A quantitative defense of linguistic methodology

Jon Sprouse[a]
Diogo Almeida[a]

[a] Department of Cognitive Sciences, University of California, Irvine

Abstract

It is often argued that the current divide between linguistic theory and other domains of language research can be traced to the unreliability of linguistic methodology, and the resulting unreliability of linguistic theory. Linguists rely upon informal acceptability judgment experiments as the primary source of linguistic data. These experiments tend to employ small samples, often composed of non-naïve participants, and generally forego the use of distracter items and inferential statistics in the analysis of the results. We present resampling analyses of a large acceptability judgment dataset that provides new quantitative evidence that traditional linguistic methodology is extremely reliable with very small samples, usually at the level of the individual participant. We also review the evidence that critics of informal linguistic experiments provide to justify their concerns, which shows that the evidence provided by critics turns out to mostly *corroborate* the reliability of informal judgments, and not undermine it. We conclude that there is no empirical, logical, or statistical reason to think that the informal experiments routinely performed by linguists are unreliable. In fact, we show evidence that these experiments might be, in some circumstances, much more powerful than formal experiments with naïve participants. Given the lack of evidence of problems with traditional linguistic methodology, we hypothesize that one potential reason for the recurrence of this debate is that the phenomena critics are particularly interested in often elicit effects that are very small and hard to detect in formal acceptability judgment tasks. This suggests that critics may be mistaking a property of specific phenomena for a property of the methodology. Taken as a whole, these results suggest that broad criticisms of linguistic theory based on the reliability of linguistic data are unfounded, and that methodological concerns should not influence the relationship between linguistic theory and other domains of language research.

Keywords: Acceptability judgments, linguistic theory, linguistic methodology, quantitative standards, experimental syntax

1. Introduction

Linguists are primarily interested in the properties of linguistic representations. To isolate these properties, linguists routinely compare two (or more) minimally different representations by eliciting a behavioral response known as an acceptability judgment from a native speaker of the language of interest (Chomsky, 1965; Schütze, 1996). Traditionally, these acceptability judgments are collected using an informal experiment (Marantz, 2005), which typically involves small sample sizes (often the linguist and a handful of naïve subjects or other linguists), and does not rely on tests of statistical significance. This traditional methodology has often been criticized by other language researchers, typically those who self-identify as psycholinguists rather than

linguists, who believe that this informal experimental methodology is unreliable. Criticisms of linguistic methodology can be found from the earliest days of modern linguistics (e.g., Hill, 1961; Spencer, 1973), with at least 5 high-profile criticisms in the past 10 years alone (Edelman & Christiansen, 2003; Ferreira, 2005; Wasow & Arnold, 2005; Gibson & Fedorenko, 2010; Dabrowska, 2010). These criticisms argue that linguistic theory itself is unsound because the data that the theory is built upon is unreliable:

> First, such judgments are inherently unreliable because of their unavoidable meta-cognitive overtones, because grammaticality is better described as a graded quantity, and for a host of other reasons. (Edelman & Christiansen, 2003, p. 60)

> The lack of validity of the standard linguistic methodology has led to many cases in the literature where questionable judgments have led to incorrect generalizations and unsound theorizing. (Gibson & Fedorenko, 2010, p. 233)

> The result has been the construction of elaborate theoretical edifices supported by disturbingly shaky empirical evidence. (Wasow & Arnold, 2005, p. 1482)

These arguments are generally offered as a possible explanation for one of the more striking sociological facts of the current state of language research: linguistic theory, which purports to be a theory of linguistic representations, exists in relative isolation from other domains of language research, such as theories of real-time language processing (Edelman and Christiansen, 2003; Ferreira, 2005). The argument seems to be that language researchers ignore linguistic theory because linguistic theory is unsound, and linguistic theory is unsound because linguistic data is unreliable:

> Researchers with higher methodological standards will often ignore the current theories from the field of linguistics. (Gibson and Fedorenko, 2010, p. 233)

Linguists have consistently offered spirited, logical arguments in response to these criticisms (e.g., Phillips & Lasnik, 2003; Marantz, 2005; Culicover & Jackendoff, 2010), and have reported the existence of numerous unpublished formal experimental results that corroborate dozens, if not hundreds, of informal experimental results (Featherston, 2009; Phillips, 2009). However, the regular recurrence of these criticisms suggests that the arguments linguists have offered so far have not been successful in convincing the critics that the results of informal experiments are reliable. Therefore, we take a new approach in this paper: we present quantitative evidence of the (striking) reliability of traditional linguistic methodology in the hopes of eliminating methodological concerns from the list of reasons for the divide between linguistics and other domains of language research.

Over the past 15 years, a large number of linguists have explored the *usefulness* (as opposed to *necessity*) of formal experimental methods for the advancement of linguistic theory (e.g., Bard, Robertson, & Sorace, 1996; Cowart, 1997; Keller, 2000; Sorace & Keller, 2005; Featherston, 2005a, 2005b; Myers, 2009; Sprouse, 2009; Sprouse & Cunningham, *under review*; Sprouse, Wagers, & Phillips, *submitted*). The pursuit of this research has created a body of experimental and statistical knowledge necessary for the execution and analysis of formal acceptability judgment experiments. Our goal in this paper is to use this body of knowledge to

address the most prominent criticisms of informal acceptability experiments. To be clear, we do not intend these arguments to replace the logical arguments that linguists have previously offered in their defense, as we believe that they are sound arguments. Instead, we hope to present a new defense in the same quantitative terms as the criticisms. The criticisms assume that the reliability of an experimental result can only be confirmed using the tools of formal experimentation and formal statistics; while this assumption is simply not true (see section 4 for a detailed discussion), it does little good to argue this point without also offering the types of formal experimental and statistical evidence that the critics assume as a prerequisite. We will present large-scale quantitative analyses of novel experiments, quantitative re-analyses of existing evidence, and a discussion of various facts from the statistical literature. Our goal is to evaluate the empirical and logical validity of the criticisms and present a comprehensive quantitative defense of linguistic methodology.

Our argument will be laid out as follows. First, the remainder of this section will be devoted to delineating the actual differences between the informal experiments of traditional linguistic methodology and the formal experiments advocated by critics, so as to circumscribe the empirical issues at the heart of this debate. Next, we will propose a quantitative definition of reliability as it is understood in the statistics literature, and present a series of resampling analyses designed to quantify the reliability of four phenomena from the linguistics literature that have been claimed to be highly variable; the results of these analyses suggest that these phenomena are in fact strikingly reliable (section 2). In order to better understand the discrepancy between the results of our reliability analyses and the claims of the critics, we will then review the empirical evidence that critics have presented in support of their claim that the results of informal experiments are unreliable. We will argue that (i) two purported pieces of evidence actually corroborate previous informal results, but were misinterpreted by the critics; (ii) two other pieces of evidence were the result of underpowered experiments; and (iii) the one systematic investigation of potential cognitive bias by linguists actually suggests that linguists are biased against their own conclusions (section 3). In the end, we can find only one reliable example of an informal experimental result not holding up under formal experimentation – a far cry from the epidemic of unreliability claimed by critics, and hardly a challenge to the dozens, if not hundreds, of formal experimental results corroborating informal experiments (Featherston, 2009; Phillips, 2009). Given that we can find no empirical reason to believe that linguistic methodology is unreliable, we will then address a recurring normative argument made by critics – that linguists should use statistical significance testing to ensure reliable results. We will evaluate several possible interpretations of this claim, and in each case draw upon dozens of articles in the statistical literature to argue that there is simply no statistical reason to routinely employ statistical significance testing given the nature of linguistic data. In fact, we will argue that the overwhelming consensus among statisticians is that replication is the most important criterion for establishing the reliability of a result – and linguistic data is one of the most replicated data types in cognitive science (section 4). Section 5 concludes with three additional case studies that may suggest a reason for the recurrence of this debate despite the overwhelming evidence that the results of informal experiments are reliable. These case studies suggest that acceptability judgments of phenomena of particular interest to the critics are in fact unreliable without large-scale formal experiments and undetectable without statistical significance testing. If the results hold for additional phenomena, this could suggest that the critics may be mistaking a property of specific phenomena for a property of the methodology. Taken as a whole, our

results suggest that broad criticisms of linguistic theory based on the reliability of linguistic data are unfounded.

The distinction between informal and formal experiments in linguistics

The first step in addressing the criticisms of traditional linguistic methodology is to clarify the similarities and differences between the traditional informal experiments that linguists routinely perform and the formal experiments that the critics advocate. Despite the fact that this topic has already been addressed by others (cf. Marantz, 2005), some of the arguments presented by critics of traditional linguistic methodology still seem to stem from a basic misunderstanding of what the traditional methodology actually entails (e.g., Gibson & Fedorenko, 2010).

Informal experiments are standardly performed in the following way: First, the linguist constructs a set of conditions to minimally contrast the relevant representational property. These conditions are carefully constructed to rule out known nuisance variables, the set of which has been incrementally accumulated by prior linguistic research. Next, the linguist constructs a set of sentences for each condition in an attempt to rule out any lexically driven extraneous factors (such as plausibility or word frequency). Finally, the linguist asks (verbally or by e-mail) a sample of native speakers (in which she and a handful of her colleagues or graduate students are normally included) to rate these items along a scale to determine whether the representational property has an effect on the relative acceptability of the two conditions. The results are then summarized in a journal article using a diacritic placed to the left of an archetypical example of each sentence type: * indicates that most subjects found most of the sentences of the condition very unacceptable, ? indicates that most subjects found most of the sentences of the condition to lie in the middle of the spectrum, and no diacritic indicates that most subjects found the sentences of the condition acceptable.

Informal and formal experiments are identical in many respects. First, they both require the careful construction of conditions that are relevant to the theoretical question and free of known nuisance variables. Second, they both require the construction of several lexicalizations of each condition to rule out lexically driven factors. In this way, the logic and construction of informal experiments is identical to the construction of formal experiments. The only differences between informal and formal experiments are in the presentation of the materials and in the analysis of the results, and often are a matter of degree. Informal experiments typically only include the experimental items, whereas formal experiments typically include unrelated distracter items in a full survey format. Informal experiments typically involve small sample sizes (1-10), whereas formal experiments typically include samples of 20 or more college students. Informal experiments are typically analyzed with non-numerical descriptive summary only (the diacritics), whereas formal experiments are typically summarized using standard descriptive statistics (mean, standard deviation) and analyzed using standard inferential statistics (*t*-test, ANOVA, linear mixed effects modeling).

It is important to note that the descriptors *informal* and *formal* relate not to the logic and construction of the linguistic experiments, which are identical for both types, but to the decisions made by the linguists regarding the presentation of materials, sample selection, and methods of analysis that will be used in the experiment. Since this is the scope of the criticism leveled against informal linguistic experiments, we will, in the next sections, investigate what impact, if any, the different decisions made in informal and formal linguistic experiments about stimulus presentation, sample selection, and method of analysis have on the conclusions the researcher

would be able to draw from each kind of experiment, as well as the empirical evidence offered by critics to justify their concerns.

2. A quantitative investigation of reliability at different sample sizes

One of the empirical questions at the heart of the debate surrounding linguistic methodology is whether large samples are a necessary condition for reliable acceptability judgment data. In section 2.1 we propose a three-part definition of reliability derived from the statistical literature. In section 2.2 we propose two methods for quantifying reliability for any pair of sentence types at arbitrary sample sizes. In section 2.3 we propose a set of desiderata for case studies of reliability. In section 2.4 we present the details of the large-scale experiment that forms the basis for our case studies. Finally, in section 2.5, we present analyses of the reliability of four phenomena from the linguistics literature that meet the desiderata, and two phenomena that have been suggested as possible confounds in informal experiments. The results of these analyses suggest that acceptability judgments are strikingly reliable even at the very small sample sizes typically used in informal experiments.

2.1 A definition of reliability

The first step in investigating the reliability of acceptability judgments is to establish a definition of what it means to reliable. The field of statistics provides us with a framework for defining reliability by delineating three desiderata for any statistical test:

(i) No false positives
If there is no difference between the experimental conditions, then the test should report no difference

(ii) No false negatives (or Reliable detection)
If there is a difference between the experimental conditions, then the test should report a difference

(iii) No sign reversals
If there is a difference between the experimental conditions that is in a specific direction, then the test should report a difference in the correct direction

In the statistics literature, these three desiderata are most commonly discussed in terms of errors (i.e., a false positive is a type 1 error, a false negative is a type 2 error, and a sign reversal is a type 3 error), but we can also think of them as three aspects of reliability. If acceptability judgments are reliable at small sample sizes, then there should be relatively few false positives when there is no difference between the target sentence types, and there should be relatively few false negatives and sign reversals when there is a difference between the target sentence types.

2.2 Two methodologies for quantifying reliability

The second step in investigating the reliability of acceptability judgments is to establish a methodology for assessing the likelihood of detecting a difference between the average

acceptability rating of two sentence types for a given sample size; we will call this the "difference detection rate." In this subsection, we describe two types of resampling simulations that provide a straightforward method for estimating difference detection rates. The simulations reported here were all based on a large dataset of judgments from 173 undergraduates described in section 2.4.

Simulation of experiments

As an example of how a resampling simulation works, suppose we wanted to assess the difference detection rates for experiments with the following characteristics: a sample size of 5 participants, one acceptability judgment per participant per experimental condition. The simulation algorithm would be:

1. Draw a random sample of 5 participants (allowing participants to be potentially drawn more than once -- this is called "sampling with replacement")
2. Randomly select 1 of the 4 judgments for each condition from each participant
3. Run a set of statistical tests on the sample evaluating each possible effect direction to induce a detection or error, depending on whether an effect is expected or not, and which effect direction is expected. For all of the results reported here, we used simple paired t-tests.
4. Repeat this procedure 1000 times, and compute the proportion of false negatives, detections (which is equal to 1-false negatives), and sign reversals.

We could then repeat this procedure for every other possible sample size from 6 to 173 to derive estimates of the detection rates at every sample size for that pair of sentence types. We could also repeat this procedure with more judgments per participant to quantify the effect of increasing the number of judgments given by each participant.

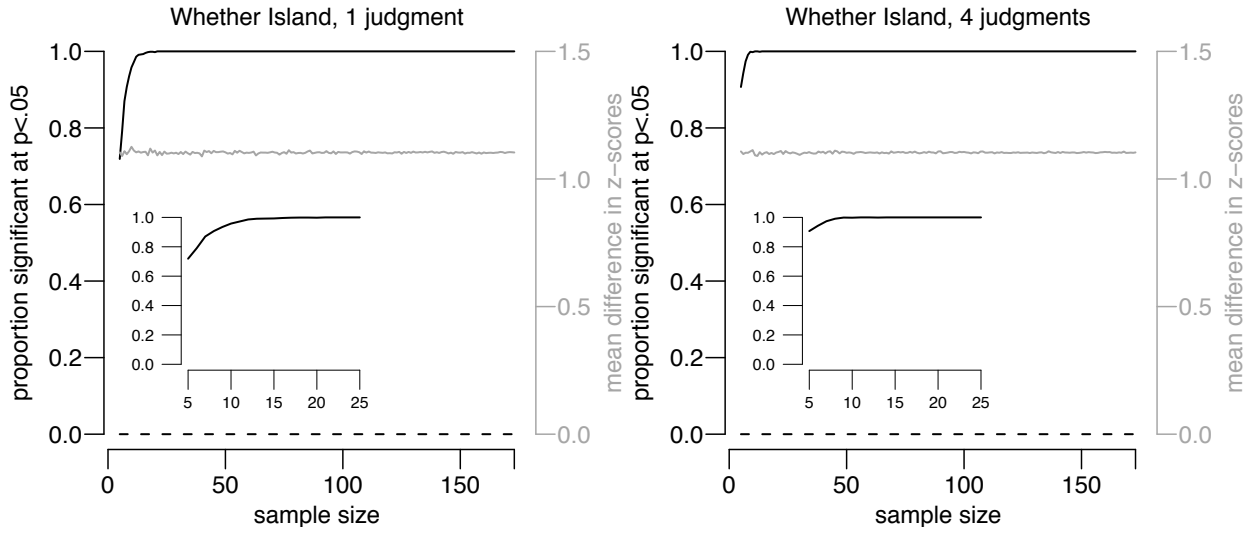
As a concrete example, let us take the following comparison between a long-distance wh-dependency out of an embedded *that*-clause and a long-distance wh-dependency out of an embedded *whether*-clause. This pair of sentence types is known in the linguistics literature as a *whether*-island effect:

(1) (a) What do you think that John bought?
(b) *What do you wonder whether John bought?

The informal experiments reported in the literature report that (1a) is more acceptable than (1b). The resampling simulations described above yield the following graphs:

Figure 1: An example of resampling simulation results for the *whether*-island effect

The left panel presents the results of a resampling simulation for samples from 5 to 173 participants with only one judgment per participant. The right panel presents the results for four judgments per participant. The x-axis is sample size, and the left-side y-axis is proportion of results at p < *.05. The solid black line is the difference detection rate. The dashed black line is the sign-reversal rate. The solid grey line reports the difference between condition means in z-units using the right-side y-axis. The inset provides a zoomed in view of the difference detection rate for sample sizes 5-25.*



The graph on the left reports 1000 random samples for each sample size from 5 to 173 with one judgment per subject. The graph on the right reports 1000 samples for each sample size from 5 to 173 with four judgments per subject. The solid black line represents the proportion of samples that resulted in a significant *p*-value ($p<.05$) in the predicted direction (1a > 1b) – this is the *detection rate*, or 1-false negative rate. Because this line asymptotes rather quickly, the inset provides a zoomed-in view for sample sizes 5 through 25. The dashed black line represents the proportion of samples that resulted in a significant *p*-value ($p<.05$) in the incorrect direction (1b > 1a) – *the sign reversal rate*. The grey line uses the right-side y-axis to report the size of the difference between the two condition means (in z-units) for each sample size. We will delay a detailed discussion of these results until section 2.5. Suffice it to say, these graphs make it clear that the detection rate for Whether islands is over 70% with only 5 subjects and a single judgment per subject, and over 90% with 5 subjects and 4 judgments per subject. The sign-reversal rate is always 0%, even with only 5 subjects, and the difference between means is the same for both small and large samples.

There is, of course, one serious drawback to the use of statistical tests like the *t*-test for identifying significant differences at each possible sample size: the power of parametric statistical tests is directly proportional to the size of the sample. This means that we cannot be sure how many of the non-significant results at small sample sizes are due to the nature of the phenomenon of interest and how many are due to the minimal sample size restrictions imposed by the *t*-test. Similarly, with very large sample sizes (over 100), *t*-tests can return significant results for very small differences. This means we cannot be sure how many of the significant

results at large sample sizes are due to the nature of the phenomenon and how many are due to the sensitivity of the *t*-test.

Simulation of individual long-run relative frequency

To overcome the sample-size requirements of the *t*-test-based resampling analyses, we also calculated an approximation of the long-run relative frequency of obtaining a judgment difference in the predicted direction from a single subject. This type of analysis attempts to answer the question: How likely is it that a single subject will report the hypothesized effect after a single judgment of each condition? This gives us a window into the reliability of the effect for a single subject without relying on statistical tests that impose their own sample-size requirements. We calculate the long-run relative frequency as follows:

1. Randomly select a single subject from the larger dataset
2. Randomly select one of the 4 responses by that subject for condition A
3. Randomly select one of the 4 responses by that subject for condition B
4. Compare the raw numerical ratings of conditions A and B, and record a success if the difference between the two conditions is in the predicted direction (e.g., A > B), otherwise record a failure.
5. Repeat steps (1-4) N times, where N is a very large number (e.g., 100,000).
6. Divide the total of successes and the total of failures (calculated in step 5) by N to derive a proportion of the possible experimental effects in each direction.

The interpretation of the results of this method requires some caution, however, due to the nature of the dataset. As section 2.4 will discuss, the experimental survey contained over 100 items. This means that participants were not simply comparing the two conditions to each other as the method above assumes. Instead, they were asked to give numerical judgments for each item as compared to a reference sentence (the standard) in a magnitude estimation task. Because there was no direct comparison of the conditions, differences in the correct direction may be slightly underestimated, and differences in the incorrect direction may be slightly overestimated. Similarly, because the raw number estimators were along a rather large scale (between 0 and 200), there is a very real question as to what size difference should count as reliable. In order to be as conservative as possible (and bias our analyses against our conclusions), we report likelihoods for a difference of greater than 10 (i.e., a success in step 4 above would require 1a > 1b by more than 10 units in the raw number ratings).

2.3 Selection criteria for our case studies

The only way to be completely certain that the thousands of data points that have been established using informal experiments are reliably replicable in formal experiments would be to actually test each one in a formal experiment. It is unrealistic to expect the field of linguistics to devote time and resources to such an endeavor, especially when the professional opinion of a substantial majority of the field is that there is no problem with the informal experimental method or the data that has been gathered by relying on it (see Marantz, 2005). Moreover, it is important to note that in cases where disagreements between the results of the two types of experiments are found, the disagreement itself cannot tell which result is the "truth," only that

there is a disagreement for some unspecified reason (see section 3 for a discussion). Given these facts, it seems clear that this methodological debate will be fought using individual case studies; therefore, the first criticism levied by opponents of a given paper will be that the case studies presented are not representative of the broader set of informal experimental results reported in the linguistics literature. In this subsection, we would like to anticipate this criticism by explicitly stating the reasons that we believe our case studies are not only appropriate, but also more representative than many of the case studies offered by critics of informal experiments.

We follow Phillips (2009) in believing that the first, and arguably most important, criterion for selecting case studies is that the phenomena must be central to linguistic theory. The logic of this criterion is straightforward: The primary concern of critics of informal experiments is that the informal methodology may have led to unsound linguistic theories; therefore, truly relevant case studies must be phenomena that could have a meaningful impact on the construction of linguistic theories. It has been argued by some critics that the importance of a phenomenon is subjective (e.g., Gibson & Fedorenko, *submitted*); this is simply untrue. Trendy topics do change over time (as one would expect); however, the phenomena that exert the most influence on the shape of the theory, such as phenomena that motivate context-sensitive grammatical rules (e.g., transformations and their constraints), have changed little over the past 50 years of linguistic theorizing. Phillips (2009) correctly observes that the major theoretical debates within linguistics are never about the status of individual phenomena – indeed, nearly all linguists agree about the phenomena that need to be explained – but rather about different theoretical approaches to explaining these phenomena.

The "importance criterion" above makes case study selection necessarily non-random; therefore, it is crucial that the selection not be overly biased toward supporting the researcher's preferred conclusion. We would like to suggest that one should choose phenomena for which there are published claims that run counter to the preferred conclusion. For critics of linguistic methodology, this would mean phenomena that have been claimed in the literature to be robust and replicable; for defenders of linguistic methodology (like us), this would mean phenomena that have been claimed in the literature to be non-reliable. This bias correction is an unstated principle of all experimental research, but it seems to us that it has been relatively disregarded in the criticism of informal judgments, as several of the critics have chosen phenomena that are known to most linguists to be controversial (see section 3 for a discussion). In fact, it is rarely mentioned in these criticisms that the fact that linguists have already identified these data points as controversial indicates that linguists are doing exactly what is expected of any experimental science – they are policing their own data. Though it may not always be evident in the published literature, practicing syntacticians often add appropriate caveats when discussing areas of the theory supported by controversial results – a practice that is shared by researchers in all areas of cognitive science.

With these two criteria in mind, we have chosen four types of island effects (Ross 1967) as our primary case studies:

(2) Whether island
  (a) What do you think [$_{CP}$ that John bought __]?
  (b) *What do you wonder [$_{CP}$ whether John bought __]?

(3) Complex NP island
  (a) What did you claim [$_{CP}$ that John bought __]?
  (b) *What did you make [$_{NP}$ the claim [$_{CP}$ that John bought __]]?

(4) Subject island
  (a) What do you think [$_{IP}$ __ interrupted the TV show]?
  (b) *What do you think [$_{IP}$ [$_{NP}$ the speech about __] interrupted the TV show]?

(5) Adjunct island
  (a) What do you think [$_{CP}$ that John forgot __ at the office]?
  (b) *What do you worry [$_{CP}$ if John forgets __ at the office]?

Island effects are undeniably central to linguistic theory, as they have generated dozens of theoretical articles over the years, and are featured in nearly every syntactic textbook. Island effects are also typical of complex syntactic violations, in that naïve participants cannot easily identify the "error" (Crain & Fodor, 1987). Crucially, several researchers have claimed that there is significant variation in the acceptability of island effects, suggesting that the prevailing syntactic analyses are incorrect (e.g., Kuno, 1973; Grimshaw, 1986; Deane, 1991; Hofmeister & Sag, 2010; but cf. Sprouse, Wagers, & Phillips, *submitted*). Moreover, the four versions of the island effects chosen here are generally considered to be the most variable, and therefore the most likely to demonstrate significant variability (cf. the less variable versions: WH-islands, Relative Clause islands, Sentential Subject islands, and *Because*-islands).

In addition to estimating the difference detection rate for known effects, we would also like to quantify the detection rate for known experimental confounds. We have chosen two versions of a confound that several authors have warned may contribute to potential false positives in informal experiments: the length of the sentences (e.g., Schütze, 1996; Cowart, 1997).

(6) Sentence Length: adding a PP (cf. the Subject island in (4))
  (a) Who thinks [$_{NP}$ the speech] interrupted the TV show?
  (b) Who thinks [$_{NP}$ the speech [$_{PP}$ about corruption]] interrupted the TV show?

(7) Sentence Length: adding an NP (cf. the Complex NP island in (3))
  (a) Who claimed [$_{CP}$ that John bought a car]?
  (b) Who made [$_{NP}$ the claim [$_{CP}$ that John bought a car]]?

To be clear, the detection rate for these two conditions is not itself a false positive, as we do expect there to be a weak effect of sentence length; however, their detection rate does provide us with an estimate of how likely it is that the sentence-length confound is contributing to any (false) positive results reported by *informal experiments that did not explicitly control for length*. It is important to note that informal experiments *do* routinely control for any number of possible confounds – in this regard, there is no difference between the materials construction phase of informal experiments and formal experiments – and, therefore, there is no reason to believe that informal experiments are more susceptible to confounds than formal experiments. We have merely included these analyses for informational purposes, and to place the reliability rate of the island effects in context.

2.4 Experimental details

176 self-reported monolingual native speakers of English (152 Female), all University of California Irvine undergraduates, participated in a Magnitude Estimation task (Stevens, 1957; Bard et al., 1996; Keller, 2000; Featherston, 2005a; Sprouse & Cunningham, *under review*) for either course credit or $5. Three participants were removed from analysis because they inverted the response scale in the acceptability task. All analyses below were run on the remaining 173 participants.

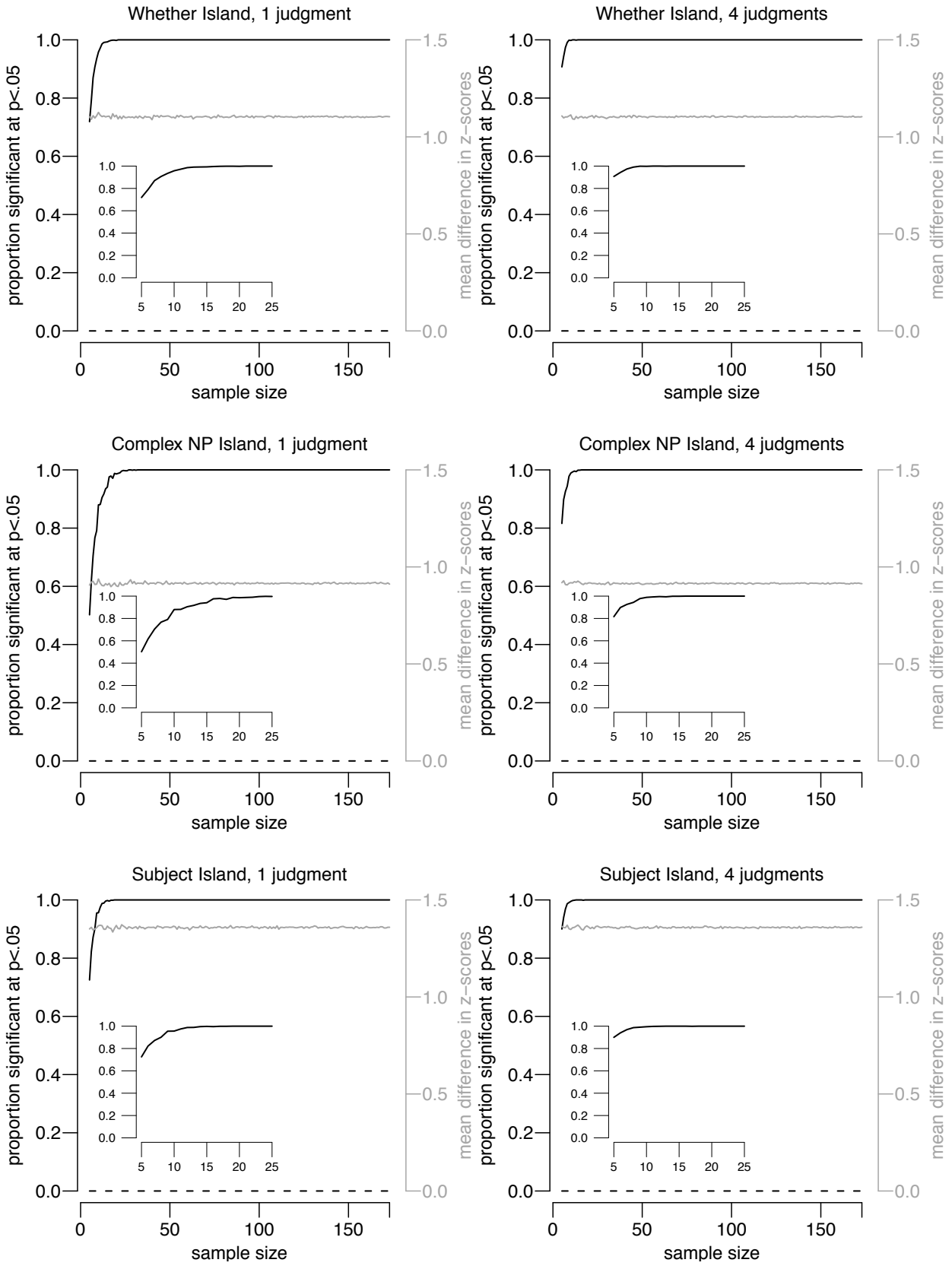
There were 24 sentence types (conditions) tested in this experiment. 16 lexicalizations of each sentence type were created, then distributed among 4 lists using a Latin Square procedure. This meant that each list consisted of 4 tokens per sentence type, for a total of 96 items. Two orders for each of the 4 lists were created by pseudorandomizing the items such that related sentence types were never presented successively. This resulted in 8 different surveys. The standard (i.e., the reference sentence) was identical for all 8 surveys, and was in the middle range of acceptability: *Who said my brother was kept tabs on by the FBI?* The standard was assigned a modulus of 100. The acceptability rating task was presented as a paper survey. The experiment began with a practice phase during which participants estimated the lengths of 7 lines using another line as a standard set to a modulus of 100. This practice phase ensured that participants understood the concept of magnitude estimation. During the main phase of the experiment, 10 items were presented per page (except for the final page), with the standard appearing at the top of every page inside a textbox with black borders. The first 9 items of the survey were practice items (3 each of low, medium, and high acceptability). These practice items were not marked as such, i.e., the participants did not know they were practice items, and they did not vary between participants in order or lexicalization. Including the practice items, each survey was 105 items long. The task directions are available on the corresponding author's website. Participants were under no time constraints during their visit.

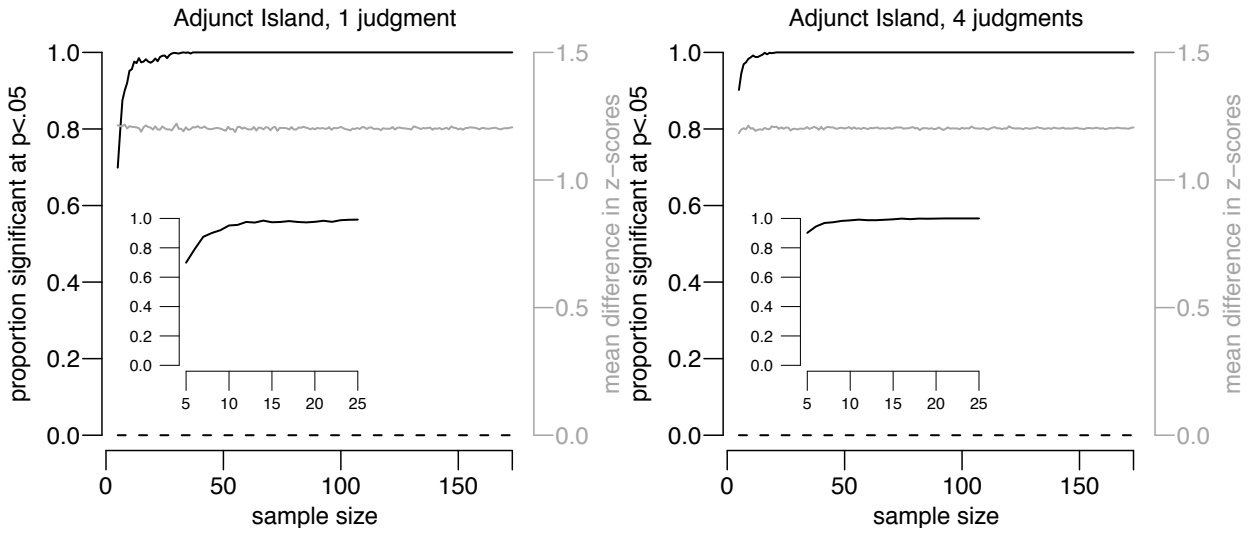
2.5 Results and Discussion

The results of the resampling analysis of the island effects are presented in Figure 2.

Figure 2: Resampling simulation results for island effects

The left panel presents the results of a resampling simulation for samples from 5 to 173 participants with only one judgment per participant. The right panel presents the results for four judgments per participant. The x-axis is sample size, and the left-side y-axis is proportion of results at $p < .05$*. The solid black line is the difference detection rate. The dashed black line is the sign-reversal rate. The solid grey line reports the difference between condition means in z-units using the right-side y-axis. The inset provides a zoomed in view of the difference detection rate for sample sizes 5-25.*


Whether Island, 1 judgment
Whether Island, 4 judgments
Complex NP Island, 1 judgment
Complex NP Island, 4 judgments
Subject Island, 1 judgment
Subject Island, 4 judgments
proportion significant at p<.05
mean difference in z−scores
sample size
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
5
10
15
20
25



The detection rate for island effects (solid black line)

With 1 judgment per participant and a sample size of 5, the detection rate for Whether, Subject, and Adjunct islands are between 70%-75% (note that 5 was the smallest sample size that we tested, even though that sample size is already smaller than the recommended minimum for the *t*-test). The detection rate for Complex NP islands was noticeably lower at this sample size (about 50%), corroborating the intuition in the linguistics literature that Complex NP islands are a relatively weak effect (Huang, 1982; Chomsky, 1986). If the number of judgments is increased to 4, the detection rate rises to over 90% for Whether, Subject, and Adjunct islands, and over 80% for Complex NP islands. These detection rates suggest that the standard informal experiments run by linguists (a sample of 5 or so colleagues, with several judgments per colleague) are easily powerful enough to detect the relevant effects.

The sign-reversal rate for island effects (dashed black line)

Whereas the detection rate tells us how likely we are to get the predicted result with small samples, the sign-reversal rate tells us how likely we are to be led astray. Sign-reversals would be absolutely devastating to theory construction, as they would lead linguists to construct theories for effects that are in fact in the opposite direction of behavioral reality. It is thus striking to note that the sign-reversal rate for every island type, every sample size, and every number of judgments per participant is 0%. There appears to be absolutely no risk of sign reversals for these effects at any sample size.

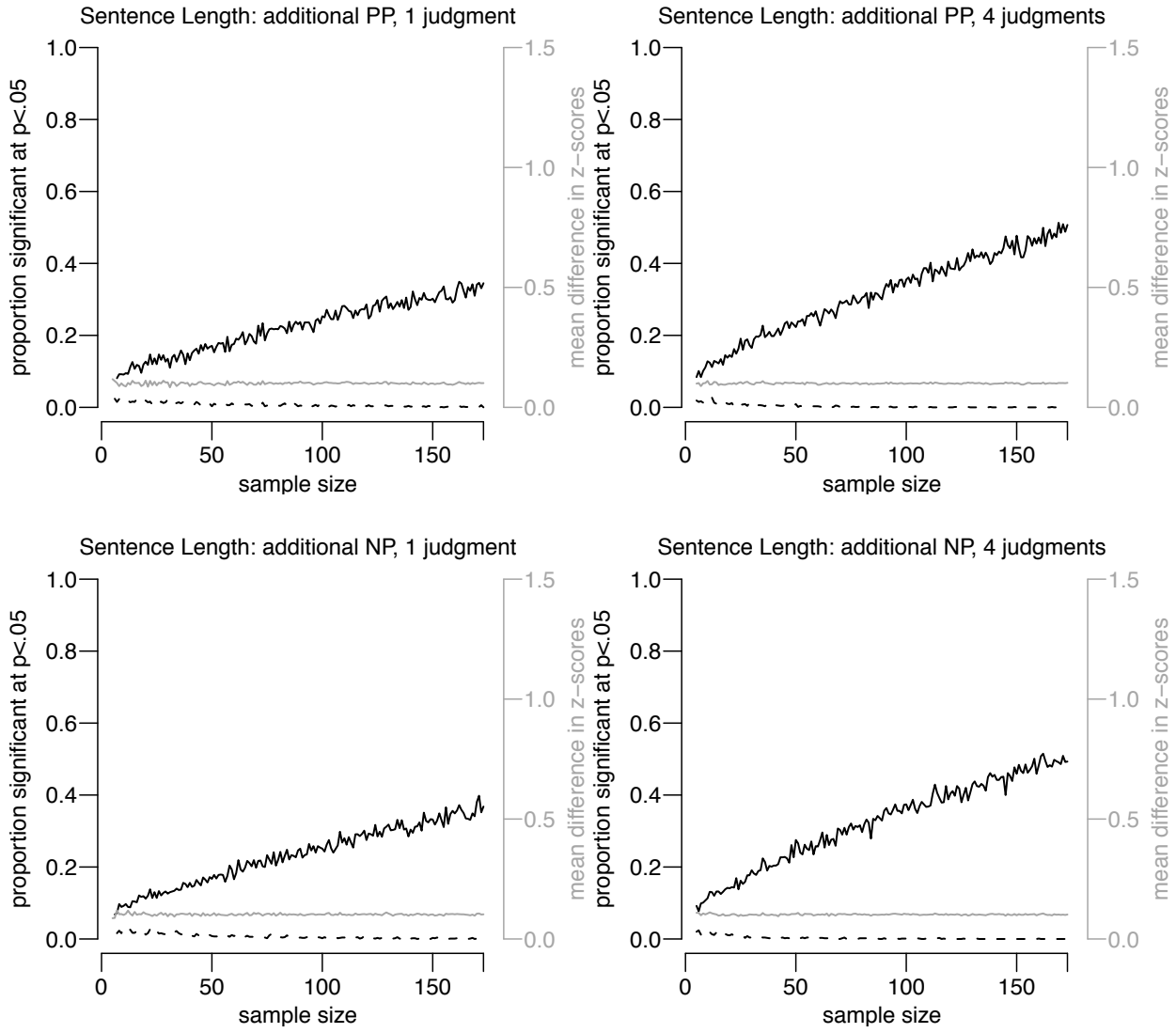
The difference between means for island effects (solid grey line)

We have included the average difference between means of each condition for each sample size (solid grey line, right y-axis) as an estimate of the effect size reliability between small samples (i.e., informal experiments) and large samples (i.e., formal experiments). The difference between means does not vary with the sample size for any of the island types. This suggests that the mean difference estimated by small samples is nearly identical to the mean

difference estimated by large samples, and that any extra statistical power gained by increasing the sample size is due to reducing the error associated with the mean estimate, and crucially is not due to an increase in the size of the effect.

Figure 3: Resampling simulation results for sentence-length confounds

The left panel presents the results of a resampling simulation for samples from 5 to 173 participants with only one judgment per participant. The right panel presents the results for four judgments per participant. The x-axis is sample size, and the left-side y-axis is proportion of results at $p < .05$. *The solid black line is the difference detection rate. The dashed black line is the sign-reversal rate. The solid grey line reports the difference between condition means in z-units using the right-side y-axis.*



The detection rate for sentence-length confounds (solid black line)

Recall that the detection rate for sentence-length confounds allows us to estimate the likelihood of a confound inducing a false positive when there is in fact no meaningful difference between two conditions. Also recall that there is no reason to believe that informal experiments will be more susceptible to such confounds than formal experiments, as both methodologies routinely control for a number of potential confounds. That being said, it is instructive to estimate the likelihood of these confounds impacting small and large sample sizes. At a sample size of 5, there is less than a 10% chance of detecting the effect of either type of sentence length, even with 4 judgments per condition. This suggests that sentence length is not likely to be influencing the results of informal experiments in any meaningful way. In contrast, with a sample size of 173, there is a 50% chance of detecting these sentence-length effects. This highlights the double-edged sword of statistical tests: increasing the sample size increases the detection rate for both experimental manipulations and confounds. This underscores the fact that statistical significance in a formal experiment should not be used as a proxy for theoretical relevance (given a large enough sample, statistically significant results are guaranteed to be found).

The sign-reversal rate for sentence-length confounds (dashed black line)

The sign-reversal rate for the sentence-length confounds is vanishingly small: it is less than 1% for sample sizes of 5. It is again reassuring to note that acceptability judgments are unlikely to yield sign-reversals (type 3 errors) at any sample size and for any contrast. In this respect, they appear to be a robust behavioral response.

Long-run relative frequency

The long-run relative frequency analysis answers the question: How likely is it that a single subject will report the hypothesized effect after a single judgment of each condition? This question will sound odd to many linguists, as informal experiments always entail multiple sentences of each condition and multiple participants in each sample. However, these results are instructive in two ways: (i) they provide a method to overcome the sample size requirements of the *t*-tests that were run as part of the resampling simulations, and (ii) they provide a response to some of the critics who insist on caricaturing informal experiments. For example, Gibson and Fedorenko (2010) claim that "the prevalent method in these fields involves evaluating a single sentence/meaning pair, typically an acceptability judgment performed by just the author of the paper, possibly supplemented by an informal poll of colleagues" (p. 233). This is simply untrue; however, if it were true, the long-run relative frequency analysis would estimate the reliability of such a methodology. As table 1 reports, this methodology would be fairly reliable: whereas a single judgment of each condition is overwhelmingly likely to go in the predicted direction for the island effects, the directionality of the sentence-length confounds is nearly a toss-up amongst the predicted direction, the unpredicted direction, and a null difference between conditions.

Table 1: Long-run relative frequency for individual judgments of island effects and sentence-length confounds

	Whether	Complex NP	Subject	Adjunct	Length-PP	Length-NP
expected direction	77%	72%	83%	79%	36%	33%
equal rating	18%	18%	11%	15%	39%	41%
opposite direction	5%	10%	6%	6%	26%	27%

Taken together, the results of the two types of simulations suggest that judgments of island effects, which have been described as highly variable in the literature, are in fact extremely robust at small sample sizes like those used in informal experiments. Furthermore, confounds such as sentence length are extremely unlikely to cause false positives at small sample sizes, and thus are unlikely to influence the results of informal experiments. It should be noted that these are exactly the findings that were casually reported by Featherston (2009) and Phillips (2009), for dozens (if not hundreds) of other linguistic phenomena. Given these results and the dozens reported by Featherston and Phillips, it seems prudent at this point to revisit the evidence that has been offered by critics of informal experiments in the literature. We turn to this in section 3.

3. The (non-)evidence driving the criticisms

In section 2 we went straight to the empirical heart of this debate, presenting quantitative evidence that acceptability judgments of linguistic phenomena are strikingly robust at small sample sizes – a result that is consistent with the claims of most theoretical linguists and the claims of Featherston (2009) and Phillips (2009) that they have corroborated dozens, if not hundreds, of informal results using formal experiments. In this section we would like to take a closer look at the purported evidence that informal experiments are unreliable. In section 3.1 we will argue that two such cases actually corroborate the results of informal experiments but were misinterpreted by the critics. In section 3.2 we will argue that two additional cases were based upon underpowered formal experiments, and that more powerful experiments corroborate the informal results (which also suggests that, in some cases, informal experiments are potentially more powerful than formal experiments). In section 3.3 we will argue that the one systematic investigation of cognitive bias among linguists actually found that linguists are biased against their own interests, as all scientists should be. In the end, we conclude that there is only one case of an informal result not holding up under formal experimentation, and that case was actually identified by linguists as controversial long before it was adopted by critics as evidence against linguistic methodology.

3.1 Variability is normal in all behavioral measures

One often-cited piece of evidence for the non-robustness of informal experimental results comes from Langendoen, Kalish-Landon, and Dore (1973), and has been re-presented in detail in Wasow and Arnold (2005) and Gibson and Fedorenko (2010, *submitted*). Langendoen et al.

investigated the claim made by Fillmore (1965) and others that the first object of a double-object construction cannot be questioned:

(8) *Who did you buy a hat?
(cf. What did you buy Mary?)

Langendoen et al. performed an answer completion task to test this claim formally. They asked 109 students to answer questions like (9) with a complete sentence:

(9) Who did you show the woman?

Their hypothesis was that if questions like (8) are indeed unacceptable, then the answers to (9) should *consistently* place the answer NP at the end of the sentence (*I showed the woman my daughter)*. Langendoen et al. reported two things: that one-fifth of the participants responded with the NP in the first object position *(I showed my daughter the woman)* and these participants were all from the metropolitan New York City area. Langendoen et al., Wasow and Arnold, and Gibson and Fedorenko all conclude that (8) is therefore not unacceptable[1]. Moreover, Wasow and Arnold and Gibson and Fedorenko conclude that informal experiments led Fillmore astray.

The problem with this case study is that the results of the formal experiment in fact *strongly support Fillmore's claim*. To see this, we only need to run a simple statistical test such as the sign test. According to Langendoen et al. (1973), one-fifth of their sample responded contrary to Fillmore's claim (~22 participants), and four-fifths (~87 participants) responded in accordance with it. A one-tailed sign test yields $p = .0000000018$ – a highly significant result. In other words, the formal experiment reported by Langendoen et al. strongly corroborates Fillmore's informal experiment; however, Wasow and Arnold (2005), and Gibson and Fedorenko (2010) all incorrectly interpret this experiment as disconfirming Fillmore's results. This is an important point: the simple act of running a formal experiment does not ensure the "truth" of the results. The results of both formal and informal experiments must be interpreted by the researcher, and as this case demonstrates, there is room for misinterpretation even in formal experiments.

We believe that Wasow and Arnold (2005), and Gibson and Fedorenko (2010, *submitted*) were misled by an assumption that appears to be widespread in formal experimental language research: for an effect to be real, then all (or nearly all) of the participants must show the effect in their individual responses. This is simply not the standard for any behavioral research, language-related or otherwise, because variability is expected in behavioral responses. Raaijmakers (2003) addresses this (often unstated) assumption in language research directly:

> The idea seems to be that a treatment effect should be present for each and every one of the subjects and items, otherwise it is not a real effect … However, it is not difficult to show that such an idea is incorrect. First, it should be noted that if one tests for a treatment effect, one tests whether the population means for the two conditions are different. It should be clear that even if only 25% of the items (or subjects for that matter)

[1] Langendoen et al. (1973) interpret the fact that a subset of the sample finds the hypothesized ungrammatical sentence acceptable as evidence that it is not a syntactic phenomenon at all, and is in fact driven by extra-syntactic factors.

> show the effect, the population means will still be different. For example, if 25% of the items have an effect of 40 ms and the remainder none, the difference between the population means for the two conditions will be 10 ms, and any valid test should give a significant result given enough power. … It is surprising (and somewhat disturbing) that such ideas appear to be relatively common, at least within the language community. (pp. 146-147)

We don't believe that any psycholinguist would impose such an unrealistic standard on reaction time, eye-tracking, or ERP data, so it is particularly interesting that linguists, or at least critics, appear to be imposing such a standard on acceptability judgment data. It is possible that this is a tacit acknowledgment of how reliable acceptability judgment data really is – linguists may be so accustomed to nearly 100% agreement among participants that an agreement rate that would be considered overwhelming in any reaction time study (80%) is considered problematic for an acceptability judgment study. Variability is normal in behavioral studies, and the variability found in linguistic studies is actually relatively small by psycholinguistic standards.

It should also be noted that this type of misinterpretation of variability is not isolated to the Langendoen et al. (1973) experiment. Wasow and Arnold (2005) also present their own formal replication of an informal acceptability judgment experiment and commit the same problematic interpretation, and once again this mistake is re-presented by Gibson and Fedorenko (2010, *submitted*). Wasow and Arnold (2005) set out to test a claim from Chomsky (1957/1975) that the complexity of a noun phrase strongly determines the position of that noun phrase within a verb-particle construction. Chomsky's claim is twofold. First, he claims that the most natural place for multi-word NPs is after the particle, therefore both (a) and (b) below should be more acceptable than both (c) and (d). Second, he claims that complex NPs (relative clauses) are less acceptable than simple NPs when they occur between the verb and particle, therefore (d) should be less acceptable than (c).

(10)	a.	The children took in all our instructions.	[3.4 out of 4]
	b.	The children took in everything we said.	[3.3 out of 4]
	c.	The children took all our instructions in.	[2.8 out of 4]
	d.	The children took everything we said in.	[1.8 out of 4]

Wasow and Arnold (2005) ran a formal experiment on 22 Stanford undergraduates to test these claims. The mean ratings of each condition are given in square brackets above. As these ratings indicate, the results match Chomsky's predictions perfectly, and the statistical tests report a significant interaction between particle position and NP type ($p < .001$). However, Wasow and Arnold (2005) interpret the results as problematic using the following logic:

> There was considerable variation in the responses. In particular, although the mean score for [condition d] was far lower than any of the others, 17% of the responses to such sentences were scores of 3 or 4. That is, about one-sixth of the time, participants judged such examples to be no worse than awkward. (p. 1491)

In other words, Wasow and Arnold (2005) assume that every response must show the effect for an effect to be real, and completely ignore the results of the statistical test that they ran. In the interest of moving this debate forward on sound empirical ground, we would like to suggest that

critics stop citing Langendoen et al. (1973) and Wasow and Arnold (2005) as evidence of the unreliability of informal experiments, as the formal experiments in these articles in fact strongly corroborate the validity of the informal experimental results.

3.2 Null results are not necessarily evidence that informal experiments are unreliable

One of the most common arguments against the use of informal experiments (e.g., Gibson & Fedorenko, 2010, *submitted*) can be schematized as follows:

1. Linguist A reports a difference between sentence-type X and sentence-type Y using an informal experiment.
2. Critic B reports a formal experiment that finds no difference between X and Y.
3. Critic B concludes that informal experiments are unreliable.

This is not a sound logical argument. The only way to reach the conclusion that informal experiments are unreliable is *if you already assume* that informal experiments are unreliable. The circularity is easy to see if we imagine that both of the experiments were formal:

1. Researcher A (using a formal experiment) reports a difference between X and Y.
2. Researcher B (using a formal experiment) reports no difference between X and Y.
3. Researcher B concludes that Researcher A is wrong.

No one would make the conclusion in 3 based on these facts alone. The null result that researcher B found could have arisen for any number of reasons that have nothing to do with the "reality" of the effect: the experiment could be underpowered (e.g., the sample is too small), the materials could contain a confound, the task could be inappropriate, etc. Absence of evidence is not evidence of absence; simple null results are never in themselves conclusive, and usually require several corroboratory experiments that systematically investigate their possible causes before they are deemed acceptable for publication. The fact that critic B in the first example is willing to interpret the null result as "true" suggests that she already assumes that informal experiments are unreliable. Furthermore, the fact that null results of this form have been published in the literature that criticizes informal linguistic experiments without the rigor normally required of null results suggests that this assumption may be more widespread than commonly assumed.

Null results due to lack of power

The logical flaw inherent in interpreting null results in favor of the formal experiment can be nicely demonstrated with an example from Gibson and Thomas (1999). Gibson and Thomas set out to investigate an informal judgment by Frazier (1985), attributed to Janet Fodor, that doubly center-embedded sentences can be made more acceptable by deleting the second VP:

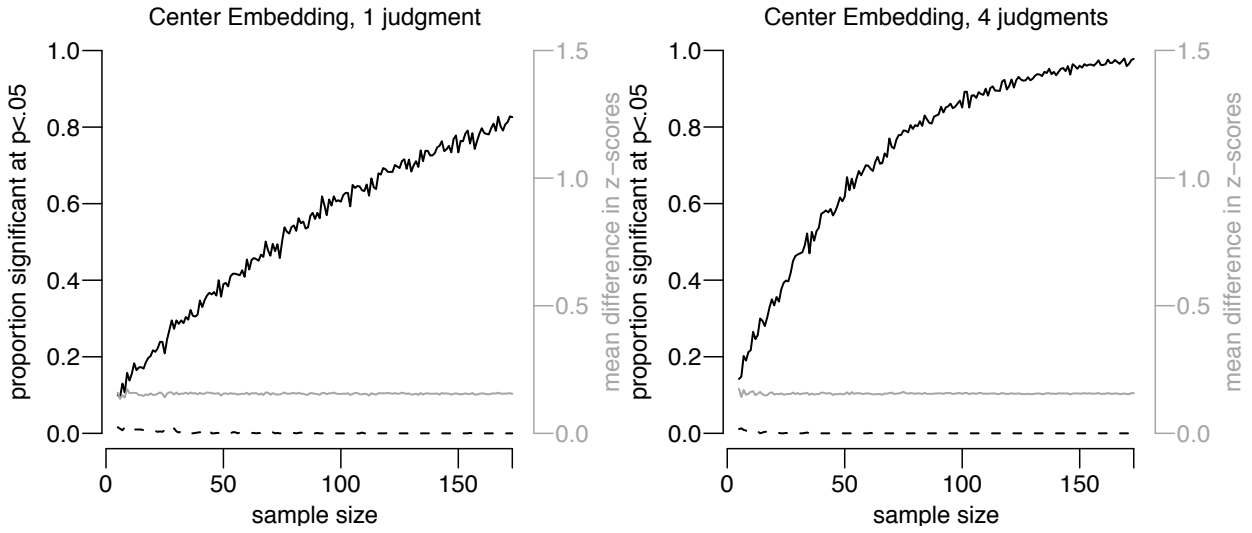
(11) a. *The ancient manuscript that the graduate student who the new card catalog had confused a great deal was studying in the library was missing a page.

b. ?The ancient manuscript that the graduate student who the new card catalog had confused a great deal was missing a page.

This phenomenon is particularly interesting given that (11a) is generally considered grammatical, but rated unacceptable due to parsing complexity (Chomsky & Miller 1963), whereas (11b) is generally considered ungrammatical, but may be perceived as more acceptable than (11a) for other parsing reasons. Gibson and Thomas ran a formal complexity rating experiment (not the same as but similar to an acceptability judgment experiment) on 40 MIT undergraduates, and found no statistically significant difference in the ratings between these two conditions (2.90 versus 2.97, out of 5). This is exactly parallel to the schema above: Fodor and Frazier report informal results that suggest a difference, whereas Gibson and Thomas report formal results that suggest no difference. According to the logic of some skeptics (e.g., Gibson & Fedorenko, 2010, *submitted*), the Gibson and Thomas formal results should be accepted as closer to the "truth" and the results from informal experiments should be discounted as spurious. However, there are clear indications that Gibson and Thomas' failure to replicate the informal experiments of Frazier and Fodor is due to lack of statistical power of the formal experiment, rather then the spuriousness of the results of the informal experiments. Sprouse and Cunningham (*under review*, p. 23) demonstrate in a series of 5 replicated experiments that the contrast between (11a) and (11b) can be detected with a sample half the size of the one used by Gibson and Thomas when one uses a magnitude estimation task with lower acceptability reference sentences, but not at all when higher acceptability reference sentences are used. This suggests that the rating task on a 5 point scale used by Gibson and Thomas might lack the sensitivity required to detect the effect with a sample size of 40. We tested this prediction by including the two conditions in the large-scale experiment reported in section 2, and the results corroborate the idea that the Gibson and Thomas experiment simply did not have enough statistical power to detect the effect.

Figure 4: Resampling simulation results for the center embedding effect

The left panel presents the results of a resampling simulation for samples from 5 to 173 participants with only one judgment per participant. The right panel presents the results for four judgments per participant. The x-axis is sample size, and the left-side y-axis is proportion of results at $p < .05$*. The solid black line is the difference detection rate. The dashed black line is the sign-reversal rate. The solid grey line reports the difference between condition means in z-units using the right-side y-axis.*



Our resampling simulations suggest that, with a sample of 40 participants, there is between a 35% and 57% chance of detecting the effect in a magnitude estimation task with a moderately acceptable reference sentence, depending on the number of judgments made by each participant. On the other hand, the resampling simulations unequivocally demonstrate that there is a reliable effect (notice the indication of an asymptote at 100% for 173 participants in the 4-token simulation).

It is important to stress that, despite the small size of the contrast between (11a) and (11b), the informal experiments of Fodor and Frazier (1985) were powerful enough to detect the difference in the predicted direction. This suggests that there may be very real power differences between the formal and informal methodologies, and that the power advantage, in some cases, rests with the informal experiments. This difference in ability to detect a hypothesized contrast may derive from the different populations that are used in each experiment: informal experiments generally rely on other linguists or linguistics graduate students for their subjects, while formal experiments generally rely on undergraduates for their subjects (a fact discussed in detail in section 3.3).

Null results due to lack of sensitivity of the experimental task

Another example of the problem of interpreting a null result in favor of the formal experiment comes from Gibson and Fedorenko (2010, *submitted*). They report that Gibson (1991) claimed that (12a) is less acceptable than (12b) based on an informal experiment:

(12) a. *The man that the woman that the dog bit likes eats fish.
b. ?I saw the man that the woman that the dog bit likes.

Crucially, Gibson and Fedorenko report that they have not been able to replicate these results using a formal experiment. Based on this failure to replicate the informal judgments in a formal experiment, they conclude that the results from the informal experiment were heavily and unduly influenced by the theoretical bias of the author (see also section 3.3):

> Without quantitative data from naïve participants, cognitive biases affect *all* researchers. (p. 233)

Once again, the assumption here is that the formal result is the "truth". However, it is also possible that the 5-point scale that they used is simply not sensitive enough to detect the small contrast in (12). To test this hypothesis, we asked 98 participants to choose which of the two sentences in (12) sounded more natural in a binary forced-choice task, counterbalancing for the order of presentation. 87 out of the 98 participants chose 12b (p = .0000000000000017 by sign test). Although further research is in order, these pilot results suggest that certain task components of their experiments may have obscured the clear preference participants have for 12b over 12a, such as the relatively coarse 5-point scale.

It should also be noted that this example once again suggests that informal experiments that use researchers as participants may have a real power advantage over formal experiments that use college students as participants (see also section 3.3).

A null result that appears to be reliable

The previous examples in section 3.2 demonstrate the need for caution in interpreting null results as conclusive, as both of those examples turned out to be the result of flaws in the experiments. The reader may, therefore, be wondering if a null result can ever be interpreted conclusively. As mentioned at the beginning of section 3, this is possible, but it requires several experiments that attempt to eliminate any experimental confounds that could be causing the null result. We have been able to find only one such example in the critical literature, and it is based on a contrast reported by Kayne (1983):

(13)	a.	*I'd like to know where who hid it.	= Kayne's (34a)
	b.	?I'd like to know where who hid what.	= Kayne's (35a)

Kayne reports that (13b) is more acceptable than (13a). Clifton, Fanselow, and Frazier (2006) tested these results in two full formal experiments with 48 participants each, as well as two follow-up experiments (with 78 and 118 participants respectively). The reason that they ran so many experiments is that they could find no difference between 13a and 13b in their first experiment, and then tried to systematically eliminate any possible confounds that could have caused the null result. Fedorenko and Gibson (in press) followed up on these results with an additional experiment (28 participants) that placed the target sentences within supportive contexts (stories), and still found no effect. Taken together, the 4 experiments from Clifton et al. and the one from Fedorenko and Gibson suggest that there really is no difference between (13a) and (13b), contrary to the claim in Kayne (1983).

The question then is whether the spurious contrast reported by Kayne (1983) is a problem for linguistic methodology. We believe that the crucial point missed by the critics that cite this example is that Clifton et al. (2006) did not choose this contrast at random – there are over 70 judgments in Kayne (1983) alone, and thousands more in the articles that have been written between 1983 and 2006 that Clifton et al. (2006) did not choose to test. Clifton et al. chose this contrast because it is known among linguists to be controversial. Fedorenko and Gibson (in press) chose this contrast because Clifton et al. had already reported a null result. In other words, linguists had already policed their own data, as all scientists are expected to do. We do not

believe any linguist would hesitate to admit that this particular contrast is likely spurious; however, we do find it disconcerting that critics have used linguists' own policing efforts as evidence that linguistic methodology is suspect.

3.3 Differences between linguists and non-linguists

The final type of argument that recurs in the critical literature involves a direct comparison between participants with advanced training in linguistic theory and naïve participants with no training in linguistics (e.g., Spencer, 1973; Bradac, Martin, Elliott, & Tardy, 1980; Dabrowska, 2010). The idea behind these studies is that because informal experiments generally rely on linguists as participants, a difference between these two groups will demonstrate that traditional linguistic methodology is unreliable. However, the logic behind these studies is eerily similar to the logic used for null results:

1. Critic A runs a formal experiment on both linguists and naïve participants that tests a contrast that has been reported in the linguistics literature (an informal experimental result).
2. The experiment demonstrates that there is a difference between the two groups.
3. Critic A concludes that linguists make bad participants, and therefore informal experimental results are unreliable.

At this point it should be clear that the critic's conclusion only follows if the critic already assumes that naïve participants are more reliable than linguists. The experiment itself merely demonstrates that there is a difference between the two groups. For example, Gibson and Fedorenko (2010) argue that the null result of the Gibson (1991) contrast (12a/12b) is an example of cognitive bias; however, as we've seen, it is actually an example of a task-sensitivity confound. Gibson and Fedorenko also argue that the Kayne (1983) contrast is an example of cognitive bias; however, as we have seen, linguists had no trouble identifying this contrast as controversial, so the cognitive bias is maximally associated with the original author. These examples demonstrate that a difference between groups is compatible with any number of interpretations, each of which requires careful, systematic investigation.

In fact, the one systematic investigation of potential cognitive bias on the part of linguists appears to us to go in the opposite direction than Edelman and Christiansen (2003), Ferreira (2005), Wasow and Arnold (2005), Gibson and Fedorenko (2010), and Dabrowska (2010) would predict. Dabrowska (2010) compared generative linguists and functional linguists in a rating study that included Complex NP islands (3b). One plausible prediction of the cognitive bias hypothesis is that generative linguists would rate the islands lower than the functional linguists because island constraints are a core part of the generative theory of syntax, but have been argued by many functionalists to be an epiphenomenon of language use (e.g., Kuno, 1973; Deane, 1991; Kluender & Kutas, 1993; Goldberg, 2007). In other words, generative linguists have a motivation to confirm the reliability of Complex NP islands, whereas functional linguists have a motivation to disconfirm the reliability of Complex NP islands. Dabrowska (2010) actually found that generative linguists' ratings were higher than the functional linguists' ratings. In short, the one systematic study of cognitive bias actually found that generative linguists were biased against their own theoretical interests.

In contrast to the critics that claim that linguists are too biased to be reliable participants, some linguists have gone so far as to argue that linguists provide cleaner judgments, free of many of the nuisance variables that affect naïve subjects such as plausibility, lexical frequency, and parsing difficulty (Newmeyer, 1983). Although we have seen no quantitative studies of this particular claim, at least two of the examples examined in section 3.2 suggest that linguists and psycholinguists can reliably detect small contrasts that are difficult to detect in samples of naïve participants. In other words, there is currently no evidence to suggest that linguists are biased participants, and at least two pieces of evidence to suggest that linguists are more sensitive than naïve participants. More research is necessary to see if this striking pattern holds for other phenomena.

4. The necessity of statistics

In the previous sections, we have argued that there is absolutely no empirical evidence of a problem with linguistic methodology: the results of informal linguistic experiments are largely corroborated by formal experiments (Featherston, 2009; Phillips, 2009), the contrasts that are of interest to linguists are strikingly reliable with very small sample sizes (section 2), and 5 out of 6 cases reported by critics actually corroborate the reliability of informal experiments. In this section we would like to address a criticism of a different sort – the claim by some critics that informal experiments are inferior to formal experiments because informal experiments are not analyzed using statistical significance tests:

> The number of subjects should be large enough to allow testing the results for statistical significance. (Wasow & Arnold, 2005, p. 1483)

> Experimental evaluations of syntactic and semantic hypotheses should be conducted with participants who are naïve to the hypotheses and samples large enough to make use of inferential statistics. (Gibson & Fedorenko, 2010, p. 233)

We can find no explicit reason offered by critics for why statistical significance testing should be used in linguistic methodology. Therefore, in this section we attempt to reconstruct possible arguments for the use of statistical significance testing and evaluate whether any of the possible arguments are empirically justified.

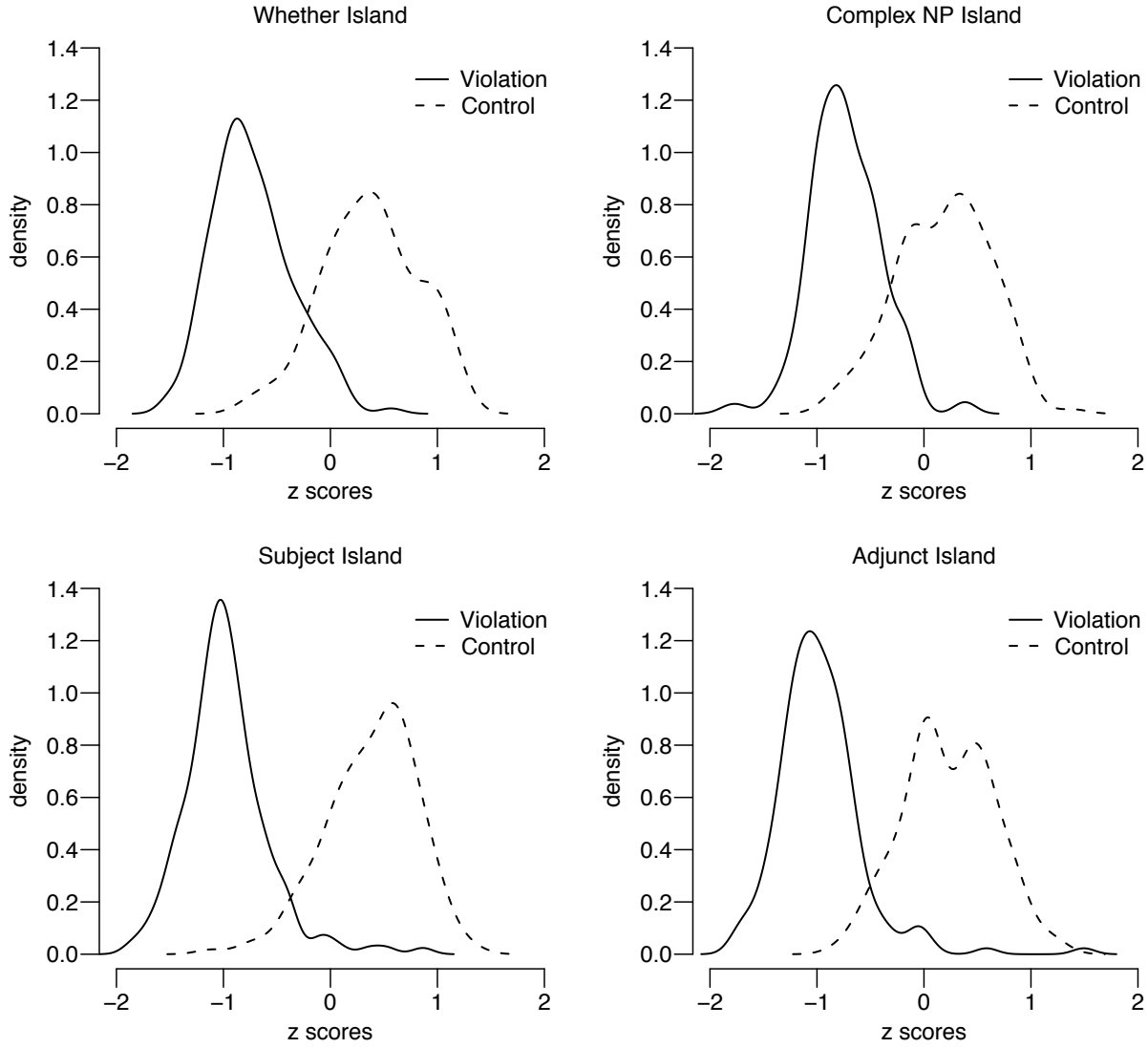
4.1 Differences between conditions

One possible reason to use statistical significance testing (henceforth SST) is to help identify *potentially* meaningful differences between conditions. In situations where the difference between two conditions is relatively small, SST can provide a criterion for separating potentially meaningful differences from differences that may have occurred by chance. If the critics' argument is that SST is necessary in linguistics to help identify potentially meaningful differences, then the critics must be assuming that the differences between conditions in linguistics are generally too small to be identified without the help of SST. And given that linguists do not believe that SST is necessary, the linguists must be assuming that the differences between conditions are generally large enough to be identified without SST.

It should be clear at this point that it is an empirical question whether contrasts in linguistics are large enough to be identified without the help of SST. In figure 5, we present the distribution of judgments for the two conditions of each island effect tested in the experiment previously discussed in section 2.

Figure 5: The distributions of judgments for island effects

The x-axis is z-score transformed rating. The y-axis is density. Density curves for the distribution of each condition were computed using the `density` *function in the base* `{stats}` *package in R. The solid line represents the violation condition. The dashed line represents the control condition.*



The separation between the two conditions for each of the islands tested is strikingly clear. There is no evidence in our case studies to suggest that linguistic effects are small enough to *require* SST, nor can we find any evidence presented in the critical literature. This is not to say that all acceptability effects are large enough to be detected without SST: we will present several

acceptability effects in section 5 that appear to require SST for detection. However, the effects reported in section 5 are phenomena from the sentence-processing literature, not linguistics.

4.2 Generalization beyond the sample

Another possible reason for critics to recommend SST for linguistic methodology could be because they believe that SST ensures that the results of an experiment will *generalize* to future experiments (that use different participants and different sentences):

> Multiple instances of the relevant construction are needed to evaluate whether an observed effect generalizes across different sets of lexical items. (Gibson & Fedorenko, 2010, p. 233)

All behavioral research is conducted on a subset of the population of interest, also called a sample. In the case of most linguistic research, the population is every *possible* native speaker of a given language, and the sample is the set of speakers that participate in the experiment. Clark (1973) argued that the items used in a language experiment (i.e., the sentences that are tested) are also just a sample of a population, in this case a subset of every *possible* sentence of that type. When a researcher reports the results of an experiment conducted on a specific sample of participants and a specific sample of sentences, they are suggesting to their colleagues that they believe that their results will *generalize* to other samples of participants and sentences.

It is true that if certain assumptions are met, statistical tests can provide a measure of generalizability for experimental results. To take a classic example, opinion polls often provide very narrow estimates of a population parameter of interest based on their poll results (i.e., the margins of error). However, psychological research never meets the critical assumption that all samples are randomly selected from the population of interest (random sampling); therefore, SST can never ensure generalizability for psychological research (Hahn & Meeker, 1993; Hunter & May, 2003; Gigenrenzer, 2004; Edgington & Onghena, 2007; Killeen, 2007; see Cohen, 1994; Nickerson, 2000; Balluerka, Goméz, & Hidalgo, 2005; Hubbard & Lindsay, 2008; Levine, 2008; Fidler & Loftus, 2009 for reviews on this and other common misconceptions about SST). No linguist, psycholinguist, or psychologist randomly samples her subjects. The subjects in any given study form a sample of convenience: college undergraduates who volunteered for the study after seeing it advertised on campus. Moreover, it is not even clear that it is possible to randomly sample subjects or items for a linguistics study – that would require first generating a list of all possible speakers of a language and a list of all possible sentences of a given type. In short, statistical tests cannot ensure generalizability in behavioral research because the random sampling assumption is never (and perhaps could never be) met, for either subjects or items. Therefore, given all of these considerations, generalizability cannot be an argument for the routine adoption of SST in linguistics.

4.3 Replication provides true generalizability

It is important to note that despite the almost complete absence of random sampling in psychological research, results from individual studies *are* generalizable (otherwise there would be little point in doing research). The generalizability is simply not supported by statistical arguments, but rather by non-statistical arguments. One argument comes from scientific

knowledge: researchers in any given field have knowledge of prior results and intuitions about how representative their sample is, how strong the effect one is investigating is, and how well the results fit with previous data and theoretical predictions. Wike and Church (1976) pointed this out in a response to Clark (1973):

> When an experimenter generalizes he utilizes not only the statistical results of his single experiment, but also the cumulative knowledge of his field and his intuition. Generality is not obtained simply by selecting p levels randomly. (Wike & Church, 1976, p. 253)

A second argument for generalizability comes from replication: each time a result is replicated, the field gains additional evidence that the result generalizes beyond the initial sample (Lykken, 1968; Wike & Church, 1976; Pollard & Richardson, 1987; Carver, 1993; Hunter & May, 1993, 2003; Hubbard & Ryan, 2000; Falk & Greenbaum ,1995; Falk, 1998a). This very basic fact about the scientific enterprise receives little attention in the literature, compared to the omnipresence of discussions of SST, as noted by Hubbard and Ryan:

> On one hand, the practice of SST, condemned by Meehl (1978) as "one of the worst things that ever happened in the history of psychology" (p. 817), is nevertheless omnipresent. On the other hand, the practice of replication, which "is almost universally accepted as the most important criterion of genuine scientific knowledge" (Rosenthal & Rosnow, 1984, p. 9), continues to receive more lip service than application. (p. 664

Replication is without a doubt the gold standard of accumulation and expansion of scientific knowledge:

> Generality is attained by a variety of techniques of replication [...] Whatever the method, it is replication that leads to the generality that Clark seeks. Generality is not achieved by using one or another statistical procedure in analyzing the results of a single experiment, it arises from the scientific community's carrying out its relentless and converging mission. (Wike & Church, 1976, p. 254)

> To the extent that we wish to generalize our results to some population we rely on replication, not on the results of a single study regardless of how the subjects are selected. (Hunter & May, 1993, p. 388)

> In practice, attempts to generalize research results are unlikely to rest on *anything* that happens in a single study, including the type of sampling carried out or the type of statistical test used. Instead, generalizability is frequently and perhaps better established via replication …, design (generalizability studies), and logical arguments based on nonstatistical judgments about subjects and variables. (Hunter & May, 2003, p. 180)

> Replications may serve as a means of establishing the reliability and the robustness of research results. Instead of measuring the quality of research by the level of significance, it would be better judged by its consistency of results in repeated experiments or in collaborative studies which are simultaneous replications. If a researcher does obtain the same result (or a result in the same direction) more than once, it strengthens the

> conclusion that the results are not due to chance … even if one lacks the means to compute exact posterior probabilities. (Falk & Greenbaum, 1995, pp. 92-93)
>
> Allen and Preiss (1993) noted that scientific knowledge comes from replication. The results from an unreplicated study, no matter how statistically significant they may be, are necessarily speculative in nature (Hubbard & Armstrong, 1994) and meaningless in themselves (Lindsay & Ehrenberg, 1993). (Hubbard & Ryan, 2000, p. 676)
>
> We appeal to inductive logic to move from the particular results in hand to a theoretically useful generalization. As I have noted, we have a body of statistical techniques, that, used intelligently, can facilitate our efforts. But given the problems of statistical induction, we must finally rely, as have the older sciences, on replication. (Cohen, 1994, p. 1002)
>
> Ideally, all experiments would be replicated before publication (…)." (Lykken, 1968, p. 159)

So it seems that the "highest methodological standard" allowing the researcher to generalize from her sample data is not the use of SST, as some critics have suggested, but rather consistent and systematic replication. If anything, SST seems to serve as a poor substitute for replication when replication is too costly (see Cohen, 1994; Nickerson, 2000; Balluerka et al., 2005; Levine, 2008; Fidler & Loftus, 2009 for reviews). Incidentally, all the resampling simulation results presented in this paper were replicated with a different sample of 176 naïve participants (see Sprouse, *submitted*).

The upshot of these considerations is the following: Acceptability judgment data is perhaps the *easiest* to replicate of all behavioral responses. Any native speaker can check a reported judgment on themselves in a matter of seconds, and on a colleague, friend, or passerby in a matter of minutes. Among linguists it is also assumed that any reader can also create variant lexicalizations to test the generalizability to other items. Not only is replication of informal linguistic judgments easy, but many of the core linguistic facts *have been replicated* thousands of times by colleagues, reviewers, editors, conference audiences, students, and readers. As Hubbard and Ryan (2000) put it:

> By attempting to essentially reproduce the findings of an earlier study, replications play a vital role in safeguarding the empirical literature from contamination by specious results. (p. 676)

The same cannot be said for the data from reaction time, ERP, MEG, fMRI or even formal linguistic experiments. We can find no empirical or logical reason in the statistical literature for linguists to be forced to adopt the routine use of inferential statistical tests when (i) the effects are clear in the data itself (see section 2; see also Bolles, 1988; Falk & Greenbaum, 1995; Haller & Krauss, 2002; Nock, Michel, & Photos, 2007), and (ii) the replication of the informal experiment is easy and cheap to conduct.

4.4 Science requires statistics

The last potential argument that we can imagine is not a statistical one, but rather a normative one: scientific inquiry requires statistics. We find it unlikely that any serious scientist would endorse this type of normativism, as it has long been known that SST is not a magic recipe for uncovering scientifically meaningful results:

> The finding of statistical significance is perhaps the least important attribute of a good experiment; it is *never* a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence - or that an experimental report ought to be published. The value of any research can be determined, not from statistical results, but only by skilled, subjective evaluation of the coherence and reasonableness of the theory, the degree of experimental control employed, the sophistication of the measuring techniques, the scientific or practical importance of the phenomena studied and so on. (Lykken, 1968, pp. 158-159)

In fact, there are literally dozens of articles (spanning almost 70 years) in the statistics literature decrying the apparent normativism surrounding the use of SST in psychology (e.g, Berkson, 1942; Jones, 1955; Kish, 1959; Eysenck, 1960; McNemar, 1960; Rozeboom, 1960; Grant, 1962; Bakan, 1966; Lykken, 1968; Signorelli, 1974; Cronbach, 1975; Guttman, 1977; Carver, 1978; Guttman, 1985; Brewer, 1985; Falk, 1986; Oakes, 1986; Nelson, Rosenthal, & Rosnow, 1986; Gigenrenzer & Murray, 1987; Rosnow & Rosenthal, 1989; Schmidt, 1992, 1996; Loftus, 1991, 1993, 1996; Carver, 1993; Shaver, 1993, Cohen, 1990, 1994; Gigerenzer, 1994; Gonzalez, 1994; Falk & Greenbaum, 1995; Hunter, 1997; Falk, 1998b; Zaknanis, 1998; Daniel, 1998; Johnson, 1999; Krueger, 2001; Hubbard & Lindsay, 2008; see Nickerson, 2000; Balluerka et al., 2005; Levine, 2008; Fidler & Loftus, 2009 for reviews).

In short, we can find no empirical, logical, or normative reason for the routine adoption of SST in linguistic methodology. To be completely clear, we do believe that SST is useful with small or controversial effects, but there is no evidence that linguistic phenomena routinely yield small or controversial effects.

5. Some final thoughts about this debate

At this point, the skeptical reader may be wondering how this debate could recur so often if there is no evidence to suggest that informal judgments are unreliable. This is a difficult question, and it is likely that there is no single answer. However, one empirical trend emerged during the investigation of these issues that may suggest a possible reason: the phenomena of interest to critics appear to be significantly less reliable than the phenomena of interest to linguists. It is well known that the majority of the critics of linguistic methodology would not self-identify as linguists, but rather as psychologists or psycholinguists. This distinction, although counterproductive in many ways, is usually a reliable indicator of differences in the phenomena of interest to the researchers: psycholinguists are generally interested in properties of the parser and the temporal dynamics of language processing, linguists are generally interested in properties of grammatical knowledge and the static properties of linguistic representations. In our data set, we have found that acceptability judgments of parsing phenomena are much less

reliable than the judgments of grammatical phenomena. This raises the possibility that critics may be interpreting the unreliability of judgments of parsing phenomena with small samples as indicative of all judgments. Indeed, if linguists were interested in parsing phenomena, we would absolutely agree with the critics that large samples and formal statistics are necessary for careful scientific work.

Obviously, testing this explanation is beyond the scope of this paper, but we do have a few relevant case studies in our dataset: the Center Embedding Illusion from section 3.2, the Comparative Illusion (e.g., Phillips, Lau, & Wagers, *in press*), and Agreement Attraction (e.g., Bock & Miller, 1991; Wagers, Lau, & Phillips, 2009).

(14) Center Embedding Illusion (Frazier, 1985; Gibson & Thomas, 1999)
  a. *The ancient manuscript that the graduate student who the new card catalog had confused a great deal was studying in the library was missing a page.
  b. ?The ancient manuscript that the graduate student who the new card catalog had confused a great deal was missing a page.

(15) Comparative Illusion (Phillips et al., *in press*)
  a. *More people have graduated law school than I have.
  b. ?More people have been to Russia than I have.

(16) Agreement Attraction (Bock & Miller, 1991; Wagers et al., 2009)
  a. *The slogan on the poster unsurprisingly were designed to get attention
  b. ?The slogan on the posters unsurprisingly were designed to get attention

Figure 6: Resampling simulation results and distributions of judgments for parsing phenomena

The left panel presents the results of a resampling simulation for samples from 5 to 173 participants with one judgment per participant. The x-axis is sample size, and the left-side y-axis is proportion of results at $p < .05$*. The solid black line is the difference detection rate. The dashed black line is the sign-reversal rate. The solid grey line reports the difference between condition means in z-units using the right-side y-axis. The right panel presents the distribution of ratings for each condition. The x-axis is z-score transformed rating. The y-axis is density. Density curves for the distribution of each condition were computed using the* `density` *function in the base* `{stats}` *package in R. The solid line represents the violation condition, which in the case of illusions is the control condition. The dashed line represents the illusion condition.*

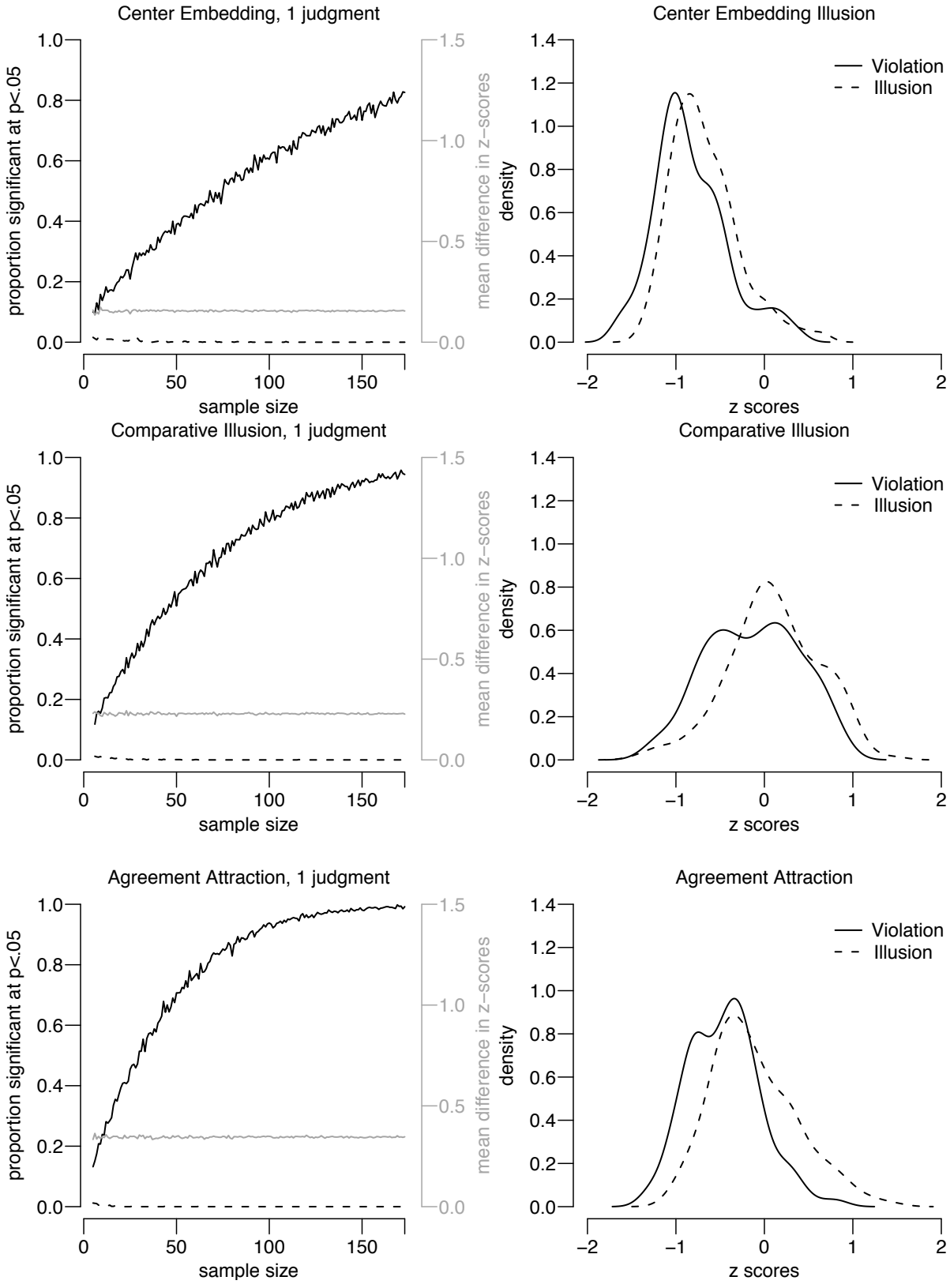

Center Embedding, 1 judgment
proportion significant at p<.05
mean difference in z−scores
sample size
Center Embedding Illusion
density
z scores
Violation
Illusion
Comparative Illusion, 1 judgment
Comparative Illusion
Agreement Attraction, 1 judgment
Agreement Attraction

Table 2: Long-run relative frequency for individual judgments of parsing phenomena and sentence-length confounds

	Center Embedding	Comparative Illusion	Agreement Attraction	Sentence Length-PP	Sentence Length-NP
expected direction	39%	40%	36%	36%	33%
equal rating	38%	39%	38%	39%	41%
opposite direction	23%	21%	26%	26%	27%

Crucially, the detection rate of the three parsing phenomena in very small sample sizes is very similar to the detection rate of the sentence-length confounds presented in figure 3, and the separation between the conditions is probably too small to be confidently established without SST. Furthermore, the likelihood of a single judgment going in the correct direction is not much more likely than the sentence-length confounds, and equally as likely as a judgment of no difference. If this pattern of results holds for judgments of other parsing phenomena, then that might begin to explain the reason critics are suspicious of informal experiments: at small sample sizes, judgments of parsing phenomena are not reliable, and critics may be confusing the properties of the phenomena with properties of the methodology.

The skeptical reader may also be wondering why it is that linguists don't simply acquiesce to the critics' requests and begin to run formal experiments. This is a sociological argument, but one worth considering given the potential value in a closer relationship between linguistic theory and other domains of psychology. Our response to this is multi-tiered. First of all, many critics are suggesting that all of the previous informal results are unreliable, with the implication that linguistic theory should be re-designed from the ground up based on new formal experiments. This would entail testing literally thousands of data points in hundreds of experiments. It is unfair to request this of a field without empirical justification. We have argued that there is simply no empirical or logical justification for this request at the present.

A more interesting question is whether linguists should begin to run formal experiments moving forward. Again, the answer boils down to empirical justification. Many linguists, the first author included, have written articles in an attempt to demonstrate the utility of formal experiments for advancing linguistic theory (e.g., Cowart, 1997; Keller, 2000; Sorace & Keller, 2004; Featherston, 2005a, 2005b; Myers, 2009; Sprouse, 2009; Sprouse, Fukuda, Ono, & Kluender, *in press*; Sprouse, Wagers, & Phillips, *submitted*). To be absolutely clear, we strongly believe that linguists should run formal experiments when the value is apparent – this could be because a particular data point is unclear, or because the hypothesis in question requires the comparison of acceptability with some other numeric quantity. However, formal experiments are more costly than informal experiments. In a field with significantly less research funding than other areas of psychology, and significantly more data points per article (50 or more is common), even $100 per experiment ($5 per participant, 20 participants) would be a burden on most researchers for (we argue) little or no empirical gain. Given this, linguists must be allowed to exercise their expert knowledge in deciding which hypotheses can benefit from formal

experiments, and which hypotheses can be quickly tested with (the strikingly reliable) informal experiments.

Conclusion

In this paper we have demonstrated that judgments of some core linguistic phenomena that have been claimed to be highly variable (island effects) are in fact strikingly robust at small sample sizes (section 2). We have argued that the number of formal results that have been reported to diverge from informal results is actually very small (perhaps only 1) compared to the number of results that converge (dozens, if not hundreds) (sections 2 and 3). We have also argued that there is no extra rigor gained by formal experiments simply because they use statistical tests, as informal experiments are held to the highest quantitative standard available – replication (section 4). Taken as a whole, we believe that these facts strongly suggest that there is no empirical, logical, or statistical evidence that informal judgments are unreliable, and in fact, significant evidence that informal judgments are strikingly reliable. As with all psychological theories, there are plenty of debates to be had regarding linguistic theory, but our results suggest that the reliability of the data is not one of them. There are likely many factors that have led to the current divide between linguistics and other domains of language research (see also Marantz, 2005; Phillips, 2009); however, our results suggest that methodology is not, or at least should not be, one of them.

Acknowledgments

This research was supported in part by National Science Foundation grant BCS-0843896 to Jon Sprouse.

References

Balluerka, N., Goméz, J., & Hidalgo, D. (2005). Null hypothesis significance testing revisited. *Methodology, 1*(2), 55-70.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423-437.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*, 32-68.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37,* 325-335.

Bock, K., & Miller, C. (1991). Broken agreement. *Cognitive Psychology, 23*(1), 45-93.

Bolles, R. C. (1988). Why you should avoid statistics. *Biological Psychiatry, 23*(1), 79-85.

Bradac, J. J., Martin L. W., Elliott, N. D., & Tardy, C. H. (1980). On the neglected side of linguistic science: multivariate studies of sentence judgment. *Linguistics, 18*, 967–995.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics, 10,* 252-268.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61,* 287-292.

Chomsky, N., & Miller, G.A. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*, vol. 2 (pp. 269-321). New York: Wiley.

Chomsky, N. (1955/1975). *The logical structure of linguistic theory*. Chicago: University of Chicago Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Clark, H. H. (1973). The language-as-a-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

Clifton, C., Jr., Fanselow, G., & Frazier, L. (2006). Amnestying superiority violations: Processing multiple questions. *Linguistic Inquiry, 37*, 51-68.

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

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

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

Crain, S., & Fodor, J. D. (1987). Sentence matching and overgeneration. *Cognition, 26*, 123–189.

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

Culicover, P. W., & Jackendoff, R. (2010). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences, 14*(6), 234-235

Dąbrowska, E. (2010). Naïve v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review, 27*(1), 1-23

Daniel, L. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(3), 23-32.

Deane, P. (1991). Limits to attention: A cognitive theory of island phenomena. *Cognitive Linguistics*, *2*, 1-63.

Edelman, S., & Christiansen, M. (2003). How seriously should we take Minimalist syntax? *Trends in Cognitive Sciences, 7*, 60-61.

Edgington, E., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman and Hall/CRC.

Eysenck, H. J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review, 67,* 269-271.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9,* 83-96.

Falk, R. (1998a). Replication—A step in the right direction. *Theory and Psychology, 8,* 313-321.

Falk, R. (1998b). In criticism of the null hypothesis statistical test. *American Psychologist, 53,* 798-799.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5,* 75-98.

Featherston, S. (2005a). Magnitude estimation and what it can do for your syntax: some wh-constraints in German. *Lingua, 115*(11), 1525-1550.

Featherston, S. (2005b). Universals and grammaticality: wh-constraints in German and English. *Linguistics*, *43*, 667-711.

Featherston, S. (2009). Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft, 28*(1): 127-132.

Fedorenko, E. & Gibson, E. (in press.) Adding a third wh-phrase does not increase the acceptability of object-initial multiple-wh-questions. *Syntax*.

Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review, 22*, 365-380.

Fidler, F., & Loftus, G. (2009). Why figures with error bars should replace *p* values. *Journal of Psychology, 217*(1), 27-37.

Fillmore, C. (1965). *Indirect Object Constructions in English and the Ordering of Transformations*. The Hague, Netherlands: Mouton.

Frazier, L. (1985). Syntactic complexity. In D. Dowty, L. Karttunen, & A. Zwicky (Eds), *Natural language processing: Psychological, computational and theoretical perspectives* (pp. 129-189). Cambridge, England: Cambridge University Press.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown.* Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Gibson, E. & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences, 14*(6), 233-234.

Gibson, E. & Fedorenko, E. (submitted). The need for quantitative methods in syntax.

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, *14*, 225-248.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33,* 587-606.

Goldberg, A. (2007). *Constructions at work*. Oxford, England: Oxford University Press.

Gonzalez, R. (1994). The statistics ritual in psychological research. *Psychological Science, 5,* 321, 325-328.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69,* 54-61.

Grimshaw, J. (1986). Subjacency and the S/S' Parameter. *Linguistic Inquiry*, *17*, 364-369.

Guttman, L. (1977). What is not what in statistics. *The Statistician, 26,* 81-107.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1,* 3-10.

Hahn, G. J., & Meeker, W. Q. (1993). Assumptions for statistical inference. *The American Statistician, 47*(1), 1-11.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychology Research Online, 7*(1).

Hill, A. A. (1961). Grammaticality. *Word, 17*, 1-10.

Hofmeister, P. & Sag, I. (2010). Cognitive constraints and island effects. *Language*, *86*, 366-415.

Huang, C.-T. (1982). Move wh in a language without wh-movement. *The Linguistic Review*, *1*, 369-416.

Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology 18*(1), 69-88.

Hubbard, R., & Ryan, P. (2000). The historical growth of statistical significance testing in psychology - and its future prospects. *Educational and Psychological Measurement, 60*(5), 661-681.

Hunter, M. A., & May, R. B. (1993). Some myths concerning parametric and nonparametric tests: *Canadian Psychology/Psychologie Canadienne, 34*(4), 384-389

Hunter, M. A., & May, R. B. (2003). Statistical testing and null distributions: What to do when samples are not random. *Canadian Journal of Experimental Psychology, 57*(3), 176-188.

Hunter, J. E. (1997). Need: A ban on the significance test. *Psychological Science, 8,* 3-7.

Jones, L. V. (1955). Statistics and research design. *Annual Review of Psychology, 6,* 405-430.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *The Journal of Wildlife Management, 63*(3), 763-772.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review, 67,* 160-167.

Kayne, R. (1983). Connectedness. *Linguistic Inquiry, 14*, 223-249.

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Doctoral dissertation, University of Edinburgh.

Killeen, P. R. (2007). Replication statistics. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 103–124). Thousand Oaks, CA: Sage.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review, 24,* 328-338.

Kluender, R., & Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and Cognitive Processes, 8*, 573-633.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56*(1), 16-26.

Kuno, S. (1973). Constraints on internal clauses and sentential subjects. *Linguistic Inquiry*, *4*, 363-85.

Langendoen, D.T., Kalish-Landon, N., & Dore, J. (1973). Dative questions: a study in the relation of acceptability to grammaticality of an English sentence type. *Cognition, 2*, 451-477.

Levine, T. R., Weber, R., Hullet, C., Park, H. S., & Lindsey, L. M. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research, 34*, 171-187

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36,* 102-105.

Loftus, G. R. (1993). Editorial comment. *Memory & Cognition, 21,* 1-3.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science,* 5, 161-171.

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

Marantz, A. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review, 22,* 429-445.

McNemar, Q. (1960). At random: Sense and nonsense. *American Psychologist, 15,* 295-300.

Myers, J. (2009). Syntactic judgment experiments. *Language and Linguistics Compass, 3*, 406-423.

Nelson, N., Rosenthal, R., & Rosnow, R. (1986). Interpretation of significance levels and effect sizes by psychological research. *American Psychologist, 41,* 1299-1301.

Newmeyer, F. J. (1983). *Grammatical theory: Its limits and its possibilities*. Chicago: University of Chicago Press.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301

Nock, M. K., Michel, B. D., & Photos, V. I. (2007). Single-case research designs. In D. McKay (Ed.), *Handbook of research methods in abnormal and clinical psychology* (pp. 337-350). Thousand Oaks, CA: Sage.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.

Phillips, C. (2009). Should we impeach armchair linguists? In S. Iwasaki, H. Hoji, P. Clancy, & S.-O. Sohn (Eds.), *Japanese/Korean Linguistics 17*. Stanford, CA: CSLI Publications.

Phillips, C., & Lasnik, H. (2003). Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Sciences, 7*, 61-62.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 10,* 159-163.

Raaijmakers, J. G. W. (2003). A further look at the 'language-as-fixed-effect fallacy.' *Canadian Journal of Experimental Psychology, 57*(3), 141-151.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276-1284.

Ross, J. R. (1967). Constraints on variables in syntax. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Schmidt, F. L. (1992). What do data really mean? *American Psychologist, 47,* 1173-1181.

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

Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61,* 293-316.

Signorelli, A. (1974). Statistics: Tool or master of the psychologist? *American Psychologist, 29,* 774-777.

Sprouse, J. (submitted). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory.

Sprouse, J. (2009). Revisiting Satiation: Evidence for an Equalization Response Strategy. *Linguistic Inquiry*, *40*, 329-341.

Sprouse, J., Fukuda, S., Ono, H. & Kluender, R. (in press). Grammatical operations, parsing processes, and the nature of wh-dependencies in English and Japanese. *Syntax*.

Sprouse, J. & Cunningham, H. (under review). Evaluating the assumptions of magnitude estimation of linguistic acceptability. *Language*.

Sprouse, J., Wagers, M., & Phillips, C. (submitted). A test of the relation between working memory capacity and island effects.

Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory and Psychology, 8*, 291-311.

Sorace, A., Keller, F. (2005). Gradience in linguistic data. *Lingua, 115*(11), 1497-1524.

Spencer, N. J. (1973). Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research, 2*(2), 83-98.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*, 153-181.

Wagers, M., Lau, E., & Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *Journal of Memory and Language*, *61*, 206-237.

Wasow, T. & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua, 115*, 1481-1496.

Wike, E., & Church, J. D. (1976). Comments on Clark's 'The language-as-fixed-effect fallacy.' *Journal of Verbal Learning and Verbal Behavior, 15*(3), 249-255

Zaknanis, K. (1998). Brain is related to behavior ($p < .05$). *Journal of Clinical and Experimental Neuropsychology, 20*(3), 419-427.