallacy repetitions within each study. Also, in order to be listed as a fallacy, the authors had to explicitly express the fallacy in no uncertain terms or they had to imply the fallacy (coded as “i”) with their analytical decisions or in their results, discussions, or in their conclusions. I did not consider it meaningful to separate fallacy totals by journal.

The Odds Against Chance and Local Type I fallacies required an explicit statement since there are few indirect ways of committing them. The Odds Against Chance Fallacy is the belief that $p$ indicates the probability that results are the result of sampling error, and the Local Type I Fallacy is the belief that a rejection of $H_0$ means Type I error is less than 5%, when $p < .05$, given $\alpha = .05$. These two were all but absent.
The Hispanic sociolinguistic research contributions provided only one (1) instance of the Odds Against Chance Fallacy, and no Local Type I Fallacies.

Unlike the above two fallacies, the Inverse Probability was explicit in 95% of the studies. The Inverse Probability is the mistaken notion that the conditional probability is the probability of the null hypothesis given the data and its assumptions, or $p(H_0 \mid D^+)$. These results are displayed in Figure 2. Where not applicable, “na” was coded.

**Figure 2: Inverse Probability**

<table>
<thead>
<tr>
<th>Level</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>1</td>
</tr>
<tr>
<td>na</td>
<td>1</td>
</tr>
<tr>
<td>y</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
</tr>
</tbody>
</table>

The Validity Fallacy is the belief that the probability that the alternative hypothesis is true is $>.95$, given $p<.05$. Like the Inverse Probability, the Validity Fallacy appeared 95% of the time as seen in Figure 3.
The Replicability Fallacy is the belief that statistically significant results imply that the study would replicate with the same results. Relatively few studies, four (4) total, implied that significant results will replicate successfully, as seen in Figure 4.

The Magnitude Fallacy is the belief that low $p$ values indicate large effects. This fallacy was explicit in 86% of the studies, and implied 5% of the time, as seen in Figure 5.
The Meaningfulness Fallacy is the belief that rejection of $H_0$ confirms $H_1$. This belief was explicit 95% of the time, and implied 5% of the time, as seen in Figure 6.

Figure 6: Meaningfulness

The Equivalence Fallacy occurs when the failure to reject $H_0: \mu_1 = \mu_2$ is understood to mean that the populations are equivalent. Although expressed explicitly only 11% of the time, it was implied 89% of the time, as shown in Figure 7.
The Quality and Success Fallacies occur when the researcher believes that achieving statistical significance speaks to the quality of the study’s design and indicates that a particular study has succeeded. This fallacy dyad occurs explicitly 59% of the time, and it was implied the remainder 41% of the time, as seen in Figure 8.

The Failure Fallacy is the belief that lack of statistical significance means the study has failed. If the author stated that a past study, not just their own, failed to provide meaningful results due to lack of statistical significant results, or if they implied the same, their study was coded under this fallacy. In total, 61% of the authors indicated clearly that lack of statistical significance led to failure to provide meaningful results for their study.
or another study. The remaining 39% made no reference to failure in their study or another study based on “unachieved” statistical significance, as seen in Figure 9.

Figure 9: Failure

<table>
<thead>
<tr>
<th>Level</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>17</td>
</tr>
<tr>
<td>y</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
</tr>
</tbody>
</table>

The Reification Fallacy occurs when a result is not considered replicated if $H_0$ is rejected in one study but not in another. Only 9% of all studies explicitly spoke of failure to replicate due to lack of statistical significance, as seen in Figure 10.

Figure 10: Reification

<table>
<thead>
<tr>
<th>Level</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>40</td>
</tr>
<tr>
<td>y</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
</tr>
</tbody>
</table>

I did not include figures for several fallacies that were nearly completely or completely categorical. Perhaps most noteworthy, the Causality and Sanctification Fallacies were present in 100% of the studies. The Causality Fallacy occurs when statistical significance is taken to mean that an underlying causal mechanism has been
identified. The Sanctification Fallacy occurs when binary thinking rules the results, for example, \( p = 0.049 \) versus \( p = 0.051 \) is considered completely different.

Likewise, the Robustness Fallacy was inferred in all studies, except one (1). This fallacy occurs when the researcher assumes that NHST is robust against outliers or violations of distributional assumptions. Considering that only one author made reference to distributional issues in NHST, this fallacy was found throughout most of the studies.

The Zero and Objectivity Fallacies were inferred in 100% of the studies. The Zero Fallacy is the belief that failing to reject a nil (null) hypothesis means that the population effect size is zero. The Objectivity Fallacy is the belief that significance testing is an objective method of hypothesis testing while other inference models are considered subjective. I coded this fallacy as inferred if the author made conclusions based on \( p \) values, as if they were the final objective in scientific statistical testing.

I also investigated the manner in which \( p \) values were reported because \( p \) value reporting practices often pointed to how authors approached significance, whether it was Fisherian or Neyman-Pearsonian, or a mashup of both. I began with the reporting of alpha, \( \alpha \). Alpha was reported 16% of time, and ignored 84% of the time. I considered mistakes like “\( p < 0.05 \) was set at the beginning” as alpha reporting, even though \( \alpha \), the Greek letter itself, was not mentioned as expected. I found that \( p \) values were reported 91% of the time, and they were set equal to some value (=) 75% of the time, and they were weighed using greater than (>) or less than (<) symbols 70% of the time. Authors attempted to explain what a \( p \) value meant 9% of the time, and the meaning of \( p \) was ignored or taken as axiomatically understood 91% of the time.
NHST is the conflation of incompatible elements of Fisher’s and Neyman-Pearson’s theories of significance. Authors had to conflate these theories very obviously to be coded under theory conflation. As seen in Figure 11, 64% of the time authors used incompatible elements from Fisher or Neyman-Pearson, as listed in Table 4, Chapter 4. This conflation was not obvious or absent 36% of the time.

Figure 11: Theory Conflation

<table>
<thead>
<tr>
<th>Level</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>12</td>
</tr>
<tr>
<td>y</td>
<td>32</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
</tr>
</tbody>
</table>

Especially relevant to Neyman-Pearson’s concept of significance, statistical power was never used nor spoken of. According to best practices, if an author uses a $p$ value, a priori power should be assessed. A posteriori power is recommended.

Results for Research Question 3: To what extent does Hispanic sociolinguistic research reflect statistical best practices as defined by the new statistics, namely, the proper use, reporting, and interpretation of effect sizes and confidence intervals?

Effect sizes were reported 64% of the time, and not reported the remaining 36% of the time, as seen in Figure 12. Effect sizes were possible, however, in all studies.
Confidence intervals appeared rarely, 7% of the time for a total of three (3) appearances, versus not appearing 93% of the time for a total of 41 studies that could have used CI but did not, as seen in Figure 13. Of the three (3) studies that offered CIs, two (2) reported them in the prose. Of those two (2) reported in prose, only one (1) time did the author attempt to interpret their CIs within the context of their study.

Results for Research Question 4: To what extent does Hispanic sociolinguistic research lend itself to meta-analysis?
None of the studies included a meta-analysis. Based on the completeness of reporting, I assessed that 30% of the studies could be used successfully in a future meta-analysis, although a better mathematician than I may be able to provide meta-analytic calculations for some of the studies I excluded. The remaining 70% lack crucial information needed for meta-analysis. These are shown in Figure 14.

Figure 14: Meta-Analysis Possible

<table>
<thead>
<tr>
<th>Level</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>31</td>
</tr>
<tr>
<td>y</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
</tr>
</tbody>
</table>

Frequencies

<table>
<thead>
<tr>
<th>Level</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>31</td>
</tr>
<tr>
<td>y</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
</tr>
</tbody>
</table>
Chapter 6

DISCUSSION

Null Hypothesis Statistical Testing

My first question in regards to NHST was whether researchers discontinued learning from their data when the first $p$ values were encountered. Indeed, this was overwhelmingly the case. Prime examples of $p$ value endgames were Blas Arroyo (2008, 2010), Díaz-Campos and Killam (2012), Ferrer (2010), E. L. Brown and Rivas (2012). Brown and Rivas (2012) state clearly that, “A primary goal of this study is to demonstrate a statistically significant effect of a nontested notion of grammatical relation probability. To show this effect with a corpus-based approach and statistical model” (p. 323). Their interpretation of significance testing even became quite perplexing: “If a statistically significant effect of this new cumulative, probabilistic measure is found...then the results bring new evidence to the study of probabilistic measures” (p. 324). Of course, the study of probabilistic measures is outside of any single study, so this may have been a statistical communication issue.

Not all of the studies reviewed focused entirely on significance testing. Lipski (2010) was remarkably restrained in his reporting of $p$ values. In fact, his careful wording buffered him from significance testing culpabilities and the possible shortcoming of Varbrul by simply reporting that “VARBRUL did not choose this factor as significant” (p. 117). Like Lipski, File-Muriel and Brown (2011) indicated when the software application reported significance. This may be a wise imitation of Lipski since File-Muriel’s study made reference to Lipski almost immediately. Levi (2004) completely
avoided significant and statistically significant cant, preferring the word “likelihood” instead. Babel (2014) used the word “likelihood” just as judiciously. M. Newman (2010) was thoughtful as well, and Montaruli, Bourhis, and Azurmendi (2011) did not focus on \( p \) values at all. In fact, Montaruli mentioned the words “statistically significant” only four times and did not make significance testing the quantitative centerpiece of her study.

Among the 44 studies, six (6), or 14%, handled statistical significance with restraint. As Kline (2013) states: “It should be a hallmark of a maturing research area that significance testing is not the primary inference method” (p. 116). Writing for linguists nearly 30 years ago, Woods (1986) stated that even at the 1% level, the test statistic value “should not be the end of the story” (p. 129). He also underscored that, “The occurrence of a significant value of a test statistic should be considered as neither necessary nor sufficient as a ‘seal of approval’ for the statistical validity of the conclusion” (p. 130).

My second question regarding NHST was the use of alpha (\( \alpha \)) as prescribed in the Neyman-Pearson theoretical model of significance. Alpha was ignored in 37 studies, but mentioned or implied in seven (7) studies. Only in Walker, García, Cortés, and Campbell-Kibler (2014) was the concept of alpha explained or set clearly as “\( \alpha = .05 \)” or another value. The remaining six (6) studies that had signs of alpha in mind also included confusion. For example, Sorenson (2013) stated that "the number of data is large enough to perform tests of statistical significance, based on a \( p \) value of 0.05" (p. 71). Sorenson did not seem to be aware that \( \alpha \) was the prescribed choice here. Similar confusion was demonstrated by Rivadeneira and Clua (2011), and González Martínez and Blas Arroyo (2011).
One of the downfalls of significance testing is that, in the Neyman-Pearson model, it can lead to dichotomous, all or nothing thinking. Perhaps at the very extreme, Sanchez (2008) set up an unconventional Goldvarb table and declared that no $p$ values would be reported. In fact, Sanchez decided that if a test result was significant for a specific territory, it would receive the number 1, if the test result was not statistically significant she would give it a 0. Sanchez’s unique “Proposed factors and significance of factor groups” table (p. 244) was populated by 1s and 0s (in reference to the territory), which is not binary thinking in a metaphorical sense, but quite literally the definition of binary! Büdenbender (2013), for her part, deemed the .05 threshold to be what ruled her study’s results, explicitly. Similar approaches were had by Shelton (2013), and González Martínez and Blas Arroyo (2011).

Another problem that crops up in statistical testing is the battle between researchers and $p$ values. While it is considered bad form to report $p$ values as approximating any cutoff value (Hatcher, 2013), the researchers in the studies reviewed nonetheless revealed wishful $p$ values. For example, Carter (2013) indicated that a result was “marginal at $p = .06$” (p. 75). Miller (2013) noted when a correlation "trends towards significance" (p. 322). Díaz-Campos and Killam (2012), in reporting an $F$-test, indicated that .052 was “just shy of being significant” (p. 94). In reality, these researchers can take fleeting solace in the fact that “failure to reject the null hypothesis in most studies is more likely to say something about the design of the study (i.e., the study did not have sufficient power) than about the phenomenon being studied” (Murphy et al., 2014, p. 26).
The APA recommends that exact $p$ values be reported. This recommendation is not new, yet some authors in the studies reviewed do not follow it. Using notation that Engman (2013) called “erroneous thinking” (p. 264), Carter (2013) decided that one star (*) would be significant at $p = .05$, and two stars (**) would be significant at $p = .01$. Shelton (2013) used these same cutoff values and did not report exact values. These arbitrary cutoff points were established in the 1920s and 1930s because researchers had to look them up in tables, and tables with every single $p$ value would be cumbersome. With today’s high-power computing, exact $p$ values are automatically offered in statistical output. While reporting ANOVA results and Bonferroni post-hoc tests in SPSS, which offers complete $p$ value reporting, Büdenbender (2013) simply inserted $p < .01$ instead of using exact $p$ values. Sanchez (2008), who did not report $p$ values, but rather 1s and 0s, clearly doesn’t follow the APA’s recommendations on $p$ value reporting. Besides not following recommendations, the work of these researchers cannot be used in meta-analysis.

We know that $p$ values are never exactly zero (Hatcher, 2013). Regardless, Blas Arroyo (2010), Díaz-Campos and Killam (2012), E. L. Brown and Rivas (2012), Rieg Alamillo (2009), Carrera-Sabaté (2006), and Rivadeneira and Clua (2011) all reported $p = 0.000$ or $p = 0.0000$. Similarly, in an age when data from corpora are being used to generate $p$ values, $p$ values can become meaningless since they are a largely a function of sample size. Examples of large corpora reporting a preponderance of statistically significant results included Wolford (2006), Hernández (2009), and Alvord (2010).
Within most of the studies reviewed, the complexity of sociolinguistic inference was reduced to a dichotomous decision based on $p$ values and their associated statistical significance. Yet the limits of statistical significance testing and what $p$ values actually mean seem to be lost on a majority of researchers. Misunderstandings take the form of explicit and implicit expressions of NHST fallacies. I report some of the fallacies that stood out.

The idea that $p$ indicates the probability that results are the result of sampling error is known as the Odds Against Chance Fallacy. An example of this fallacy is explicit from a highly productive Spanish scholar who takes a boldly prescriptivist position on linguistic questions, Blas Arroyo (2010), who states: “For purposes of exposition, only $p$ values that indicate the probability that the differences might or might not be due to chance (significance level: $p < 0.05$) will be offered” (p. 631, own translation)\(^1\). Likewise, Ferrer (2010) indicates: “We use the $t$-test to determine the statistical significance of the differences between the means, that is, to know the probability that the difference found is due to chance” (p. 486).

Most authors reduced significance testing to a binary decision based on two values that were considered completely different when the difference was minuscule, typically $p = .049$ versus $p = .051$. This is the Sanctification Fallacy, and it was explicit or implicit to varying degrees in nearly every study reviewed. González Martínez and Blas Arroyo (2011) make their decision to sanctify $p$ values very clear: “we will simply

\(^1\) “Por razones expositivas se ofrecerán únicamente los valores $p$ que indican la probabilidad de que las diferencias puedan ser o no debidas al azar (nivel de significación: $p. < 0.05$)” (Blas Arroyo, 2010, p. 631)
indicate significant or nonsignificant when the test shows a discrepancy between two variables or groups of variables” (p. 667, own translation).2

Many resources state that $p$ stands for probability. The authors of the studies reviewed, with some exceptions, appear to believe that $p$ stands for probability of the hypothesis (either the null or alternative) given the data. This is the Inverse Probability Fallacy. $P$ stands for the conditional probability, that is the probability of the data, or outcomes even more extreme, given that the null hypothesis is true and all other assumptions for the test are true, or $p(Data + | H_0 \text{ and all other assumptions})$. This represents the likelihood of a result when the null hypothesis is exactly true. In addition, the sampling method must be random, scores must be independent and perfectly reliable, all distributional requirements must be met (e.g., normality, homoscedasticity), and that there must be no source of error besides measurement or sampling error, which we should recognize as nearly impossible to achieve (Kline, 2013).

Other explicit examples of fallacies include the Magnitude Fallacy which occurs when low $p$ values are understood to indicate large effects. Hawkins (2005), for example, referred to “Understanding the statistical impact and magnitude of these [demographic and economic] factors,” and offered interpretations of beta-coefficients, which is a type of effect size, but one could hope for more interpretation of effect sizes when they are generated from metric data, and more contextualization of ES cutoff suggestions for their study.

2 “simplemente señalaremos cuándo la prueba señala significativa o no significativa la discrepancia entre dos variables o conjuntos de variables” (González Martínez & Blas Arroyo, 2011, p. 667)
If a researcher believes that the probability that the alternative hypothesis is true is $> .95$, given $p < .05$, they might commit the Validity Fallacy. Mayberry (2011) provided a clear example of this fallacy – while perhaps conflating point estimates (point thinking) with confidence intervals (estimation thinking) – when she stated that statistical analyses were “obtained after submitting the data for all of the 472 verbs to a chi-square test at 95% level of confidence” while only providing $p$ values with a .05 cutoff.

Other fallacies appeared less explicitly, such as the Reification Fallacy which occurs when a result is not considered replicated if $H_0$ is “rejected” in one study but “accepted” in another. File-Muriel and Brown (2011) provide a possible example of Reification: “Word length, which has been reported as significant in past studies…was not found to be a significant predictor of any of the three dependent variables in the present study” (p. 237); however, simply noting the difference may not mean that the authors absolutely reify the differences based solely on statistical tests.

Another example of a less explicit fallacy is the Robustness Fallacy which occurs when NHST is considered to be robust against outliers or violations of distributional assumptions. Although Díaz-Campos and Killam (2012) provided a clear warning against this fallacy in language that was reminiscent of introductory statistics books, they did not report testing for these assumptions: “The assumptions of any general linear model include that the residuals from this type of analysis are independent observations, identically normally distributed, with common variance” (p. 92). Díaz-Campos and Killam (2012) were not alone. Nobody appeared to confirm $p(\text{Data} \mid H_0$ and all other assumptions) while performing their statistical testing. Ensuring assumptions are tenable
seems far less niggling when we consider that $p$ values are point estimates that are not
drawn on empirical sample distributions, but on sampling, or perhaps resampling,
distributions. Thus they reflect outcomes never observed and require many assumptions
about the unobserved data. Remaining unaware of distributional assumptions precludes
real and necessary discussions around non-normal distributions, which are much needed
in linguistics (Gries, 2015).

One of the test statistics we use and report often in sociolinguistics is chi-square.
SLA often uses and reports $t$-tests. The chi-square and $t$-test test statistics provide correct
$p$ values if and only if the null is true. This means that we know and can anticipate the
chi-square and $t$-test test statistic distribution when our results are not significant. If they
are significant, we are no longer dealing with an expected distribution: We may now
have a noncentral chi-square distribution or noncentral $t$! This new distribution based on
(often cheerfully reported) statistical significance will include varying degrees of skew
and kurtosis that change predictability. Therefore, what many linguists have been doing is
rejecting the null and declaring statistical significance, then unknowingly proceeding
with statistical inferences as if they were based on statistically insignificant null
distributions for chi-square and $t$. We are supposed to examine and base our results on
new distributions, possibly skewed (left or right) and kurtodic (platykurtic or leptokurtic).
Some researchers might not be aware that they are supposed to be testing for noncentral
distributions when they “successfully” reject the null hypothesis and make the (cheerful)
claim of statistical significance. The importance of knowing these distributional
assumptions is driven home by Cumming (2012) when he addresses noncentral \( t \) distributions:

It’s important because, in the world, there often is a real effect and so alternative hypotheses often are true. In addition, statistical power is a probability calculated assuming that an alternative point hypothesis is true, so we generally need noncentral \( t \) to calculate power. Also, noncentral \( t \) is the sampling distribution of Cohen’s \( d \) and so it’s needed to calculate CIs for \( d \). (p. 264)

Cumming (2012) brings up statistical power in relation to noncentral \( t \) distributions. Power, one of the most salient features of NHST, was almost completely absent from the studies reviewed, as often seen in SLA (Plonsky, 2013, 2014). Just as noncentral \( t \) is needed to calculate power, noncentral chi-square distributions often reveal themselves in the power analyses of statistical tests in which the null distributions are (often asymptotically) a chi-squared distribution, as appears in important tests like the likelihood ratio tests known to sociolinguists. In the Hispanic sociolinguistic studies reviewed, these noncentral distributions were not discussed, and statistical power was never reported. Power, however, was mentioned once by Walker (2014) when he noticed no statistically significant effects and suggested that they were “simply due to lack of statistical power” (p. 185). Yet Walker does not report on performing power analyses.

For those who put their full faith into NHST, statistical power is relevant because it tells us “the probability that a test will correctly identify a genuine effect” (Ellis, 2010, p. 52). Technically, it is the probability of correctly rejecting a false null hypothesis; that is, the power tells us the probability that when we reject the null hypothesis and conclude
we have significant results, there actually is an “effect” in the population (Ellis, 2010; Good & Hardin, 2012; Hatcher, 2013). Power is not the following, taken from a linguistics reference: “statistical significance [emphasis in original] denotes whether a particular result is powerful enough to indicate a more generalizable phenomenon” (Mackey & Gass, 2012, p. 85). Calculations of statistical power and statistical significance are separate and unique in both their calculations and meaning, although they dovetail to provide a more complete story during the data discovery phase, and they help inform us about distributions. Conflating statistical significance and statistical power is not a good idea.

In NHST, a priori power calculations are performed at the planning stage when sample sizes are considered. When we contemplate how many participants are required in our sociolinguistic studies to support our hypothesis, essentially we are asking about statistical power because we want to detect a significant effect or significant relationship if they exist in the real world. In the Neyman-Pearson paradigm of hypothesis testing – the one most adhered to in the studies reviewed – power is crucial because it is inversely related to the probability of making a Type II error, that is, the failure to reject a false null hypothesis. When this occurs, we accept the null when we should not. If we make the error of claiming nonsignificance, and we ignore nonsignificant results, as often occurs in the studies reviewed, then we are ignoring part of the story that our data is trying to tell. Unfortunately, Type II error rates are commonly ignored in quantitative studies within the social and behavioral sciences (Schmidt & Hunter, 2014). When error types are reported, they are almost invariably Type I. Yet if an effect exists in our study – what we
often hope for – the probability of making a Type I error is zero (Ellis, 2010). If indeed an effect exists, and Type I error is zero, then Type II error is the only type we can possibly find. Given that Type II errors occur only when our study lacks statistical power (Ellis, 2010), and we do not test for it, Type II error can go as high as 95%, and we will never know it. Alarmingly, “[e]rror rates of 50% or higher have been shown to be the usual case in many research literatures” (Schmidt & Hunter, 2014). If we have not tested for it, and Type II error is truly 50% or higher in linguistics as it is in other fields, then we will proceed with the conclusion that the statistically insignificant variable or variant is practically unimportant when the very opposite might be true. Over the years, we will build a research record that, quite possibly, ignores substantively important phenomena and contains an unappealingly large number of unresolved conflicting and “more research is necessary” results. When untested power and volatile $p$ values are mixed together, what takes shape is the perfect storm in which we really do not know which differences and magnitudes are real and important to report. This situation is not ideal for reliable theory building, the ultimate goal of every high-integrity empiricist and theorist, and the academic discipline they belong to. It is this type of disquieting reality that has led quantitative methodologists to claim that the traditional approach to data analysis using NHST “makes it virtually impossible to reach correct conclusions in most research areas” (Schmidt & Hunter, 2014, p. 9).

In the studies reviewed herein, with notable exceptions, researchers in Hispanic sociolinguistics seem to set up and test null hypotheses in “an extremely mechanical way reminiscent of compulsive hand washing,” as Gigerenzer (2000, p. 283) put it rather
derisively. Yet power analyses, clearly among the desiderata of Neyman-Pearson hypothesis testing, was overlooked or not reported. Thus, as explained, we cannot be certain of the results that were reported in the studies reviewed, especially for those studies that only gave credence to results that were statistically significant!

Although power is crucial for those who see $p$ values as the endgame, it is important to recognize that it is only relevant if we put all of our eggs in the NHST basket; that is, statistical power is only relevant within the context of significance testing (Cumming, 2012). That being the case, statistical power is attached to a problematic statistical testing routine in which the magnitude of a population effect, what we actually want to know, is not addressed. One alternative is to shift our focus from null hypothesis testing and power analyses to precise parameter estimation (PPE) (Bonett & Wright, 2011; Lai & Kelley, 2011; Lance & Vandenberg, 2014), or remove $p$ values from center stage and adopt the more informative and transparent group of statistical methods of the new statistics.

It would be unfair to impute total responsibility to applied researchers of Hispanic sociolinguistics for misinterpreting the outcomes of their statistical tests. Statistical tests rely on many assumptions that are far-fetched in most studies, not just sociolinguistic studies. Classical parametric tests depend on distributional assumptions, such as normality (discussed) and homoscedasticity (from Greek meaning “having the same scatter”) which are probably untenable in many analyses. Robust tests ease some distributional assumptions, but their $p$ values may still be generally incorrect in actual data sets, especially when samples are not random (Kline, 2013). Achieving a random
sample that is stochastic in sociolinguistics can be a methodological challenge (e.g.,
snowball sampling), yet the expected distributions and $p$ value results depend on random
sampling. “What would the $p$ value mean when computed for a nonrandom sample? The

When a researcher wades into the NHST quagmire, and decides to stay there
throughout their study, they are likely to get stuck somewhere between Fisher’s
significant test and Neyman-Pearson’s hypothesis test when they peer at their $p$ values
and attempt to make sense of them. The least subtle friction comes from trying to
harmonize NHST as a gradient phenomenon and a binary phenomenon at the same time
(Engman, 2013). This is not logically possible, and even if the struggle to resolve logical
opposites never rises to the level of consciousness, the researcher will instinctively look
for a way out. Too often the way out results in one NHST fallacy or another.
Compounding the confusion, a sociolinguist can pick up any number of references and
find an incorrect definition for $p$, or any number of Intro Stat conceptual mashups (Dixon
& O'Reilly, 1999) published in authoritative language, even among authors who write on
quantitative methodologies in linguistics.

A continued overreliance on significance tests, and ongoing misunderstandings
beg the question: What do our resources in linguistics and sociolinguistics tell us in
regards to statistical testing? As it turns out, quantitative linguistics is like other social
and behavioral science disciplines in that it is populated by two broad categories of
researchers when it comes to significance testing. The first category of researchers point
out NHST flaws and attempt to provide solutions; the second category of researchers, far
larger than the first, simply ignore the first, or continue in the broader NHST research tradition because it has advanced many careers, and bucking tradition can be impolitic, if not unwise. As also quoted by Perry (2011, p. 96):

To the [researcher] who practices this religion, Statistics refers to the seeking out and interpretation of $p$ values. Like any good religion, it involves vague mysteries capable of contradictory and irrational interpretation…And it provides Salvation: Proper invocation of the religious dogmas of Statistics will result in publication in prestigious journals. This form of Salvation yields fruit in this world (increases in salary, prestige, invitations to speak at meetings) and beyond this life (continual reference in the citation indexes). (Salsburg, 1985, p. 220)

The NHST polemic – if the voiced/ignored opposition can be called that – is not new in linguistics. Nearly 30 years ago, in *Statistics for Language Studies*, Woods (1986) wrote, “The value of statistical hypothesis testing as a scientific tool has been greatly exaggerated” (p. 127). Yet even Woods (1986) is not immune to NHST theory conflation when he states, “A hypothesis test simply gives an indication of the strength of evidence (in a single experiment) for or against a working hypothesis” (p. 127). The strength of evidence is Fisherian when it refers to a single experiment. The “working hypothesis” is Neyman-Pearson’s alternative hypothesis, which is part of their binary behavioristic approach to significance that applies only to ongoing, identical repetitions of an experiment, not to any single experiment. Kline (2013) refers to this belief that there is evidence for the working hypothesis when the null is rejected as the Meaningfulness Fallacy.
In *Research in Applied Linguistics: Becoming a Discerning Consumer*, Perry (2011) also invokes the Meaningfulness Fallacy. Writing on a hypothetical researcher, he states, “In this case the null hypothesis was rejected, giving support to her researcher hypothesis” (p. 181). He invokes the Causality Fallacy as well:

At most, a correlational analysis can find if there is a potential causal relationship before going to the more arduous task of doing a full blown experiment. If there is no statistically significant correlation, a causal relationship can be ruled out right from the start. (p. 183)

While my objective was to review trends, this notion that a causal relationship can be ruled out from the start due to failure to fit a line in a statistically significant manner is particularly questionable if assumptions are not checked, and nothing is said of sample size. Significance is largely a function of sample size, and nonsignificance should always be considered inconclusive: “as Fisher himself pointed out, nonsignificant results are inconclusive” (Schneider, 2015, p. 423). For Fisher, we only choose between rejecting a null hypothesis and suspending judgement (Glaser, 1999; Martin & Bridgmon, 2012). It is somewhat disconcerting to find contrary advice appearing in a book written for linguists and designed to disabuse statistical misconceptions. A causal relationship can exist without successfully fitting a straight line via ordinary least squares methods. Real data, in sufficient quantity, and especially data taken from human beings, and even more so when it is sociolinguistic and diachronic, often appear as curves (Osborne, 2015). Unless assumptions are checked, a causal relationship cannot be ruled out “right from the start.” Following this linear OLS line of thinking, in their guide on research methods in
second language acquisition, Mackey and Gass (2012) tell their readers that, “statistically nonsignificant results must be ignored [emphasis in original] in research studies” (p. 85). Here the authors indicate that researchers should not compare across statistical critical values (typically above and below .05) and only focus on results that are statistically significant.

The majority of researchers of Hispanic sociolinguistics I reviewed continue to pay special attention to variables that resulted in values labeled as statistically significant while downplaying, or completely eschewing those variables that do not cross critical values into statistical significance. This well-entrenched habit is not good. In 1986, from a book for linguists, we read: “It is a misguided strategy to abandon an otherwise attractive line of research because a statistically significant result is not obtained” (Woods, 1986, p. 127). Writing for researchers of SLA, Larson-Hall (2010) clarifies that, in reality, a “study with a resulting $p$ value of $p = .08$ might have much more practical significance than one whose $p$ value was $p = .003$. Plonsky (2015) comes forward to affirm Woods (1986) and Larson-Hall (2010): “This practice, that of suppressing results with $p > 0.05$, is as common in our field as it is unfortunate” (Plonsky, 2015, p. 241).

For those who opine that we simply cannot report and interpret every $p$ value, the point here is that we should not be paying differing $p$ values radically different amounts of attention. If we feel compelled to report them, we should report all $p$ values to three digits as suggested by the APA, making them parenthetical, or putting them in footnotes. Perhaps even better, researchers could include $p$ values in their Methods sections with information on sample size estimation, power analyses, data cleaning and imputation, that
is, along with other preparatory and data integrity steps. If this seems like a lot, it is. Of course, we have a more elegant solution to \( p \) value endgames: the new statistics. Our quantitative interpretive efforts, to a greater degree, should focus on confidence intervals, and our cheers of success or jeers of “failure” should surround effect sizes and what they mean within the context of our studies.

Compared to applied linguistics, variationist sociolinguists have fewer books aimed at statistical analyses. Paolillo (2002) is one of our few references. In his book on statistical models and methods for analyzing linguistic variation – referred to by (Tagliamonte, 2007) as “perhaps the most detailed with a focus on statistical terms and explanations” (p. 202) – Paolillo (2002) states:

The hypothesis testing procedure does not allow us to test the research hypothesis directly; rather, we propose an alternative hypothesis known as the null hypothesis, which is represented by a model in which the proposed relationship is lacking. We then try to show that the null hypothesis is inadequate to explain the distribution of the data. If we successfully refute the null hypothesis, we can then take comfort in our research hypothesis, represented by a model including the proposed relationship, as a more plausible explanation of the data. (p. 113)

There are two issues with this statement: (1) it is confusing, and (2) it is incorrect. First, Paolillo refers to the alternative hypothesis as the null hypothesis. Indeed the null hypothesis is an alternative to the research hypothesis, quite literally, but traditionally the research hypothesis is the alternative hypothesis. This could be confusing. When we write
for sociolinguists we need change our terminology to what is actually used and understood within our research domain from what can be found in some statistics books, even recently as is this case for distribution naming mentioned here (Angrist & Pischke, 2015). Second, we cannot take comfort in the research hypothesis when we “successfully refute the null hypothesis.” This is known the Valid Research Hypothesis Fallacy (Carver, 1978) or the Meaningfulness Fallacy by Kline (2013) which refers to the false belief that there is a greater probability that the research hypothesis is true when the null is rejected. In addition, one should not conceive of statistical tests as measures of success or failure of a study. It is simply a volatile finding (Gigerenzer & Marewski, 2015), part of the discovery process that can inform the researcher how to proceed if they decide to ignore the volatility of \( p \) values, and the host of other NHST problems. Decades of battling these \( p \) values has given \( p \) Brobdingnagian proportions in the minds of researchers, and it should not be encouraged.

Also from Paulillo (2002):

Typically, the \textbf{significance level} [emphasis in original]...for social science research is 0.05, by which it is meant that we accept a 5% risk of incorrectly rejecting the null hypothesis. This is usually symbolized by \( p \leq 0.05 \), where \( p \) stands for probability. (p. 114)

Two issues appear in this segment: (1) the theoretical approaches to statistical testing are confused, and (2) it is incorrect. First, the theories of Fisher and Neyman-Pearson are conflated. It appears that Paolillo is attempting to describe the Neyman-Pearson Type I error, while invoking Fisher’s \( p \) value. The correct symbol is alpha (\( \alpha \))
(Neyman-Pearson), and it is a point estimate, meaning it does not vary along a gradient as it does for Fisher (although we recognize that a coefficient could land on a continuum split at alpha). For Neyman-Pearson, the test statistic crosses $\alpha$ or it does not; therefore, $\leq$ as a possible range of values is wrong by definition. Second, the set significance level, $\alpha$ not $p$, is not the percentage of risk of incorrectly rejecting the null hypothesis. Although difficult to glean given the theoretical mashup, this is redolent of the Local Type I Error Fallacy (Kline, 2013). This fallacy is the mistaken belief that $p < .05$, given $\alpha = .05$, means that the likelihood that the decision just taken to reject $H_0$ is a Type I error is less than 5%.

Given the difficulties of reconciling disparate theoretical approaches to significance, there may be little wonder why Tagliamonte (2007) is so cautious when she states, almost forced into tautology, “Significance of one or a set of specific factor groups may lead to one interpretation. Significance of another factor group (or set) may lead to another” (p. 207).

Other sociolinguists have also attempted to explain $p$ values. Walker (2010), in his book written for sociolinguists, states that the “lower the $p$ value, the less likely it is that the observed distribution is due to chance… The cutoff point depends on how sure you want to be that the results are not due to chance.” Like Walker (2010), Paltridge and Phakiti (2010), writing for applied linguists, define probability as, “[t]he degree to which a statistical finding is likely to occur by chance” (p. 355). Adding to this consensus, Mackey and Gass (2012) state: “If a result is non-significant, this means that we cannot be certain whether it occurred in the particular sample only because of chance” (p. 85).
First, it is good to keep in mind that \( p \) values are not drawn on the observed, empirical distribution, but rather on the hypothetical, unobserved sampling distribution when the null is true. Second, it is illogical to view \( p \) as measuring the probability of chance because a \( p \) value is calculated against a true null that is already set to a 100% chance of being true. In the world of probability that lies between 0 and 1, the null being true is already set to 1 (Perezgonzalez, 2015; Schneider, 2015). Added to sampling error, \( p \) is measuring the plausibility of data coming from an underlying mechanism that generates data randomly, by 100% chance, not chance itself. Kline (2013) considers this fallacy, which he calls the Odds Against Chance Fallacy, to be the most pervasive, and it appears in our references and in our research as if it were axiomatically true. To his great credit, Kiesling (2011), in *Linguistic Variation and Change*, a variationist, approximates a proper definition for statistical significance: “[it] usually means something like ‘This pattern is very unlikely to have come about by chance – there must be some reason why the linguistic variable patterns this way’” (p. 45). If Kiesling (2011) had added “by complete chance,” his definition of significance would come very close to what I call the Random Process Coefficient, a cognitively appropriate name for Fisherian \( p \) values that justly deemphasizes them as simple findings (Gigerenzer & Marewski, 2015).

If even considered, NHST should be used with full knowledge of its liabilities, and it should not be relied upon to make claims, and it should not be used to judge the usefulness of a researcher’s efforts, nor eligibility for publication for that matter. Complete elimination of NHST is not necessarily if it is properly deemphasized. NHST answers those rare research questions that in fact require a dichotomous answer, and it is
hidden many operations “under the hood.” For example, NHST operations are nested in maximum likelihood estimation (MLE) used in logistic regression, and it can be found in common data discovery tests such as Levene’s Test for homogeneity of variance, or in post-hoc tests such Bonferroni’s adjustments for pairwise comparisons (found more commonly in SLA research). Kline (2013) considers some of these “canary in the coalmine” (p. 107) tests to be “illegitimate” and “indefensible” (pp. 107-108), especially in the case of logistic regression, which is commonly considered a nonparametric technique regardless of its “canaries.”

Confidence Intervals

In the studies reviewed, the use and interpretation of confidence intervals among Hispanic sociolinguists did not inspire confidence. Confidence intervals appeared very rarely. They appeared in figures without interpretation, with one exception. Regardless of their requirements, confidence intervals are more informative than significance tests. Instead of centering on a hypothetical zero value of the nil (null) hypothesis, they are correctly centered on the observed value. They also give us a correct picture of the extent of uncertainty, especially in small sample studies.

Outside of Hispanic sociolinguistics, recognition of the superiority of confidence intervals and point estimates of effect sizes over NHST has grown exponentially (Schmidt & Hunter, 2014). Years ago, the report of the Task Force on Statistical Inference of the American Psychological Association (APA) (Wilkinson & Inference, 1999) stated that researchers should report effect size estimates and confidence intervals. More recently, the sixth edition of the APA’s Publication Manual stated that it is almost
always necessary for primary studies to report effect size estimates and confidence
intervals (American Psychological Association, 2010). Despite these developments, most
articles reviewed here still focus squarely on significance tests as some sort of NHST
theoretical blend, which persists regardless of having been discredited rather thoroughly.
Perhaps, as Orlitzky (2011) argued, the problem is that the evidence against NHST has
not been institutionalized. It seems like most articles that discredit significance testing
have directed their efforts towards convincing individual researchers to change their
statistical practices as opposed to attempting to promote broader, more systematic,
institutional change. Individual researchers within sociolinguistics may find it
challenging to go against what has become an institutionalized practice in the field,
perhaps “due to fear of deviating from normative practices” (Sharpe, 2013), and because
they must conform to what is expected by journal editors who act as gatekeepers to
publishing. To their frustration, sociolinguistic researchers may discover that they know
more about best practices in statistics than journal editors who review their work.

While coordinated efforts to institutionalize change are important, we also need to
promote ongoing statistical education at all levels to handle the complexity of new
statistical methods. As Sharpe (2013) remarks: “resistance says less about substantive
researchers, their awareness, education, pressures or mindsets, and more about the
properties of a statistical innovation such as its complexity” (p. 575). This applied equally
to journal editors, researchers, professors, and students.
Effect Sizes

We observe *effects* in our sample, or in our classrooms, or in our language labs. We observe *effect sizes* in the populations we study, in the real world (Ellis, 2010). Since the studies reviewed here were interested in inferring from samples to populations to assess what might be occurring in the real world in terms of magnitudes of effects, reporting effect sizes for phenomena of interests were certainly possible throughout. A total of 64% of the authors reported effect sizes given my criteria, thus 36% did not. I used very generous criteria for effect size inclusion during coding. If the author gave an effect size of any sort, or I could figure out how to arrive at a standardized effect size from the data presented, then that author’s study contributed to the total of effect sizes reported. Even if I did not find effect size reporting statistically rigorous, I added it. For example, in Varbrul, factor groups (variables) are ranked using the range between factors (variants), but these ranges can be influenced by outliers, and different factor groups (variables) can have different numbers of factors (variants). Still, I recognized these range calculations as a type of effect size. More critical reviewers might be justified in lowering my effect size finding of 64%. Clearly, new statistics adherents would be unhappy with my finding, and even less happy with those of a more critical reviewer.

Among those who reported effect sizes, T-shirt definitions of small, medium, and large were very closely followed. In general, effect sizes were estimated for results that were statistically significant, ignoring effect sizes for statistically nonsignificant results. Values for effect sizes were also reported only as point estimates, which ignores the fact that effect sizes are subject to sampling error as well. Effect sizes were most often listed
without justifying how the statistically generated effect size coefficients were justified in
relation to fixed benchmarks, nor how they were sensible within the context of their
studies and related academic record. This approach is reminiscent of the dichotomous
thinking that prevails in NHST: “If people interpreted effect sizes [using fixed
benchmarks] with the same rigidity that \( \alpha = .05 \) has been used in statistical testing, we
would merely be being stupid in another metric” (Thompson, 2001, pp. 82-83). Indeed,
we should interpret effect sizes in the context of our own research domain (Durlak,
2009), because the magnitude of our effect sizes, the reliability of our measures, the
distributional assumptions of our data, and the effect size types selected are all impacted
by our research domain (Grissom & Kim, 2012), that is, sociolinguistics in general and
Hispanic sociolinguistics in particular. Only then can we accurately discuss how
standardized effect sizes include other quantities or characteristics, such as
unstandardized effect sizes, error variance, and sample base rates. “These are all crucial
aspects in study planning and must not be overlooked” (Kline, 2013, p. 159).

**Meta-Analysis**

Narrative, qualitative syntheses of previous studies were very well done among
the 44 studies reviewed; however, there was no evidence of meta-analysis. Yet our
primary goal as sociolinguistic researchers is addressed by meta-analysis as a tool! If our
goal in sociolinguistics is to produce cumulative knowledge, this means that we should
develop theories that explain sociolinguistic phenomena that are under examination.
Ultimately, we need to be able to accurately and precisely calibrate the relationships that
exist between sociolinguistic variables for phenomena of interest. Without calibration,
can we say that we have the reliable raw quantitative materials out of which to construct theories?

In the studies reviewed, Hispanic sociolinguists continue to rely on NHST with small sample sizes or very large ones (in corpora). We are also using different statistics software applications. When empirical sociolinguistic research relies on small sample sizes and NHST, our study results may appear more conflicting than when larger sample sizes are consistently used, or when researchers process their data using different statistical programs. We should not be surprised when different results are found for the same research phenomena when one researcher uses Varbrul and another uses SPSS, even when deploying the same statistical method, especially when sample sizes are different. As we have seen, even with the same sample sizes, variability can impact $p$ values. Meta-analysis can bring clarity to the sociolinguistic research problems that we previously considered only through the lens of significance testing, as Plonsky (2014) has demonstrated for applied linguistics. Meta-analysis can integrate findings across sociolinguistic studies to reveal the simpler patterns of relationships that underlie our research literature, thus providing a basis for sociolinguistic theoretical development. Meta-analysis can correct for the effects of sampling error, measurement error, and other artifacts that produce the illusion of dead-end, conflicting, or “more research is necessary” findings (Schmidt & Hunter, 2014) as do appear in Hispanic sociolinguistic studies.

Sociolinguistics is not alone. Most research literatures show conflicting findings of this sort. Considering the volatility of $p$ values, it is far too common for some studies
to find statistically significant relationships when other do not, regardless of study design and implementation similarities. Schmidt and Hunter (2014) report that in much of the research literature in most areas of the behavioral and social sciences, this split is approximately 50–50. If sociolinguistics follows this trend, we may have a few issues to address in terms of developing accurate theories and reliable cumulative knowledge.

Meta-analysis is being applied widely outside of sociolinguistics to solve problems related to artifact disturbances in research synthesis. D. Brown (2014), Lo and Lo (2014), and Plonsky (2014) are demonstrating how this can be done in applied linguistics. Outside of linguistics, thousands of meta-analytic articles have been published in areas such as psychology, education, medicine and other disciplines, and excellent references are available, as well as software for handling these sometimes sophisticated meta-analytic processes. Meta-analyses in sociolinguistics, however, are few and far in between, although some appear at the crossroads of psycholinguistics and sociolinguistics, such as the meta-analysis of the effects of speakers’ accents on interpersonal evaluations by Fuertes, Gottdiener, Martin, Gilbert, and Giles (2012). My own search results did not turn up meta-analyses in Hispanic sociolinguistics.

Sample Sizes

Considering that \( p \) values are largely dependent on sample size, and many empirical studies reviewed had small sample sizes or large sample sizes, compounded with unquestioned reverence for \( p \) values, the door is wide open to questions regarding how many subject or cases are needed for any particular sociolinguistic study to provide meaningful results. In many cases, the sample sizes in the studies reviewed seemed to be
too small to detect effects, or so large that tiny $p$ values peppered the research results. Defining an appropriate sample size is crucial in statistics, but numerous rules of thumb have been calculated and then used as simple heuristics by researchers when determining needed sample sizes. Many times these rules of thumb specify minimum sample sizes for particular statistical analyses. Other times they specify a ratio of the number of participants needed relative to other aspects of the analyses, for example, the ratio of the number of participants to the number of variables, as is often done in sociolinguistics. However, these rules of thumb are problematic. For example, the diversity of opinions on rules of thumb often allow researchers to justify any sample size. If a student of linguistics asks their professor, “How many participants do you think I will need in order to get reliable results from multiple regression?” Since we can turn to any number of citations from authoritative figures, we turn to heuristics from the literature citing a rule of thumb that we have come across in the literature when choosing a sample size.

Even in multiple regression, a relatively well-developed topic, there is no agreement as to what constitutes a “small sample size” (Ryan, 2013). Yet a lot of sociolinguistic research is correlational, which may have even less agreement. That being the case, what can be considered a “small sample size”? This definition is especially important to understanding the extent of uncertainty that needs to be handled statistically (keeping in mind that statistics is often defined as the scientist’s way for handling variability [sample] or error [population]). As it turns out, confidence intervals – an important part of best statistical practices – can be used to define “small sample size” for both correlational and experimental studies.
According to Schmidt and Hunter (2014), for correlation, if we want the confidence interval for the correlation coefficient to have a width of ±.05, the minimum sample size is approximately 1,538 for small population correlations. If we want a sample size of 1,000 to suffice, the population correlation has to be at least .44. If we accept this standard of accuracy for correlational studies, as are most sociolinguistic studies, “small sample size” includes all studies with less than a thousand persons. It is worth noting that logistic regression requires quite a large sample size because maximum likelihood (ML) estimates are less powerful than ordinary least squares (OLS) found in simple and multiple linear regression (Best & Wolf, 2014). In an analysis using various sample sizes (N = 50, 100, 250, and 500), “[e]ven in relatively large samples, serious misestimation of odds ratios was common” in logistic regression (Osborne, 2015, p. 336).

Using the same CI calculation routine as above, for studies that use Cohen’s \(d\), as do many SLA studies, small effect sizes will require a sample size of 3,076. If we want a larger effect size then our sample size must be greater than 3,076. As Schmidt and Hunter (2014) explain, if the difference between the population means is found to be .30 standard deviations, then the minimum sample size that is accurate to within ±.05 of .30 will be 6,216. Given this standard of accuracy, for experimental studies then, what we call “small sample size” starts at 3,000 and goes up from there.

According to best practices in sample size calculations, then, in order to meet the level of statistical certainty desired for 95% confidence intervals:

- Sociolinguistic correlational studies should include about 1,538 participants.
Experimental/pseudo-experimental SLA studies should include about 3,076 participants.

Compounding sample size problems, top linguistics journals may prefer research that makes theoretical contributions. The upshot is that some theories may never really get tested sufficiently to make claims because they are not replicated enough. If theories are not reaffirmed through replication, or if they rest on “conflicting results” or “more research is necessary,” then the ultimate result will not be optimal theoretical development.

Given the stark reality of statistically insufficient sample sizes and, at times, insufficient replication, sociolinguists should set out to design their studies in such a way that they can be replicated and eventually contributed to meta-analysis. Meta-analysis can aggregate studies and meet minimum sample size requirements. Thus, no single sociolinguistic study should be conceived of as being statistically sufficient for making definitive conclusions regarding observed phenomena. We should start our studies with meta-analytic thinking, and keep replication in mind. While attempts at replication of SLA research can be found, empirical sociolinguistic research often resists replication by design (e.g., a snapshot in time of a specific group), and it falls quite short of optimal sample size. As such, the statistics deployed cannot support a high level of certainty in claims of causation, especially in an environment that is rarely more than correlational. Within this environment, especially when ignoring statistical assumptions, the definitions of “weak” and “strong” correlations (never adjusted for the sociolinguistic topic!) may be more adventitious than truthful.
Statistical Monoculture?

In aggregate, the three journals reviewed do not indicate that Hispanic sociolinguistics is a statistical monoculture. However, there are heavy biases for Varbrul (logistic regression with fixed effects) and Analysis of Variance (ANOVA). SLA seems to have a greater affinity for ANOVAs (and t-tests) than sociolinguistics. In either case, the heavy roles Varbrul and ANOVA are playing in linguistics are not optimal.

ANOVA. Second only to Varbrul, ANOVAs seem to be popular with Hispanic sociolinguists in the studies reviewed. This is not a particularly sanguine development. There seems to be a desire for statistical solutions to provide all or nothing conclusions based on averaging speakers or cases. It starts with the dichotomous thinking that attaches itself to NHST. It then takes shape in t-tests and ANOVAs in which variables and their story of variability are averaged to a mathematical mean (usually) and then two means are compared (t-tests), or three or more are compared (ANOVs). The researcher then wants to know if there is a difference between the groups, yes or no. Hispanic sociolinguists who are fortunate enough to work with metric variables would do well to be aware of the drawbacks of such binary thinking. Plonsky (2015) weighs in:

We see this all the time in applied linguistics: Researchers force continuously-scaled independent variables (e.g., motivation) into categorical ones (low, high), without any theoretical justification for doing so, just to make them more amendable to an ANOVA-type analysis. We insist on looking for differences rather than asking about the extent of the relationship between variables. I view...
this practice as related to the dichotomous view of data/analyses…embodied by

NHST. (p. 239)

The comments of Plonsky (2015) focus our attention on the dichotomous mindset
that starts with NHST: “The moral of the story is that if we give up significance tests, we
give up categorical claims asserting that something surprising and important has
occurred” (Abelson, 1997). This moral applies equally to any number of statistical
analyses pursued by analysts who are ensnared by the same mindset.

**Varbrul Analyses.** Variable Rule Analyses, or Varbrul (Cedergren & Sankoff,
1974), and its later implementation as Goldvarb (Rand & Sankoff 1990), offered
sociolinguists a methodology and statistical tool to measure binary categorical outcomes
using logistic regression before many other disciplines. Within statistics, binary logistic
regression is not considered overly complex, but it was a somewhat new statistical test in
1974, and certainly more complex than other tests like Pearson’s chi-square, Pearson’s
product moment correlation $r$, $t$-tests and others. With its implementation, the designers
of Goldvarb handed down the Varbrul methodological routine that has been followed
(mostly) carefully for the last 40 years, essentially unchanged. Theses, dissertations, book
chapters, academic articles, and conference presentations still rely on Goldvarb and its
methodology to process and interpret variationist data. In recent years, however, a
handful of sociolinguistic scholars have questioned its continued use, some quite
pointedly.

Oblique and frontal attacks on Varbrul come from sociolinguists with some of the
strongest backgrounds in statistics. Varbrul’s critics suggest that we replace Varbrul
(Goldvarb) and its inflexible fixed effects logistic regression with more mainstream applications like R, SPSS, or SAS because they allow for multilevel, mixed (variable) effects modeling, among many other options (Gorman & Johnson, 2013; Gries, 2013a; Johnson, 2009, 2010; Roy, 2011, submitted). These are excellent suggestions, but they will require some sophisticated training in statistics.

On the good side, Varbrul as a methodology provides much needed guidance in the use and communication of statistics for research. However, Varbrul, or Goldvarb, has one central problem: It is not extensible. Limited functionality, however, does not mean that Goldvarb is an unworthy instructional tool for students of sociolinguistics. It may one day be a legacy program, like others, but 40 years of research has been conducted using Varbrul, and it is still used today. Researchers of Hispanic sociolinguistics who continue to use it, however, might reconsider using factor group rankings as indicators of effect size since ranges may include extremes, or outliers (the *raison d'être* of “robust” statistics). Also, it is clearly not advised to discount phenomena actually observed in the real world by the researcher only because it does not appear to be statistically significant. In addition, care must be taken not to use the language of causation (e.g., “constraints”) in studies that are purely correlational from a statistical perspective (Macaulay, 2009). It will also be a good idea to keep abreast of literature that points to serious problems with small sample sizes in logistic regression. Osborne (2015), referred to earlier, is not alone in calling attention to problems with small sample sizes in logistic regression; Osborne (2015) affirms Best and Wolf (2014) who performed a sample size simulation study leading to the same conclusions: Small sample sizes yield highly unreliable results in
logistic regression. Note that these authors discuss sample size problems in logistic regression as a statistical method itself, regardless of what statistical application we use, Goldvarb, Rbrul, as espoused by Johnson (2009), or any other software application.

**Mixed Effects Modeling.** Although it is beyond the scope of the present thesis, fixed effects versus mixed effects modeling will need to be addressed. Mixed effects, or multilevel modeling, is of interest to both sociolinguists and SLA researchers (Cunnings, 2012). Complaints about fixed effects modeling vis-à-vis mixed effects modeling have arisen in response to 40 years of Varbrul’s fixed effects implementation of logistic regression in variationism. The objective is to process mixed effects due to limitations (incorrect results) of fixed effects. In Stroup (2012) the criterion for determining a fixed versus random effect, or “factor” using sociolinguistic terminology, is whether it can be treated as a statistically random, or stochastic, sample taken from a population. If the researcher adheres to the technical definition of “random,” speakers cannot be considered random in most sociolinguistic analyses because they are not usually sampled from a community in a stochastically. Stochastic sampling of speakers has been a methodological challenge for many years (Bayley et al., 2013; Diaz-Campos, 2011; Wodak, Johnstone, & Kerswill, 2011), yet logistic regression requires independence of observations, that is, it requires “random” as technically defined in statistics. Statistical independence and fixed effects issues were in view before the statistical method and representation of variable rules (Varbrul) were developed by Cedergren and Sankoff in 1974. Clark (1973), for example, presented a critique of fixed-effects issues quite clearly
Scientific Sociolinguistics?

While reviewing thirty years of references on quantitative methodologies in linguistics to ferret out the accuracy of the advice that we have been given in terms of best statistical practices in linguistics, I began to question where sociolinguistics as a field resides in the continuum between the so-called “hard” and “soft” sciences. If linguistics is the scientific study of language, is it reasonable to expect (socio)linguists to behave like other scientists? As Coupland (2007) remarks, variationist sociolinguistics adopted a decidedly more “scientific,” empirical, quantitative turn towards the “hard” sciences. Yet Hedges (1987), the innovator behind Hedges $g$ (effect size), wrote that social scientists “know intuitively that there is something ‘softer’ and less cumulative about our research results than about those of the physical sciences” (1987, p. 443). Hedges was referring to the lack of meta-analyses in the social sciences vis-à-vis hard sciences. Meta-analyses were shown here to be almost completely missing in sociolinguistics as a discipline, and completely missing in Hispanic sociolinguistics (although search results are not always perfect). If we do find that NHST is the primary quantitative focus for Hispanic sociolinguistic research beyond the studies reviewed here, to the exclusion of estimation and meta-analytic thinking, can we say that our implementation of statistics in Hispanic sociolinguistics is making us a “more hard” science, or is it working against us?

The invidious distinction between “hard” and “soft” sciences is clearly drawn in academic training with clear references to which discipline belongs to which category.
Physics and chemistry, for example, are in the hard sciences category because they investigate phenomena with a high degree of objectivity and certainty, unlike soft sciences like psychology, sociology, linguistics, or sociolinguistics. To borrow Neyman-Pearson’s notion, the differences can be behavioristic in that the hard sciences - characterized as more rigorous, replicable and generalizable – should be approached with respect, admiration, and belief while the soft sciences should be approached with dubiety and cynicism. This implies that sociolinguistics as a soft science is less rigorous, replicable, and generalizable. My findings support this implication.

As described by Yngve and Wasik (2006) in *Hard-science Linguistics*, those who use scientific methods are respected and believed while others are not. In 1974, sociolinguistics took a leap towards the hard sciences by developing and promulgating Varbrul as the mathematical implementation of the variable rule theoretical approach developed by William Labov in the late 1960s (Tagliamonte, 2006, 2011). Besides a true interest in answering variationist questions, there would also be advantages to being recognized as a serious, or hard, science. Sociolinguists could focus on those elements that would make it a quantifiable, and therefore, veritable science (Holmes & Hazen, 2014). Insofar as sociolinguistics deserves to be labeled as a harder science, it would be due to the statistical methods that sociolinguists choose when studying language and behavior. Given the new statistical methods and statistical admonitions surrounding us, what can we conclude about quantitative Hispanic sociolinguistics 40 years after the introduction of Varbrul?
In sociolinguistics, the distinction between hard and soft science may play two roles. First, it may influence how others view what sociolinguists do and the value that they place on it. Second, it may influence how sociolinguists view themselves. It is conceivable that, to elevate the “hard science” status of sociolinguistics, some researchers have turned to mechanized NHST rituals and sophisticated statistical routines without questioning them. The great risk is that if statistical methods are misunderstood or misapplied, the value of the researcher’s work may actually become obscured or misdirected, ultimately leading to crabbed and misconstrued statistical language that may countervail efforts to build an accurate scientific record, hinder theory generation, and confuse novice sociolinguists. Wolfram (1993) comments:

My own research encounters and experience in teaching students to conduce variation analysis has taught me that the presentation of elegant-appearing summaries of the results found in published versions of variation studies are often far removed from the laboratory in which such analyses are conducted. In the process, some of the critical procedural and analytical decisions may be disregarded or camouflaged, making analysis seem much cleaner than it actually is. (p. 194)

If quantitative sociolinguistics is to be the sort of scientific sociolinguistics envisioned in the 70s, in “the Spirit of ‘74” Gorman (2009, p. 1), our reflectors need to turn toward the new statistics within the frequentist paradigm for making statistical inference, and consider Bayesian inference more often.
Research Integrity

It remains clear that research integrity requires periodic review of the methods we use, regardless of tradition. A cornerstone of sociolinguistic research is the creation of new knowledge that is often generated through quantitative research methods (Bayley et al., 2013; Bayley & Lucas, 2007; Gorman & Johnson, 2013; Podesva & Sharma, 2014; Tagliamonte, 2006, 2007, 2011). Such knowledge is disseminated through specialized journals after it is filtered through a rigorous peer review process to ensure that only the highest quality sociolinguistic research gets disseminated. Reviewers, and the editors they serve, operate as content and methodological gatekeepers who are accountable for the integrity of the research they publish. At the same time, many sociolinguistic scholars perform research in a “publish or perish” environment shared across academic disciplines (Susan & Kitche, 2013), and the importance of publishing is conveyed to graduate students. While nobody involved in academic research and publishing should be expected to have a command of every data analysis method used in modern social science research, we should “maintain a broad and evolving awareness of methodological developments” (Hancock & Mueller, 2010, p. xiii) in statistics outside of sociolinguistics, and make honest attempts to evaluate the quantitative, analytical methods that remain at the core of many of our intellectual contributions.

Null Hypothesis Statistical Testing

Hispanic sociolinguistics is not likely to eliminate NHST in the near future, although clearly and demonstrably, “[i]mportant decisions should not be taken simply on the basis of a statistical hypothesis test (Woods, 1986, p. 127). This being case, it is wise
to remain vigilant of NHST fallacies and recognize the simple purpose of a $p$ value, and the validity of the arguments against NHST that are far more developed outside of linguistics. When a low $p$ value is encountered, we can simply interpret it as a “degree of evidence,” “degree of surprise,” or “degree of belief” that some real world process is not generating that quantity of collected data in an entirely random way. If there appears to be some evidence that the data are not likely to be due to a 100% randomizing process, then some degree of belief can be ascribed to the notion that somehow the constellation of data points may be related among themselves through some sort of shared, not 100% randomizing process.

For those who still find $p$ values worthwhile, caution should be used in setting some threshold to declare a test statistic as “significant.” Indeed, setting a threshold that lends itself to binary decision making without justification should not be considered best practices in the $p$ value discovery process reported in sociolinguistics. Sir Ronald Fisher, who introduced the idea of $p$ values, wrote:

Attempts to explain the cogency of tests of significance in scientific research by hypothetical frequencies of possible statements being right or wrong seem to miss their essential nature. One who ‘rejects’ a hypothesis as a matter of habit, when $p \geq 1\%$, will be mistaken in not more than 1% of such decisions. However, the calculation is absurdly academic. No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. (Fisher, 1956, p. 42)
In terms of teaching NHST and statistical critical thinking in the formation of new sociolinguistic researchers, higher bars can be reached. For example, within the Neyman-Pearson concept of statistical testing, most language researchers have typically been taught a lot about Type I error and very little about Type II error and statistical power. Thus, they are unaware that the Type II error rate is very large in the typical study. Budding sociolinguists may believe that the error rate is the alpha level used (typically .05 or .01). By defining Type I error rates at 5%, researchers may not realize that the Type II error rate is typically left free to go as high as 95%, as discussed earlier. The question is which error rate applies to a given study. We do not know if we do not test for it. If a language researcher does not know whether the null hypothesis is true or false, then they also do not know whether the relevant error rate is Type I or Type II; that is, they do not know if their error rate is 5% or some value as high as 95%. As discussed earlier, error rates of 50% or higher have been shown to be the usual case in many research literatures (Schmidt & Hunter, 2014).

In his tutorial for teaching data testing procedures, Perezgonzalez (2015) reasons through Fisher’s approach and Neyman-Pearson’s approach highlighting their incompatibilities in today’s NHST, and he concludes that the “best option would be to ditch NHST altogether” (p. 10) and refer to Fisher and of Neyman-Pearson in the rare instances that it is appropriate. In line with the new statistics, Perezgonzalez (2015) draws attention to the need for exploratory data analysis, effect sizes, confidence intervals, meta-analysis, Bayesian applications and, “chiefly, honest critical thinking” (Perezgonzalez, 2015, p. 10).
Bayesian inference makes sense for sociolinguistics, and for linguistics in general. Frequentists see probability as a measure of the frequency of repeated events, interpreting parameters as fixed, and data as random. Bayesians see probability as a measure of the degree of certainty about values, interpreting parameters as random and data as fixed (Kaplan, 2014). Every study reviewed here was frequentist. As frequentists, we determine the probability of obtaining the observed data (or in NHST, we include the probability of data more extreme than that observed), given that particular conditions exist, or the probability of the data given the hypothesis, and many assumptions: \( p(\text{Data} \mid H_0) \). Bayesians determine the probability that particular conditions exist given the data that have been collected, or the probability of the hypothesis given the data: \( p(H_0 \mid \text{Data}) \). In most cases, this what we are truly interested in. Gudmestad, House, and Geeslin (2013) argue for more Bayesian analyses in SLA and sociolinguistics, and demonstrate how it can be done. When we consider Bayesian analyses, however, we “should be wary of chasing yet another chimera: an apparently universal inference procedure” (Gigerenzer & Marewski, 2015, p. 1). The spectrum of frequentist and Bayesian inference options should all remain in the sociolinguist’s statistical toolbox, and we should apply frequentist and Bayesian inferential methods as appropriate for our sociolinguistic analyses.

At present, most Hispanic sociolinguists appear to be quite fettered to NHST within the frequentist paradigm. Pursing studies as frequentists is not a problem (we simply have fewer statistical tools to choose from), but our reliance on statistical significance tests in interpreting sociolinguistic studies may indeed lead to false
conclusions about what our study results mean. Without regard to their area of expertise, linguists would do well to discontinue reliance on NHST. New statistics books are introducing confidence intervals before statistical testing, calling attention to NHST problems, and redirecting student attention to the methods espoused by authors of the new statistics (Nussbaum, 2014; Vogt et al., 2014). The recommendation to demote NHST will continue until it is fulfilled. In the meantime, Hispanic sociolinguists can be the first to be among “those in the know,” and among those who are contributing to the field with high integrity research that lends itself to replication and meta-analytic thinking, in other words, contributing research that lends itself to valid and reliable theory building.

**Toward a Cumulative Quantitative Hispanic sociolinguistics**

Most empirical articles I reviewed made conclusions based on \( p \) values. Reliance on NHST continues, but there are at least four reasons why statistical reform is needed. First, the central causal role of NHST in our sociolinguistic studies demands scrutiny, especially with increased recognition of research-integrity issues. Second, over the past decade, which corresponds to the research period in this thesis, many helpful resources have become available to support the practical use of estimation (ES and CI, or ESCI) and meta-analysis. Third, psycholinguistics, corpus linguistics, and applied linguistics are all advancing in terms of the statistical integrity of their research. Fourth, the APA’s (2010) *Publication Manual* included unequivocal statements that quantitative interpretation should be based on estimation.
To remain viable, quantitative sociolinguistics must adopt best practices in statistics, and there is no reason why Hispanic sociolinguists, who have tremendous import in the Americas and Europe, cannot act as leaders in this regard. The following comments of Plonsky (2013) apply equally to sociolinguistics:

methodological infirmity not only hinders progress in the development of theory but may also negatively affect the reputation and legitimacy of SLA as a discipline and limit its potential to contribute to parent fields such as linguistics, education, and psychology. (p. 656)

Clearly, as researchers of Hispanic sociolinguistics, we want to accelerate progress in the development of sociolinguistic theory, deploy the quantitative methods in a manner consistent with the notion of “scientific sociolinguists,” bolster our legitimacy as a discipline, and enhance our potential to contribute to both sibling and parent fields. We also want our quantitative conclusions to be correct and meaningful.

New Directions

Within the frequentist paradigm, the new statistics is the new direction for Hispanic sociolinguists, and for linguistics in general. The New Statistics, as authored by (Cumming, 2012, 2014), and supported by many scholars like Kline (2013), emphasizes research integrity. Integrity is achieved by study pre-specification, avoidance of inappropriate data-analytic practices (such as ignoring assumptions required for accurate test results), complete reporting, and encouragement of replication. In addition, (Cumming, 2012, 2014) joins a long list of statisticians, most notably psychometricians from the American Psychological Association (APA), who call attention to the “severe
flaws” of NHST, and the need to move from dichotomous point thinking to estimation thinking based on effect sizes, confidence intervals, and meta-analysis. Cumming (2014) offers an eight-step “new-statistics strategy for research with integrity” (p. 7) that linguists and sociolinguistics can adopt, summarized as follows:

1. Formulate research questions in estimation terms such as “How large is the effect?” or “To what extent?”

2. Identify the effect sizes that will best answer the research questions.

3. Declare full details of the intended procedure and data analysis.

4. After running the study, calculate point estimates and confidence intervals for the chosen effect sizes.

5. Make one or more figures, including confidence intervals where error bars to depict 95% CIs.

6. Interpret the ESs and CIs. In writing up results, discuss the ES estimates, which are the main research outcome, and the CI lengths, which indicate precision. Consider theoretical and practical implications, in accordance with the research aims.

7. Use meta-analytic thinking throughout. Think of any single study as building on past studies and leading to future studies. Present results to facilitate their inclusion in future meta-analyses. Use meta-analysis to integrate findings whenever appropriate.

8. Report. Make a full description of the research, preferably including the raw data, available to other researchers. This may be done via journal publication or posting to some enduring publicly available online repository. Be fully transparent about every step, including data selection, cleaning, and analysis.
My findings suggest that journal editorial policies and disciplinary customs in Hispanic variationist sociolinguistics include quantitative routines from the past, but they may be changing. Varbrul is not the only software application being used, nor is it the only methodology being explored, and not every scholar has put full faith in null hypothesis statistical testing. I am optimistic that current and future Hispanic sociolinguists will learn, explore, and enjoy the enormous benefits of adopting new quantitative methods to analyze variation. Indeed, multilevel modeling, resampling techniques, and even Bayesian statistics may one day dominate the journals of sociolinguists. For now, however, we need to inform and help each other understand the limits of statistical testing, and find the best path to navigate the new statistics, ensuring that we always maintain a broad and evolving awareness of the trends in statistical best practices, even – or perhaps especially – outside of linguistics. Only in this way will we build the highest integrity academic record in quantitative sociolinguistics in general, and in Hispanic variationist sociolinguistics in particular.
REFERENCES


Engman, A. (2013). Is there life after P<0.05? Statistical significance and quantitative sociology. *Quality and Quantity, 47*(1), 257-270. doi: 10.1007/s11135-011-9516-z


Hatcher, L. (2013). *Advanced statistics in research: Reading, understanding, and writing up data analysis results* Saginaw, MI: Shadow Finch Media.


APPENDIX A

METHODS USED
<table>
<thead>
<tr>
<th>Method</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>22</td>
<td>50%</td>
</tr>
<tr>
<td>n</td>
<td>13</td>
<td>30%</td>
</tr>
<tr>
<td>y</td>
<td>9</td>
<td>20%</td>
</tr>
<tr>
<td>Chi-square</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>22</td>
<td>50%</td>
</tr>
<tr>
<td>n</td>
<td>16</td>
<td>36%</td>
</tr>
<tr>
<td>y</td>
<td>6</td>
<td>14%</td>
</tr>
<tr>
<td>z test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>42</td>
<td>95%</td>
</tr>
<tr>
<td>y</td>
<td>2</td>
<td>5%</td>
</tr>
<tr>
<td>t test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>40</td>
<td>91%</td>
</tr>
<tr>
<td>y</td>
<td>4</td>
<td>9%</td>
</tr>
<tr>
<td>ANOVA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>30</td>
<td>68%</td>
</tr>
<tr>
<td>y</td>
<td>14</td>
<td>32%</td>
</tr>
<tr>
<td>Reg Regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>43</td>
<td>98%</td>
</tr>
<tr>
<td>y</td>
<td>1</td>
<td>2%</td>
</tr>
<tr>
<td>Mixed Effects Regr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>41</td>
<td>93%</td>
</tr>
<tr>
<td>y</td>
<td>3</td>
<td>7%</td>
</tr>
<tr>
<td>DFA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>43</td>
<td>98%</td>
</tr>
<tr>
<td>y</td>
<td>1</td>
<td>2%</td>
</tr>
<tr>
<td>PCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>43</td>
<td>98%</td>
</tr>
<tr>
<td>y</td>
<td>1</td>
<td>2%</td>
</tr>
<tr>
<td>Factor analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>44</td>
<td>100%</td>
</tr>
<tr>
<td>Cluster analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>43</td>
<td>98%</td>
</tr>
<tr>
<td>y</td>
<td>1</td>
<td>2%</td>
</tr>
<tr>
<td>LR Fixed Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>22</td>
<td>50%</td>
</tr>
<tr>
<td>y</td>
<td>22</td>
<td>50%</td>
</tr>
<tr>
<td>LR Mixed Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>40</td>
<td>91%</td>
</tr>
<tr>
<td>y</td>
<td>4</td>
<td>9%</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronbach alpha</td>
<td>1</td>
<td>2%</td>
</tr>
<tr>
<td>probit</td>
<td>1</td>
<td>2%</td>
</tr>
<tr>
<td>Journal</td>
<td>Year</td>
<td>Author(s)</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>-----------</td>
</tr>
<tr>
<td>LVC</td>
<td>2006</td>
<td>Carrera-Sabaté, Josefina</td>
</tr>
<tr>
<td>LVC</td>
<td>2006</td>
<td>Cutillas-Espinosa, Juan A. Hernández-Campoy, Juan Manuel</td>
</tr>
<tr>
<td>LVC</td>
<td>2006</td>
<td>Dunlap, Carolyn</td>
</tr>
<tr>
<td>LVC</td>
<td>2006</td>
<td>Wolford, Tonya E.</td>
</tr>
<tr>
<td>LVC</td>
<td>2007</td>
<td>Travis, Catherine E.</td>
</tr>
<tr>
<td>LVC</td>
<td>2008</td>
<td>Blas Arroyo, José Luis</td>
</tr>
<tr>
<td>LVC</td>
<td>2008</td>
<td>Hackert, Stephanie</td>
</tr>
<tr>
<td>LVC</td>
<td>2008</td>
<td>Schwenter, Scott A. Torres Cacoullos, Rena</td>
</tr>
<tr>
<td>LVC</td>
<td>2008</td>
<td>Sanchez, Tara</td>
</tr>
<tr>
<td>LVC</td>
<td>2009</td>
<td>Reig Alamillo, Asela</td>
</tr>
<tr>
<td>LVC</td>
<td>2010</td>
<td>Aaron, Jessi E.</td>
</tr>
<tr>
<td>LVC</td>
<td>2010</td>
<td>Lipski, John M.</td>
</tr>
<tr>
<td>LVC</td>
<td>2011</td>
<td>Copple, Mary T.</td>
</tr>
<tr>
<td>LVC</td>
<td>2011</td>
<td>File-Muriel, Richard J. Brown, Earl K.</td>
</tr>
<tr>
<td>LVC</td>
<td>2012</td>
<td>Poplack, Shana Dion, Nathalie</td>
</tr>
<tr>
<td>Year</td>
<td>Authors</td>
<td>Title</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>2012</td>
<td>Brown, Esther L. Rivas, Javier</td>
<td>Grammatical relation probability: How usage patterns shape analogy</td>
</tr>
<tr>
<td>2013</td>
<td>Miller, Karen</td>
<td>Acquisition of variable rules: /s/-lenition in the speech of Chilean Spanish-speaking children and their caregivers</td>
</tr>
<tr>
<td>2014</td>
<td>Claes, Jeroen</td>
<td>A cognitive construction grammar approach to the pluralization of presentational <em>haber</em> in Puerto Rican Spanish</td>
</tr>
<tr>
<td>2014</td>
<td>Shin, Naomi L.</td>
<td>Grammatical complexification in Spanish in New York: 3sg pronoun expression and verbal ambiguity</td>
</tr>
<tr>
<td>2014</td>
<td>Walker, Abby García, Christina Cortés, Yomi Campbell-Kibler, Kathryn</td>
<td>Comparing social meanings across listener and speaker groups: The indexical field of Spanish /s</td>
</tr>
<tr>
<td>2015</td>
<td>Blas Arroyo, José Luis Vellón Lahoz, Javier</td>
<td>The refuge of a dying variant within the grammar: Patterns of change and continuity in the Spanish verbal periphrasis <em>haber de</em> + infinitive over the past two centuries</td>
</tr>
<tr>
<td>2006</td>
<td>Porcel, Jorge</td>
<td>The paradox of Spanish among Miami Cubans</td>
</tr>
<tr>
<td>2008</td>
<td>Barrett, Rusty</td>
<td>Linguistic differentiation and Mayan language revitalization in Guatemala</td>
</tr>
<tr>
<td>2008</td>
<td>Newman, Michael Trenchs-Parera, Mireia Ng, Shukhan</td>
<td>Normalizing bilingualism: The effects of the Catalanian linguistic normalization policy one generation after</td>
</tr>
<tr>
<td>2009</td>
<td>Hernández, José Esteban</td>
<td>Measuring rates of word-final nasal velarization: The effect of dialect contact on in-group and out-group exchanges</td>
</tr>
<tr>
<td>2009</td>
<td>Aaron, Jessi E.</td>
<td>Coming back to life: From indicator to stereotype and a strange story of frequency</td>
</tr>
<tr>
<td>2009</td>
<td>Angermeyer, Philipp S.</td>
<td>Translation style and participant roles in court interpreting</td>
</tr>
<tr>
<td>2010</td>
<td>Newman, Michael</td>
<td>Focusing, implicational scaling, and the dialect status of New York Latino English</td>
</tr>
<tr>
<td>JS</td>
<td>2011</td>
<td>Montaruli, Elisa Bourhis, Richard Y. Azurmendi, Maria-Jose</td>
</tr>
<tr>
<td>JS</td>
<td>2012</td>
<td>Delforge, Ann Marie</td>
</tr>
<tr>
<td>JS</td>
<td>2013</td>
<td>Carter, Phillip M.</td>
</tr>
<tr>
<td>JS</td>
<td>2014</td>
<td>Babel, Anna M.</td>
</tr>
<tr>
<td>JS</td>
<td>2014</td>
<td>Lev-Ari, Shiri San Giacomo, Marcela Peperkamp, Sharon</td>
</tr>
<tr>
<td>Hisp</td>
<td>2010</td>
<td>Alvord, Scott M.</td>
</tr>
<tr>
<td>Hisp</td>
<td>2010</td>
<td>Bishop, Kelley Michnowicz, Jim</td>
</tr>
<tr>
<td>Hisp</td>
<td>2010</td>
<td>Blas Arroyo, José Luis</td>
</tr>
<tr>
<td>Hisp</td>
<td>2011</td>
<td>Clegg, Jens H.</td>
</tr>
<tr>
<td>Hisp</td>
<td>2011</td>
<td>Mayberry, María</td>
</tr>
<tr>
<td>Hisp</td>
<td>2011</td>
<td>González Martínez, Juan Blas Arroyo, José</td>
</tr>
<tr>
<td>Hisp</td>
<td>2011</td>
<td>Rivadeneira, Marcela J. Clua, Esteve B.</td>
</tr>
<tr>
<td>Hisp</td>
<td>2012</td>
<td>Díaz-Campos, Manuel Killam, Jason</td>
</tr>
<tr>
<td>Hisp</td>
<td>2012</td>
<td>Tight, Daniel G.</td>
</tr>
<tr>
<td>Year</td>
<td>Author(s)</td>
<td>Title</td>
</tr>
<tr>
<td>------</td>
<td>------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>2013</td>
<td>Büdenbender, Eva-María</td>
<td>“Te conozco, bacalao”: Investigating the influence of social stereotypes on linguistic attitudes</td>
</tr>
<tr>
<td>2013</td>
<td>Shelton, Michael</td>
<td>Spanish rhotics: More evidence of gradience in the system</td>
</tr>
<tr>
<td>2013</td>
<td>Sorenson, Travis</td>
<td>Voseo to tuteo accommodation among Salvadorans in the United States</td>
</tr>
<tr>
<td>2014</td>
<td>Weyers, Joseph R.</td>
<td>The tuteo of Rocha, Uruguay: A study of pride and language maintenance</td>
</tr>
</tbody>
</table>