Power in acceptability judgment experiments and the reliability of data in syntax.

Jon Sprouse
Department of Cognitive Sciences
University of California, Irvine

Diogo Almeida
Department of Linguistics and Languages
Michigan State University

Abstract

There has been a consistent pattern of criticism of the reliability of acceptability judgment data in syntax for at least 50 years (e.g., Hill 1961), culminating in several high-profile criticisms within the past ten years (e.g., Edelman and Christiansen 2003, Ferreira 2005, Wasow and Arnold 2005, Featherston 2007, Gibson and Fedorenko 2010a, 2010b). One of the fundamental claims of these critics is that traditional acceptability judgment collection methods, lead to an intolerably high number of false negative results (i.e., low statistical power). We empirically assessed this claim by re-testing 95 phenomena that span the full range of effect sizes in syntactic data, estimated from two recent large scale acceptability surveys (Sprouse and Almeida, to appear, and Sprouse, Schutze, and Almeida, submitted) using two different tasks (magnitude estimation, which is commonly used in formal experiments, and forced-choice, which is commonly used in traditional methods), and using resampling simulations to empirically estimate false negative rates (statistical power) for each phenomenon at every sample size between 5 and 100 participants. Contrary to the claims of critics, these results suggest that traditional methods have a remarkably low false negative rate by experimental psychology standards, and in fact are more sensitive than formal experiments at detecting differences between conditions. We discuss the implications of these results for questions about the reliability of syntactic data, as well as the practical consequences of these results for the methodological options available to syntacticians.

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

1. Introduction

It is well-known that acceptability judgments form a substantial component of the empirical foundation of (generative) syntactic theories (Chomsky 1965, Schütze 1996). For example, in a recent survey of US-English data points published in *Linguistic Inquiry* from 2001 through 2010, Sprouse, Schütze, and Almeida (submitted) estimated that 77% were derived from some sort of acceptability judgment (the remaining 23% were based on some form of meaning/ambiguity judgment). It is also well-known that the vast majority of those acceptability judgments were collected relatively informally, that is without the formal collection protocols that are familiar from experimental psychology (e.g., Sprouse, Schütze, and Almeida found that fewer than 5% of the syntax-related articles published in LI 2001-2010 contained explicit discussion of formal experiments). The informality with which acceptability judgments are traditionally collected has led to a steady stream of methodological criticisms since the earliest days of generative syntax

(e.g., Hill 1961, Spencer 1973), culminating in a particularly dramatic increase in methodological discussions over the past 15 years, presumably due to the relative ease with which formal acceptability judgment can be constructed, deployed, and analyzed using freely available software and internet-based participant pools. (Bard et al. 1996, Keller 2000, 2003, Edelman and Christiansen 2003, Phillips and Lasnik 2003, Featherston 2005a, 2005b, 2007, 2008, 2009, Ferreira 2005, Sorace and Keller 2005, Wasow and Arnold 2005, den Dikken et al. 2007, Alexopoulou and Keller 2007, Fanselow 2007, Newmeyer 2007, Culbertson and Gross 2009, Myers 2009, Phillips 2009, Bader and Häussler 2010, Dąbrowska 2010, Gibson and Fedorenko 2010a, 2010b, Fedorenko and Gibson 2010, Culicover and Jackendoff 2010, Gross and Culberton 2011, Weskott and Fanselow 2011, Gibson et al. 2011, Sprouse 2007a, 2007b, 2008, 2009, 2011a, 2011b, Sprouse, Fukuda, Ono, and Kluender 2011, Sprouse, Wagers, and Phillips 2012, Sprouse and Almeida 2011, Sprouse and Almeida, to appear, Sprouse, Schütze, and Almeida, submitted).

One oft-repeated claim in this literature is that traditional methods are somehow *unreliable*, resulting in the construction of ill-supported syntactic theories (e.g., Edelman and Christiansen 2003, Ferreira 2005, Wasow and Arnold 2005, Gibson and Fedorenko 2010a, 2010b). Although the word *unreliable* has an intuitive meaning to most readers of these criticisms, formally there are (at least) two types of unreliability that are relevant to the evaluation of data collection methods. This is because there are two possible states of the world: (i) there is a difference between the relevant experimental conditions, i.e., an effect, or (ii) there is no difference between the conditions, i.e., no effect; and there are two possible results of the experiment: (i) the experiment reports a difference between the conditions, i.e., a positive result, or (ii) the experiment reports no difference between the conditions, i.e., a negative (or null) result.[1] This leads to four possible outcomes for any given experiment:

Table 1: Four possible outcomes for any given experiment

State of the world	Result of the experiment	Type of result	Outcome
Difference	Difference	True positive	Correct
No difference	No difference	True negative	Correct
No difference	Difference	False positive	Type I error
Difference	No difference	False negative	Type II error

Intuitively, a reliable experiment would be one that minimizes both false positives (Type I error) and false negatives (Type II errors). Conversely, an experiment would be unreliable if it produces an intolerably high rate of false positives, an intolerably high rate of false negatives, or both.

With this schematic of reliability, it is possible to classify criticisms of traditional acceptability judgment methods in syntax into two types. The first type claims that traditional

[1] There is a third type of unreliability: a positive result in the opposite direction. These results are sometimes known as sign-reversals or Type III errors. We will not have much to say about this type of unreliability given that it is exceedingly rare in acceptability judgment experiments (e.g., Sprouse and Almeida, to appear, and Sprouse, Schütze, and Almeida, submitted, found only two examples out of 511 phenomena tested).

methods lead to an intolerably high false positive rate, and relatedly, that formal experiments would lead to a lower false positive rate (e.g., Ferreira 2005, Wasow and Arnold 2005, Gibson and Fedorenko 2010b). The second type claims that traditional methods lead to an intolerably high rate of false negatives, and relatedly, that formal experiments would lead to a lower false negative rate (e.g., Bard et al. 1996, Keller 2000, Featherston 2007). Claims of the first type have been investigated in detail by Sprouse and Almeida (to appear), who formally tested all 469 data points in a popular syntax textbook (Adger 2003) and found that the maximum possible false positive rate (i.e., conservatively assuming that all negatives results in the formal experiments are true negatives rather than false negatives) for Adger (2003) is 2%. Similarly, Sprouse, Schütze, and Almeida (submitted) formally tested a random sample of 292 acceptability judgment data points from LI 2001-2010, allowing them to estimate a maximum false positive rate for LI 2001-2010 at 95%. Therefore in this paper we will focus on claims of the second type in an attempt to provide a (more) complete picture of the reliability of traditional methods relative to formal methods.

Because the vast majority of the data in syntactic theory has been collected using traditional methods (and crucially not also collected using formal methods), there are only a few methods available to us to address the question of relative false negative rates in traditional and formal methods. One possibility is to systematically re-test any negative results that have been reported using traditional methods using formal methods in order to identify potential false negatives. This procedure has been conducted on a few topics so far, with some previously undetected differences being reported (e.g., Keller 2000, Featherston 2005a, 2005b, Sprouse et al. 2011). The problem with these studies is that the new differences that have been reported have been relatively small, and the sample sizes in the formal experiments have been relatively large. Given that there is almost always a true small difference between two non-identical sentences, these results are ambiguous between the conclusion that traditional methods failed to detect the difference, and the conclusion that the original authors decided that the difference was too small to be theoretically relevant (see also Fanselow 2007 and Myers 2009). This then highlights a real limitation of all experimental methods that is often overlooked in the literature: with a large enough sample size, non-identical conditions will (almost) always manifest some sort of difference; the problem is that neither the experiment nor the statistical tests can tell us which differences are theoretically relevant, and which are theoretically irrelevant. It is up to the researcher to use her theoretical knowledge to make that determination (see also Cohen 1994).

Given the logical limitations inherent in a case-by-case re-evaluation of individual results (and the time it would take to re-test every claim in the literature), we will present a different method for assessing the reliability of syntactic methods with respect to false negatives. We re-tested 190 sentence types, forming 95 two-condition phenomena that span the full range of effect sizes observed in two large scale surveys of syntactic data (Sprouse and Almeida, to appear; Sprouse et al, submitted). Of these, 48 phenomena were taken from Adger's (2003) textbook *Core Syntax* as tested by Sprouse and Almeida (to appear), and 47 were taken from articles in *Linguistic Inquiry* (2001-2010) as tested by Sprouse, Schütze, and Almeida (submitted). We tested these 95 phenomena using two different tasks: the magnitude estimation task, which is commonly used in formal experimental tasks (e.g., Bard et al. 1996, Keller 2000, Featherston 2005a, 2005b, Alexopoulou and Keller 2007, Sprouse et al. 2011, Sprouse, Wagers, and Phillips 2012), and the forced-choice task which is commonly used in traditional methods (Bard et al. 1996, Schütze 1996, Marantz 2005, Sprouse and Almeida 2011). We then ran re-sampling simulations on the results to empirically derive estimates of statistical power for each

phenomenon at every sample size between 5 and 100 participants for each type of experiment (forced-choice and magnitude estimation). This type of quantitative comparison directly addresses two questions about the reliability of syntactic data:

1. Is it true, as some critics have claimed, that traditional methods (exemplified by the forced-choice experiments) are likely to lead to a high rate of false negatives?

2. Is it true, as some critics have claimed, that traditional methods (exemplified by the forced-choice experiments) lead to more false negatives than formal experiments (exemplified by the magnitude estimation experiments)?

The rest of this article is organized around these two questions. In section 2, we present a brief introduction to the concept of *false negative rate*, which is more commonly discussed as statistical power in the statistical literature. In section 3, we discuss the idea of effect sizes, and describe the method we used to choose the 95 phenomena that form the case studies for these simulations (taken from Adger 2003 via Sprouse and Almeida, to appear, and LI 2001-2010 via Sprouse, Schütze, and Almeida, submitted). Section 4 presents the details of the four experiments: (i) 48 phenomena from Adger (2003) using the forced-choice task, (ii) the same 48 phenomena from Adger (2003) using the magnitude estimation task, (iii) 47 phenomena from LI (2001-2010) using the forced-choice task, (ii) the same 47 phenomena from LI (2001-2010) using the magnitude estimation task. Section 5 describes the resampling simulations that we used to empirically estimate the false negative rates for all 95 phenomena for every possible sample size between 5 and 100 participants. Section 6 is a general discussion of the results of the four experiments and resampling simulations, with subsections devoted to the false negative rate of traditional methods, a comparison of the false negative rate between traditional methods and formal experiments, a comparison of the false negative rate of traditional methods to the false negative rate in other areas of psychology, and a general discussion of the costs and benefits for both traditional and formal methods. Section 7 concludes.

2. Statistical power

The *false negative rate* is more commonly called *statistical power* in the statistical literature. Statistical power is a recasting of the *false negative rate* as the likelihood of an experiment to detect a difference between conditions when one truly exists. Statistical power is normally expressed as a percentage. For example, if the false negative rate is 20%, then the experiment has an 80% probability of detecting a difference when a difference truly exists, and we say that the experiment has 80% statistical power. Mathematically, if the *false negative rate* of an experiment is β, then the statistical power of that experiment is 100% - β (or if one prefers proportions: 1-β).

The consequences of statistical power are relatively intuitive; however, it is important to note that statistical power itself is the result of the interaction of several distinct aspects of an experiment. Contributing factors include: the task (some tasks are more sensitive than others), the size of the sample of participants (larger samples yield more powerful experiments than smaller samples), the number of responses collected per participant per condition (more responses lead to higher power), the size of the difference (or *effect* size) that one wishes to detect (larger differences are easier to detect than smaller differences), and the false positive rate that one is willing to tolerate (if all other aspects of an experiment are held constant, fewer false

positives leads to more false negatives). In order to profitably compare the power of two different types of experiments, we must either systematically manipulate or carefully control each of these factors of the experiment. For the present study, we manipulated the task (forced-choice versus magnitude estimation), we manipulated the samples size by using resampling simulations (e.g., a single sample of 150 participants can be used to simulate samples from 0 to 100 participants), we held the number of responses per participant per condition constant at 1 (because this is the smallest possible number, our power estimates will be minimum estimates), we varied the size of the difference to be detected across the full spectrum of effect sizes observed in the literature (operationalized as the full span of effect sizes observed in the two large scale surveys of Sprouse and Almeida, to appear, and Sprouse, Schütze, and Almeida, submitted), and we held the maximum rate of false positives that we are willing to tolerate at the consensus level of 5% (familiar from experimental psychology) by setting the criterion for significance at $p<.05$.

Because power is expressed as a numerical value between 0 and 100%, the criterion at which an experiment may be considered "well-powered" may vary from field to field, or even from researcher to researcher. As a matter of convention, many fields of experimental psychology and the social sciences have adopted the suggestion by Cohen (1962, 1988, 1992) that the target power rate should be at or above 80%. This suggestion is based on the following logic: (i) most experimenters conventionally tolerate a false-positive (Type I error) rate of 5%, (ii) false positives are approximately 4 times more troublesome than false negatives (Type II errors), (iii) power is mathematically equal to 1-β, where β is the false negative rate (Type II error rate), therefore (iv) β should be set at 20%, and consequently (v) power should be set at 80% (Cohen, 1992). For the purposes of addressing concerns that traditional methods lead to an intolerably high false negative rate (i.e., low power), it seems appropriate to adopt the consensus criterion of 80% from experiment psychology, as it allows a direct comparison between traditional methods (as represented by forced-choice experiments) and the explicit "best practice" guidelines that behavioral experiments in experimental psychology strive to achieve. To the extent that traditional experiments (as represented by forced-choice experiments) reach 80% power under routine circumstances, the widespread assumption that traditional methods are "under-powered", coarser, or less sensitive than formal experiments will need to be revised.

Although the 80% power recommendation may be appropriate for comparing the results of traditional methods with other experiments in experimental psychology and the social sciences, it is important to note that it may not ultimately be the most appropriate criterion for syntactic theory. Recall that Cohen arrived at this figure by assuming that false positives (Type I errors) are 4 times more troublesome than false negatives (Type II errors). It is an open question whether syntacticians would share this assumption. For example, many syntactic theories seek to capture both the differences between conditions and the invariances between conditions. False negatives, which fail to indicate a difference when there is in fact one, may therefore have more of an effect on the theorizing of syntacticians than it would on other experimental psychologists, since this detection failure could be interpreted as indicating that no difference exists between the relevant conditions[2], a result that could have important theoretical implications. Arriving at a

[2] Standard null hypothesis significance testing is inappropriate for establishing invariances, therefore establishing theoretically relevant invariances would require more than a decrease in the tolerated false negative rate. One possibility is the adoption of Bayesian statistical tests,

consensus regarding the correct ratio of false positives to false negatives is a complex problem. On the one hand, it would be easy to simply assume that both types of errors are equally problematic and therefore should be equally minimized. As one concrete example, one could assume a 5% false positive rate ($p<.05$) and a 5% false negative rate (95% power). However, this would in a very real sense be holding syntax to a higher statistical standard than other fields of experimental psychology and the social sciences. As a field, syntacticians may agree that this is a warranted step in pursuit of the goals of syntactic theory, but it would substantially alter the nature of the methodological debates that have occurred to date.

3. Effect sizes and the choice of phenomena for the study

In order to ensure that the comparison between forced-choice and magnitude estimation experiments is robust, we chose a set of 95 phenomena that span a wide-range of phenomena (taken from both a textbook (Adger 2003) and journal articles (LI 2001-2010), and a wide-range of effect sizes. In order to choose the phenomena we calculated a measure of effect size known as Cohen's *d* (Cohen, 1988) for the 104 significant two-condition phenomena tested from Adger (2003) by Sprouse and Almeida (to appear) and for the 139 significant two-condition phenomena tested from LI (2001-2010) by Sprouse, Schütze, and Almeida (submitted). Cohen's *d* is calculated by subtracting the mean rating for each condition, and then dividing the difference between means by the mean of the standard deviation of each condition. In other words, Cohen's *d* is the ratio of the difference between means to the mean standard deviation. Therefore a Cohen's *d* of less than one indicates that the difference between means is smaller than the mean of the standard deviations, and a Cohen's *d* greater than one indicates that the difference between means is larger than the mean of the standard deviations.

By measuring effect size as a ratio of the difference between means to the mean of the standard deviations, Cohen's *d* allows us to compare any two effect sizes to each other, even if they are measured on different scales (e.g., reading times and acceptability judgments). In other words, Cohen's *d* is a standardized measure of effect size. One of the advantages of standardized measures of effect size is that researchers can specify rules of thumb for interpreting them. For example, Cohen (1988, 1992) suggested the following criteria for the intuitive interpretation of *d* values: a *d* greater than 0.2 and less than 0.5 is considered a "small" effect, a *d* greater than 0.5 but less than 0.8 is considered a "medium" effect, and a *d* greater than 0.8 is considered a "large" effect. Here is what Cohen (1992) said about the intent behind these criteria:

> Because the ES indices are not generally familiar, I have proposed as conventions, or operational definitions, "small", "medium," and "large" values of each ES index to provide the user with some sense of its scale. It was my intent that medium ES represent an effect of a size likely to be apparent to the naked eye of a careful observer, that small ES be noticeably smaller yet not trivial, and that large ES be the same distance above medium as small is below it. I also made an effort to make these conventions comparable across different statistical tests. (Cohen, 1992, p. 99)

---

which allow direct testing of invariances (for a review see Gallistel 2009, for an introduction to performing Bayesian statistical tests see Kruschke 2011).

To make the idea of effect sizes more tangible, here we present six example phenomena at a range of effect sizes for the reader to judge for herself (Cohen's *d* is in parentheses).

(3) I'd planned to have finished, and finished I have. (.41)
*I'd planned to have finished, and have finished I did.

(4) What Julie became was fond of the book. (.73)
*What Julie did of the book was become fond.

(5) I believed she might be pregnant. (1.05)
*I believed she may be pregnant.

(6) I worry if the lawyer forgets his briefcase at the office. (1.58)
*What do you worry if the lawyer forgets at the office?

(7) He has known him. (2.23)
*Him has he known.

(8) I shaved myself. (3.61)
*Myself shaved me.

With this in mind, we can look at the distribution of effect sizes for the 104 (two-condition) phenomena from Adger (2003) as tested by Sprouse and Almeida (to appear) and the 139 (two-condition) phenomena from LI 2001-2010 as tested by Sprouse, Schütze, and Almeida (submitted):

Figure 1: The distribution of effect sizes (Cohen's *d*) for the 104 significant, two-condition phenomena from Adger's *Core Syntax* (2003) as tested by Sprouse and Almeida (to appear) using the magnitude estimation task.

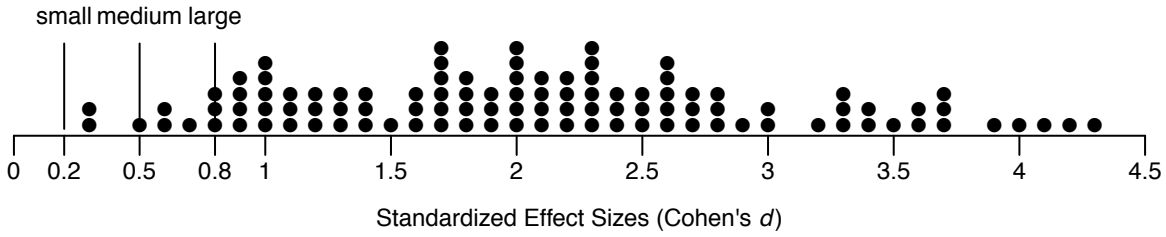
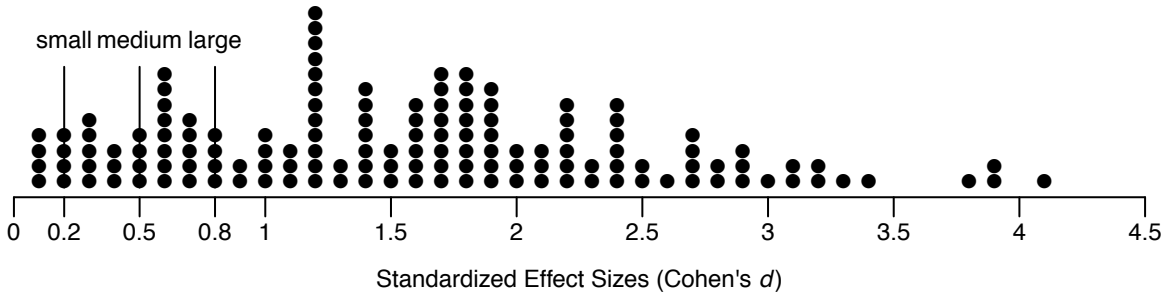


Figure 2: The distribution of effect sizes (Cohen's *d*) for the 139 significant, two-condition phenomena from *Linguistic Inquiry* (2001-2010) as tested by Sprouse, Schütze, and Almeida (submitted) using the magnitude estimation task.



Approximately 2% of the two-condition phenomena from Adger (2003) are considered "small" by Cohen's (1988, 1992) criteria, 4% are considered "medium", and 94% are considered "large". Approximately 12% of the phenomena from LI (2001-2010) are considered "small", 12% are considered "medium", and 76% are considered "large".

Because the distribution of effect sizes in Adger (2003) contained relatively few small and medium effects, we decided to choose all of the small and medium effects from Adger, and a sample of large effects that span the entire range of large effect sizes (0.8-4.4). This allows us to estimate the power of both forced-choice and magnitude estimation experiments at every sample size between 5 and 100 participants for a wide-range of effect sizes. And because the distribution of effect sizes for the sample from LI (2001-2010) contained substantially more small and medium effect sizes relative to the textbook data from Adger (2003), we chose the 47 phenomena with the smallest effect sizes. The increased number of smaller effects in the set of LI phenomena allows us to make more robust generalizations about the relative power of the two types of experiments for small and medium effect sizes (where potentially controversial data points are more likely to be found). The distribution of the effect sizes for the two sets of phenomena (48 from Adger (2003) and 47 from LI (2001-2010) as derived by the four experiments (see sections 4 and 5) are schematized in Figure 3 and Figure 4:

Figure 3: The distribution of effect sizes (Cohen's *d*) for the 47 phenomena from Adger's *Core Syntax* (2003) that we chose for the experiments here (effect sizes were derived from the results of Experiment 2).

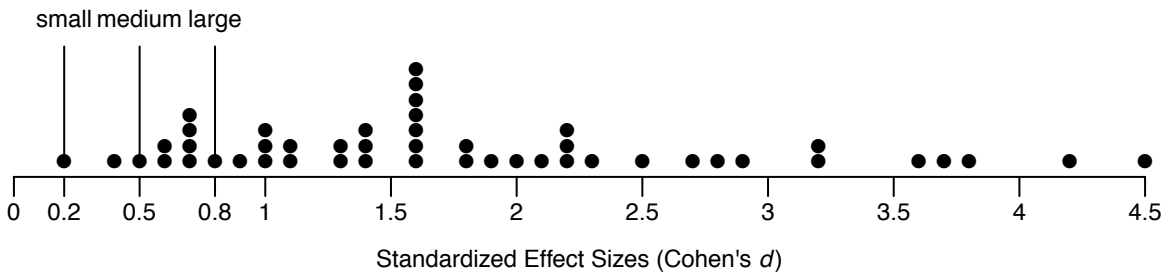
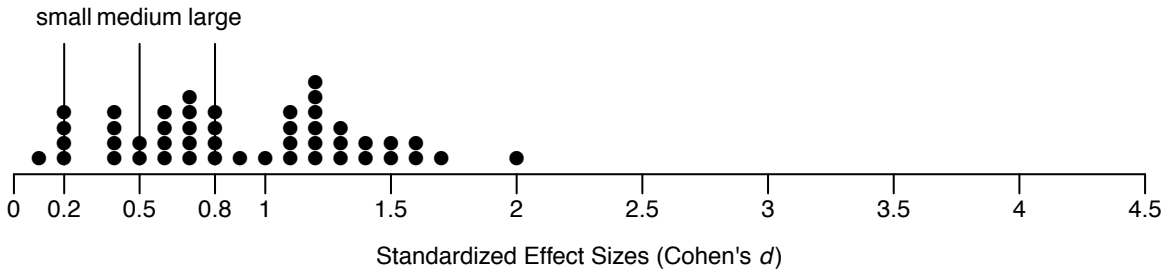


Figure 4: The distribution of effect sizes (Cohen's *d*) for the 47 phenomena from *Linguistic Inquiry* (2001-2010) that we chose for the experiments here (effect sizes were derived from the results of Experiment 4).



A full list of the phenomena chosen for our experiments are presented in Appendix A (48 phenomena from Adger 2003) and Appendix B (47 phenomena from Linguistic Inquiry 2001-2010).

4. The experiments

We tested the two sets of phenomena using both the forced-choice and magnitude estimation tasks in four experiments using four distinct samples of participants:

Table 2: The four experiments

Experiment	Phenomena	Task	Participants
Experiment 1	Adger's Core Syntax (2003)	Forced-Choice	152
Experiment 2	Adger's Core Syntax (2003)	Magnitude Estimation	152
Experiment 3	Linguistic Inquiry (2001-2010)	Forced-Choice	144
Experiment 4	Linguistic Inquiry (2001-2010)	Magnitude Estimation	144

In this section we provide additional details about each of the experiments.

4.1 Participants

152 participants completed Experiment 1, 152 participants different participants completed Experiment 2, 144 participants completed Experiment 3, and 144 different participants completed Experiment 4. Participants were recruited online using the Amazon Mechanical Turk (AMT) marketplace, and paid $3.00 for their participation in a magnitude estimation experiment (Experiments 2 and 4), and $2.00 for their participant in a forced-choice experiment. Participant selection criteria were enforced as follows. First, the AMT interface automatically restricted participation to AMT users with a US-based location. Second, we included two questions at the beginning of the experiment to assess language history: (1) Were you born and raised in the US? (2) Did both of your parents speak English to you at home? These questions were not used to

determine eligibility for payment so that there was no financial incentive to lie. 4 participants were excluded from Experiment 1 for answering 'no' to one of the language history questions. 12 participants were excluded from Experiment 2 for either answering 'no' to one of the language history questions or for obvious attempts to cheat (e.g., entering 1 in every response box). No participants were excluded from Experiments 3 and 4.

4.2 The tasks

In the magnitude estimation task (Stevens 1957, Bard et al. 1996), participants are presented with a reference sentence, called the *standard*, which is pre-assigned an acceptability rating, called the *modulus*. The standard is generally chosen such that it is in the middle range of acceptability. Participants are asked to use the standard to estimate the acceptability of the experimental items. For example, if the standard is assigned a modulus of 100, and the participant believes that an experimental item is twice as acceptable as the standard, the participant would rate the experimental item as 200. If a participant believes the experimental item is half as acceptable as the standard, she would rate the experimental item as 50. In this way, the standard can act as a sort of unit of measure for estimating the acceptability of the target sentences. However, it should be noted that this assumption has been challenged by recent research showing that the participants do not actually use the standard to make ratio judgments of the target sentences, which suggests that participants may simply treat the ME task as a scaling task similar to the Likert scales (Sprouse 2011b).

In the forced choice task, participants are asked to directly compare two (or more) sentence types (e.g., 2a and 2b) and decide which sentence is more acceptable.

(2) a. What do you think that John bought?
b. What do you wonder whether John bought?

The relevant sentence types are designed to be as structurally and lexically similar as possible, and should ideally differ in only one structural dimension, which should be the structural property of interest. For instance, in the case of (2), the structural property of interest is the type of CP that forms the embedded clause; however, the nature of this manipulation requires that the sentences differ lexically at the matrix verb and the first word of the embedded clause (*wonder whether* versus *think that*), as well as semantically with respect to the sentential force of the embedded clause (declarative versus interrogative). All other properties, both lexical and structural, are identical. In short, the comparison should be as close to a syntactic minimal pair as possible.

4.3 Materials for Experiments 1 and 2

The materials for Experiments 1 and 2 were identical to the materials constructed for the original Sprouse and Almeida (to appear) experiments: eight lexicalizations of each sentence type were constructed by varying (i) content words and (ii) function words that are not critical to the structural manipulation as described in the text of Adger (2003).

For the forced choice task (Experiment 1), the 8 lexicalizations were distributed among 8 lists by pairs, such that each pair of related lexicalizations appeared in the same list. Next, the order of presentation of each pair was counterbalanced across the lists, such that for every pair,

four of the lists included one order, and four lists included the other order. This minimized the effect of response biases on the results (e.g., a strategy of 'always choose the first item'). Finally, the order of the pairs in each list were randomized, resulting in eight surveys containing 48 randomized and counterbalanced pairs (96 total sentences).

For the magnitude estimation experiment (Experiment 2), the 8 lexicalizations were distributed among eight lists using a Latin Square procedure. Each list was pseudorandomized such that related conditions did not appear sequentially. This resulted in eight surveys of 96 pseudorandomized items. Nine additional "anchoring" items (three each of acceptable, unacceptable, and moderate acceptability) were placed as the first nine items of each survey. These items were identical, and presented in the identical order, for every survey. Participants rated these items just like the others; they were not marked as distinct from the rest of the survey in any way. However, these items were not included in the analysis as they served simply to expose each participant to a wide range of acceptability prior to rating the experimental items (a type of unannounced "practice"). This resulted in eight surveys that were 105 items long.

4.4 Materials for Experiments 3 and 4

The materials for Experiments 3 and 4 were identical to the materials constructed for the original Sprouse, Schütze, and Almeida (submitted) experiments: eight lexicalizations of each sentence type were constructed by varying (i) content words and (ii) function words that are not critical to the structural manipulation as described in the text of LI (2001-2010). We originally chose 50 phenomena from LI (2001-2010) for these experiments; however, after the experiments were conducted we identified problems with the materials for three of the phenomena, leaving 47 phenomena for the final analysis.

The distribution procedure for the forced-choice experiment (Experiment 3) was identical to the procedure used for Experiment 1, except that two versions of each list were created, for a total of 16 100-item surveys. The distribution procedure for the magnitude estimation experiment (Experiment 4) was identical to the procedure used for Experiment 2, except that six anchoring items were used, resulting in 8 surveys that were 106 items long.

4.5 Presentation

For the magnitude estimation experiments (Experiments 2 and 4), participants were first asked to complete a practice phase in which they rated the lengths of 6 horizontal lines on the screen prior to the sentence rating task in order to familiarize them with the ME task itself. After this initial practice phase, participants were told that this procedure can be easily extended to sentences. No explicit practice phase for sentences was provided; however, the nine/six unmarked anchor items (which were not included in the analysis) did serves as a sort of unannounced sentence practice. There was also no practice for the forced choice experiment (Experiments 1 and 3), as the task is generally considered intuitively simple. The surveys were advertised on the Amazon Mechanical Turk website (see Sprouse 2011a, for evidence of the equivalence of data collected using AMT when compared to data collected in the lab), and presented as web-based surveys using an HTML template available on the first author's website. Participants completed the surveys at their own pace.

4.6 The number of judgments per condition

Each participant rated only one pair for each phenomenon in the forced-choice experiments (Experiments 1 and 3) or one token of each condition in the magnitude estimation experiments (Experiments 2 and 4). From the perspective of both traditional methods and formal experiments, this number is quite low. We chose to only test one token of each condition per participant for several reasons. First, this is the lower limit of possible experimental designs. This means that the power estimates that we derive will provide a lower bound for such experiments. By simply increasing the number of tokens per condition to 2 or 4, syntacticians can easily increase the power at any given sample size. Second, only including one token per condition allowed us to test all of the phenomena from each source in a single survey without risking fatigue on the part of the participants (i.e. the total survey length was always very close to 100 items). Because it is standard practice to z-score transform magnitude estimation responses prior to analysis (Sprouse and Almeida 2011, Schütze and Sprouse 2011), it is useful to test all related phenomena in a single survey so that the z-score transformation is based upon the same sentence types for every participant. Finally, some critics (e.g., Gibson & Fedorenko 2010b) have suggested that traditional methods are predicated upon a single judgment per condition. While in our experience this is false (see also Marantz 2005), incorporating that claim into our design allows us to address the concerns of critics of traditional methods directly.

4.7 Basic results

Before conducting the resampling simulations, we first conducted statistical analyses on each phenomenon using the full set of results from each experiment. For the forced-choice experiments (Experiments 1 and 3) we conducted two-tailed sign tests (with manual verification of the direction of the result) to determine if, for each phenomenon, participants chose the condition that is predicted to be more acceptable more often than they chose the condition that is predicted to be less acceptable. For the magnitude estimation experiments (Experiments 2 and 4), responses were z-score transformed and then analyzed using two-tailed paired *t*-tests (with manual verification of the direction of the result). We chose two-tailed (i.e., non-directional) versions of the statistical tests so that our results are as conservative as possible (a two-tailed *p*-value of .05 is equivalent to a one-tailed *p*-value of .025). All 95 phenomena were significant in the predicted direction according to the statistical analyses on the full samples. The full results of the sign tests and paired *t*-tests are presented in Appendix A (Experiments 1 and 2) and Appendix B (Experiments 3 and 4).

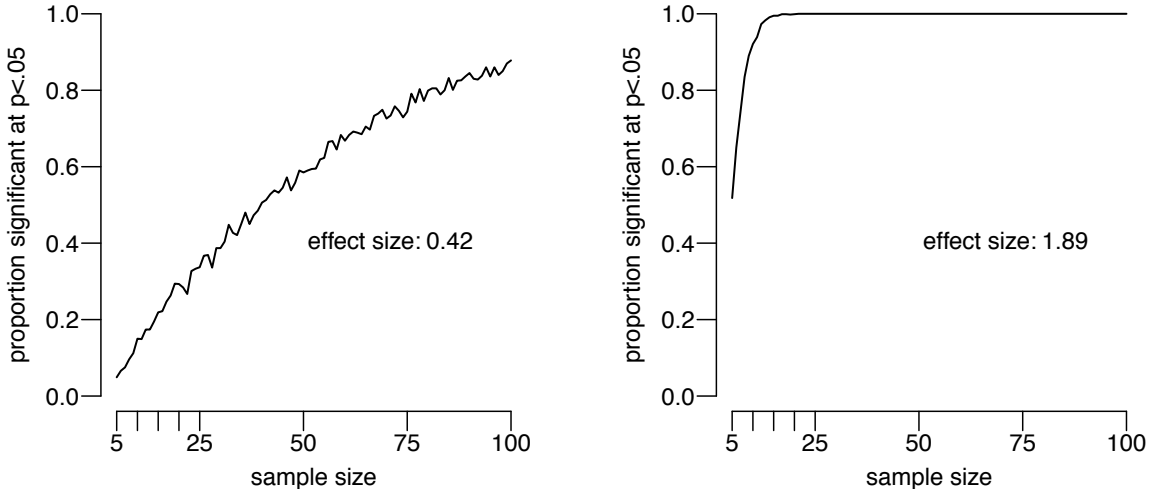
5. The resampling simulations

In order to empirically estimate the statistical power of each experiment type for each phenomenon at every sample size between 0 and 100 participants, we performed resampling simulations on each sample. In essence, these resampling simulations treated our large samples (N=148, N=140, N=144, N=144) as full populations, and sampled from them to determine the statistical power (operationalized here as a *rate of detection*) at each sample size that is possible with the population (5 to 100). For example, imagine that we are interested in the *whether* island effect illustrated by the pair of sentences in (2). To establish a rate of detection for a sample size of 5, we could perform the following procedure:

1. Draw a random sample of 5 participants (allowing participants to be potentially drawn more than once; this is called *sampling with replacement*)
2. Run a statistical test on the sample (for the magnitude estimation experiments, we used two-tailed paired t-tests; for the forced choice experiment, we used two-tailed sign tests).
3. Repeat steps 1 and 2 1000 times to simulate 1000 experiments with a sample size of 5.
4. Calculate the proportion of simulations (out of the 1000) that resulted in a significant result (i.e., a two-tailed *p*-value that is less than .05).

This procedure would tell us the rate of detection of the *whether* island effect for samples of size 5. We can then repeat this procedure for samples of size 6, 7, 8… 100 to derive a complete relationship between sample size and detectability for *whether* islands. Finally, we can repeat this procedure for all 95 phenomena to derive power relationships (operationalized here as empirically derived rates of detection) for the full range of effect sizes in Adger (2003) and the full set of small to medium effects from LI (2001-2010). As a concrete example, Figure 5 presents the results of these resampling simulations for two of the phenomena tested in Experiment 2:

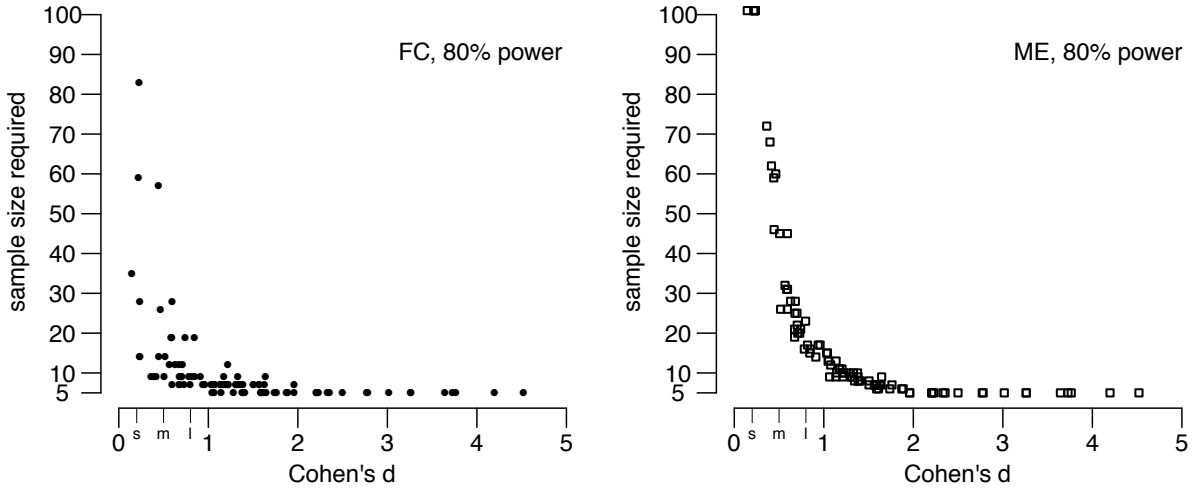
Figure 5: The results of the resampling simulation for two phenomena from Experiment 2. The x-axis represents sample sizes from 0 to 100. The y-axis represents the proportion of simulations that resulted in a significant t-test at $p < .05$. The effect size (Cohen's *d*) is labeled within each graph. Participants only judged one token per condition (e.g., a sample size of 5 represents only 5 judgments).



The detection rate graphs in Figure 5 provide complete power estimates for each phenomenon for sample sizes between 5 and 100. However, deriving general patterns from these results is difficult, as it requires comparing individual graphs for all 95 phenomena across the two types of experiments (forced-choice and magnitude estimation). In order to summarize the data in a more useful format, we can choose a target level of statistical power, such as the 80% target rate suggested by Cohen (1962, 1988, 1992), and identify the sample size required to reach that power rate for each of the 95 phenomena for each of the experiment types. We can then plot

the relationship between effect size (Cohen's *d*) and the sample size required to reach 80% power. As a concrete example, Figures 6 presents power curves for all 95 phenomena (combining the phenomena from Adger (2003) and LI (2001-2010)) with effect size on the x-axis and sample size required to reach 80% power on the y-axis. The figure in the left panel reports the sample size required to reach 80% power for all 95 phenomena (the 48 from Adger (2003) and the 47 from LI (2001-2010) under the forced-choice experiments (Experiments 1 and 3). The figure in the right panel reports the sample size required to reach 80% for all 95 phenomena under the magnitude estimation experiments (Experiments 2 and 4):

Figure 6: The sample size required to reach 80% power for all 95 phenomena (48 from Adger (2003) and 47 from LI (2001-2010) using forced-choice (left panel) and magnitude estimation (right panel) experiments. The effect size along the x-axis (Cohen's *d*) are from the magnitude estimation experiments (Experiments 2 and 4) for both panels. The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font.



In the next section we will use various combinations of these figures to address the two primary questions driving this study: (1) Is it true, as some critics have claimed, that traditional methods are likely to lead to a high rate of false negatives? (2) Is it true, as some critics have claimed, that traditional methods lead to more false negatives than formal experiments?

6. General Discussion

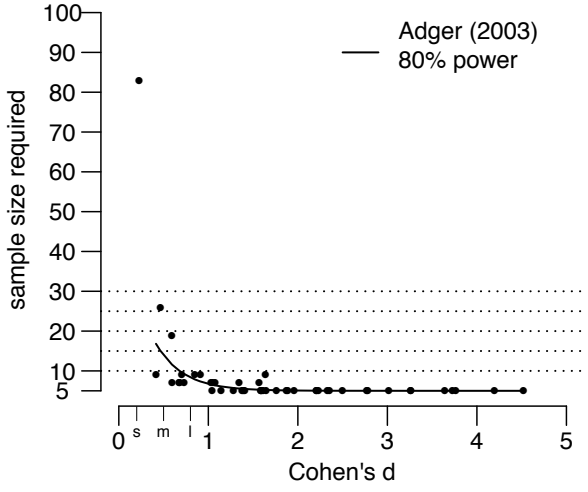
6.1 The false negative rate of traditional methods

The four experiments described in section 4, coupled with the resampling simulations described in section 5, give us the ability to analyze the relationship between the false negative rate (i.e., statistical power), sample size, and effect size for three different data sets: the phenomena from Adger (2003), the phenomena from LI (2001-2010), and the combination of the two sets of phenomena. To our knowledge there have been no concrete predictions about the false negative rate of traditional methods relative to sample size and effect size in the literature to date. However, it is possible to formulate a set of good-faith competing predictions. For example, if

traditional methods lead to an intolerably high number of false negatives, then forced-choice experiments with sample sizes that are typical of traditional methods (which we have intentionally left unspecified at this point) will fail to reach 80% for a large number of phenomena (which we have also left unspecified at this point) in each data set; instead, a larger-than-typical sample size will be required to reach 80% power for a large number of phenomena in each data set. If traditional methods lead to a tolerable number of false negatives, then forced-choice experiments will reach 80% power for a large number of phenomena in each data set with sample sizes that are typical of traditional methods.

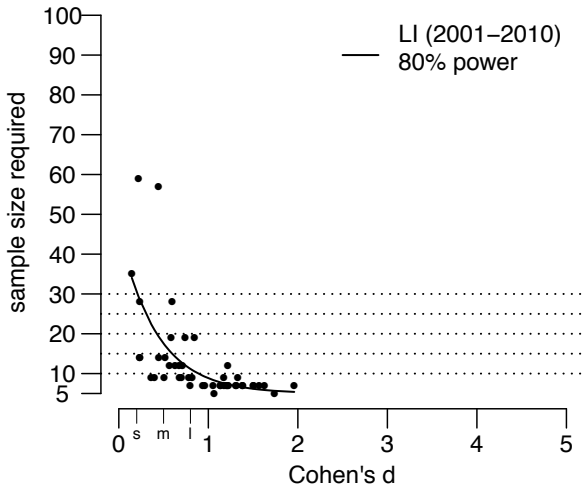
To assess these competing predictions we can create statistical power curve graphs like those above for each of the data sets (Adger (2003), LI (2001-2010), and the combined set), and add the following lines: (i) a non-linear trend-line representing a central tendency for the relationship between effect size and the sample size required to reach 80% power, and (ii) dotted horizontal lines representing potential "typical" sample sizes. We can then calculate the minimum effect size that is detectable at each potential typical sample size by finding the intersection of the non-linear trend lines and the horizontal dotted lines. We can also calculate the percentage of phenomena in syntactic theory that would likely be detectable at each sample size by comparing the minimum detectable effect size to the full distribution of phenomena in Adger (2003) and LI (2001-2010) (plotted in Figures 1 and 2). Figures 7, 8, and 9 present the data points, trend lines, and potential typical sample size lines (at 10, 15, 20, 25, and 30), and the corresponding tables present the minimum detectable effect sizes and percentage of phenomena detected at each sample size.

Figure 7: The sample size required (y-axis) to reach 80% power for each of the 48 effect sizes (x-axis) tested from Adger (2003). The solid line is a non-linear trend line based on 47 of the points (one outlier was removed for causing a sub-optimal fit). The dotted lines represent potential typical sample sizes (10, 15, 20, 25, 30). The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font. The minimum effect size detectable at each sample size and the percentage of phenomena detectable at each sample size are reported in the corresponding table.



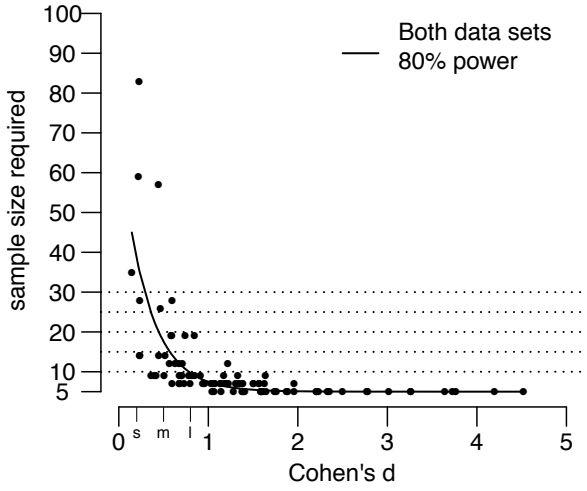
Sample size	Min. effect size	% phenomena
10	.68	95%
15	.46	98%
20	.34	99%
25	.25	99%
30	.19	99%

Figure 8: The sample size required (y-axis) to reach 80% power for each of the 47 effect sizes (x-axis) tested from LI (2001-2010). The solid line is a non-linear trend line for the points. The dotted lines represent potential typical sample sizes (10, 15, 20, 25, 30). The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font. The minimum effect size detectable at each sample size and the percentage of phenomena detectable at each sample size are reported in the corresponding table.



Sample size	Min. effect size	% phenomena
10	.89	73%
15	.60	82%
20	.42	89%
25	.30	94%
30	.21	95%

Figure 9: The sample size required (y-axis) to reach 80% power for the combined set of 95 effect sizes (x-axis). The solid line is a non-linear trend line for the points. The dotted lines represent potential typical sample sizes (10, 15, 20, 25, 30). The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font. The minimum effect size detectable at each sample size and the percentage of phenomena detectable at each sample size are reported in the corresponding table.



Sample size	Min. effect size	% phenomena
10	.78	84%
15	.57	89%
20	.45	93%
25	.36	94%
30	.29	97%

Figures 7, 8, and 9 present the graphical and numerical results of the experiments and simulations; however, the interpretation of these results crucially depends on what one believes are "typical" sample sizes for traditional methods, and what one believes constitutes a large number of detectable phenomena. As for the former, some who are critical of traditional methods have suggested that traditional methods involve absurdly small sample sizes and an absurdly low number of judgments per participant per condition (e.g., one participant and one judgment, see Gibson and Fedorenko 2010b). It is for this reason that we limited our participants to only one judgment per condition, thus providing an absolute minimum power estimate for each phenomenon. However, in our experience, careful syntacticians always ask a number of friends, students, and colleagues for judgments before submitting an article for publication, and, at least for well-represented languages, reviewers, editors, and conference audiences often provide additional judgments of phenomena prior to publication (see also Marantz 2005 and Phillips 2009). Furthermore, careful syntacticians often ask for multiple judgments per condition from each participant. Given the relatively informal nature of traditional methods, there is likely to be quite a bit of variation from researcher to researcher both in the choice of sample sizes and number of judgments during data collection, and in the interpretation of what constitutes a typical sample size and a typical number of judgments. For us, we believe that 10 judgments (in the form of 10 participants each giving one judgment, 5 participants each giving 2 judgments, etc) is likely the absolute minimum that will be tested by careful syntacticians prior to publication, therefore we have used 10 participants (each giving 1 judgment) as our minimum sample size in Figures 7, 8, and 9.

As for the question of what constitutes a "large" number of phenomena that are detectable at typical sample sizes, again we imagine that the answer will vary from researcher to researcher. This question can basically be restated as the following: What percentage of phenomena do you want to be detectable with typical methods (e.g., typical sample sizes), and what percentage of phenomena do you want to be detectable with more sensitive methods (e.g., larger sample sizes)? To our knowledge, there are no rules of thumb in experimental psychology that we can use to answer this question. This is likely a positive state of affairs; after all, it makes no sense for the methodology to dictate the size of the effects that the researcher considers when constructing a theory. Instead, the predictions of the theory should dictate which effects (regardless of size) the researcher should be looking for. Therefore we invite readers to make this call for themselves. For us, if 80% of phenomena are detectable at typical sample sizes, then we are inclined to conclude that the traditional methods are well-powered for general use in syntactic theory, as this would mean that for 4 out of 5 predictions, syntacticians can use typical traditional methods; for the other prediction, syntacticians would need to increase the sensitivity of the traditional methods (in most cases, by simply increasing the sample size or the number of judgments per participant).

With these issues in mind, we can look again at the results presented in the tables in Figures 7, 8, and 9 and see that at our minimum sample size of 10 participants each giving only 1 judgment, 95% of phenomena in Adger (2003) would be detected with 80% power, approximately 73% of phenomena in LI (2001-2010) would be detected with 80% power (based on the distribution of effect sizes in the full sample tested by Sprouse, Schütze, and Almeida, submitted), and 84% of the combined data set would be detected with 80%. By increasing the sample size to 20 participants each giving only 1 judgment, the coverage rates increase to 99%, 89%, and 93%. This suggests to us that most researchers would consider traditional methods to

be a well-powered tool for the detection of a large range of phenomena in syntactic theory even at relatively small sample sizes and relatively few judgments per participant.

6.2 The relative false negative rates of traditional methods and formal experiments

In the previous section we suggested that a reasonable conclusion from these results, at least based on our interpretation of typical sample sizes and large coverage, is that traditional methods are a well-powered tool for the detection of effects in syntactic theory. However, it is logically possible for traditional methods to be a well-powered tool, while at the same time being less powerful than formal experiments. If it is indeed the case that formal experiments lead to fewer false negatives than traditional methods (i.e., formal experiments are more powerful), then one could still make the argument that the field should consider replacing traditional methods with formal experiments whenever possible. In other words, the conclusion of (some) critics of traditional methods would still be valid, albeit for reasons other than the ones that the critics offered (e.g., Ferreira 2005, Wasow and Arnold 2005, Gibson and Fedorenko 2010b, Gibson et al. 2011). To address this issue, we can compare the relative power of traditional methods (as represented by forced-choice experiments) against the relative power of formal experiments (as represented by magnitude estimation experiments).

Figure 10: The sample size required (y-axis) to reach 80% power for the set of 48 effect sizes (x-axis) from Adger (2003) for both forced-choice (solid dots) and magnitude estimation (empty squares) experiments. The solid line is a non-linear trend line for the forced-choice results. The dashed line represents a non-linear trend line for the magnitude estimation results. The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font. The left panel includes both points and trend lines; the right panel presents the trend lines in isolation.

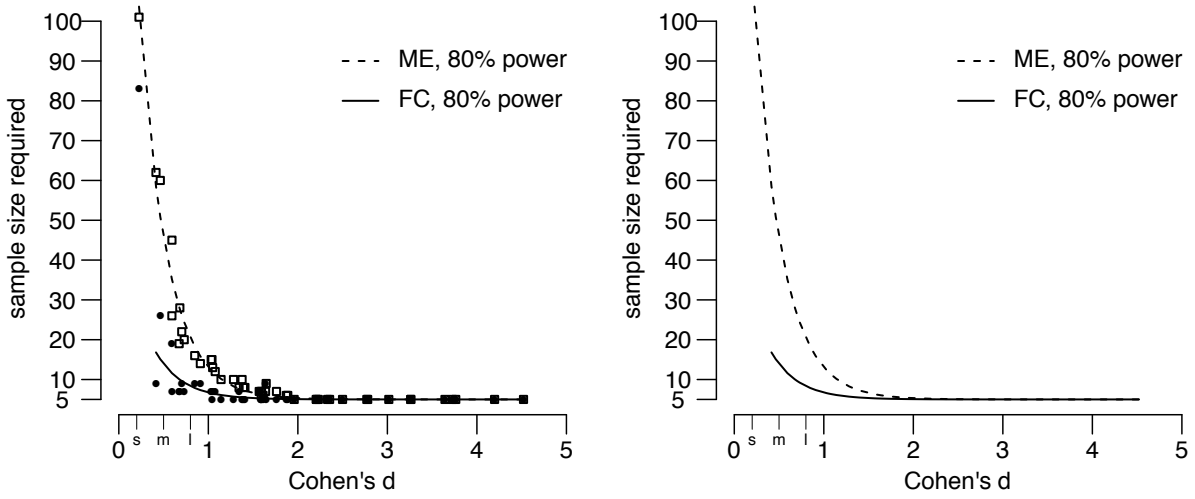


Figure 11: The sample size required (y-axis) to reach 80% power for the set of 47 effect sizes (x-axis) from LI (2001-2010) for both forced-choice (solid dots) and magnitude estimation (empty squares) experiments. The solid line is a non-linear trend line for the forced-choice results. The dashed line represents a non-linear trend line for the magnitude estimation results. The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font. The left panel includes both points and trend lines; the right panel presents the trend lines in isolation.

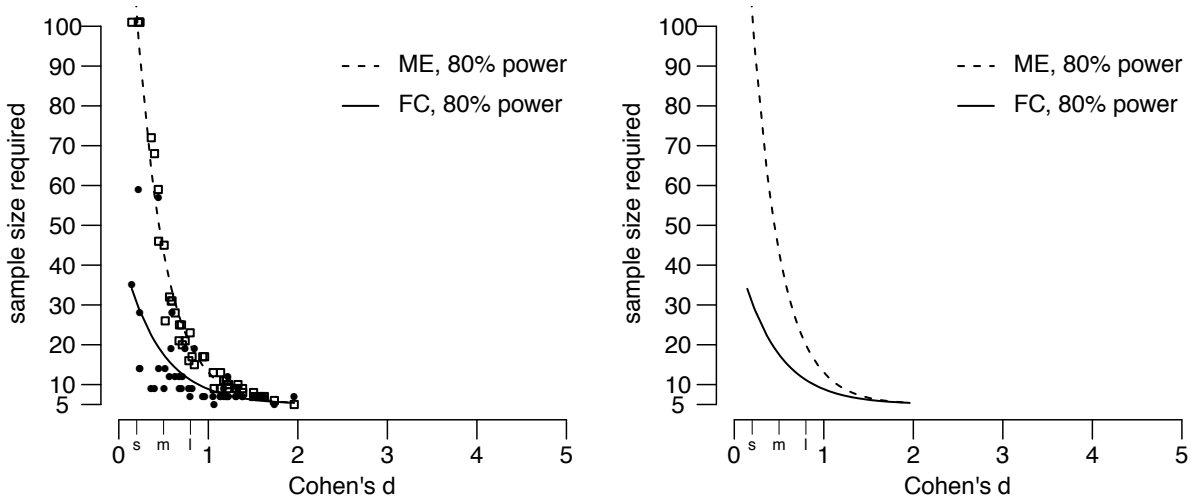
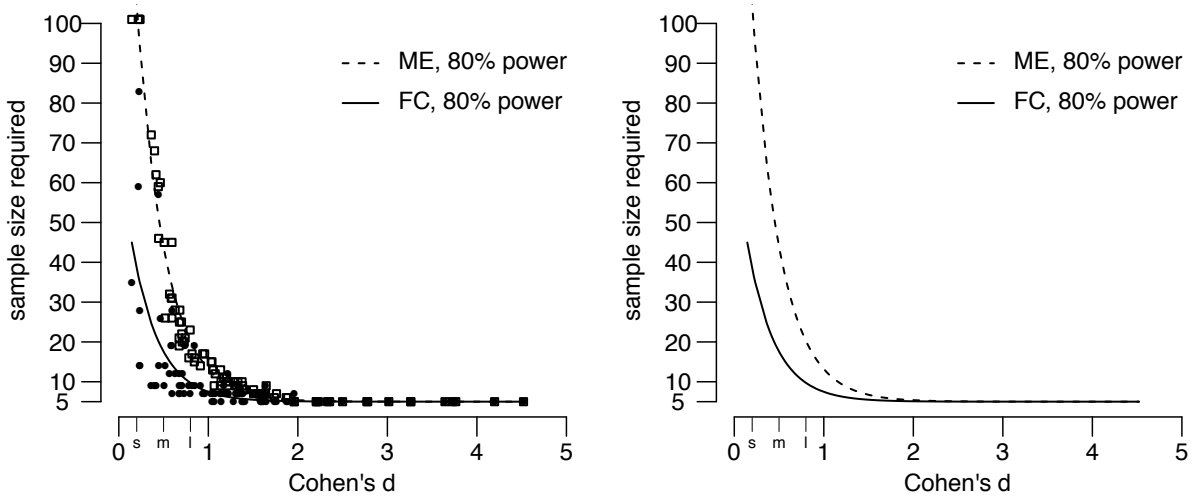


Figure 12: The sample size required (y-axis) to reach 80% power for the combined set of 95 effect sizes (x-axis) for both forced-choice (solid dots) and magnitude estimation (empty squares) experiments. The solid line is a non-linear trend line for the forced-choice results. The dashed line represents a non-linear trend line for the magnitude estimation results. The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font. The left panel includes both points and trend lines; the right panel presents the trend lines in isolation.



It is clear from Figures 10, 11, and 12, that although the two types of experiments are identical for effect sizes greater than 2 (i.e., very large effects by Cohen's (1988, 1992) definition), they diverge with smaller effect sizes. Crucially, this divergence is in the opposite direction than predicted by the claims of critics of traditional methods: traditional methods (as represented by the forced-choice task) are in fact more powerful than formal experiments (as represented by magnitude estimation experiments), as smaller sample sizes are required to detect effect sizes below 2 using the forced-choice task. One possible criticism of these comparisons is that part of the difference between traditional methods and formal experiments is a difference in typical sample sizes: whereas traditional methods tend to use small sample sizes (perhaps a minimum of 10 participants each providing one judgment), formal experiments tend to use larger sample sizes (perhaps a minimum of 20 participants; but cf. Myers 2009a, 2009b for a proposed type of formal experiments that use small sample sizes). To the extent that sample size is part of the distinction between these two methods, it is perhaps unfair to compare them at identical sample sizes. Instead, we could compare traditional methods at their typical sample size (e.g., 10) and formal experiments at their typical sample size (e.g., 20) to see which method has the power advantage at their respective typical sample size.

Figure 13: The difference in power between experiments using the FC task with 10 participants and experiments using the ME task with 20 participants (one judgment per condition per participant). Each line represents a phenomenon. The x-axis represents the effect size (in Cohen's *d*), the y-axis represents the difference in power (in percentages) between the two experiments, with advantages for the FC experiment as positive percentages, and advantages for the ME experiment in negative percentages.

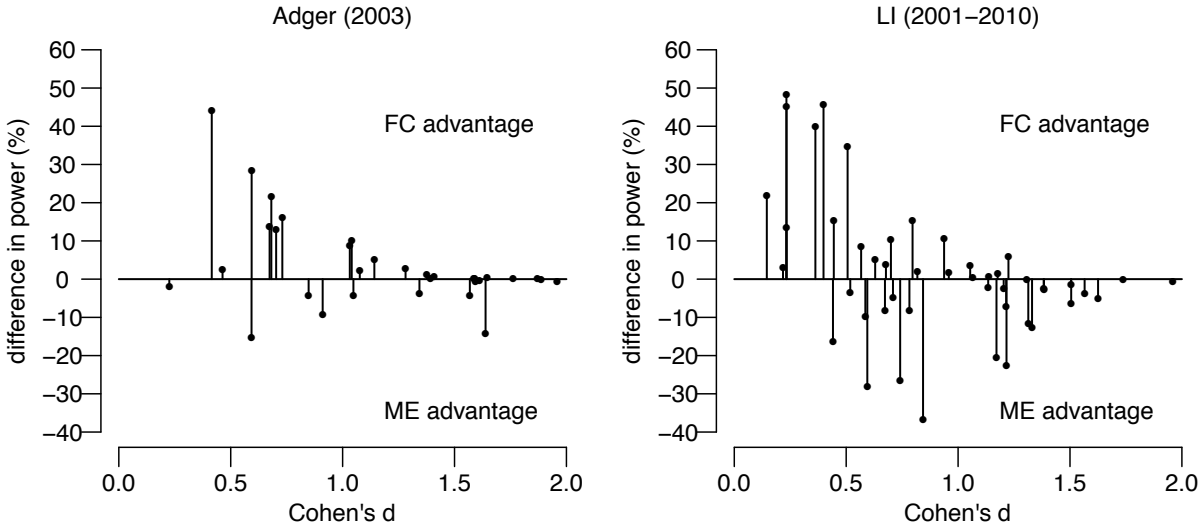


Figure 13 presents the difference in power between the forced-choice task (FC) with 10 participants and the magnitude estimation (ME) task with 20 participants for every phenomenon in the two data sets with an effect size below 2 (because both tasks reach 100% power at these sample sizes for effect sizes above 2). Three patterns are visible in these graphs. First, it is not the case that one task is universally more powerful than the other at typical sample sizes: there are phenomena that are better detected with each of the tasks at their typical sample sizes.

Second, the overall advantage is with the FC task and 10 participants: for the Adger phenomena, there is a cumulative increase of 113% power across the phenomena with effect sizes smaller than 2, and for the LI phenomena, there is a cumulative increase of 93% power across the relevant phenomena. Finally, these graphs suggest that the phenomena with the smallest effect sizes (i.e., the phenomena that are the hardest to detect) are more easily detectable with the FC task and 10 participants than the ME task and 20 participants. In other words, experiments using the FC task appear to be more powerful than experiments employing the ME task at typical sample sizes when the phenomena of interest have relatively small effect sizes.

These results may be surprising from the perspective of claims that are critical of traditional methods; however it should be noted that these results are consistent with the design of forced-choice and magnitude estimation tasks. In the FC task, participants are asked to directly compare two conditions to each other and choose the more acceptable condition. In contrast, in the ME task, participants are asked to rate conditions relative to a third sentence (the standard), which means that the only comparison that can be made between two conditions is mediated by their numerical rating relative to the standard. In other words, the FC task is designed to detect differences between conditions, but not quantify the magnitude of those differences; the ME task is designed to assign numerical ratings to conditions, but not directly detect differences between conditions. In the analyses presented above, we used the detection of a significant difference as the criterion for success (of course other criteria are possible depending on one's theoretical assumptions; see Sprouse and Almeida, to appear, and Sprouse, Schütze, and Almeida, submitted, for discussion). As such, these results simply confirm the intentional design differences between forced-choice and magnitude estimation tasks: the former are explicitly designed to detect simple differences between conditions, the latter are explicitly designed to assign numerical ratings to conditions. In this way, each task comes with its own set of costs and benefits – an issue that we discuss in detail in section 6.4.

6.3 The relative power of judgment experiments compared to the rest of experimental psychology

On one hand, the power advantages of the FC task may be surprising to some readers given the prevailing claims in the methodological literature. One the other hand, some readers may be unsurprised by this result, as some linguists have claimed that the effect sizes of phenomena studied in syntactic theory are substantially larger than the effect sizes studied in the branches of psychology that primarily rely on formal experimental methodologies (the existence of this claim is noted in Schütze 1996 and Gibson and Fedorenko 2010b, but it is difficult to track down first-hand claims in peer-reviewed print). We have already partially substantiated this claim in Figure 2: the distribution of randomly sampled phenomena from LI (2001-2010) suggests that approximately 12% of the phenomena from LI (2001-2010) are considered "small", 12% are considered "medium", and 76% are considered "large." Going back to Cohen's claim that he chose the "medium" criterion such that effect would be "apparent to the naked eye of a careful observer," this distribution suggests that nearly 9 out of 10 phenomena in syntactic theory should be visible to linguists without the need for formal statistical testing. Though we know of no exhaustive studies of the distribution of effect sizes in experimental psychology, the prevalence of formal statistical testing in experimental psychology suggests to us that the proportion of "medium" and "large" phenomena in experimental psychology is substantially smaller than 9 out of 10.

However, even if there are no exhaustive surveys of effect sizes typical of the experimental psychology literature, there have been a series of studies that have attempted to assess the median statistical power of experiments published in the psychology literature. Cohen (1962) demonstrated that in the 1960 volume of the *Journal of Abnormal and Social Psychology* the median power of the experiments was 46% for the average effect size of the phenomena under investigation (i.e., not much different than a coin toss). A follow-up study by Sedlmeier and Gigenrenzer (1989) for the 1984 issue of same journal found virtually identical results (44% median power for the average effect size of the phenomena of interest). In a review of the 1993 and 1994 volumes of the *British Journal of Psychology*, Clark-Carter (1997) reported a slightly larger average 59% power for the average phenomena of interest. Finally, Bezeau and Graves (2001) reported a mean of 50% power for "medium" effect sizes (*d* between 0.5 and 0.8) in their review of three neuropsychology journals, although they note that the average effect size in the neuropsychology literature seemed to be substantially larger than the ones studied in other branches of experimental psychology, a similar finding to the one reported here for acceptability data in theoretical syntax. Given that syntacticians employing traditional methods rarely report sample sizes, a comparable study is impossible for syntactic journals. However, we can approximate such as study by imposing our own minimum sample size estimate. For example, if one assumes that all of the English-language studies published in LI (2001-2010) had a sample size of 10 (with only 1 judgment per participant per condition), then the statistical power for the average effect size (a median of 1.57 and a mean of 1.58 for the 146 phenomena randomly sampled by Sprouse, Schütze, and Almeida, submitted) would be 94.5%. Since 94.5% power for the average phenomenon of interest is higher than Cohen's minimum recommendations of 80% and much higher than the observed 46%-59% reported in some branches of experimental psychology, it seems reasonable to conclude that traditional acceptability judgment experiments are a well-powered methodology for investigating the phenomena of interest to syntacticians.

6.4 The costs and benefits of traditional methods and formal experiments

The decision about which methodology to use can only be made by weighing the costs and benefits of each methodology relative to the research question at hand. Critics of traditional methods in syntax have suggested that there may be (at least) two heavy costs to the use of traditional methods: a high rate of false positives and a high rate of false negatives. Given that some critics have advocated the nearly universal adoption of formal experiments (e.g., Ferreira 2005, Featherston 2007, Gibson and Fedorenko 2010a, 2010b), we can only conclude that they assume that these costs are high enough to outweigh any benefit that traditional methods may have. However, the results of the present studies, together with the results of Sprouse and Almeida, to appear, and Sprouse et al., submitted, suggest that these concerns have been overstated: not only do traditional methods produce low false positive rates (Sprouse and Almeida, to appear, Sprouse, Schütze, and Almeida, submitted), they also produce a very low false negative rate (i.e., they have high statistical power) according to the standards used in experimental psychology, and in fact have a lower false negative rate than formal experiments with numerical rating tasks. At best, this suggests that the critics' suggestion for the universal adoption of formal experiments is unwarranted; at worst, the critics' suggestion may be detrimental to the field, as the higher false negative rate (i.e., lower statistical power) for formal experiments with numerical rating tasks means that larger samples will be required to detect the

same set of phenomena, potentially slowing the rate of discovery in the field (see also Sprouse et al., submitted).

The clear message here is that science is not a recipe that one can simply follow to uncover all and only the "real" phenomena. Instead, researchers need to be aware of the impact that their methodological choices could have on their results so that they can make an informed decision based on the goals of their particular research question. There are several benefits of traditional methods that have been catalogued before (e.g., Culicover and Jackendoff 2010): they are relatively quick to deploy, they are generally free, and they are very portable (requiring only pen and paper). Sprouse and Almeida (to appear) and Sprouse, Schütze, and Almeida (submitted) have suggested that they have a very low false positive rate. To that we can now add that they also have relatively high statistical power (relative to formal experiments with numerical tasks, and relative to experiments in the broader field of psychology). The costs of traditional methods are a bit more complex. Traditional methods tend to be ill-suited for numerical rating tasks because numerical rating tasks generally require sample sizes that are larger than the sample sizes used for traditional methods (Sprouse and Almeida 2011, Schütze and Sprouse 2011). Therefore if the hypothesis in question requires numerical ratings, traditional methods will likely be inadequate. Traditional methods tend not to be analyzed using statistical tests, which provide a type of confidence in the results. If there is no other way to establish confidence in the results, such as replication (which may in fact be the only way to establish the generalizability of the results beyond the original sample: Balluerka et al. 2005, Hubbard and Lindsay 2008, and many others), the lack of statistical tests in traditional methods may cause readers to be less confident in the results. Finally, there is a clear sociological cost to the use of traditional methods in syntax: whereas many syntacticians believe that traditional methods are reliable, researchers in fields that are used to formal experiments may erroneously believe that traditional methods are unreliable because they lack many of the properties of formal experiments (e.g., Ferreira 2005, Gibson and Fedorenko 2010a, 2010b).

The benefits of formal experiments are relatively straightforward as well. Formal experiments are often necessary for the reliable collection of numerical ratings, so they are the best choice for hypotheses about the *size* of the difference between conditions (e.g., Sprouse et al. 2011), hypotheses about the source of gradient acceptability (e.g., Keller 2000, Featherson 2005b), and comparisons between acceptability and other cognitive measures (e.g., Sprouse, Wagers, and Phillips 2012). As mentioned above, formal experiments also tend to be analyzed using statistical tests, which can provide a type of confidence in the results when replication is difficult or costly. And formal experiments are more likely to be seen as reliable to researchers in fields that rely exclusively on formal experiments (Ferreira 2005, Gibson and Fedorenko 2010a, 2010b). The costs of formal experiments have rarely been discussed in the literature. First and foremost, formal experiments are much more expensive than traditional methods. In the laboratory, participants are routinely paid $5 for the completion of a 100 item magnitude estimation survey; on Amazon Mechanical Turk the same survey would cost $3.30 per participant ($3 to the participant, $.30 to Amazon). A 100-item survey can maximally test 50 two-condition phenomena (one rating per condition per participant), which is probably enough for a medium-length syntax article. Using the results of the current studies as a guideline, experiments with numerical rating tasks should probably be designed to collect at least 40 observations per condition (and more if possible). This could mean 40 participants rating each condition once for a cost of $200 in the laboratory and $132 on Amazon Mechanical Turk, or 20 participants rating each condition twice, which would halve the cost at the expense of halving the

number of phenomena that can be tested in a single experiment (25 instead of 50). While these prices are cheap by experimental psychology standards, they are much more expensive than traditional methods (which tend to be free). It also generally takes more time to recruit participants for formal experiments; however Amazon Mechanical Turk is neutralizing this cost: 80 participants can be collected per hour on Amazon Mechanical Turk (Sprouse 2011a). Finally, the results of this study suggest that formal experiments may be less powerful than traditional methods unless special care is taken in selecting an appropriate sample size.

7. Conclusion

In this study we tested 97 phenomena (from Adger 2003 via Sprouse and Almeida, to appear, and from Linguistic Inquiry 2001-2010 via Sprouse, Schütze, and Almeida, submitted) that span the range of effect sizes in syntactic theory using both the forced-choice and magnitude estimation tasks on relatively large sample sizes (at least 140 participants per sample). We then ran resampling simulations on these results to empirically estimate the statistical power for each phenomenon at every possible sample size between 5 and 100 participants. These simulations allowed us to address two of the primary criticisms that have been levied against traditional acceptability judgment collection methods in syntax: (i) that traditional methods lead to an intolerably high false negative rate, and (ii) that formal experimental methods will lead to a lower (and more tolerable) false negative rate.

The results of this study suggest that, contrary to the claims of critics, traditional methods are in fact a well-powered methodology for syntax, detecting over 80% of phenomena with 10 observations or fewer. In fact, traditional methods are more powerful than formal experimental methods, at least with respect to detecting difference between conditions, as traditional methods require substantially fewer participants to reach 80% power (the suggested power level in experimental psychology; see Cohen 1988, 1992; see also Sprouse and Almeida 2011 for similar findings). Although these results may appear surprising given the prevalence of the criticisms of traditional methods in the literature, they in fact make sense in light of the distribution of effect sizes estimated by the 146 phenomena that Sprouse, Schütze, and Almeida (submitted) randomly sampled from Linguistic Inquiry (2001-2010). This distribution suggests that the phenomena of interest to syntacticians tend to be substantially larger than those of interest to other areas of psychology (nearly 90% of phenomena are large enough to be visible to the "naked eye"; Cohen 1992), and that traditional methods will lead to at least 94.5% power for the average effect size (as compared to 46%-59% for the average effect size in other areas of psychology). Taken as a whole, these results strongly suggest that it is unlikely that traditional methods have led to an intolerable false negative rate in syntax. And to the extent that previous studies have demonstrated that traditional methods have not led to an intolerable false positive rate either (Sprouse and Almeida, to appear, Sprouse, Schütze, and Almeida, submitted), we believe that traditional methods should be seen as valid, reliable, and well-powered experiments for the investigation of the phenomena of interest to syntacticians.

Of course, this is not to say that there is no place for formal experiments in the syntactician's toolkit, such as quantifying the *size* of the difference between conditions (e.g., Sprouse et al. 2011), investigating the source of gradient acceptability (e.g., Keller 2000, Featherson 2005b), and comparing acceptability to other cognitive measures (e.g., Sprouse, Wagers, and Phillips 2012). The inevitable conclusion is that science cannot be reduced to a simple recipe, no matter how attractive one particular method, be it formal experiments or

traditional methods, may appear. There are costs and benefits to every methodology. Syntacticians should be allowed to decide which methodology best suits their scientific goals, even if the resulting methodology appears superficially dissimilar to the methodologies in neighboring fields.

References

Adger, David. 2003. *Core Syntax: A Minimalist Approach*. Oxford University Press.

Bader, Marcus, and Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46:273–330.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72:32–68.

Bezeau, S., and R. Graves. 2001. Statistical Power and Effect Sizes of Clinical Neuropsychology Research. *Journal of Clinical and Experimental Neuropsychology* 23:399–406.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clark-Carter, D. 1997. The account taken of statistical power in research published in the British Journal of Psychology. *British Journal of Psychology* 88:71–83.

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

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences, 2nd ed*. Hillsdale, NJ: Erlbaum

Cohen, J. 1992. Statistical Power Analysis. *Current Directions in Psychological Science* 1:98–101.

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.

Culbertson, Jennifer, and Steven Gross. 2009. Are linguists better subjects? *British Journal for the Philosophy of Science* 60:721–736.

Culicover, Peter W., and Ray Jackendoff. 2010. Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences* 14:234–235.

Dąbrowska, Ewa. 2010. Naïve v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27:1–23.

den Dikken, Marcel, Judy Bernstein, Christina Tortora, and Raffaella Zanuttini. 2007. Data and grammar: means and individuals. *Theoretical Linguistics* 33:335–352.

Edelman, Shimon, and Morten Christiansen. 2003. How seriously should we take Minimalist syntax? *Trends in Cognitive Sciences* 7:60–61.

Fanselow, Gisbert. 2007. Carrots – perfect as vegetables, but please not as a main dish. *Theoretical Linguistics* 33:353–367.

Featherston, Sam. 2005a. Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua* 115:1525–1550.

Featherston, Sam. 2005b. Universals and grammaticality: Wh-constraints in German and English. *Linguistics* 43:667–711.

Featherston, Sam. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33:269–318.

Featherston, Sam. 2008. Thermometer judgments as linguistic evidence. In *Was ist linguistische evidenz?*, ed. by C. M. Riehl and A. Rothe. Aachen: Shaker Verlag.

Featherston, Sam. 2009. Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft* 28:127–132.

Ferreira, Fernanda. 2005. Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review* 22:365–380.

Gallistel, Randy. 2009. The importance of proving the null. *Psychological Review* 116:439–53.

Grewendorf, Günther. 2007. Empirical evidence and theoretical reasoning in generative grammar. *Theoretical Linguistics* 33:369–381.

Gibson, Edward, and Evelina Fedorenko. 2010a. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14:233–234.

Gibson, Edward, and Evelina Fedorenko. 2010b. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes.*

Gross, Steven, and Jennifer Culbertson. 2011. Revisited linguistic intuitions. *British Journal for the Philosophy of Science.*

Haider, Hubert. 2007. As a matter of facts – comments on Featherston's sticks and carrots. *Theoretical Linguistics* 33:381–395.

Hill, A. A. 1961. Grammaticality. *Word* 17:1–10.

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

Kruschke, John A. 2011. *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York: Academic Press.

Marantz, Alec. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* 22:429–445.

Myers, James. 2009a. The design and analysis of small-scale syntactic judgment experiments. Lingua, 119:425–444.

Myers, James. 2009b). Syntactic judgment experiments. *Language and Linguistics Compass* 3:406–423.

Newmeyer, Frederick J. (2007). Commentary on Sam Featherston, 'Data in generative grammar: The stick and the carrot.' *Theoretical Linguistics* 33:395–399.

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

Phillips, Colin, and Howard Lasnik. 2003. Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Sciences* 7:61–62.

Rouder, Jeffrey N., Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* 16:225–237.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

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

Sprouse, Jon. 2007a. A program for experimental syntax. Doctoral dissertation, University of Maryland.

Sprouse, Jon. 2007b. Continuous Acceptability, Categorical Grammaticality, and Experimental Syntax. *Biolinguistics* 1:118–129.

Sprouse, Jon. 2008. The differential sensitivity of acceptability to processing effects. *Linguistic Inquiry* 39:686–694.

Sprouse, Jon. 2009. Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*. 40:329–341.

Sprouse, Jon. 2011a. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43:155–167.

Sprouse, Jon. 2011b. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87:274–288.

Sprouse, Jon. and Diogo Almeida. (2011). The role of experimental syntax in an integrated cognitive science of language. In *The Cambridge Handbook of Biolinguistics*, ed. by Kleanthes Grohmann and Cedric Boeckx.

Sprouse, Jon. and Diogo Almeida. (to appear). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*.

Sprouse, Jon; Shin Fukuda; Hajime Ono; and Robert Kluender. 2011. Grammatical operations, parsing processes, and the nature of wh-dependencies in English and Japanese. *Syntax* 14:179–203.

Sprouse, Jon, Matt Wagers, and Colin Phillips. 2012. A test of the relation between working memory capacity and island effects. *Language*.

Sorace, Antonia, and Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115:1497–1524.

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

Stevens, Stanley Smith. 1956. The direct estimation of sensory magnitudes: loudness. *The American journal of psychology* 69:1–25.

Wasow, Thomas, and Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115:1481–1496.

Weskott, Thomas, and Gisbert Fanselow. 2011. On the Informativity of Different Measures of Linguistic Acceptability. *Language* 87:249–273.

Wetzels, Ruud; Dora Matzke; Michael D. Lee; Jeffrey N. Rouder; Geoffrey J. Iverson; and Eric-Jan Wagenmakers. 2011. Statistical evidence in experimental psychology: An empirical comparison using 855 t-tests. *Perspectives on Psychological Science* 6:291–298.

Jon Sprouse
University of California, Irvine
3151 Social Science Plaza A
Irvine, CA 92697-5100
jsprouse@uci.edu

Diogo Almeida
Michigan State University
A-614 Wells Hall
East Lansing, MI 48824-1027
diogo@msu.edu

Appendix

Table 3: Example sentences for each of the 96 conditions tested in the present experiments. The ID column indicates the chapter, example number, and diacritic from Adger (2003). The Mean column represents the mean rating for each condition (after the by-participant z-score transformation), SD represents the standard deviation, Diff. Means represents the difference between the means of the two conditions (a type of intuitive effect size), and Cohen's *d* is one type of standardized effect size.

ID	Example sentence	Mean	(SD)	Diff. Means	Cohen's *d*
ch9.122.g	What did Sandy give to whom?	-0.05	(0.70)	0.16	0.23
ch9.123.*	Who did Sandy give what to?	-0.21	(0.72)		
ch5.84.g	I'd planned to have finished, and finished I have.	-0.26	(0.65)	0.26	0.41
ch5.84.*	I'd planned to have finished, and have finished I did.	-0.51	(0.58)		
ch6.108-109.g	Elliot could quickly have freed the animals.	0.30	(0.74)	0.33	0.46
ch6.106.*	Elliot quickly may free the animals.	-0.02	(0.68)		
ch8.6.?	That the company would go bankrupt was claimed by the stockholders.	-0.38	(0.59)	0.39	0.70
ch8.7.*	The company would go bankrupt was claimed that by the stockholders.	-0.77	(0.52)		
ch3.152.g	What Julie became was fond of the book.	-0.32	(0.62)	0.42	0.73
ch3.153.*	What Julie did of the book was become fond.	-0.74	(0.55)		
ch10.84-86.g	What did Peter listen to a speech about?	-0.01	(0.75)	0.43	0.59
ch10.72.*	What did Peter listen to Darren's speech about?	-0.44	(0.69)		
ch8.3.g	What she thought was that the poison was neutralized.	0.12	(0.72)	0.52	0.58
ch8.3.*	What she thought that was the poison was neutralized.	-0.40	(1.03)		
ch3.149.g	Julie became fond of the book.	0.64	(0.73)	0.53	0.68
ch3.148.*	Julie became fond.	0.12	(0.81)		
ch8.56.g	That the answer is obvious upset Helen.	-0.06	(0.68)	0.57	0.91
ch8.58.*	That whether the world is round is unknown upset Helen.	-0.63	(0.58)		

Table 3 (continued)

ID	Example sentence	Mean	(SD)	Diff. Means	Cohen's *d*
ch8.62.g	That Nina had won surprised the coach.	-0.05	(0.69)	0.58	0.85
ch8.64.*	That Nina had won surprised the coach insulted her.	-0.63	(0.67)		
ch5.139.g	Terry has never driven a car.	1.06	(0.52)	0.65	1.02
ch5.140.*	Terry never has driven a car.	0.41	(0.75)		
ch3.73.g	Parents of students want to annoy teachers.	0.04	(0.67)	0.67	1.07
ch3.74.*	Parents of a student wants to annoy teachers.	-0.63	(0.57)		
ch2.02.g	The pigs grunt.	0.61	(1.11)	0.68	0.67
ch2.04.*	The pigs grunts.	-0.07	(0.91)		
ch5.25-26.g	I believed she might be pregnant.	0.63	(0.66)	0.73	1.05
ch5.27-28.*	I believed she may be pregnant.	-0.10	(0.74)		
ch9.84-85.*	I wondered how did Lewis survive.	-0.10	(0.57)	0.77	1.55
ch9.105.*	Mary thinks how Lewis survived.	-0.88	(0.42)		
ch8.23,25.g	What Kelsey wondered was whether the store had the DVD in stock.	0.33	(0.66)	0.81	1.14
ch8.24,26.*	What Kelsey wondered whether was the store had the DVD in stock.	-0.47	(0.75)		
ch7.105.g	The therapist's analysis of Morticia was flawed.	0.85	(0.73)	0.83	1.04
ch7.105.*	The therapist's analysis of Morticia's was flawed.	0.03	(0.86)		
ch9.83.g	I wondered if we could leave early.	1.02	(0.56)	0.83	1.41
ch9.83.*	I wondered could we leave early.	0.18	(0.62)		
ch5.19-20.g	I thought he was honest.	1.05	(0.60)	0.88	1.27
ch5.21-22.*	I thought he is honest.	0.17	(0.76)		
ch5.39.g	Tom said he won the trophy and won the trophy he did.	-0.13	(0.60)	0.88	1.60
ch5.38.*	Tom said he won the trophy and won the trophy he.	-1.01	(0.49)		

Table 3 (continued)

ID	Example sentence	Mean	(SD)	Diff. Means	Cohen's *d*
ch2.68.g	We all thought him to be unhappy.	0.26	(0.77)	0.89	1.35
ch2.70.*	We all thought he to be unhappy.	-0.63	(0.52)		
ch8.184,186.g	I expected there to be a problem.	0.41	(0.75)	0.94	1.34
ch8.185,187.*	I persuaded there to be a problem.	-0.53	(0.66)		
ch10.94,96.g	That Peter loved Amber seemed to be known by everyone.	-0.01	(0.69)	0.97	1.58
ch10.95,97.*	Who did that Peter loved seem to be known by everyone?	-0.98	(0.53)		
ch10.118.g	I worry if the lawyer forgets his briefcase at the office.	0.23	(0.71)	1.02	1.58
ch10.121.*	What do you worry if the lawyer forgets at the office?	-0.79	(0.57)		
ch8.105.g	For him to do that would be a mistake.	0.58	(0.63)	1.12	1.76
ch8.105.*	For to do that would be a mistake.	-0.54	(0.64)		
ch8.19-20.g	Marcy wondered if the meeting would start on time.	1.23	(0.69)	1.12	1.38
ch8.21-22.*	Marcy wondered if that the meeting would start on time.	0.11	(0.93)		
ch5.9.g	What Brian must do is call Susan.	0.21	(0.64)	1.15	2.15
ch5.9.*	What Brian does is must call Susan.	-0.94	(0.40)		
ch10.69.g	I believed the claim that Philip would visit the city of Athens.	0.54	(0.68)	1.16	1.86
ch10.70.*	Which city did you believe the claim that Philip would visit?	-0.62	(0.56)		
ch7.104.g	A book of mine is on the desk.	0.62	(0.80)	1.17	1.63
ch7.103.*	A book of my is on the desk.	-0.55	(0.62)		
ch9.28-29.g	Which poem did Harry recite?	1.16	(0.52)	1.37	1.82
ch9.32-33.*	Which the poem did Harry recite?	-0.21	(0.93)		
ch3.15.g	Extremely frantically, Jerry danced at the club.	0.16	(0.67)	1.38	2.31
ch3.16.*	Frantically at, Jerry danced extremely the club.	-1.22	(0.51)		

Table 3 (continued)

ID	Example sentence	Mean	(SD)	Diff. Means	Cohen's *d*
ch3.57.g	Humans love to eat those pigs.	0.36	(0.65)	1.39	2.47
ch3.63.*	Peter is those pigs.	-1.02	(0.46)		
ch8.102.g	Alicia planned for him to attend college.	0.84	(0.71)	1.43	2.22
ch8.103.*	Alicia planned for he to attend college.	-0.59	(0.58)		
ch6.58.g	He has known him.	0.49	(0.86)	1.50	2.23
ch6.58.*	Him has he known.	-1.01	(0.42)		
ch5.8.g	George may seek Isabelle.	0.99	(0.75)	1.53	1.62
ch5.8.*	George seek may Isabelle.	-0.54	(1.11)		
ch3.77.g	It rained.	0.95	(0.79)	1.53	1.96
ch3.79.*	The weather rained.	-0.59	(0.78)		
ch3.118-120.g	The bookcase ran.	0.95	(1.04)	1.58	1.58
ch3.118-120.*	The thief ran.	-0.63	(0.95)		
ch7.30-33.g	This man needs a taxi.	1.22	(0.60)	1.63	2.17
ch7.30-33.*	The this man needs a taxi.	-0.40	(0.88)		
ch6.98.g	The boy was killed by Stan.	0.86	(0.67)	1.78	3.18
ch6.100.*	There arrived by Stan.	-0.91	(0.42)		
ch5.43-44.g	She tried to leave.	1.12	(0.55)	1.78	2.74
ch5.49.*	She tried to do leave.	-0.66	(0.74)		
ch7.3.g	The letters are on the table.	1.07	(0.64)	1.85	2.78
ch7.4.*	Letters the are on the table.	-0.78	(0.69)		
ch3.92-94.g	Andy demonized David.	1.06	(0.69)	1.96	3.23
ch3.113.*	Andy demonized up the river.	-0.89	(0.51)		

Table 3 (continued)

ID	Example sentence	Mean	(SD)	Diff. Means	Cohen's *d*
ch5.144.g	Jason hasn't arrived.	1.10	(0.54)	1.97	3.76
ch5.145.*	Jason not arrived.	-0.87	(0.50)		
ch9.4.g	What did Carl buy?	1.15	(0.62)	1.97	3.68
ch9.12.*	Something did Carl buy.	-0.82	(0.44)		
ch4.37.g	I shaved myself.	0.89	(0.62)	1.99	3.61
ch4.38.*	Myself shaved me.	-1.09	(0.47)		
ch5.135-136.g	Ryan did not fly the airplane.	1.29	(0.93)	2.18	2.91
ch5.133-134.*	Ryan not flew the airplane.	-0.90	(0.52)		
ch5.31.g	Dale loved Clare.	1.31	(0.56)	2.29	4.49
ch5.36.*	Dale do loved Clare.	-0.97	(0.45)		
ch3.33a.g	Julie and Jenny arrived first.	1.08	(0.57)	2.32	4.19
ch3.33d.*	It was Jenny arrived that Julie and first.	-1.24	(0.53)		