# 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 Psychology
New York University, Abu Dhabi

## 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 & Christiansen 2003, Ferreira 2005, Wasow & Arnold 2005, Featherston 2007, Gibson & 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), and that this can be remedied by the use of more formal methods of data collection. We empirically assessed this claim by conducting a series of experiments designed to derive comprehensive estimates of statistical power for different types of acceptability judgment experiments. We tested 47 phenomena (94 sentence types) from a random sample of phenomena in Linguistic Inquiry (2001-2010) that span a large range of effect sizes (Cohen's *d* 0.15-1.96), using all four major judgment tasks normally used in syntactic research (magnitude estimation, Likert scale, yes-no, and forced-choice), and four samples each of 144 participants. We then ran re-sampling simulations to empirically estimate statistical power for every combination of effect size, sample size (5-100), and task. The results provide the first comprehensive evaluation of statistical power in acceptability judgments, which can be used to (i) evaluate the statistical power of previously published studies, (ii) plan appropriately powered studies in the future, and most importantly, (iii) establish a common vocabulary for assessing whether any definition of the more traditional methods can be seen as a well-powered experiment in its own right. We discuss the relative power of the four types of experiments, the relative power of acceptability judgment experiments to experiments in other domains of psychology, and the empirical coverage of each experiment type.

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 from articles that appeared in *Linguistic Inquiry* from 2001 through 2010, Sprouse, Schütze, & Almeida (*submitted*) estimated that 77% were derived from acceptability judgments (the remaining 23% were based on meaning/ambiguity judgments). 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 & Christiansen 2003, Phillips & Lasnik 2003, Featherston 2005a, 2005b, 2007, 2008, 2009, Ferreira 2005, Sorace & Keller 2005, Wasow & Arnold 2005, den Dikken et al. 2007, Alexopoulou & Keller 2007, Fanselow 2007, Newmeyer 2007, Culbertson & Gross 2009, Myers 2009, Phillips 2009, Bader & Häussler 2010, Dąbrowska 2010, Gibson & Fedorenko 2010a, 2010b, Fedorenko & Gibson 2010, Culicover & Jackendoff 2010, Gross & Culberton 2011, Weskott & Fanselow 2011, Gibson et al. 2011, Sprouse 2007a, 2007b, 2008, 2009, 2011a, 2011b, Sprouse, Fukuda, Ono, & Kluender 2011, Sprouse, Wagers, & Phillips 2012, Sprouse & Almeida 2012, Sprouse and Almeida *in press*, Sprouse, Schütze, & 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 & Christiansen 2003, Ferreira 2005, Wasow & Arnold 2005, Gibson & Fedorenko 2010a, 2010b). Our goal in this paper is to formalize precisely what it would mean for traditional methods to be unreliable compared to more formal experimental methods, and to empirically assess to what extent this claim is true. As a first step toward this goal, this section presents (i) a framework for defining the differences between traditional methods and more formal methods, (ii) a framework for defining the two primary types of (un)reliability in experimental methods, (iii) a brief review of previous empirical work on the first type of (un)reliability (Type I errors), and (iii) our strategy for empirically evaluating the second type of (un)reliabilty (Type II errors).

## 1.1 Two types of experiments: The distinction between *informal* and *formal* experiments in syntax

The first step in addressing the criticisms of traditional methods in syntax is to clarify the similarities and differences between the traditional methods that syntacticians routinely employ and the more formal methods that critics of traditional methods often advocate. Developing a concrete framework for describing these differences will allow us to identify the properties of the two methods that could potentially contribute to differences in reliability, and will allow us to experimentally manipulate those properties to assess whether they do in fact affect reliability.

The first place to look for differences between the two methods is with their underlying logic. In this case, the underlying logic of the two methods is identical. First, the researcher constructs a set of conditions to minimally contrast the relevant syntactic property. These conditions are carefully constructed to rule out known nuisance variables, the set of which has been incrementally accumulated by prior research. Next, the researcher constructs a set of sentences for each condition in an attempt to rule out any lexically driven extraneous factors (such as sentence plausibility or glaring word frequency imbalances). Finally, the researcher asks a sample of relevant native speakers to rate these items along a scale to determine whether the syntactic manipulation has an effect on the relative acceptability of the conditions. The results are then submitted to the scrutiny of the researchers' peers and eventually summarized in a journal article. Successful replications boost confidence in the findings, while failed replications lower it. In short, the logic of traditional methods is precisely the logic of the *experimental method*: theoretically relevant factors are directly manipulated in order to test causal relationships. Although it is not uncommon for the differences between the two methods to be described as a difference between *observation* (traditional methods) and *experimentation* (more formal methods), this is clearly incorrect. Observational methods (by definition) do not involve the direct manipulation of theoretically relevant factors (e.g., corpus analysis), whereas traditional methods clearly do involve experimental manipulations. This has been pointed out several times in the literature:

> Gathering of native speaker judgments is a trivially simple kind of experiment… Any good linguistics study involves carefully constructed materials, appropriate control items, and robust and replicable results. (Phillips & Lasnik 2003:61)

> The linguist presenting examples of this sort has already performed an experiment on him/herself or one or more informants. (Marantz 2005:343)

> [I]nformal judgment collection is itself a form of experimentation. … [I]nformal methods and full-fledged experimentation lie on a continuum, rather than representing radically different types of data sources…(Myers 2009a:426)

> Psychologists and other scientists deal with such 'nuisance' variables through the use of carefully constructed experimental designs, and, surprisingly perhaps, so do theoretical linguists. (Myers 2009b:413)

It is important to eliminate this false dichotomy as it can obscure the true differences between the two methods – a topic that we turn to next.

Despite sharing the fundamental logic of the experimental method, there are differences between the two methods that could potentially affect reliability; however, the differences are generally matters of degree. For instance, the informal experiments of traditional methods are normally conducted as the need arises, often in several independent interviews using only the experimental items and their controls (i.e., no distractor items), and are conducted verbally or by e-mail. These interviews are generally conducted on relatively small groups of consultants (e.g., 1-10), although several such small groups may be consulted repeatedly over the course of the study. The consultants generally include non-naïve participants, often including the researcher herself. Finally, the results are reported using only non-numerical descriptive summaries (the diacritics: ?, ??, ?*, *). The more formal experiments recommended by critics of traditional methods tend to rely on one or a few large planned written surveys, which are applied in the classroom, lab, or over the internet. These surveys generally include distractor items, and are presented to a larger group (e.g.,, > 20) of naïve participants. The results are typically summarized using standard descriptive statistics (mean, standard deviation) and analyzed using standard inferential statistics (*t*-test, ANOVA, linear mixed effects models).

In sum, because the two methods share a fundamental logic (the experimental method), we can formalize the differences between them using the set of dimensions over which experiments can vary: the task and/or response scale used, the number of participants recruited, the number of experimental items constructed per condition, and even the phenomena that are investigated with each experiment.

## 1.2 Two types of unreliability, and two types of criticisms of traditional methods

Although the word *unreliable* has an intuitive meaning, 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:

[1] There is a third type of unreliability: a positive result in the opposite direction. Such results are sometimes described as sign-reversal 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 & Almeida *in pres*, and Sprouse, Schütze, & Almeida *submitted* found only two examples out of 511 phenomena tested).

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 unacceptably high number of false positives, an unacceptably high number of false negatives, or both. With this definition of reliability, it is possible to classify criticisms of traditional acceptability judgment methods into two types. The first type of criticism claims that traditional methods lead to an unacceptably high false positive rate, and relatedly, that formal experiments would lead to a lower false positive rate (e.g., Ferreira 2005, Wasow & Arnold 2005, Gibson & Fedorenko 2010b). The second type of criticism claims that traditional methods lead to an unacceptably high false negative rate, 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 & Almeida (*in press*), who formally tested all of the data points in a popular syntax textbook (Adger 2003) and found that the maximum possible false positive rate (i.e., following the critical literature in 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 5%. The strongest possible interpretation of these results by a critic of traditional methods would be that traditional methods yield false positives between 2-5% of the time. However, it is important to note that this interpretation *assumes* that the false negative rate of *formal* experimental methods is 0%. If any of the negatives returned by the formal experiments are false negatives, then the estimated false positive rate for traditional methods would be lower than 2-5% (see Sprouse et al. *submitted* for discussion).

When it comes to claims of the second type, there are currently to our knowledge no systematic studies that have reported (i) estimates of the true *false negative rate* for traditional methods, nor (ii) estimates of the true *false negative rate* for the types of formal methods advocated in the literature. The reason for this gap is relatively straightforward: calculating a true *false negative rate* requires collecting a list of all of the true differences in the world, and comparing it to the list of currently detected differences. If we had a list of all of the true differences, we would not need experiments any longer. One could try to estimate these lists from the existing literature, but this is fraught with difficulties. For example, one could use formal methods to systematically re-test any negative results that have been reported using traditional methods in order to identify potential false negatives. This procedure has been conducted on a few topics so far, with some previously unreported differences being reported (e.g., Keller 2000, Featherston 2005a, 2005b, Sprouse et al. 2011). A major problem with re-tests of this sort is that negative experimental results are inherently ambiguous, because experimental detection is jointly determined by two different properties: the sample size and the size of the difference under investigation. We can never know why certain differences were unreported in earlier studies: one possibility is that traditional methods failed to detect the differences, but another possibility is that the original authors did detect the differences, but decided that they were too small to be theoretically relevant (theoretical relevance is sometimes called *practical significance* in the statistics literature to highlight this logical shortcoming of *statistical significance*, see Shaver 1993, Cohen 1994, and Nickerson 2000; see also Fanselow 2007 and Myers 2009).

While it may not be possible to estimate the true *false negative rate* of any experimental method, there is a notion in the statistics literature that provides a closely related type of information and can in principle be calculated or estimated for any given experiment: *statistical power*. Statistical power is the likelihood of an experiment to detect a true positive if a difference truly exists between the conditions being tested (see section 2). Statistical power can be straightforwardly calculated for various types of experiments, as long as one specifies the effect sizes of the phenomena of interest and the number of participants recruited. As such, statistical power provides important information about false negatives (i.e., failure to detect a true positive), and can be used as a proxy measure for the false negative rate in the field. Unfortunately, given that there is neither a single definition of traditional methods in the field, nor a culture of reporting information about the collection methods in published syntax articles, it is virtually impossible to calculate the *statistical power* of traditional methods directly. However, it *is* possible to estimate the statistical power of the formal methods used in Experimental syntax, which may offer a solution to the problem of estimating statistical power for traditional methods - a topic we turn to next.

## 1.4 A comprehensive investigation of statistical power of formal methods.

One way to circumvent the impossibility of calculating a direct measure of statistical power for traditional methods is to conduct a comprehensive assessment of statistical power for every possible type of acceptability judgment experiment. By calculating statistical power estimates for the full spectrum of sample sizes, effect sizes, and judgment tasks, there is no longer any need for a single consensus definition of what constitutes "traditional methods" in syntax. Instead, researchers can use the comprehensive statistical power information provided in this study to evaluate whether one or more of the combinations of experimental properties are good proxies for their own definition of traditional methods, to see which types of experiments are likely well-powered (i.e., have acceptably low false negative rates in the long run), and which are under-powered (i.e., have unacceptably high false negative rates in the long run). In this way, the previously intractable question of calculating the statistical power of traditional methods reduces to the more tractable question of whether various definitions of traditional methods likely fall in the class of well-powered experiments, or whether they fall in the class of under-powered experiments. Although researchers may still differ on the answer to this question depending on their preferred definition of traditional methods, a comprehensive analysis of statistical power for all types of judgment experiments at the very least allows such disagreements to be expressed in concrete terms (i.e., as explicit differences between experiment types) and makes the empirical consequences (i.e., the precise differences in statistical power) clear.

Assessing the statistical power for the full spectrum of sample sizes, phenomena, and tasks is not a small endeavor (although not as resource-intensive as a case-by-case retest of every data point in syntax). The first step is to identify a set of phenomena with "true differences" that span the full range of effect sizes that are likely to arise in syntactic theory. Sprouse, Schütze, & Almeida (*submitted*) randomly selected 150 two-condition phenomena (300 sentence types) from

*Linguistic Inquiry* 2001-2010 in their assessment of the false positive rate of traditional methods. From this set of 150 phenomena, we chose a subset of 50 phenomena (100 sentence types), which, based on their previous publication in *Linguistic Inquiry* and their use in the Sprouse et al. study, are now known to be observable using both traditional methods and formal experiments. The 50 phenomena were chosen such that they span the most informative range of effect sizes (see section 3). We then tested these 50 phenomena on four distinct samples of 144 participants in order to assess every possible sample size of interest to syntacticians (see section 4). Each of the four samples of participants completed a different acceptability judgment task: magnitude estimation (ME), Likert scale (LS, 7-point in this study), yes-no (YN), and two alternative forced choice (FC). 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 (see section 5). This type of quantitative comparison provides a wealth of information about the nature of acceptability judgments that has historically been unavailable: it allows us to empirically estimate the statistical power of each task for every possible effect size in syntax (see section 6), which in turn allows us to assess the statistical power of previous studies, to better predict the statistical power of future studies, and most importantly in light of the historical criticism of traditional methods in syntax, to assess the likelihood of a lack of detection of true differences for different definitions of traditional methods (see section 7).

Before moving on to a discussion of the experiments, it should be noted that there have been a few high profile comparisons of acceptability judgment tasks that have previously touched upon the topic of statistical power (though not all use that precise term). For example, in their seminal introduction of magnitude estimation to the field of syntax, Bard et al. (1996) presented a comparison of the results between ME and LS for several sentence types in order to demonstrate that the continuous response scale of ME tasks allows participants to report more levels of acceptability than the ordinal response scale of LS tasks. Weskott & Fanselow (2011) presented a comparison of ME, LS, and YN results for three phenomena (two two-sentence and one three-sentence phenomena) in order to assess the claim that ME is more sensitive than LS and YN tasks. They found that at sample sizes of 24 and 48 participants all three tasks yield statistically significant results for those particular phenomena. Similarly, Bader & Haüssler 2010 presented a comparison of ME and YN tasks for 16 sentence types (forming one 2x2 factorial design and two 2x3 factorial designs) in order to construct a signal detection model of acceptability judgments. In the process they found that the two tasks yielded similar patterns of acceptability, and at sample sizes of 24 and 36 participants, both tasks yielded statistically significant results for those phenomena. The present study builds on and extends these previous results in several ways. First, instead of investigating a single sample size and using a categorical criterion (statistical significance or not), the current studies use resampling simulations to assess the statistical power for a large range of possible sample sizes (5 to 100 participants), which we believe covers every possible sample size that linguists are likely to encounter in evaluating or constructing experiments. Second, instead of focusing on a few phenomena of a particular effect size, the current studies investigate the full range of effect sizes that were observed in the random sample of *Linguistic Inquiry* data points in Sprouse et al. *submitted*. This allows for a nearly exhaustive assessment of the interaction of statistical power with the effect size of the phenomena under consideration. Similarly, although the set of 50 phenomena tested here were not chosen at random, they were selected without theoretical bias from a random sample of 150 phenomena from LI (2001-2010), which suggests that the results will be relatively representative of cutting edge syntactic research. Third, the resampling simulations (over 20 million randomly selected samples) allow us to empirically estimate the statistical power for every combination of effect size, sample size, and task, which provides more information than the categorical question of whether a given experiment yielded a significant result. Finally, the simultaneous comparison of all four acceptability tasks across a full range of effect sizes and sample sizes allows us to evaluate almost any conceivable experiment, from previously published traditional collection studies to the design of future formal experiments.

## 2. Statistical power

As mentioned in section 1, the *statistical power* of different experiments is a good proxy for the false negative rate because statistical power is the likelihood that an experiment will detect a difference between conditions when one truly exists. Statistical power is normally expressed as a percentage. For example, if an experiment has an 80% probability of detecting a difference when a difference truly exists, then we say that the experiment has 80% statistical power.

The notion of statistical power is 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 (as we will see, some tasks are more sensitive than others), the size of the sample of participants (larger samples are more likely to yield significant results, e.g. Meehl 1967, Cohen 1994, Nickerson 2000), 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 (all else being equal, 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 will lead to more false negatives). In order to profitably compare the power of 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 (ME, LS, YN, FC), we manipulated the samples size by using resampling simulations (e.g., a single sample of 140 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 spectrum of effect sizes observed in the random sample of phenomena from *Linguistic Inquiry* (Sprouse et al. submitted), and we held the maximum rate of false positives that we are willing to tolerate at the consensus level of 5% (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). We will make use of Cohen's power suggestion in this study because we believe that the "best practice" guidelines of experimental psychology provide a common vocabulary for evaluating syntactic methods relative to other domains of cognitive science. However, it is important to note that Cohen's suggestion may not ultimately be the most appropriate power level for syntactic theory, as it rests on the assumption that false positives are 4x more costly than false negatives. Many syntactic theories seek to capture both the differences between conditions (positives) and the invariances between conditions (negatives). This means that false negatives may be erroneously incorporated into syntactic theories as invariances. Arriving at a 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.[2]

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

In order to maximize the representativeness of these results, we chose 50 phenomena from the larger set of 150 phenomena that were randomly sampled from *Linguistic Inquiry* 2001-2010 (Sprouse et al. *submitted*) as follows. First, we calculated a measure of effect size known as Cohen's *d* (Cohen 1988) for the full 150 phenomena. Cohen's *d* is mean difference between conditions divided by the mean difference between the standard deviations of each condition. In other words, Cohen's *d* is the ratio of the mean difference to the mean standard deviation. Cohen's *d* is a *standardized* measure of effect size, which means it allows us to compare any effect size to any other, even if the two effects are measured on different scales (e.g., reading times and acceptability judgments). 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

[2] It should also be noted that 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, which allow direct testing of invariances (for a review see Gallistel 2009, for an introduction to performing Bayesian statistical tests see Kruschke 2011).

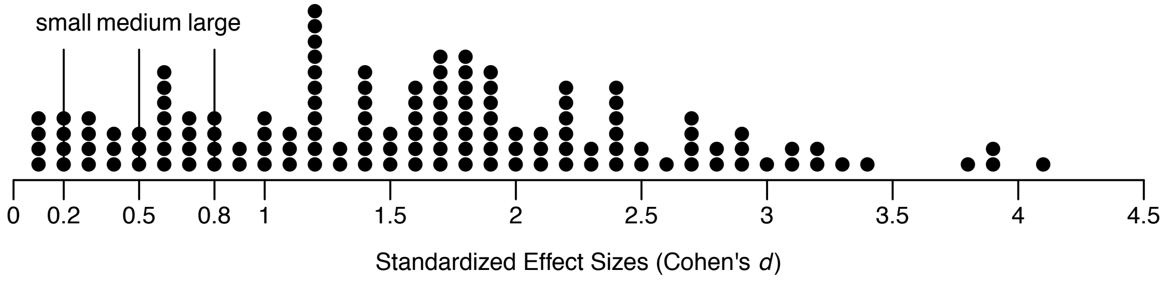
0.8 is considered a “large” effect. Here is what Cohen (1992) said about the intent behind these criteria:

> 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)

To make the idea of effect sizes more tangible, we list example sentences for each of the phenomena tested in this study along with their Cohen’s *d* in the appendix.

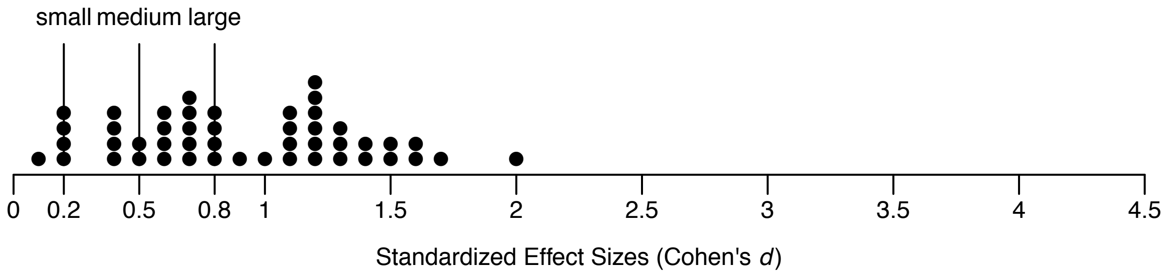
After calculating effect sizes (Cohen’s *d*) for all 150 phenomena, we then plotted the distribution of effect sizes for the 139 phenomena that were significant in the Sprouse et al. (*submitted*) study:

Figure 1: 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, & Almeida (*submitted*) using the magnitude estimation task (effect sizes were derived from the results of Sprouse et al. *submitted*).



The distribution of this random sample of 139 phenomena suggests that approximately 13% of the phenomena from LI (2001-2010) are considered “small” (*d* < 0.5), 8% are considered "medium" (0.5 < *d* <0.8), and 79% are considered "large" (*d* > 0.8), with a margin of error of ±5%. Because the distribution of effect sizes in LI spans a range that includes extremely large effect sizes (many are greater than 2), and because very large effect sizes are likely to lead to a ceiling effect in statistical power (100%), we decided to restrict our subset of 50 to the smaller half of the range (0 < *d* < 2). This 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). After running the experiments we noticed potential confounds in the materials for three of phenomena (see Sprouse et al. *submitted* for discussion), which reduced our sample of phenomena to 47. The distribution of the effect sizes for the remaining 47 phenomena from LI (2001-2010) are presented in Figure 2. A full list of the phenomena with example sentences are provided in the appendix.

Figure 2: The distribution of effect sizes (Cohen's *d*) for the 50 phenomena from *Linguistic Inquiry* (2001-2010) that were chosen for the experiments here (effect sizes were derived from experiment 1 discussed in section 4).



## 4. The experiments

### 4.1 Participants

144 participants completed each of the four experiments (576 participants total). Participants were recruited online using the Amazon Mechanical Turk (AMT) marketplace, and paid $3.00 for their participation in the ME experiment, $2.50 for the LS and YN experiments, and $2.00 for the FC 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. No participants were excluded from the ME and FC experiments based on these questions. However, 4 participants were excluded from the LS experiments, and 5 participants were excluded from the YN experiment 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).

## 4.2 The tasks

In the ME 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* (which we set at 100). Participants are asked to indicate the acceptability of target sentences as a multiple of the acceptability of the standard by providing a rating that is a multiple of the modulus. However, it should be noted that recent research suggests that participants do not actually use the standard to make ratio judgments of the target sentences (Sprouse 2011b). In the (7-point) LS task, each target sentence is presented with a series of 7 radio buttons labeled 1-7, with 1 labeled "least acceptable" and 7 labeled "most acceptable". Participants are asked to use the radio buttons to indicate their acceptability judgments. In the YN task, each target sentence is presented with a pair of radio buttons labeled "yes" and "no". Participants are asked to use the radio buttons to indicate whether the sentence is acceptable or not. In the (two-alternative) FC task, target sentences are presented in vertically arranged pairs, with each sentence in the pair followed by a single radio button. Participants are asked to indicate which of the two sentences is more acceptable by selecting the radio button next to that sentence. In the current FC experiment, the pairs were lexically matched so as to form minimal pairs that varied only by the syntactic property of interest.

## 4.3 Materials

The materials for all four experiments were identical to the materials constructed for the original Sprouse et al. (*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). This led to 8 lexically matched sentence sets for each phenomenon. 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.

For the ME, LS, and YN experiments, 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 100 pseudorandomized items. Six additional "anchoring" items (two each of acceptable, unacceptable, and moderate acceptability) were placed as the first six 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 106 items long.

For the FC experiment, 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'). Next, two copies of each list were created, resulting in 16 lists. Finally, the order of the pairs in each list were randomized, resulting in 16 surveys containing 50 randomized and counterbalanced pairs (100 total sentences).

## 4.4 Presentation

For the ME experiment, 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 six unmarked anchor items did serves as a sort of unannounced sentence practice. There was also no explicit practice for the LS, YN, and FC experiments, as these tasks are generally considered relatively intuitive. 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.

## 4.5 The number of judgments per condition

Each participant rated only one token of each condition in the ME, LS, and YN experiments, and only one pair per phenomenon in the FC experiment. From the perspective of both traditional collection methods and more 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 lowest limit of possible experimental designs. This means that the power estimates that we present here 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 (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 & Almeida 2012, Schütze & Sprouse 2012), 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 2010a, 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 also allows us to address the concerns of critics of traditional methods directly.

## 5. The resampling simulations

In order to empirically estimate the statistical power of each experiment type for each phenomenon at every sample size between 5 and 100 participants, we performed resampling simulations on each sample. In essence, these resampling simulations treated our large samples (N=144, N=140, N=139, N=144) as full populations, and sampled from them to estimate 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, 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 ME and LS experiments, we used two-tailed paired *t*-tests; for the YN and FC experiments, 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 that particular phenomenon 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 that phenomenon. Finally, we can repeat this procedure for all 47 phenomena to derive power relationships (operationalized as empirically estimated rates of detection) for effect sizes between 0 (very small) and 2 (very large) in LI (2001-2010). Even though resampling simulations of this sort are relatively rare in the experimental syntax literature, they are relatively common in other areas of experimental psychology. Resampling simulations form the basis of several approaches to statistical significance testing, such as the bootstrap, randomization, and permutation tests, and as such, their properties are well understood (e.g., Efron and Tibshirani 1993, Edgington and Onghena 2007).

It should also be noted that we intentionally chose to use single-random-effect frequentist tests like the *t*-test and the sign test (McNemar's exact test) instead of multiple-random-effects tests like linear mixed-effects models. The reasons for this are straightforward. Fitting linear mixed-effects models requires substantially more computational time than standard statistical tests. This is not a problem if the number of analyses is small, but the current studies required 20 million simulations. Using standard statistical tests, these simulations took more than one week of computation time; if we had used linear mixed-effects models, they would likely have taken more than a month. Given this cost, it only makes sense to use linear mixed-effects models if the shortcomings of standard tests outweigh the costs. In this case, we do not believe they do. Linear mixed-effects models are designed to accommodate a specific analysis decision (treating items as a random effect) that is not necessarily warranted by the design of the current experiments: our items were constructed to be representative of the set of possible items (not random, which could include very unrepresentative items), our items were lexically matched across conditions, and participants only rated one item per condition (Wike and Church 1976, Cohen 1976, Keppel 1976, Smith 1976, Wickens and Keppel 1983, Raijmaakers 2003). Furthermore, the potential risk that is supposedly addressed by treating items as random samples of a larger population, (which, again, does not match our experimental design) is that, without it, standard statistical tests could have an inflated false positive (Type I error) rate when materials are random samples of larger populations (Clark 1973) but are still treated as fixed effects. Given that all of the phenomena tested are known to be true positives, there cannot be an increase in false positives, but rather an increase in true positive detection – otherwise known as an increase in power. So the only possible concern is that our choice of statistical tests may overestimate the power for these tasks and phenomena. Again, in the unlikely event that this is true for our experimental design, previous simulations of the difference between tests that only include subjects as a random effect (such as the *t*-test and ANOVA) and tests that include items as a random effect (such as minF′ and linear mixed effects models) suggest that this difference will be very small for designs like ours that use well-balanced and matched items (e.g., Wickens and Keppel 1983).

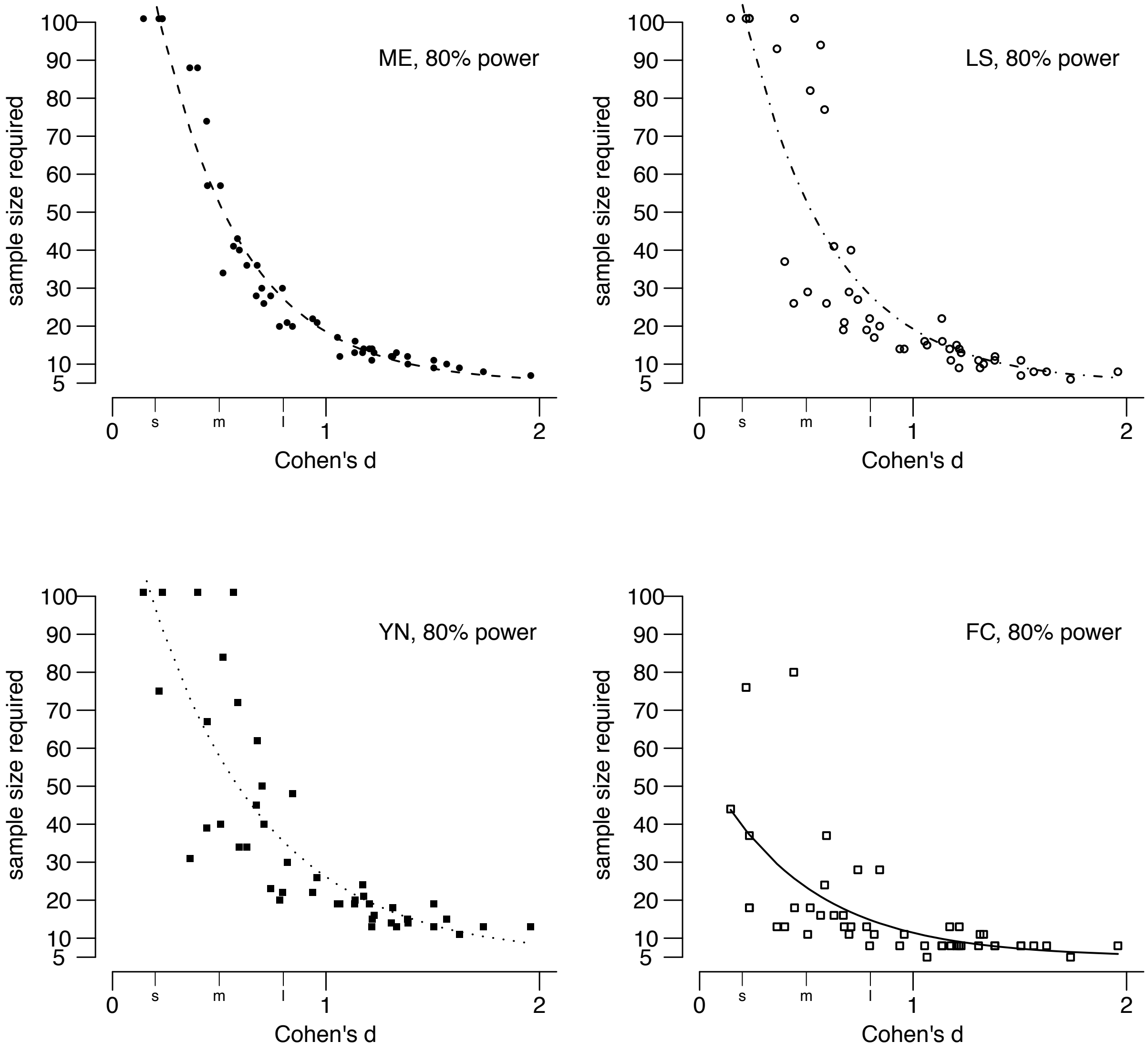
## 6. The statistical power of acceptability judgment experiments

There are several ways to summarize the data contained in the 20 million resampling simulations conducted in this study. In this section we will present several alternate summaries, with each highlighting different aspects of the relationship between effect size, sample size, power, and task.

### 6.1 Highlighting the relationship between effect size and sample size

One way of visualizing statistical power is to choose a target level of statistical power, such as the 80% target rate suggested by Cohen (1962, 1988, 1992), and then identify the sample size required to reach that power rate for each of the 47 phenomena for each of the task types. We can plot this relationship between effect size (Cohen's *d*) on the x-axis and the sample size required to reach 80% power on the y-axis. Figure 4 presents four graphs of this type, one for each task. The effect sizes were taken from the ME experiment to provide a standard effect size for each phenomenon across all four tasks. We have added non-linear trend lines using the NLS (non-linear least squares) function in R (using a logarithmically decreasing function) to better illustrate the relationship between effect size and sample size. By highlighting sample size as a function of effect size for a specific target power level, these figures provide an intuitive sense of power – one could simply look at the effect size of interest (e.g., 0.8) and see what sample size would be required for a well-powered experiment (e.g., 25 for ME, 15 for FC).

Figure 4: The sample size (y-axis) required to reach 80% power as a function of effect size (x-axis) and task (the panels) for all 47 phenomena from LI (2001-2010). The effect sizes along the x-axis (Cohen's *d*) are from the ME experiment for all four tasks. The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font.



We can also use these figures to compare the relative power of the four tasks. Figure 5 is identical to Figure 4 in all respects, except that the curves for the four tasks are plotted together, and we have removed the points (leaving only the non-linear trends) to increase legibility (graphs containing the full sets of data points will be available on the first author's website as supplementary materials).

Figure 5: The sample size (y-axis) required to reach 80% power as a function of effect size (x-axis) and task (separate lines) for all 47 phenomena from LI (2001-2010). The effect sizes along the x-axis (Cohen's *d*) are from the ME experiment. The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font.

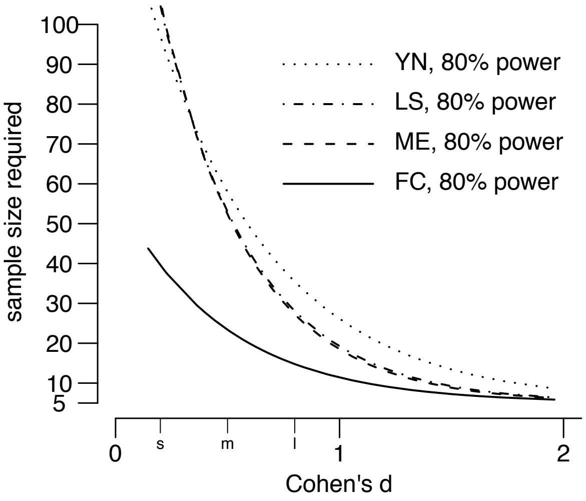
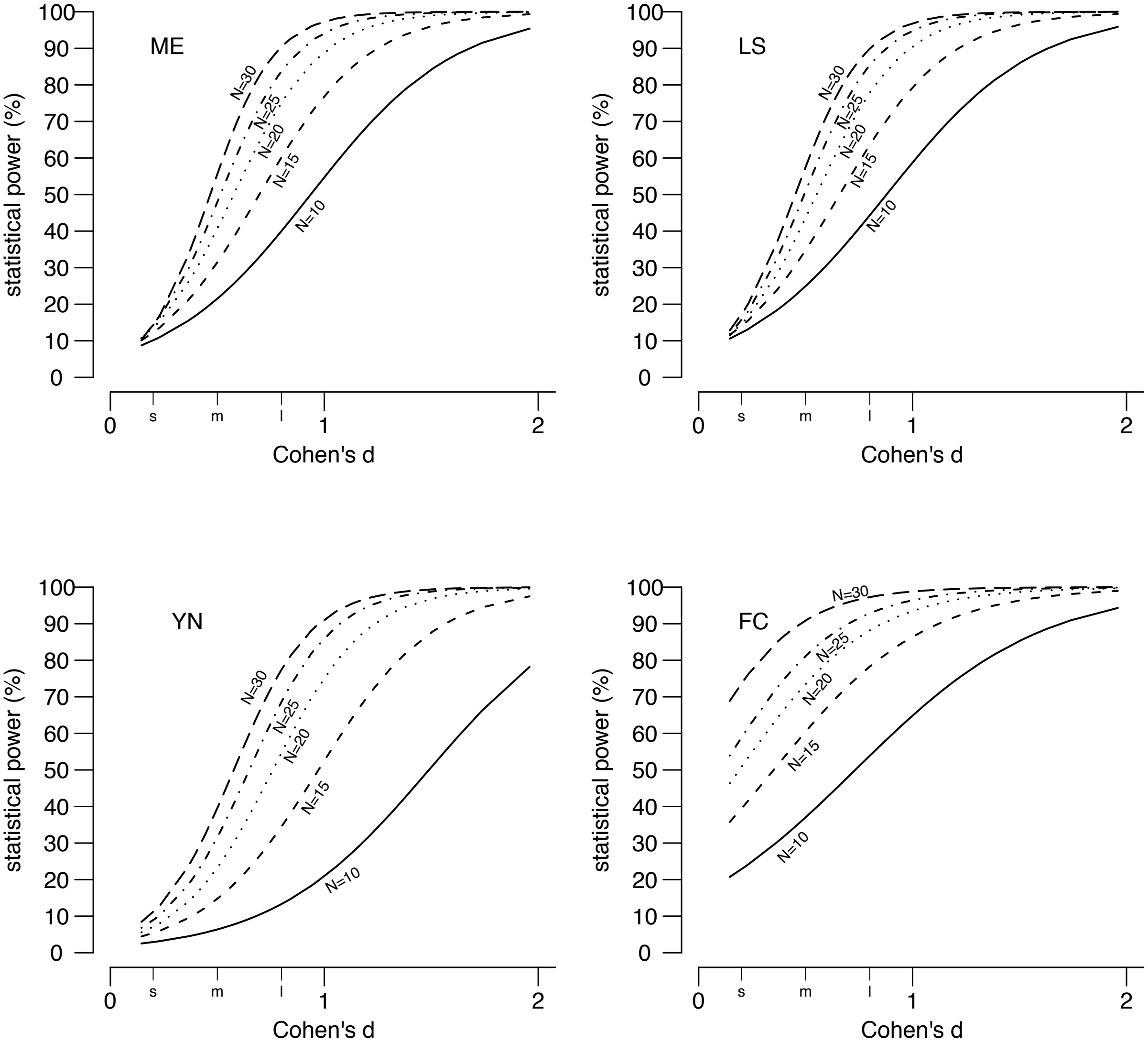


Figure 5 reveals several interesting patterns. First, the sample size required to reach 80% power for ME and LS appear to track each other very closely, suggesting that there is little difference in power between these two tasks. This accords well with the comparison of ME and LS for several phenomena in German by Weskott & Fanselow (2011), as well as the conclusion by Sprouse (2011b) that the inability of participants to meet the cognitive assumptions of ME likely means that they treat ME as a modified LS task. There also appears to be a power disadvantage for YN, as the sample size required to reach 80% power is highest for it. This is not unexpected: the YN task only allows two response options (yes and no). If the two conditions for a phenomenon are on the same side of the category boundary (both yes or both no), or not clearly on opposite sides (weakly yes and weakly no), then the YN task will be less sensitive to the difference. The most striking pattern in Figure 5 is the relatively high power of the FC task: for nearly every effect size, and especially for smaller effect sizes, the FC task requires a substantially smaller sample to reach 80% power. Again, this is not unexpected, as the FC task is explicitly designed to reveal differences between conditions by presenting the two conditions of each phenomenon as a pair, and forcing participants to report a difference between the two by choosing one condition over the other. The other tasks take an indirect approach to detecting differences: each condition is rated along a response scale, and then those responses are compared to one another (see also Gigerenzer & Richter 1990 and Gigerenzer, Krauss & Vitouch 2004 for a discussion of the comparative merits of forced-choice tasks over simple ratings tasks, and Labov 1996 for the comparative merits of forced-choice tasks over simple YN tasks).

### 6.2 Highlighting the relationship between effect size and statistical power

Another useful way of summarizing these results is to identify a target sample size, such as 20 participants, and then calculate the statistical power achieved at that sample size for each of the 47 phenomena for each of the task types. We can then plot the relationship between effect size (Cohen's *d*) on the x-axis and the statistical power achieved on the y-axis. Figures 6 presents four graphs of this type, one for each task, with five sample sizes plotted per task (10, 15, 20, 25, and 30 participants). The effect sizes were again taken from the ME experiment to provide a standard effect size for each phenomenon across all four tasks. We have also again opted to display the non-linear trend lines (sigmoidal this time) without plotting the raw data points in an effort to reduce visual clutter. By highlighting statistical power as a function of effect size for a series of sample sizes, these figures provide an intuitive sense of the effect of incrementally increasing sample size – one can simply look at the effect size of interest (e.g., 0.8) and see how much power would be achieved at each sample size.

Figure 6: The statistical power (y-axis) achieved as a function of effect size (x-axis), sample size (the lines), and task (the panels) for all 47 phenomena from LI (2001-2010). The effect sizes along the x-axis (Cohen's *d*) are from the ME experiment for all four tasks. The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font.

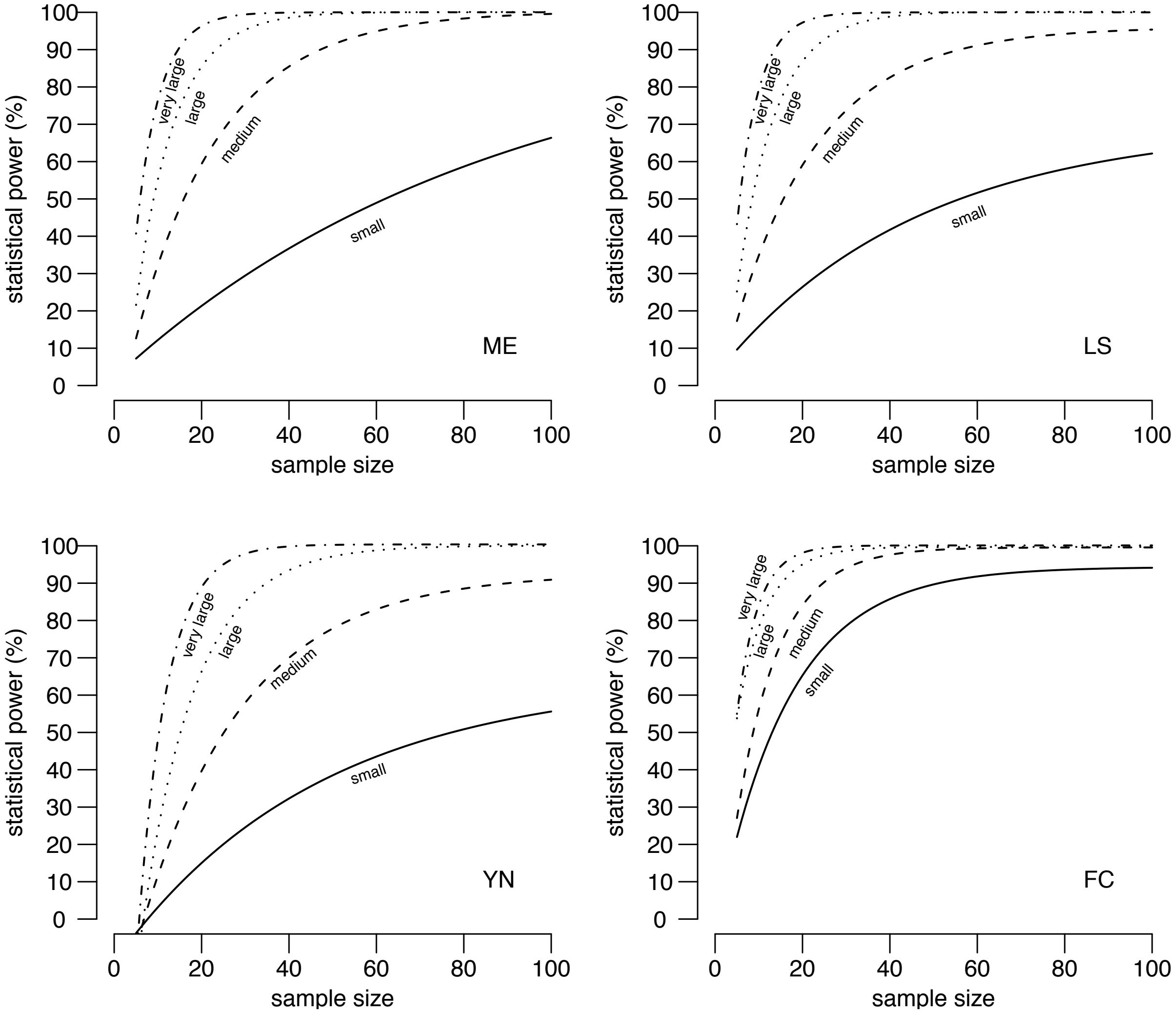


Given that this is simply a different way of visualizing the same information, it is not surprising that the patterns in Figure 6 are similar to those in Figures 4 and 5. The ME and LS power curves are nearly identical, suggesting that there is no noticeable power difference between these two tasks. The YN curves appear to be the weakest of the group, especially at smaller sample sizes (e.g., 10 and 15), suggesting again that the YN task may be slightly less sensitive than the ME and LS tasks at low sample sizes. The FC curves once again appear substantially more powerful than the other tasks, again suggesting that FC may be the most sensitive task for detecting differences between conditions. One additional piece of information in Figure 6 that is worth noting is the relative increase afforded by each 5 participant increase in sample size. For all tasks, the largest increase is between 10 and 15, with successively smaller increases as sample sizes increase. In other words, extra participants are more valuable at smaller sample sizes than they are at larger sample sizes. This accords well with the diminishing returns suggested by the central limit theorem.

### 6.3 Highlighting the relationship between sample size and statistical power

One final way of summarizing this data is to collapse a range of effect sizes together, such as all of the "small" effects (Cohen's *d* between 0.2 and 0.5), and then calculate the statistical power achieved at every sample size (5 to 100) for each of the 47 phenomena for each of the tasks. We can then plot the relationship between sample size on the x-axis and the statistical power achieved on the y-axis for that range of effect sizes. Figures 7 presents four graphs of this type, one for each task, with four ranges of effect sizes: small (0.2-0.5), medium (0.5-0.8), large (0.8-1.1), and very large (>1.1). The effect sizes were again taken from the ME experiment to provide a standard effect size for each phenomenon across tasks. We have also again opted to display the non-linear trend lines without the plotting the raw data points in an effort to reduce visual clutter. By highlighting statistical power as a function of sample size for a series of effect sizes, these figures provide an intuitive sense of the effect of varying sample size – one can simply look at the effect sizes of interest (e.g., medium) and see how much power would be achieved at each sample size.

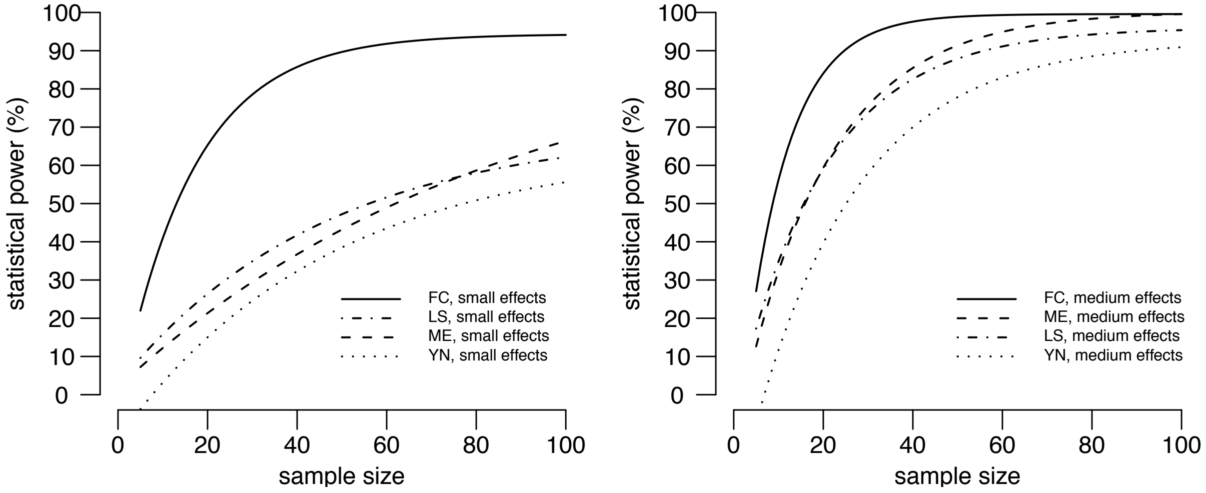
Figure 7: The statistical power (y-axis) achieved as a function of sample size (x-axis), ranges of effect size (the lines), and task (the panels) for all 47 phenomena from LI (2001-2010). The effect sizes are from the ME experiment for all four tasks.

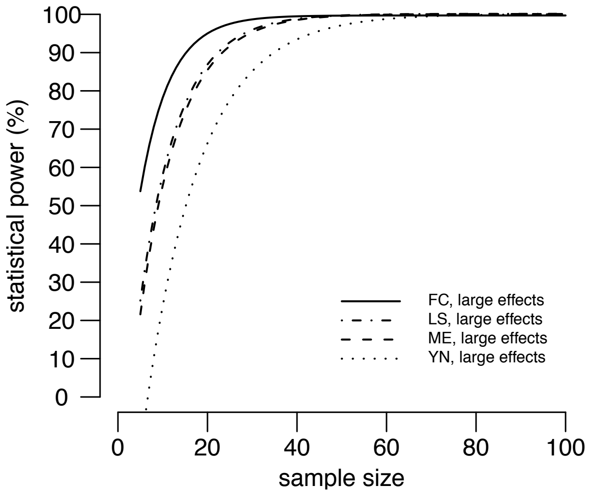


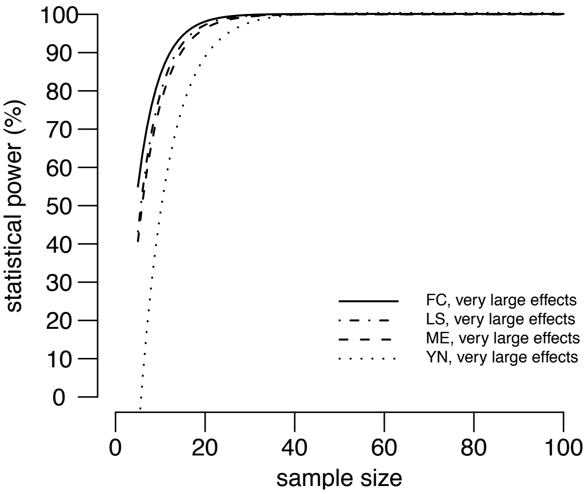
By highlighting the different power curves for each range of effect sizes, Figure 7 reveals an interesting difference among the effect sizes: the difference in power between small effects and medium effects at any given sample size is larger than the difference in power between medium and large effects at the same sample size. Similarly, the difference in power between medium and large effects at any given sample size is larger than the difference in power between large and very large effects. This means that even though the difference between small and medium effects is a Cohen's *d* of 0.3, and the difference between medium and large effects is also a Cohen's *d* of 0.3, the scale is non-linear with respect to sample size and power – larger samples are needed to increase power for small effects than are needed to increase power by the same amount for larger effects (at least until the asymptote). Again, this is not unexpected, as this information is visible in the previous figures as well, and the fact that larger effects require smaller samples to achieve adequate power is well understood. Nonetheless, Figure 7 provides a clear demonstration of the different relationships between sample size and power across the range of effect sizes.

In addition to comparing all four effect size ranges for a single task in each panel, we can also compare all four tasks for single effect size range in each panel in order to directly compare the power of the tasks:

Figure 8: The statistical power (y-axis) achieved as a function of sample size (x-axis), tasks (the lines), and effect sizes (the panels) for all 47 phenomena from LI (2001-2010). The effect sizes are from the ME experiment for all four tasks.







Much like the previous direct comparisons across tasks, the relationship between sample size and statistical power reveals a substantial advantage for the FC task over the other tasks. This advantage is largest for small effect sizes, and decreases as effect sizes increase (with effectively no advantage for FC with very large effect sizes).

### 6.4 Using this information to evaluate previous and future experimental designs

Before discussing these results in relation to criticisms of traditional judgment collection in syntax, it may be worthwhile to discuss how these results can be used by practicing syntacticians to evaluate previously published studies and plan future studies. By testing 47 phenomena that (i) span a large range of effect sizes (Cohen's *d* of 0-2), (ii) were taken from a random sample of phenomena from the most recent ten years of *Linguistic Inquiry*, and (iii) were tested using all four acceptability judgment tasks, we can be fairly confident that the statistical power results discussed in this section are representative of the data collected in the field of syntax. As such, these results can be consulted to evaluate the statistical power of most experimental designs syntacticians are likely to employ. For previously published experimental studies this is particularly straightforward: one can simply calculate the Cohen's *d* for the phenomena of interest from the means and standard deviations reported in the study to determine how the statistical power would change as a function of sample size. For studies that do not have published means and standard deviations, or for planning a new study where the means and standard deviations are unknown, the situation is a bit more complicated. One possible approach would be to informally compare the phenomena with unknown effect sizes to the known phenomena in this study, perhaps by asking a few colleagues whether they believe the difference is larger or smaller than those in the appendix. Although this method is not precise, it should be possible to arrive at a relatively accurate, albeit coarse, effect size (i.e., small, medium, large, very large) with very few such informal judgments, at which point Figures 7 and 8 could be used to estimate statistical power by sample size. In this way, these results can be used to assess whether previously published or planned studies have sufficient statistical power to detect the phenomena of interest. Furthermore, given that the 47 phenomena are a subset of a randomly chosen sample of syntactic phenomena and span a large range of effect sizes, they can be used as filler items in studies that require items of a specific rating or phenomena of a specific effect size to properly balance out an experiment. Example items that can be used this way are provided in the appendix.

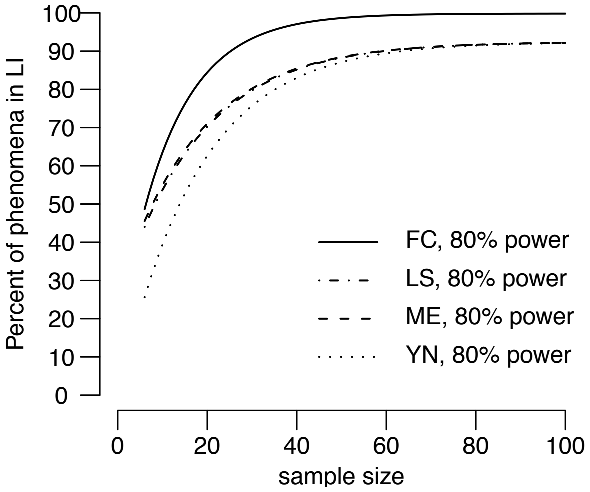
## 7. Assessing the reliability of traditional judgment collection methods

As discussed in section 1, the question of whether traditional collection methods have led to an unacceptably high number of false positives has been investigated extensively (see Sprouse & Almeida *in press* and Sprouse et al. *submitted*). The question of whether traditional methods have led to an unacceptably high number of false negatives is a more difficult question, as the two methods generally available for assessing false negatives (retesting previous experiments or calculating statistical power) are either logically or logistically problematic. One way to circumvent these problems is to assess the statistical power for the most widely used types of *formal* acceptability judgment experiments. By calculating statistical power estimates for the full spectrum of sample sizes, phenomena, and judgment tasks, the field can evaluate any particular definition of traditional methods to see whether they are reliable with respect to false negatives, or unreliable with respect to false negatives. In this way the previously intractable question of calculating the false negative rate of traditional methods reduces to the more tractable question of whether traditional methods likely fall in the class of well-powered experiments, or whether they fall in the class of under-powered experiments. Although researchers may still disagree on the answer to this question, the results presented in this study make the terms (sample sizes, tasks, etc.) and consequences (statistical power) of those disagreements clearer.

## 7.1 The class of well-powered experiments

One way to identify the class of reliable experiments is to choose a target power level, say Cohen's (1988) suggestion of 80%, and then calculate the percentage of phenomena in *Linguistic Inquiry* (2001-2010) that each task would detect at that power level at each sample size (using the distribution of effect sizes sampled by Sprouse et al. *submitted*). Figure 9 presents the results of such a comparison.

Figure 9: The percentage of phenomena (y-axis) in *Linguistic Inquiry* (2001-2010) that would be detectable with 80% power by each task (lines) as a function of the sample size (x-axis) assuming only 1 judgment per participant per condition. The accompanying table specifies the percentages precisely. Based on the size of the random sample, these numbers have a margin of error of ±5 (Sprouse et al. *submitted*).



| Sample | FC | YN | ME | LS |
|---|---|---|---|---|
| 6 | 35% | 15% | 29% | 28% |
| 10 | 69% | 39% | 58% | 57% |
| 15 | 78% | 58% | 70% | 69% |
| 20 | 83% | 64% | 74% | 74% |
| 25 | 89% | 72% | 77% | 76% |
| 30 | 91% | 75% | 79% | 79% |
| 35 | 94% | 77% | 81% | 81% |
| 40 | 95% | 80% | 82% | 82% |

Given the information in Figure 9, the question of what constitutes a reliable method reduces to a question of how many phenomena should be detectable with a given method. If one believes that a reliable method should have an 80% chance to detect 70% of phenomena in LI, then FC experiments with a sample of 10 participants (and only one judgment per participant) would be reliable. Using the same definition, YN experiments would require around 25 participants (only one judgment per condition) to be reliable. If one believes that a reliable method should have an 80% chance of detecting 80% of phenomena in LI, then FC experiments with 15-20 participants (again, only one judgment per participant) would be reliable, whereas YN experiments would require 40 participants to be reliable. It is important to note that these are minimum estimates for the sample sizes in question because each participant only provided one judgment per condition. Increasing the number of judgments per participant per condition will increase the power of the experiment, and therefore either reduce the sample size required for the same empirical coverage, or increase the empirical coverage for a given sample size. It is also interesting to note that the ME and LS tasks require larger sample sizes than the FC task to achieve the same empirical coverage at a given power level (e.g., 30 for ME versus 15-20 for FC for coverage of 80% of phenomena at 80% power). ME and LS are commonly recommended by critics of traditional methods – a suggestion that, if adopted uncritically, could lead to lower-powered experiments than might have occurred with FC.

Although the definition of a reliable experiment will likely vary from researcher to researcher, we believe it is possible to infer a consensus definition of reliable experiments from the common practices of formal experimentalists. In our experience, formal experimentalists in syntax tend to recruit between 25 and 40 participants per experiment. To the extent that our experiences are representative, it would suggest that formal experimentalists are likely comfortable with experiments that would detect 76%-82% of the phenomena in LI (at 80% power). Under the default assumption that traditional methods and formal methods should be held to the same standards of reliability, this would suggest that formal FC experiments with 15 participants (each providing one judgment per condition) would be considered reliable by formal experimentalists. Although the reporting standards in the field make it difficult to verify, we believe that the long path taken by most syntax projects from initial data collection by the author to peer review and publication (at which point the phenomena under investigation becomes available for scrutiny of the whole field), ensures that data points published in the literature are judged substantially more than 15 times (at least for languages that are well-represented among professional linguists). To the extent one wishes to generalize the power of formal FC experiments with one judgment per participant to informal FC experiments, this would suggest that informal FC experiments are in fact a reliable tool for syntacticians. Moreover, to the extent that informal FC experiments are widely employed in the literature, this would suggest that a large portion of traditional experiments are reliable (for example, Bard et al. 1996 note that the examples they discuss from Haegeman 1991 were presented in forced-choice pairs, and Myers 2009b complains that linguists overuse forced-choice pairs over full-factorial designs).

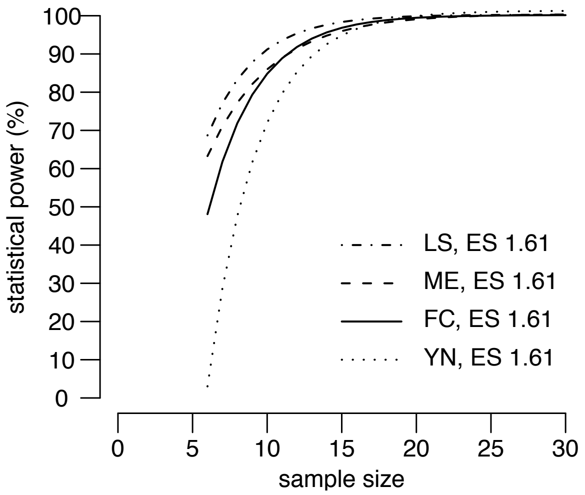
### 7.2 A comparison with other domains of experimental psychology

The fact that FC tasks with relatively small samples can nonetheless be well-powered may be surprising to some readers given the prevailing criticisms of traditional methods in the literature. However, some readers may be unsurprised by this result, as it has been claimed that the effect sizes of phenomena studied in syntactic theory are substantially larger than the effect sizes studied in other domains of experimental psychology (the existence of this claim is noted in Schütze 1996 and Gibson and Fedorenko 2010b but it is difficult to track down the original sources). We have already partially substantiated this claim in Figure 1: the distribution of randomly sampled phenomena from LI (2001-2010) suggests that approximately 13% of the phenomena from LI (2001-2010) are considered "small", 8% are considered "medium", and 76% are considered "large." Going back to Cohen's statement 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.

Even though 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 & 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 & 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 also 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 an estimate by calculating the statistical power of each method for the median effect size in LI for a range of likely sample sizes used in traditional methods (both the mean and the median effect size are 1.61 for the phenomena randomly sampled by Sprouse et al. *submitted*). Figure 10 reports both a graph and table of the statistical power for each task for an effect size of 1.61 for sample sizes from 6 to 30 (with only one judgment per participant).

Figure 10: The statistical power (y-axis) for the median effect size (Cohen's *d* of 1.61) reported in Linguistic Inquiry (2001-2010) that would be detectable by each task (lines) as a function of the sample size (x-axis) assuming only 1 judgment per participant per condition. The accompanying table specifies the statistical power precisely.



| Sample | FC | YN | ME | LS |
|---|---|---|---|---|
| 6 | 52% | 08% | 64% | 69% |
| 10 | 77% | 72% | 86% | 91% |
| 15 | 94% | 96% | 97% | 99% |
| 20 | 99% | 100% | 99% | 100% |
| 25 | 100% | 100% | 100% | 100% |
| 30 | 100% | 100% | 100% | 100% |

As Figure 10 illustrates, even with only 10 participants reporting a single judgment each, all four methods outperform the power analyses from other domains of experimental psychology. The two tasks most likely to be used in traditional methods, FC and YN, reach 77% and 72% power respectively for the median effect size in LI (1.61), which is substantially higher than the 46%-59% power observed for other branches of experimental psychology despite the fact that these branches exclusively use formal experiments. It seems reasonable to conclude from this that the phenomena of interest to syntacticians are substantially larger in size than the phenomena of interest to other types of experimental psychologists, and that traditional acceptability judgment methods, although informal compared to the methods of other fields, are nonetheless a set of well-powered methodologies for investigating the phenomena of interest to syntacticians.

### 7.3 A pluralistic approach to methodology

We believe that 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 substantial 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 & 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 & Almeida *in press* and Sprouse et al. *submitted*, suggest that these concerns have been overstated: not only do traditional methods produce low false positive rates, but they also seem to be well protected against false negatives, at least according to the standards used in experimental psychology, both because of the comparatively larger effects sizes investigated by syntacticians, and because some of the tasks routinely used by syntacticians, such as the forced-choice task, yield remarkably clear data with relatively few subjects. This suggests that several traditional methods are in fact well-powered methodologies for investigating the phenomena of interest to syntacticians.

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 & Jackendoff 2010): they are relatively quick to deploy, they are generally free, and they are very portable (requiring only pen and paper), and therefore very easy to replicate. Sprouse & Almeida (*in press*) and Sprouse, Schütze, & Almeida (*submitted*) have suggested that they also have a very low false positive rate. To that we can now add that they also have relatively high statistical power. 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 & Almeida 2012, Schütze & Sprouse 2012). 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, Goméz, & Hidalgo 2005, Hubbard & 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 appear to be relatively informal compared to methods in other fields (e.g., Ferreira 2005, Gibson & 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, & 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. Formal experiments are also more likely to be seen as reliable to researchers in fields that rely exclusively on formal experiments (Ferreira 2005, Gibson & Fedorenko 2010a, 2010b). There are, however, very real costs to running formal experiments, many of which have not 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, assuming that all 50 phenomena could be presented in a single survey without causing problems due to overt similarity across the conditions. 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, although Amazon Mechanical Turk is neutralizing this cost: 80 participants can be collected per hour on Amazon Mechanical Turk (Sprouse 2011a).

## 8. Conclusion

In this article we set out to evaluate the statistical power of judgment collection methods in syntax in an effort to address current concerns about the reliability of the data underlying syntactic theories. Because traditional methods likely cannot be reduced to a single definition, we presented empirically derived estimates of the statistical power for every possible type of acceptability judgment experiment: four tasks (ME, LS, YN, FC), a range of sample sizes (5-100), and a set of phenomena that were randomly sampled from Linguistic Inquiry (2001-2010) and that span a substantial range of effect sizes (Cohen's *d* 0.15-1.96). The results provide the first comprehensive evaluation of statistical power in acceptability judgments that we are aware of, which can be used to (i) evaluate the statistical power of previously published studies, (ii) plan appropriately powered studies in the future, and (iii) establish a common vocabulary for assessing the statistical power of any definition of traditional methods.

Although our goal was to provide the data necessary for readers to evaluate their own definition of traditional methods, there are several patterns that emerge naturally from these results. For example, FC experiments are substantially more powerful than other experiments at detecting differences between conditions; ME and LS experiments are approximately equally powered; and YN experiments appear to have the lowest statistical power. The power of FC experiments is very impressive: FC experiments with as few as 10 participants each providing only one judgment can detect 70% (±5) of phenomena in LI with Cohen's suggested power of 80% (increasing to 80% of phenomena with 15 participants at the same power level). In fact, all of the acceptability judgment tasks substantially outperform the mean statistical power of experiments in other domains of experimental psychology with only 10 participants (each providing one judgment), suggesting that the effect sizes of the phenomena of interest to syntacticians are substantially larger than the effect sizes of phenomena investigaged in other domains of psychology. This is corroborated by the effect sizes observed in the random sample of phenomena from LI (2001-2010) in Sprouse et al. *submitted*: 87% of effect sizes were medium or larger according to Cohen's (1988) rule of thumb, which means 87% of phenomena in LI are large enough to be visible to a trained observer without the use of statistics.

While the variability of definitions of traditional methods makes it difficult to draw conclusions that will be accepted by every researcher, we believe that the results of this study suggest that, contrary to the claims of critics, traditional methods are often well-powered for the detection of 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. There are many advantages to formal experiments, 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, & Phillips 2012). However, 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, and therefore syntacticians should be allowed to decide which methodology best suits their scientific goals. It is our hope that the statistical power estimates in this study provide information that syntacticians can use to aid those decisions.

## References

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

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

Balluerka, Nekane, Juana Goméz, & Maria Dolores Hidalgo. 2005. Null hypothesis significance testing revisited. *Methodology* 1, 55–70.

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

Bezeau, Scott, & Roger 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, Herbert H. 1973. The Language-as-Fixed-Effect Fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12, 335–359.

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

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

Cohen, Jacob. 1976. Random means random. *Journal of Verbal Learning and Verbal Behavior* 15, 261–262.

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

Cohen, Jacob. 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, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

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

Culicover, Peter W., & 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, & Raffaella Zanuttini. 2007. Data and grammar: Means and individuals. *Theoretical Linguistics* 33, 335–352.

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

Edgington, Eugene, & Patrick Onghena. 2007. *Randomization tests, 4th ed*. Boca Raton, FL: Chapman & Hall/CRC.

Efron, Bradley, & Robert Tibshirani. 1994. *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

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

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

Gigerenzer, Gerd & Hans Richter. 1990. Context effects and their interaction with development: Area judgments. *Cognitive Development* 5, 235–264.

Gigerenzer, Gerd, Stefan Krauss & Oliver Vitouch. 2004. The null ritual: What you always wanted to know about significance testing but were afraid to ask. In *The Sage handbook of quantitative methodology for the social sciences*, ed. by D. Kaplan. Thousand Oaks, CA: Sage.

Gross, Steven, & Jennifer Culbertson. 2011. Revisited linguistic intuitions. *British Journal for the Philosophy of Science* 62, 639–656.

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

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

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

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

Keppel, Geoffrey. 1976. Words as random variables. *Journal of Verbal Learning and Verbal Behavior* 15, 263–265.

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.

Meehl, Paul E. 1967. Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34, 103–115.

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.

Nickerson, Raymond S. 2000. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* 5:241–301.

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

Raaijmakers, Jeroen. G. 2003. A further look at the "Language-as-Fixed Fallacy". *Canadian Journal of Experimental Psychology* 57, 141–151.

Shaver, James P. 1993. What statistical significance testing is, and what it is not. *The Journal of Experimental Education* 61, 293–316.

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

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

Smith, J. E. Keith. 1976. The Assuming-Will-Make-It-So Fallacy. *Journal of Verbal Learning and Verbal Behavior* 15, 262–263.

Sprouse, Jon. 2007a. A program for experimental syntax. Ph.D. 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, & Diogo Almeida. 2012. 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, & Diogo Almeida. (*in press*). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*.

Sprouse, Jon, Shin Fukuda, Hajime Ono, & 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, & Colin Phillips. 2012. A test of the relation between working memory capacity and island effects. *Language* 88, 82–123

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

Spencer, Nancy. 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, & Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115, 1481–1496.

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

Wickens, Thomas D. & Geoffrey Keppel. 1983. On the choice of design and of test statistic in the analysis of experiments with sampled materials. *Journal of Verbal Learning and Verbal Behavior* 22, 296–309.

Wike, Edward L., & James D. Church. 1976. Comments on Clark's "The Language-as-Fixed-Effect Fallacy". *Journal of Verbal Learning and Verbal Behavior* 15, 249–255.

Jon Sprouse
jsprouse@uci.edu

Diogo Almeida
diogo@nyu.edu