Glossa a journal of general linguistics

Sprouse, Jon and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: a journal of general linguistics* 2(1): 14. 1–32, DOI: https://doi.org/10.5334/gjgl.236

RESEARCH

Design sensitivity and statistical power in acceptability judgment experiments

Jon Sprouse[1] and Diogo Almeida[2]

[1] Department of Linguistics, University of Connecticut, US

[2] Psychology Program, Division of Science, New York University, Abu Dhabi, UAE

Corresponding author: Jon Sprouse (jon.sprouse@uconn.edu)

Previous investigations into the validity of acceptability judgment data have focused almost exclusively on *type I errors* (or false positives) because of the consequences of such errors for syntactic theories (Sprouse & Almeida 2012; Sprouse et al. 2013). The current study complements these previous studies by systematically investigating the *type II error rate* (false negatives), or equivalently, the *statistical power*, of a wide cross-section of possible acceptability judgment experiments. Though type II errors have historically been assumed to be less costly than type I errors, the dynamics of scientific publishing mean that high type II error rates (i.e., studies with low statistical power) can lead to increases in type I error rates in a given field of study. We present a set of experiments and resampling simulations to estimate statistical power for four tasks (forced-choice, Likert scale, magnitude estimation, and yes-no), 50 effect sizes instantiated by real phenomena, sample sizes from 5 to 100 participants, and two approaches to statistical analysis (null hypothesis and Bayesian). Our goals are twofold (i) to provide a fuller picture of the status of acceptability judgment data in syntax, and (ii) to provide detailed information that syntacticians can use to design and evaluate the sensitivity of acceptability judgment experiments in their own research.

Keywords: Acceptability judgments; statistical power; experimental syntax; linguistic methodology

1 Introduction

Acceptability judgments form a substantial component of the empirical foundation of (generative) syntactic theories (Chomsky 1965; Schütze 1996). In a recent survey of US-English data points from articles that appeared in *Linguistic Inquiry* from 2001 through 2010, Sprouse et al. (2013) estimated that 77% were derived from some form of acceptability judgment (the remaining 23% were judgments about meaning). The vast majority of judgments in syntactic theory tend to be collected relatively informally, and specifically without any form of explicit quantitative analysis. As the use of formal experimental methods (sometimes called experimental syntax following Cowart 1997) has grown in popularity, many researchers have begun to investigate the quantitative properties of acceptability judgments, in an attempt to better understand the empirical foundation of the field (for an annotated bibliography see Sprouse 2013). Our goal in this article is to add one more critical piece of quantitative information to this growing body of knowledge: an empirical estimation of the sensitivity of formal acceptability judgment experiments in detecting theoretically interesting contrasts between different sentence types. We operationalize the notion of sensitivity by estimating and evaluating the *rate of statistical detection of acceptability rating differences* in a series of resampling simulations based on a large dataset of real pairwise comparisons where putatively real differences of different sizes exist. This *rate of detection* can be understood as an *empirical estimate* of statistical power (and will be referred as such) observed in different kinds of acceptability judgment experiments.

Because there is no one method for collecting and analyzing acceptability judgments, we have conducted a set of experiments and simulations to cover a wide range of possible experimental designs, fully crossing four of the most popular acceptability judgment tasks (yes-no, two-alternative forced-choice, Likert scale, and magnitude estimation), a set of 50 phenomena taken from *Linguistic Inquiry* (2001–2010) that span the largest observable range of effect sizes in a large random sample of theoretical manipulations found in syntax articles (Sprouse et al. 2013), a range of potential sample sizes from 5 to 100 participants (obtained through resampling simulations out of a sample of 144 per task), and two distinct approaches to hypothesis testing (null hypothesis testing and Bayesian hypothesis testing). The result is a database of information regarding the *rate of statistical detection* (our proxy measure of *statistical power*) that covers a substantial portion of possible experimental designs in syntax. Our two goals in this paper are to (i) provide a fuller picture of the status of acceptability judgment data in syntax (i.e., a complement to the validity experiments in Sprouse et al. 2013 and Sprouse & Almeida 2012), and (ii) provide detailed information that syntacticians can use to design and evaluate the sensitivity of acceptability judgment experiments.

The recent trend of formalizing the (historically informal) collection of acceptability judgments allows syntacticians to begin to quantitatively evaluate how well syntactic experiments achieve the goal of distinguishing a hypothesis of interest from a theoretically uninteresting hypothesis. Neyman & Pearson (1928; 1933) provide one of the most influential frameworks for thinking about the outcome of statistical hypothesis testing (as discussed in more detail in section 2). The first step is to construct two statistical hypotheses to test: the null hypothesis (H_0), which states that there is no quantifiable difference between two (or more) experimental conditions, and the alternative hypothesis (H_A), which states that there is a difference of a certain minimum size between two (or more) experimental conditions. In syntax, experimental conditions tend to be sentence types, so H_0 would posit no difference in acceptability between two or more sentence types, and H_A posit that there is a difference of a certain size between two or more sentence types. From this partitioning of the hypothesis space, it follows that there are two states of the world, H_0 is true or H_A is true, and there are two outcomes of a hypothesis test, H_0 is claimed to be true or H_A is claimed to be true. This results in four combinations, two of which are correct outcomes of the test, and two of which are errors, as shown in Table 1.

A type I error, or false positive, occurs when the hypothesis test favors H_A, but H_0 is in fact true. A type II error, or false negative, occurs when the hypothesis test favors H_0, but H_A is in fact true. Many of the recent investigations of the properties of syntactic experiments have studied the prevalence of type I errors (Gibson & Fedorenko 2010; Sprouse & Almeida 2012; Gibson & Fedorenko 2013; Sprouse et al. 2013). However, *sensitivity*, specifically in the guise of *statistical power* in the Neyman-Pearson approach, is fundamentally defined in terms of type II errors. *Statistical power* is the probability that a hypothesis test will favor H_A when H_A is in fact true, therefore it is the complement of type II errors (power = 1 – type II errors). It is typically expressed as a percentage. For example, if a

	H_A is true of the world	H_0 is true of the world
H_A is favored by the test	positive correct decision	type I error
H_0 is favored by the test	type II error	negative correct decision

Table 1: Four possible outcomes for a hypothesis test under the Neyman-Pearson approach.

hypothesis test has 80% power, then it will favor H_A 80% of the time that H_A is true, and favor H_0 20% of the time that H_A is true.

Given that statistical power is in many ways the complement of recent investigations into type I error, the first goal of this project is to simply extend the knowledge of the field regarding the properties of syntactic experiments. There is a robust and growing literature across experimental disciplines exploring the consequences of low statistical power (e.g., Ioannidis 2005; Simmons et al. 2011; Button et al. 2013). There are at least three consequences of low statistical power that are potentially relevant for the growing literature on formal experimentation in syntax. The first is the most obvious: given the definition of statistical power, a low powered experiment is potentially a waste of time and resources. A low powered experiment has a low probability of doing what experiments are intended to do – detect evidence for the theoretically interesting hypothesis (H_A). The second consequence is that low power makes it more difficult to interpret null results as evidence in favor of the null hypothesis. To the extent that null hypotheses (i.e., predicted *invariances* between sentence types) are relevant for the construction of syntactic theories (which we believe they are), this is potentially problematic for theory construction. Furthermore, null hypotheses are critical for the elimination of type I errors through replication, arguably making investigations of type I error incomplete without an investigation of statistical power. The third consequence is much less obvious: low statistical power studies have the potential to increase the number of type I errors (false positives) in the literature. This is not an obvious consequence of low power (after all, power is defined solely in terms of type II errors), but as we will see in section 2, it follows directly from the mathematics of statistical hypothesis testing and the fact that current academic practices result in strong publication bias towards results that present *nominally* statistically significant results (Button et al. 2013). Therefore, previous investigations of type I error (such as Sprouse et al. 2013) are potentially incomplete without a comprehensive investigation of how statistical power is generally taken into consideration in the design and evaluation of experiments in the field.[1] Given recent concerns about replicability of data in both psychology (Open Science Collaboration 2015) and linguistics (e.g., Gibson & Fedorenko 2013; Sprouse et al. 2013), and these three consequences of low statistical power, we believe it is important to broadly investigate statistical power in syntactic experiments.

The second goal of this project is to provide actionable information for syntacticians to use both when they construct their own experiments and when they evaluate the experiments of others. Syntacticians can use the graphs and appendices in this paper, as well as the raw data provided on the first author's website, to estimate the statistical power for a substantial portion of possible experimental designs in syntax (in terms of tasks, phenomena, and sample sizes). It is our hope that this will take some of the guesswork out of the design of future experiments, as well as provide concrete baselines for editors, reviewers, and readers to critically evaluate the sensitivity/power of acceptability judgment experiments.

The article is structured as follows. Section 2 provides a detailed review of three major approaches to hypothesis testing, and the concept of statistical power within each (readers already familiar with statistical power should feel free to skip this section). Section

[1] There is a fourth consequence that is sometimes discussed (e.g., Button et al. 2013), but it appears to be less of a problem for syntax than other literatures: low power can lead to exaggerated effect sizes, sometimes called the winner's curse (Ioannidis 2008). Exaggerated effect sizes can mean that replications of published effects are likely to show shrinking effect sizes (potentially causing researchers to lose confidence in previously published results), and can mean that any replications that are predicated upon previously reported sample sizes are less likely to return significant results (again, potentially causing researchers to lose confidence in previously published results). However, because syntactic theories (currently) do not often directly predict effect sizes, this appears to be less of a problem for syntax.

3 describes the approach to empirically estimating statistical power that we used in the current study. Section 4 presents the results of our investigation. Section 5 discusses the consequences of our results for our twin goals: (i) exploring the status of acceptability judgment data in the field, and (ii) providing estimates that syntacticians can use to design and evaluate judgment experiments. Section 6 concludes. The appendix provides detailed power results in a tabular form.

2 Statistical power and its consequences

This section is primarily a high-level review of the formal concept of *statistical power*, and the mathematical reasons that it has so many (not always obvious) consequences for hypothesis testing. We first review three major philosophical approaches to hypothesis testing: Fisher Hypothesis Testing (FHT), Neyman-Pearson Hypothesis Testing (NPHT), and Bayesian Hypothesis Testing (BHT). This background is necessary to formally define statistical power (in NPHT), and to highlight the role that statistical power plays in all three approaches to hypothesis testing. These sections also allow us to establish the first two consequences of low statistical power: low powered experiments are less likely to favor H_A when H_A is true, and low powered experiments make it difficult to interpret null results as evidence for H_0. We then review a concept from testing theory called *positive predictive value* (PPV), which provides the link necessary to demonstrate the third consequence of low statistical power: low power can inflate the type I error rate when a field uses a criterion for publication such as $p < .05$. We provide this review for readers who may not be familiar with the details of statistical power, as there is currently no comprehensive discussion of statistical power in the experimental syntax literature. However, nothing in this section is specific to syntax, and therefore can be skipped by readers who are already familiar with these issues.

2.1 Fisher Hypothesis Testing

Ronald A. Fisher was the first to develop a unified approach to null hypothesis testing in the early 20th century (Fisher 1955; Hubbard 2004). Although Bayesian statistics technically pre-dates Fisher by over 150 years, and even though Fisher in many ways developed his null hypothesis approach as a response to what he perceived as short-comings in the Bayesian approach (Fisher 1925), in the history of modern approaches to hypothesis testing, Fisher's approach deserves a privileged position. It is a direct precursor to the Neyman-Pearson approach, it is the foil for modern developments in Bayesian statistics, and it is in many ways the default method of null hypothesis testing in various domains of cognitive science (see Gigerenzer 2004; Hubbard 2004 for reviews).

Under Fisher Hypothesis Testing (FHT), there is only one hypothesis under consideration: the theoretically uninteresting hypothesis called the null hypothesis (abbreviated H_0), which for syntax is very often the claim that there is no difference in acceptability between two (or more) sentence types. Statistical tests in FHT assume that H_0 is true, and return the probability of obtaining the observed experimental result, or a result that is more extreme, under this assumption. This probability is called a *p*-value. In other words, a *p*-value is the probability of the observed data (or a result more extreme) given the null hypothesis. For the mathematically inclined, this means that the symbol *p* is a shorthand for the full conditional probability statement $p(\text{data} \mid H_0)$, where the pipe (|) symbol is read "given that". Low *p*-values indicate that the result is relatively unlikely under the null hypothesis, and high *p*-values indicate that the result is relatively likely under the null hypothesis.

FHT does not specify an algorithm for using the information that is provided by *p*-values to decide between two hypotheses. In FHT, *p*-values are a *direct measure of the strength* *of evidence against the null hypothesis*, and it is up to the researcher to decide what to do with this information. The implication is that if a *p*-value is low enough, the researcher can draw the disjunctive conclusion that "either the null hypothesis is false or a very rare event occurred". FHT does not pre-specify what threshold the researcher should use to decide which conclusion to draw. Although Fisher himself made suggestions about using .05 or .01 as potential thresholds to be used heuristically and primarily for convenience (Fisher 1925), he interpreted *p*-values as gradient *measures of evidence* against the null hypothesis, such that a *p*-value of .049 and a *p*-value of .051 are roughly equal in evidential value, and such that smaller *p*-values can be said to be stronger evidence against the null hypothesis than larger *p*-values. Fisher intended researchers to combine this information with non-statistical information to reach a conclusion about the validity of rejecting the null hypothesis (Lykken 1968 goes so far as to call statistical significance the "least important attribute" of a good experiment). Furthermore, while FHT implies that a decision to reject the null hypothesis should be taken as evidence in favor of a theoretically interesting alternative hypothesis, FHT cannot make any probability statements about this alternative hypothesis, nor can FHT make any statements about the strength of evidence for this alternative hypothesis. This is because no alternative hypothesis was considered during the statistical hypothesis test, only the null hypothesis was considered. In a very real sense, the alternative hypothesis is left as an implied hypothesis covering all theories that are not the null hypothesis.

The limitations of FHT became most apparent in the case of high *p*-values. High *p*-values arise in two scenarios: (i) when the null hypothesis is true of the world, or (ii) when an alternative hypothesis is true of the world, but the test lacks the sensitivity to detect the evidence for the alternative hypothesis. FHT cannot distinguish these two cases. FHT cannot determine a probability for the truth of the null hypothesis because it *assumes* that the null hypothesis is true when calculating the *p*-values. And FHT cannot determine a probability that it failed to detect the evidence for the alternative hypothesis because no alternative hypothesis was specified, and no criterion for its detection was adopted. This means that all FHT can say about high *p*-values are that they are uninformative.

The uninformative nature of high *p*-values is a well-known limitation for FHT, but more importantly for our current study, scenario (ii) above reveals the precarious status of statistical power in FHT. On the one hand, the concept of statistical power (the ability of a test to favor the alternative hypothesis when the alternative hypothesis is true) is clearly relevant for hypothesis testing in general. If a test has low statistical power (i.e., it is not sensitive to evidence for the alternative hypothesis), the researcher gains no information about the world. But on the other hand, the design of FHT makes it difficult (potentially impossible) to formally calculate statistical power.

2.2 Neyman-Pearson Hypothesis Testing

Neyman-Pearson Hypothesis testing (NPHT) and FHT are both null hypothesis approaches to hypothesis testing. Both NPHT and FHT use the same null hypothesis statistical tests, and both NPHT and FHT use *p*-values to evaluate hypotheses. However, NPHT and FHT differ explicitly in their underlying goals. Whereas FHT seeks only to provide a *measure of evidence* against the null hypothesis, leaving the question of how to use that evidence to the researcher, NPHT seeks to provide an *explicit decision-making algorithm* for deciding between the theoretically uninteresting null hypothesis (H_0) and an explicit theoretically interesting alternative hypothesis (abbreviated H_A). Because errors are possible with any decision, NPHT further seeks to control the maximum number of errors that would be made over the long run if one were to replicate a specific hypothesis test an infinite number of times. Put more succinctly, the output of FHT is a *statement about the strength* *of evidence against the null hypothesis*; and the output of NPHT is *a statement about the maximum probability of making decision errors over the long run, given a pre-specified decision rule*. Despite their superficial similarities, FHT and NPHT are philosophically different approaches to hypothesis testing, the former privileging the role of epistemically-laden *inductive inferences* and the latter the minimization of long-run *error probabilities* (see Hubbard 2004 for a comparison of FHT and NPHT).

NPHT formalizes the decision procedure as follows. First, the researcher defines a null hypothesis (H_0), just like in FHT. Then the researcher defines an alternative hypothesis (H_A), such as "the population means of condition x and condition y will differ by at least z units" (the specification of the minimum effect size will become relevant when we discuss statistical power below). The researcher then defines a significance level for deciding between H_0 and H_A. The significance level is typically defined in terms of either *p*-values or the test statistics used to derive *p*-values (the choice is equivalent for our purposes, so we will focus on *p*-values). If the *p*-value of the observed data is below the significance level, the alternative hypothesis (H_A) is chosen; if the *p*-value is above the significance level, the null hypothesis (H_0) is chosen.

Because we can never know the true underlying state of the world, it is possible, and indeed guaranteed, that the NPHT procedure will lead to errors for some proportion of hypothesis tests. As we saw in Table 1 in section 1, the two explicit hypotheses in NPHT lead to four possible outcomes of the hypothesis test: a positive correct decision when the test favors H_A and H_A is true, a negative correct decision when the test favors H_0 and H_0 is true, a type I error when the test favors H_A but H_0 is true, and a type II error when the test favors H_0 but H_A is true. NPHT explicitly seeks to minimize both types of errors. This is accomplished in NPHT by setting a (distinct) upper limit on the probability of making each type of error *over the long run*, and using those limits in the construction of the hypothesis test itself. The critical idea is that a researcher can *never* know if a single decision is correct or not (that would be perfect knowledge, which we do not have). But if a researcher sets an upper limit on errors, and if that researcher uses that limit for every experiment, then as the number of experiments approaches infinity, the number of errors will approach the upper limits that the researcher set. Therefore, if the two error rates are set sufficiently low, NPHT can help researchers have confidence in the decisions that they make, provided they strictly follow the decision criteria imposed by the NPHT framework.

The upper limit for the type I error rate is called *alpha* (α), or the α-level. α enforces the upper limit on type I errors by directly determining the significance level that is used as the criterion for decisions between H_0 and H_A. The relationship between α and the significance level is given by the following equation: significance level $= 1 - (1 - \alpha)^c$, where c is the number of statistical comparisons made in the hypothesis test. As a concrete example, let's imagine that our goal is to achieve a 5% long run maximum type I error rate. Therefore, we set α to .05. By the equation above, the significance level is $1 - (1 - .05)^1$, which is .05. This means that for every experiment in which *p* is less than .05 (the observed data has less than a 5% chance of occurring under the null hypothesis), the researcher will choose H_A, and for every experiment in which *p* is greater than .05, the researcher will choose H_0. Now, let's assume that the null hypothesis is true of the world (the only scenario in which a type I error is possible). And let's imagine that a researcher performs an infinite number of replications of a given experiment. Because the null hypothesis is true, and because *p*-values are calculated by assuming that the null hypothesis is true, 5% of those (infinite) experiments will yield a *p*-value of .05 or less, and 95% of the experiments will yield a *p*-value of .05 or more. For the 5% below .05 (the α-level), the researcher will make a type I error. Thus the long run type I error rate is 5%, exactly what the researcher specified by setting α to .05. By setting α ahead of time (and using it for every experiment), when NPHT leads a researcher to choose H_A, that choice can be paired with the maximum probability of type I errors over the long run (the α-level).

In NPHT the upper limit for the type II error rate is called *beta* (β). Whereas the experimenter can directly implement an upper limit on the type I error rate by setting the appropriate α-level, the experimenter cannot directly implement β. Instead, β is a consequence of the combination of the α-level, the minimum size of the effect as defined in H_A, the measurement error associated with the task, and the size of the sample being tested. This is because of what a type II error is: a type II error occurs when H_A is true, but the value obtained in the experiment is within the range that we have decided (a priori) to call evidence for H_0. This means that type II errors are a sort of "complement" of type I errors inside of the area where the distributions of the two hypotheses overlap. Figure 1 attempts to demonstrate this graphically. Each hypothesis has a distribution of potential test statistics. The minimum effect size, the measurement error, and the sample size all combine to determine the locations, widths, and overlap of the distributions (this is why these properties affect β). The α-level divides the distribution of H_0 into two decisions, significant and non-significant, and therefore sets the maximum type I error rate. The type II error rate is the portion of the distribution of H_A to the left of the α-level that also overlaps with the H_0 distribution. This matches the definition of a type II error: a type II error occurs when the experimenter decides to call a result non-significant because it falls to the left of the α-level, but the result is actually part of the distribution of H_A (it is in the red zone of H_A).

What this relationship means in practice is that, after choosing the desired β, in order to actually enforce β over the long run, a researcher must choose an α-level, define the minimum effect size of the alternative hypothesis, know the measurement error of the task, and then calculate the correct sample size to achieve the desired β. Assuming the experiment is designed using all of these values (i.e., all of these choices and calculations must be made before the experiment is run), NPHT ensures that *over the long run* the maximum type II error rate will be equal to β. This means that when NPHT leads a researcher to choose H_0, that choice can be paired with the maximum probability of type II errors over the long run (the β-level).

So what is *statistical power* in NPHT? Recall that we defined statistical power as a measure of how often a given hypothesis test favors H_A when the H_A is true. NPHT simply converts this to a probability statement: statistical power is the probability of a

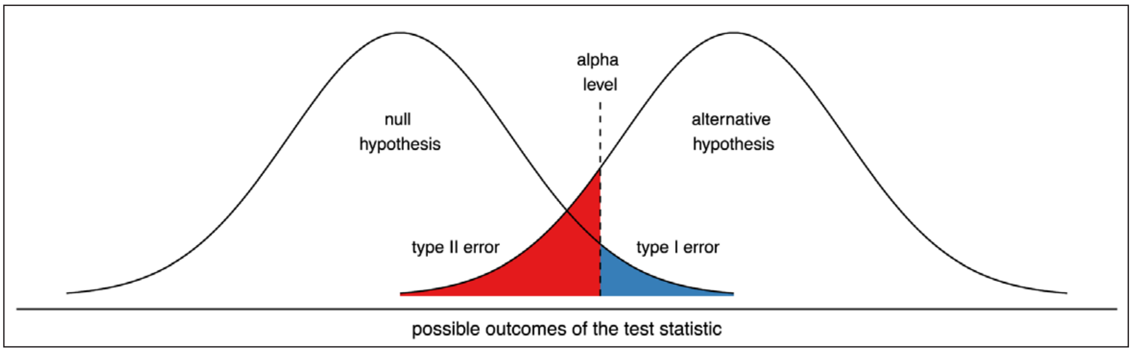

Figure 1: An illustration of the relationship between type I error and type II error for a hypothetical distribution of test statistics for a null hypothesis and an alternative hypothesis for a continuous range of possible outcomes of the test statistic. The type II error rate can be thought of as a "complement" of the type I error rate in the area where the null and alternative distributions overlap (the details of which is determined by minimum effect size, measurement error, sample size, and of course, α-level).

hypothesis test choosing H_A when H_A is true. And crucially we already have that number in our NPHT calculations. β is the probability of choosing H_0 when H_A is true (making a type II error). Statistical power is simply the other (correct) decision: choosing H_A when H_A is true. Because the two decisions cover the full range of possible outcomes, they are simply the complement of each other. Therefore, statistical power equals 1 – β. Cohen (1988) famously recommended setting β to .2 (a maximum type II error rate of 20%). This recommendation means that experiments should have 80% power, that is, an 80% chance of choosing H_A when H_A is true. The 20% type II error rate (80% power) recommendation reflects a common belief (that both Neyman and Pearson and Cohen endorse) that type II errors are less costly than type I errors. It is an open question whether this asymmetry holds in every field. One could imagine that type II errors might be more costly in syntax, because grammars are predicated upon both differences between sentence types and the lack of differences between sentence types. We know of no explicit discussion of this in the field.[2]

With all of this in mind, it is now possible to see how the first two consequences of low statistical power can arise. The first consequence is that low powered experiments are potentially a waste of time and resources as they have a low probability of detecting evidence for the theoretically interesting hypothesis (H_A). This follows directly from the definition of statistical power. For example, an experiment with 50% power is only going to be able to find statistical evidence (as defined by the NPHT theory) against the null hypothesis (when it is indeed false) 50% percent of the time. At such power level, running the experiment amounts to doing the equivalent of an expensive and time consuming coin toss. The second consequence is that low powered experiments render null results (i.e., no statistically significance difference between two conditions) uninformative about the status of the null hypothesis. As shown in Table 1, a null result may appear under one of two circumstances: (i) the null hypothesis is true of the world or (ii) the null hypothesis is false, but the experiment did not have enough statistical power to reject it at a particular α-level. A well-powered experiment can help disambiguate between these two alternatives by providing the researcher with information that would allow her to adjudicate whether (ii) is a plausible reason for the null result. In other words, an adequately powered experiment may license, under the right conditions, the indirect inference that certain observed null results may count as evidence *for* the null hypothesis (i.e., that there is no difference of a pre-specified, theoretically interesting magnitude between two conditions). This follows directly from the relationship between statistical power and β, where β is the maximum type II error rate over the long run. In order to make the argument that null results are evidence for the null hypothesis under NPHT, one must also report β. The lower the β, the more credible a conclusion for the null hypothesis seems. Because statistical power is the complement of β, low statistical power means high β, and therefore less support for concluding positively for the null hypothesis.

2.3 Bayesian Hypothesis Testing

The fundamental assumption of Bayesian Hypothesis Testing (BHT) is that most researchers ultimately want to know how likely a specific hypothesis is to be true given the experimental results that they obtained. In other words, what most researchers really want to calculate is p(H | data), where H can be any of the various hypotheses under consideration (H_0, H_A, some other H_A, etc.). Null hypothesis approaches to testing, like

[2] In fact, for NPHT, if it is ever determined that type II errors are more costly than type I errors, that would imply that H_0 is the theoretically more interesting hypothesis, so the position of H_0 and H_A should be swapped in the hypothesis test: the new H_0 should be that there is a difference between conditions of a certain minimum size, and the new H_A should be that there is no difference between conditions.

FHT and NPHT, do not provide this information. They explicitly provide p(data | H_0), and then build inferential processes around that probability to help researchers choose between competing hypotheses. But crucially, in null hypothesis approaches, p(H | data) is never calculated. Although the order of terms around the pipe (|) operator seems like a small difference, it is important to note that the resulting probabilities are very different pieces of information. A classic example is the difference between p(living-in-LA | being-a-movie-star), that is, the probability of living in LA given that you are a movie star, and p(being-a-movie-star | living-in-LA), which is the probability of being a movie star given that you live in LA. The former probability is relatively high because the US film industry is highly concentrated in LA, whereas the latter probability is relatively low because nearly 4 million people live in LA but relatively few would be considered movie stars. BHT seeks to provide the information that most researchers actually desire, and leverages Bayes Theorem (Bayes 1764) to derive the relevant probabilities (for an accessible introductory textbook, see Kruschke 2011).

There are a number of excellent introductions to Bayes Theorem and Bayesian statistics (e.g., Kruschke's 2011 textbook), so here we simply review a few basics. First, equation (1) is Bayes Theorem, with each of its components labeled with curly braces for clarity:

(1) Bayes Theorem with components labeled

$$\underbrace{p(H|D)}_{Posterior} = \frac{\overbrace{p(D|H)}^{Likelihood} \times \overbrace{p(H)}^{Prior}}{\underbrace{p(D)}_{Evidence}}$$

The first component, the posterior probability, is the probability of the hypothesis that the researcher is interested in, assuming the observed data. The likelihood is the probability that the hypothesis in question would generate the data that is observed. The prior is the probability of the hypothesis in question before the data is considered. And the evidence is the probability of obtaining the observed data in general (under all possible hypotheses). If the likelihood, prior, and evidence are known, Bayes Theorem will provide the posterior probability of the hypothesis given the observed data. Although Bayes Theorem looks complicated, it is actually a straightforward consequence (theorem) of the definition of conditional probability (an axiom of probability theory). It is a very short derivation requiring only basic algebra, but we won't go through it here because there are many demonstrations available in the literature and on the internet.

Given that the posterior probabilities returned by Bayes Theorem appear to be exactly what researchers are interested in when evaluating hypotheses, and given that Bayes Theorem predates null hypothesis testing by over 150 years, one might wonder why null hypothesis testing approaches exist at all. It turns out that there are two components of Bayes Theorem that are often difficult to calculate: the likelihood and the prior. While the likelihood is generally easy to calculate for the null hypothesis, the same is rarely true of alternative hypotheses. It has only been in the last few decades that modern (personal) computers have allowed researchers to estimate likelihoods for alternative hypotheses using sophisticated simulation methods. Because of this newly acquired computational power, there is a robust and growing literature investigating the use of such simulation methods for applications of Bayes Theorem to experimental research (see Wagenmakers et al. 2016 for a short argument in favor of Bayesian inference and recommended readings; see Mulder & Wagenmakers 2016 for an introduction to a special issue dedicated to Bayes factors in psychological research; see Rouder et al. in press for a specific Bayesian analysis of factorial experimental designs). In a similar vein, in many cases the prior is difficult to calculate because there is often little or no previous quantitative information about the hypothesis of interest. Therefore, in many cases the specification of the prior is left to the subjective beliefs of the researcher. This subjectivity means that two researchers may, in principle, end up with very different posterior probabilities for the same hypothesis under Bayes Theorem. Fisher himself was well aware of Bayes Theorem and these two potential difficulties. He explicitly developed his null hypothesis approach to testing to take advantage of the ease with which the likelihood of the null hypothesis can be calculated (computer simulations were not available in the early 20th century), and to attempt to eliminate what he saw as the subjectivity surrounding the choice of prior probability for an alternative hypothesis (Fisher 1925).[3]

Despite our use of a single label for Bayesian hypothesis testing, BHT is not a unified approach to hypothesis testing. There are different approaches to methods of simulating likelihoods, and there are different types of information that can be derived from Bayes Theorem in addition to the posterior probabilities. One piece of information that has become increasingly popular in the experimental literature is the Bayes Factor (Jeffreys 1939/1961). Bayes Factors (BF) measure how much more likely the data are under one hypothesis compared to the other. For example, if a researcher were to compare an alternative hypothesis to the null hypothesis, a BF of 10 would indicate that the data are ten times more likely under the alternative hypothesis than under the null hypothesis. In this way, Bayes Factors provide a measure of the strength of evidence for one hypothesis over another. Bayes Factors are particularly popular for the analysis of experimental results for two reasons. First, they are relatively easy to calculate. Though they cannot get around the problem of specifying the likelihood of the alternative hypothesis, there are a number of tools that are now available that make the process a bit simpler for experimentalists (e.g., the Bayes Factor package for R by Morey & Rouder 2015). Second, Bayes Factors eliminate the need to specify a prior probability of each hypothesis because the posterior probability of the hypothesis is not being calculated (only the probability of the data under each hypothesis). This can be seen in the equation used to extract Bayes Factors from Bayes Theorem. The derivation starts with the full Bayes Theorem (equation 1 above), but applied twice: once to H_0 and once to H_A. The second step is to arrange the two equations into a ratio: place the token of Bayes Theorem for H_A over top of the token of Bayes Theorem for H_0 as a fraction. If we were to calculate this ratio completely, it would give us the ratio of the posterior probability of H_A (given the data) to the posterior probability of H_0 (given the data). We aren't interested in this full calculation, so instead of trying to calculate it, we instead begin to simplify the terms in the equation. The important step for this is shown in (2) below, where the right-hand side of the H_0 equation is inverted – this is because the right-hand side of the H_0 equation is a fraction, and dividing by a fraction is equal to multiplying by the inverse of that fraction:

(2) The ratio of posterior probabilities for H_A and H_0

$$\frac{p(H_A \mid D)}{p(H_0 \mid D)} = \frac{p(D|H_A) \times p(H_A)}{p(D)} \times \frac{p(D)}{p(D|H_0) \times p(H_0)}$$

The next step is to simplify this equation by eliminating p(D) in both the numerator and denominator, and group like terms (likelihoods with likelihoods, priors with priors):

[3] There is much discussion in the literature about this characterization of Bayesian approaches as subjective and to what extent that subjectivity is a positive or negative. There is also discussion in the literature about how successful Fischer actually was in expelling subjectivity from hypothesis testing: although *p*-values do not involve any subjective probabilities, the decision about how to leverage *p*-values is generally subjective.

(3) Deriving Bayes Factors from Bayes Theorem

$$\underbrace{\frac{p(H_A \mid D)}{p(H_0 \mid D)}}_{\textit{Posterior Odds}} = \underbrace{\frac{p(D|H_A)}{p(D|H_0)}}_{\textit{Bayes Factor}} \times \underbrace{\frac{p(H_A)}{p(H_0)}}_{\textit{Prior Odds}}$$

The result is an equation (3) with three odds ratios: the odds ratio of posterior probabilities (posterior odds), the odds ratio of the probability of the data under each hypothesis (the Bayes Factor), and the odds ratio of the prior probabilities (prior odds). From equation (3) it is easy to see that Bayes Factors are independent of the (potentially subjective) priors because they are a separate term in the equation. It is also easy to see that the posterior odds ratio can be calculated from Bayes Factors if one is willing to specify the two relevant priors. In this way, Bayes Factors simultaneously provide information that is useful on their own (the ratio of the probabilities of the data under each hypothesis), and can be used to calculate the posterior odds (the ratio of the probabilities of the two hypotheses under the data) if one so desires.

As an odds ratio, BF can be any value between 0 and ∞. A BF of 1 indicates that the data is equally likely under each hypothesis. This suggests that the experiment is inconclusive, as it does not discriminate between the two hypotheses. A BF greater than 1 indicates that the data is more likely under the alternative hypothesis than the null hypothesis. A BF below 1 indicates that the data is more likely under the null hypothesis than the alternative hypothesis.[4] In this way BFs make a three-way distinction: they can reveal evidence for H_0, evidence for H_A, or that the experiment is uninformative relative to the two hypotheses. This contrasts with FHT, which can only make a two-way distinction: a low *p*-value is evidence against H_0, a high *p*-value is uninformative. This contrasts with NPHT, which makes a different two-way distinction: a *p*-value below the significance level leads to a decision in favor of H_A with a limit on the long-run type I error, and a *p*-value above the significance level leads to a decision in favor of H_0 with a limit on the long-run type II error.

Much like *p*-values in FHT, there are no explicit decision rules for interpreting BFs. This also means that there can be no explicit definition of statistical power when using BFs for BHT. However, Jeffreys (1939/1961) did suggest some guidelines for interpreting BFs that have been generally accepted in the experimental literature as shown in Table 2.

BF	Interpretation
<1/100	extreme evidence for H_0
1/100 to 1/10	strong evidence for H_0
1/10 to 1/3	substantial evidence for H_0
1/3 to 1	anecdotal evidence for H_0
1 to 3	anecdotal evidence for H_A
3 to 10	substantial evidence for H_A
10 to 100	strong evidence for H_A
>100	extreme evidence for H_A

Table 2: Jeffreys (1939/1961) guidelines for interpreting Bayes Factors (specifically BF_{10}).

[4] Crucially, this description assumes that H_A is in the numerator as in equation (3). It is also possible to put H_0 in the numerator. In that case, the interpretation of the odds ratio would be inverted as well (a BF greater than 1 would indicate that the data is more likely under H_0 than H_A). The directionality of the ratio can be indicated by a subscript of 10 for BFs with alternative in the numerator, and 01 for BFs with the null in the numerator: BF_{10} vs BF_{01}.

From these guidelines, Jeffreys suggests a conventional cutoff of 3 for deciding that there is substantial evidence for H_A (or 1/3 for H_0). Therefore, much like the conventional cut off of $p < .05$ for FHT, there is an intuitive concept of statistical power when using BFs for BHT: statistical power is the probability of a test returning a BF greater than 3 when there is in fact strong evidence for H_A.

2.4 Positive predictive value and the type I error rate for the field

Positive predictive value (PPV) is another probability that can be useful when assessing how well a test works (any kind of test, from hypothesis tests to medical diagnostics). PPV is the probability that a positive result of a test (i.e., a statistically significant result or a positive result in a diagnostic test) reflects a true positive result. It is therefore a measure of how informative a positive result truly is. Like any probability, PPV can be directly calculated from frequency counts: if one has run a test a number of times, and has independent knowledge about which results were true positives (the test favored H_A and H_A was true) and which outcomes were false positives (the test favored H_A and H_0 was true) then one can calculate PPV as a simple ratio of true positives results to all of the positives results:

(4) The frequency definition of PPV

$$PPV_{freq} = \frac{true\ positive\ results}{true\ positive\ results + false\ positive\ results}$$

One way to think about this is that while type I and type II error rates are each calculated using one of the columns in Table 1, PPV is calculated using the top row. The converse notion, negative predictive value (NPV), can similarly be defined using the bottom row in the table: the proportion of negative correct decisions (a conclusion of no difference when there is no real difference) divided by the total number of negative decisions (correct decisions + type II errors). An ideal test would minimize both PPV and NPV, but just as there is a tension between type I error rates and type II error rates, there is a tension between PPV and NPV.

Though PPV is easy to calculate using frequency counts, we rarely have real counts of true positive and false positive results. Since PPV is itself a probability, we can define it directly in terms of probabilities. We give the definition in (5) before explaining how it works:

(5) The probability definition of PPV

$$PPV_{prob} = \frac{(1-\beta)\times R}{(1-\beta)\times R+\alpha}$$

The term (1 – β) in equation (5) should look familiar: it is simply the statistical power of the test. The term α should also look familiar: it is the maximum type I error rate of the test (over the long run). The only new component is R. R is the odds (before the study is run) that a given alternative hypothesis is true out of all of the alternative hypotheses that will be tested using that test. For example, if there were only 100 alternative hypotheses in the world that are going to be tested using a specific test (an underestimate, to be sure), and 20 of those alternative hypotheses are true in the world, then R would be 20/80 or .25 (crucially R is an odds ratio, not a probability, so it is 20/80 and not 20/100).

To see that PPV_{freq} and PPV_{prob} are identical, we can continue with our example where there are 100 alternative hypotheses being tested in the world, and 20 are true. Let's further assume that our test has 80% power and a 5% maximum type I error rate. To use

$\mathrm{PPV}_{\mathrm{freq}}$, we need to calculate the number of true positive results (80% power x 20 true H_{A} = 16) true positive results) and the number of false positive results (5% type I error x 80 false H_{A} = 4). Plugging these numbers into $\mathrm{PPV}_{\mathrm{freq}}$ we get:

(6) An example using $\mathrm{PPV}_{\mathrm{freq}}$

$$PPV_{freq} = \frac{true\ positive\ results}{true\ positive\ results + false\ positive\ results} = \frac{16}{16+4} = .8$$

So the PPV for the test is 80%, meaning that 80% of positive results are true positives (and 20% are false positives). If we instead use the equation for $\mathrm{PPV}_{\mathrm{prob}}$, we have R = .25, because the odds of HA being true are 20/80, β = .2, because the test has 80% power, and α = .05, because we are imposing a maximum type I error rate of 5%. Plugging these numbers into $\mathrm{PPV}_{\mathrm{prob}}$ we get:

(7) An example using $\mathrm{PPV}_{\mathrm{prob}}$

$$PPV_{prob} = \frac{(1-\beta)\times R}{(1-\beta)\times R+\alpha} = \frac{(1-\ .2)\times\ .25}{(1-\ .2)\times\ .25+\ .05} = .8$$

The probability definition of PPV (5) reveals the not so obvious relationship between lower statistical power and a higher rate of false positive results. If R and α are held constant, lower statistical power will lead to lower PPV. And because PPV is a measure of the proportion of true positives, lower PPV means higher false positives. Therefore, if R is constant and α is consistently adopted in a scientific field of study, lower statistical power means more false positives for the field. R is by definition a constant for a field over the long run, as it is dictated by the scientific facts under investigation. α can in principle be varied (e.g., publishers could require that low powered studies use a stricter α-level), but in practice, each field tends to have a conventional significance level that will lead to publication (e.g., $p < .05$ for single comparisons). When α is held constant for publication, as it is in most fields, low power means a higher proportion of published false positives.

3 The design and procedure for empirically estimating statistical power

Our goal in this project is to empirically estimate the *rate of statistical detection* (our proxy measure of *statistical power*) that would obtain for the widest possible range of experimental scenarios in theoretical syntax. To that end, we tested four judgment tasks, across a wide range of effect sizes reported in the theoretical syntax literature, and across a wide range of potential sample sizes, and then used resampling simulations to provide empirical estimates of the observed *rate of statistical detection* for each combination of task, effect size, and sample size in our data sets. In this section, we review each of the components of the design in detail.

3.1 Four judgment tasks

There are at least four distinct judgment tasks that are routinely used in the syntax literature. As it is possible that each task may result in different levels of sensitivity to the experimental manipulation (e.g., through different levels of response variability), we decided to test all four in this study. Each task was deployed in a completely separate experiment, such that participants only ever completed one task. In the *yes-no task* (YN), 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) *forced-choice task* (FC), 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. In the (7-point) *Likert scale task* (LS), 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 *magnitude estimation task* (ME), 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).

3.2 The phenomena, and therefore effect sizes, tested from Linguistic Inquiry (2001–2010)

In order to maximize the span of effect sizes for which we could estimate statistical power, we chose 50 two-condition phenomena from the larger set of 150 two-condition phenomena that were randomly sampled from all of the articles published in Linguistic Inquiry between 2001 and 2010 for the large-scale replication study by Sprouse et al. (2013). The 50 phenomena selected here were experimentally replicated in the Sprouse et al. (2013) study, and therefore are assumed to be phenomena for which an H_A is true, a requirement to assess statistical power. In order to choose phenomena that span the range of effect sizes in the literature, we first calculated a standardized measure of effect size known as Cohen's *d* (Cohen 1988) for the full set of 150 phenomena on one of the scale tests (Magnitude Estimation) used by Sprouse et al. (2013). Cohen's *d* is calculated by first subtracting one condition mean from the other, and then dividing the difference by the pooled standard deviation of the two conditions. Cohen's *d* is considered a *standardized* measure of effect size because 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* of 0.2 is considered a "small" effect, a *d* of 0.5 is considered a "medium" effect, and a *d* of 0.8 is considered a "large" effect. Here is what Cohen (1992) said about the intent behind these criteria:

> Because the [effect size] 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: 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.

In the top row of Figure 2 we plot the distribution of effect sizes for the 139 phenomena that were replicated in Magnitude Estimation task in the Sprouse et al. (2013) study in terms of both directionality (the effects were in the predicted direction) and statistical significance (they passed the conventional $p < .05$ criterion):

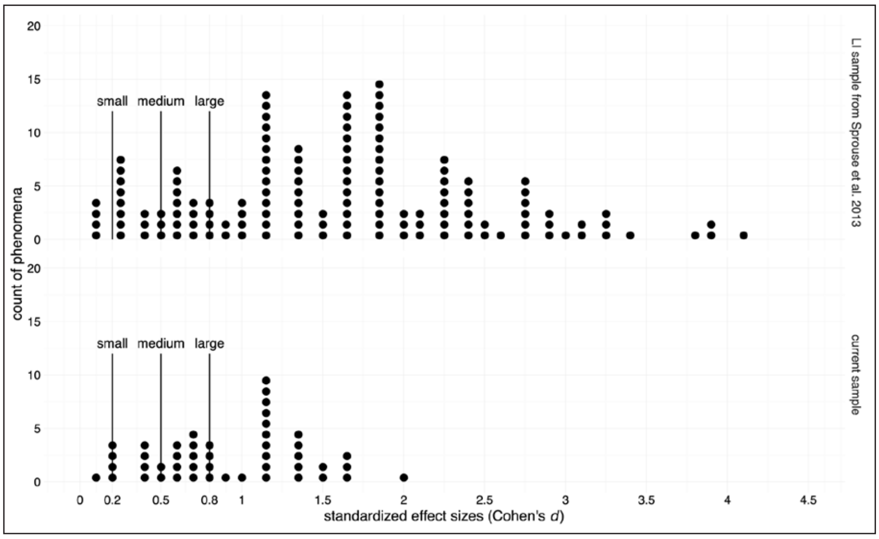

Figure 2: The top row displays the distribution of effect sizes (Cohen's *d*) for the 139 significant, two-condition phenomena from Linguistic Inquiry (2001–2010) as tested by Sprouse, Schütze & Almeida (2013) using the magnitude estimation task (magnitude estimation was chosen because it yields a continuous response measure, which, a priori, is more amenable to calculating a continuous measure like Cohen's *d*). The bottom row displays the distribution of effect sizes (Cohen's *d*) for the 47 phenomena from Linguistic Inquiry (2001–2010) that were tested in the current experiments. The effect sizes are from the magnitude estimation experiment. The vertical lines in both rows mark Cohen's (1992) suggestions for small, medium, and large effect sizes.

As Figure 2 makes clear, the vast majority of phenomena randomly sampled from Linguistic Inquiry are "large" by Cohen's criteria, with 110 phenomena (79%) yielding a d greater than or equal to 0.8. Counting the "small" and "medium" categories is more difficult because it depends on where the category boundaries are placed. But one possibility is to use the values suggested by Cohen as boundaries: 13% of the phenomena from Linguistic Inquiry are below the "medium" threshold ($d < 0.5$), and 8% are between the "medium" and "large" thresholds ($0.5 < d < 0.8$). Because the distribution of effect sizes in Linguistic Inquiry 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%) with very small sample sizes, we decided to restrict our test set in this paper to a subset of 50 phenomena from the smaller half of the range ($0 < d < 2$). We believe this range will generally be more useful to researchers seeking to design their own studies or evaluate existing studies. Each of the 50 phenomena were tested in each of the four tasks (one experiment per task, with each experiment containing 100 items: one token each of the two conditions per phenomenon). After running the experiments in this study we noticed typos in the materials for three of the phenomena (see the materials posted at www.sprouse.uconn.edu for the Sprouse et al. 2013 paper for a detailed discussion of any hypothetically possible problems with the materials). We thus analyzed the remaining 47 phenomena. The distribution of the effect sizes for the remaining 47 phenomena from Linguistic Inquiry (2001–2010) are presented in the bottom row of Figure 2. A full list of the phenomena with example sentences are provided in the appendix.

One issue merits some clarification before we move on to the other components of the analysis. Because we selected 47 real-world phenomena for analysis, it may appear as though we are attempting to estimate the post-hoc statistical power for these 47 real-world phenomena under different tasks (and different sample sizes). While post-hoc power analyses are sometimes performed in the literature to gather some amount of information about the test of a given phenomenon, post-hoc power tests technically do not yield statistical power as Neyman-Pearson defined the term. As discussed in section 2.2, statistical power can only be used to control the probability of (type II) decision errors if it is defined a priori and incorporated into the design of the hypothesis test as a whole (e.g., the choice of task, significance criterion, and sample size). So we would like to be clear that we are not trying to make any claims about the statistical power of these particular tests. Instead, we are taking the 47 effect sizes that we obtained using these phenomena, and the 47 patterns of variability that we obtained using these phenomena, and using them to estimate the *a priori* rate of statistical detection that would obtain if one wanted to test *minimum effect sizes* that are equal to these 47 empirical effect sizes using the judgment tasks discussed in section 3.1 (and assuming that those experiments will be subject to the empirical variability that we observed in our experiments). Put differently, we are attempting to estimate *a priori* power for effects that have statistical properties similar to the properties of these 47 phenomena; we are not attempting to estimate the post-hoc power of these particular phenomena.

3.3 The materials

The materials for all four experiments were identical to the materials constructed for the original Sprouse et al. (2013) 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 Linguistic Inquiry (2001–2010). This led to 8 lexically matched sentence sets for each phenomenon.

For the ME, LS, and YN experiments, the 8 lexicalizations were distributed among eight lists using a Latin Square procedure, ensuring that participants did not see more than one sentence from each lexically-related set. Each list was pseudorandomized such that the two conditions from a single phenomenon 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" used to help the participant establish the scale prior to rating experimental items). This resulted in eight surveys that were 106 items long.

For the FC experiment, the 8 lexicalizations were maintained in pairs based on the two-condition phenomena. The lexically-matched pairs comprising a phenomenon were distributed among 8 lists. Next, the order of presentation of each phenomenon pair was counterbalanced across the lists, such that for every phenomenon 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 total lists. Finally, the order of the pairs in each of the 16 lists was randomized, resulting in 16 surveys containing 50 randomized and counterbalanced pairs (100 total sentences).

It should be noted that within this design 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 informal collection methods and more formal experiments, this number is quite low. We chose to only test one token of each condition per participant for two reasons. First, this is the lowest limit of possible experimental designs. This means that the estimates of rate of statistical detection that we present here will provide an absolute lower bound for such experiments. By simply increasing the number of tokens per condition to 2 or 4, syntacticians can easily increase the power of their experiments 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 106 for YN, LS, and ME, and 100 for FC). Because it is useful to z-score transform judgments made on scales (e.g., LS and ME) to eliminate some forms of scale bias, it important 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.

3.4 Presentation of the experiments

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 (AMT) 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.

3.5 Participants and sample sizes

Statistical power is proportional to sample size because sample size is one factor that determines the width of the sampling distributions, and therefore the overlap of the two sampling distributions of the two hypotheses (see Figure 1). In order to have the ability to estimate power for a wide range of sample sizes, we recruited 144 participants for each of the four experiments (144 per task, for 576 participants total). Participants were recruited online using the Amazon Mechanical Turk 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 (the differences in pay were based on previously observed differences in the amount of time it takes to complete each task). 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). These large sample sizes allowed us to treat our samples as mini populations for the resampling simulations. For the resampling simulations, we sampled (with replacement) from each population in order to estimate power for sample sizes from 5, one of the lowest sample sizes that can return a significant result, to 100, a likely upper bound for most acceptability judgment experiments.

3.6 The resampling simulations

We empirically estimated the *statistical detection rate* (our proxy measure for statistical power) observed in each experiment type for each phenomenon at every sample size between 5 and 100 participants, by performing resampling simulations on each sample. In essence, these resampling simulations treated our large samples 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 (see section 3.7 for the choice of tests).
3. Repeat steps one and two 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 test statistic beyond the pre-established criterion for significance (or α; see section 3.7 for the choice of criteria). This proportion is an empirical estimate of statistical power for a sample size of 5.

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… up to 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 0.15 (very small) and 1.96 (very large) in Linguistic Inquiry (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 & Tibshirani 1993; Edgington & Onghena 2007).

3.7 Statistical tests

In addition to the choice of task, effect size, and sample size, statistical power depends on the choice of statistical test that is chosen as part of the hypothesis test. On the one hand, we wanted these simulations to provide the widest range of possible information for syntacticians. Therefore, we decided to test both a null hypothesis test and a Bayesian test for every simulation. On the other hand, the large number of combinations of test properties (task, effect sizes, and sample sizes) meant that we needed to run around 18 million simulations (requiring substantial computational time). Thus, we decided to use statistical tests that were relatively fast to compute: for LS and ME we ran repeated-measures *t*-tests and Bayes Factor calculations from Rouder et al. (2009); for FC and YN, we ran repeated-measures sign-tests and Bayes Factor. For LS and ME, we ran the statistical tests on the z-score transformed ratings, as z-scores help to remove some forms of scale bias (Featherston 2005; Sprouse & Almeida 2012; Sprouse et al. 2013).[5]

[5] The z-score transformation is applied to each participant separately. First, the mean of all of the participant's ratings is calculated. The mean rating is then subtracted from each raw score, in effect centering the scale around 0, and making the mean an anchor point to facilitate standardization. Next, the standard deviation of the participant's raw scores is calculated. Each difference from the mean is then divided by the standard deviation. This standardizes the scale, as the units for each participant are now standard deviations from the mean.

One may wonder whether our choice of *t*-tests and sign-tests might be a problem, as these tests only treat subjects as random effects, and crucially do not allow items to be treated as random effects. This limitation of *t*-tests and sign-tests has two potential implications. First, when items are not treated as random effects, it is not possible to statistically evaluate how well the experimental effect generalized across the items (because the items are basically averaged together in the statistical test). Second, when items are not treated as random effects, it is possible for variation between items to be confounded with the experimental effect (i.e., when one item only appears in one condition, and a second item only appears in a second condition, the difference between conditions includes the random differences between items; e.g., Clark 1973). One avenue for dealing with these two issues is to use mixed effects models that allow for the specification of items as random effects (e.g., Baayen et al. 2008).

Our response to this is one of weighing costs and benefits (see also Cohen 1976; Keppel 1976; Smith 1976; Wike & Church 1976; Wickens & Keppel 1983; Raijmaakers 2003). The cost of mixed effects models is that they are substantially slower to fit than *t*-tests and sign-tests. In the context of our 18 million simulations, it would have added a month or more of computational time. So the question is whether the benefits are important (or even necessary) given the goals of our project. When it comes to evaluating the generalization of the effects across items, we would say that while this is an interesting question about any given phenomenon, this question is not in the scope of the current project. We are not interested in the properties of the specific phenomena that we used in this study, but rather the power that would obtain under different experimental designs if one were interested in testing minimum effect sizes that are equal to the observed effect sizes of these phenomena (assuming similar variability to the variability observed with these phenomena). However, we should point out that we did use multiple items per condition in our design (8 tokens per condition), as did Sprouse et al. (2013) for the larger set of Linguistic Inquiry phenomena. Therefore, multiple items did contribute to the power estimates reported here (which is at least a step in the direction of establishing statistical generalizability, as any effects that were driven by only a few items would presumably show lower power estimates).

When it comes to the concern about confounding item variability with the experimental effect, it is important to note that the empirical concern is that such a scenario will fail to precisely control the two error rates (type I and type II): if item variability accidentally adds to the experimental effect, a type I error would be more likely to occur, and if item variability accidentally detracts from the experimental effect, a type II error would be more likely to occur. Because all of the phenomena in our study are assumed to be true effects, by hypothesis there can be no true type I errors. So for us the concern is really to what extent our power estimates are accurate: if item variability adds to the experimental effect, we will overestimate power; if item variability detracts from the experimental effect, we will underestimate power. We decided that such estimate problems were likely to be small in the current study. For one, we used lexically matched materials distributed using a Latin Square procedure, which reduces much (but not all) of the concern about item variability confounding the experimental effect (e.g., Wickens & Keppel 1983; Raaijmakers 2003). Perhaps more importantly, in the discussion sections below, we either interpret the differences in power among tasks, effect sizes, and sample sizes *relatively to each other*, not absolutely, or we interpret the estimates as rough guides, not specific criteria. Thus, small perturbations in the power estimates one way or another are less likely to influence the overall interpretation of the results. That being said, the data sets for both projects (this one, and Sprouse et al. 2013) are freely available online, therefore any researchers interested in the generalizability of effects across items in these phenomena can use those results to investigate this question.

3.8 Decision criteria

As discussed in section 2, statistical power is only quantifiable relative to an explicit decision criterion for choosing between competing hypotheses. For this study we simply adopted the conventional criteria that tend to be used for publication in the cognitive sciences: $p < .05$ for *t*-tests and sign-tests, and BF > 3 for Bayes Factors. There has been much discussion recently surrounding the use of these criteria for publication, especially as regards *p*-hacking and the file-drawer problem (e.g., Simmons et al. 2011). We believe this discussion is healthy for the cognitive sciences in general and should continue, although we note that it is at best unclear how these concerns impact generative syntax research, which by and large has eschewed the use of inferential statistics as regular research practice. However, given that these criteria still play a central role in current practices for hypothesis testing and for publication in the cognitive sciences, they seemed like a natural starting point for a study of this type.

3.9 Comparison to previous investigations of sensitivity

Before moving on to a discussion of the results of our simulations, 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 sensitivity (though most did not use the term statistical power). 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 extends these results in several ways. First, instead of investigating a single sample size and/or using a categorical criterion (statistical significance or not; a larger sample size 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. (2013). 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 47 phenomena tested here were not chosen at random, they were selected without theoretical bias from a random sample of 150 phenomena from Linguistic Inquiry, which suggests that the results will be relatively representative of effect sizes in cutting edge syntactic research. Third, the resampling simulations (over 18 million randomly selected samples) allow us to empirically estimate the rate of statistical detection 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 experimental survey of acceptability judgments, from previously published informal collection studies to the design of future formal experiments.

4 The statistical power of acceptability judgment experiments

The resampling simulations result in 18,048 rate of detection estimates (our proxy measure for statistical power): 4 tasks x 47 phenomena x 96 sample sizes. Therefore, interpreting the results requires some amount of summarization. We present two different approaches to summarizing the results in this section, and summarize the implications of the results.

4.1 The relationship between sample size and mean estimated power

The first approach we can take is to ask how much the rate of statistical detection depends on different sample sizes. Statistical power is always relative to a given effect size, therefore in order to make the relationship between sample size and power clearer, in this approach we group the 47 effect sizes that we investigated into four categories following Cohen's (1992) criteria: small effects have a *d* less than 0.5, medium effects have *d* between 0.5 and 0.8, large effects have a *d* between 0.8 and 1.1, and extra large effects have a *d* greater than 1.1.

We can then plot the relationship between sample size (x-axis) and the resulting power (y-axis) for each category of effect size. The result is the 4x4 grid of power curves in Figure 3.

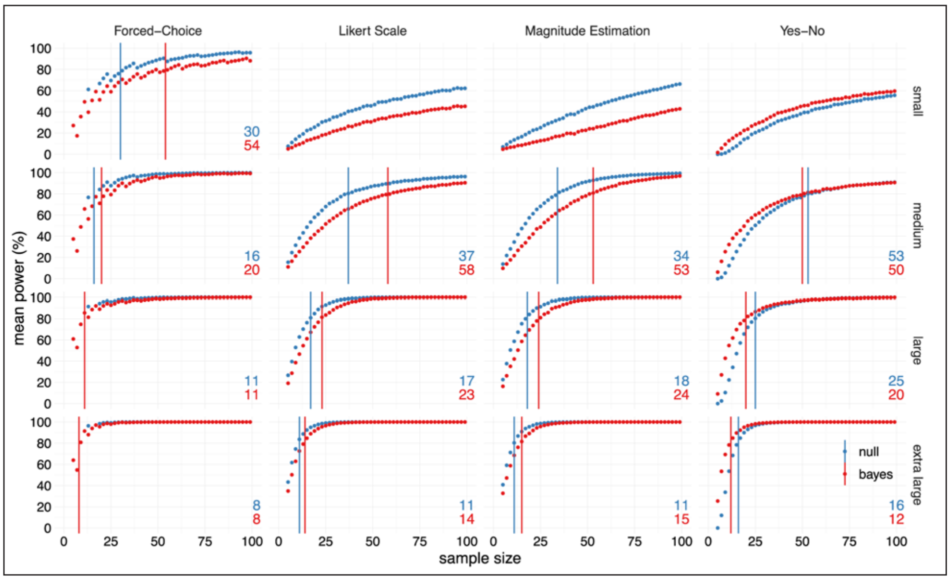

Figure 3: Power curves for null hypothesis tests (blue dots) and Bayes Factor tests (red dots) displaying the relationship between sample size and estimated power for each task (columns) and each category of effect size (rows). Each point represents the empirical estimate of power (the percentage of simulations (out of 1000) that was below the significance threshold of *p* < .05 for null hypothesis tests, and BF > 3 for Bayes Factors) averaged over all of the phenomena that belong to each category. The vertical lines represent the sample size that first reaches 80% power (or above): blue for null hypothesis, red for Bayes Factors. The colored numbers in the lower right hand corner indicate the precise sample sizes where 80% power is first reached in our simulations. Cells with only one line indicate that the 80% power threshold is obtained with the same sample sizes in the two statistical analyses. Cells with no line did not reach 80% power with sample sizes less than or equal to 100. For increased clarity, only the even-numbered sample sizes (6 to 100) are plotted.

Figure 3 serves three purposes. First and foremost, it is a representation of most of the information revealed by the resampling simulations. It presents the estimated power that one would obtain for sample sizes from 5 to 100, for all four different tasks, for both approaches to hypothesis testing, and for four categories of effect sizes. The only (minor) loss of information occurs because of the division of effect sizes into four categories (instead of 47 distinct effect sizes). Second, Figure 3 presents an explicit comparison of the power of the two approaches to hypothesis testing. Our results suggest that, for equivalent sample sizes, null hypothesis approaches reach rates of statistical detection that are slightly higher than those obtained by Bayes Factors for FC, LS, and ME, but the reverse seems to be true for YN. These differences are most pronounced for LS and ME, and most pronounced for small and medium effect sizes. For large and extra large effect sizes, the two approaches appear to be substantially identical in the observed relationship between sample size and rate of statistical detection. Third, Figure 3 can be used to estimate the sample size at which a given statistical power will be reached. We have made the point at which 80% power is reached explicit in the plots (the vertical lines) because this is the standing recommendation in the statistical literature (Cohen 1988), but readers can substitute any power level of interest. The vertical lines make it easy to assess which tasks reach 80% power with smaller sample sizes: FC appears to require the smallest sample sizes, LS and ME are nearly identical, and YN requires the largest sample sizes. Once again, these differences are most pronounced for small and medium effect sizes, and least pronounced with large and extra large effect sizes.

Figure 4 makes the comparison among the tasks more explicit by plotting the tasks together in the same cell. Once again, it is clear that FC has a power advantage over the other tasks, especially for small and medium sized effects. This advantage makes sense given that the forced-choice task explicitly asks participants to judge whether there is a difference between two (lexically matched) conditions (see Gigerenzer & Richter 1990 for a similar point in a non-linguistic domain). Similarly, Figure 4 makes it clear that LS and ME exhibit roughly the same relationship between sample size and rate of statistical detection, with a slight advantage to LS under Bayes Factors for smaller effect sizes. Again, this rough identity and small advantage for LS makes some sense given previous research into LS and ME showing that participants do not really perform ME as it is intended (Sprouse 2011b) and that LS judgments exhibit less variability than ME (Weskott & Fanselow 2011), most likely because LS offers fewer response options than ME (7 vs infinite). Finally, Figure 4 reveals an interesting confluence of effects that lead to YN performing better than LS and ME under Bayes Factors, particularly for smaller effect sizes. It appears that YN receives a 5% power increase under Bayes Factors, while LS and ME simultaneously receive a 10% power decrease. These two changes conspire to give YN the second highest detection rate under Bayes Factors (FC is still first) for smaller effect sizes.

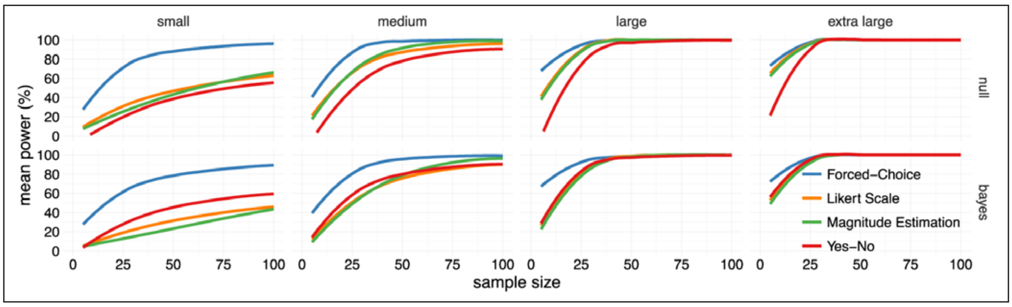

Figure 4: Power curves for null hypothesis tests and Bayes Factor tests displaying the relationship between sample size and estimated power, organized by effect size category (columns), with all four tasks plotted together. For clarity, only the (loess) fitted lines are plotted (no data points).

4.2 The sample size required to reach 80% power for each effect size tested

The second approach we can take is to ask how large of a sample is necessary to produce a well-powered experiment for a given effect size under investigation. For expository purposes we assume Cohen's (1988) suggestion that 80% power is necessary for a well-powered experiment. We can then plot effect sizes along the x-axis and the sample size required to reach 80% power along the y-axis for each task, as in Figure 5.

The primary value of Figure 5 is that the fitted lines can be used for very quick estimation of the sample size needed for a well-powered experiment at any given effect sizes. However, these plots also reveal that there is quite a bit of variation in sample sizes required for smaller effect sizes. This is likely partly because of sampling error (the effect sizes were all calculated from one experiment, the ME experiment, not for each task separately), and partly due to differences in measurement error (each combination of task and phenomenon likely has slightly different levels of noise in the measurements). This means that syntacticians should take both the fitted estimate and the variability in the points into consideration when estimating the sample sizes required for well-powered experiments. Finally, Figure 5 also continues to reveal that Bayes Factors have slightly lower detection rates than null hypothesis tests for FC, LS, and ME.

5 Implications of these results for research in syntax

The experiments and resampling simulations here provide a wealth of information about the rate of statistical detection of syntactic experiments (summarized in the figures in section 4, and the appendix). In this section, we discuss the two ways that this information may be useful to the field: evaluating the validity of judgment data in syntax, and designing/evaluating judgment experiments.

5.1 The validity of judgment data

One of the central questions in experimental syntax over the past two decades has been the question of whether the informally collected judgments that constitute the majority of evidence are type I errors or not (e.g., Sprouse & Almeida 2012; Gibson & Fedorenko 2013; Sprouse et al. 2013). The primary tool in this discussion has been replication studies designed to estimate the positive predictive value (PPV) of the field. Replication is one way to begin to sort out true positive results from type I errors, as true results will by definition be more likely to replicate than type I errors (especially over multiple replications). However, this logic only holds if the replication studies have high statistical power. If the replications have low statistical power, then any null results that arise in the replications are more likely to be type II errors (favoring a H_0 when H_A is true), and therefore can't be

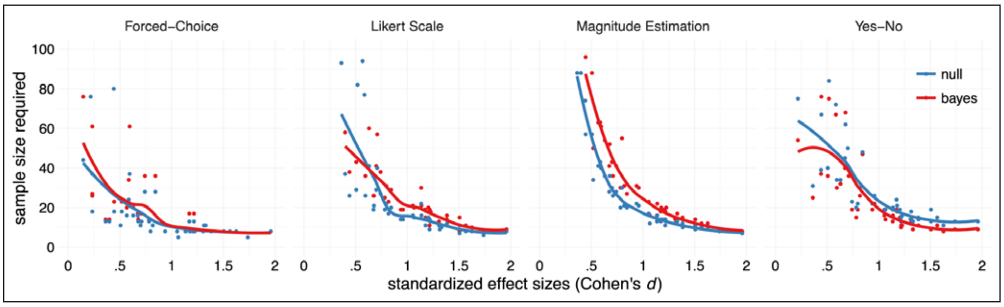

Figure 5: Power curves for null hypothesis (blue) and Bayes Factor (red) tests displaying the relationship between effect size and sample size at 80% power for each task. Both the specific values and fitted curves (loess) are plotted. Effect sizes are calculated relative to magnitude estimation (hence the fitted line for magnitude estimation has the best fit). Cohen's suggestions for small (0.2), medium (0.5), and large (0.8) effect sizes are indicated on the x-axis.

used to infer that the original result was a type I error. In short, it all comes down to the logic of interpreting null results as evidence for the null hypothesis discussed above: null results can only reasonably be taken as evidence for H_0 if statistical power is high. Despite this, to the best of our knowledge, there has been no explicit discussion of power in the replication debates inside of linguistics. The two large scale replication projects (Sprouse & Almeida 2012; Sprouse et al. 2013) estimate type I error rates to be between 1% and 12%, depending on the sample, the experimental method used, and the criterion for what counts as a successful replication. But these rates can only be interpreted in the context of high statistical power. With low power, the null replications that are interpreted as type I errors could be type II errors. As the current study demonstrates, this could be especially problematic for replications of small effect sizes, particularly for experiments using LS and ME (which are typical in the experimental syntax literature), as small effect sizes do not reach even 80% power using LS and ME with 100 participants, which in our experience is quite large for an experimental syntax sample size. The caveat to the latter statement is that our experiment provides the absolute lowest bound for power, as only one item per experimental condition was used, which means that, if multiple items per condition can realistically be used in an experiment, statistical power can be very much improved. Syntacticians interested in interpreting the various claims about replicability of syntactic data can use the results of this study to evaluate whether any purported null replications have sufficient power to be interpreted as evidence for the null hypothesis.

Given the current interest in replicability in psychology and linguistics, it may also be interesting to explore how data in syntax compares to data in psychology, both in terms of type I errors and statistical power (type II errors). As previously mentioned, the two large scale replication studies in syntax find type I error rates between 1% and 12% (i.e., replication rates in the range of 88% to 99%). The journal Science has published a large scale replication attempt by a large international team of psychologists and cognitive scientists using 100 articles selected from four leading journals in different sub-areas (Open Science Collaboration 2015). This study reports aggregate replication rates between 36% and 68% depending on the criteria adopted for what counts as a successful replication. The replication rates for the experiments categorized as part of the subfield of cognitive psychology (arguably a subfield linguistics is a part of, at least in the generative tradition) were between 48 and 92%, again depending on the criteria of what counted as a successful replication. This suggests that syntax compares rather favorably to the rest of the field of psychology in terms of estimated type I error rates. When it comes to statistical power (or type II error rates), there have been several studies about the *a priori* power that experiments in the psychology literature possess. 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 very different from 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 of 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. Comparing these results with syntactic data is a bit tricky because the information necessary to estimate statistical power for informal experiments in syntax is rarely published. Nonetheless we can use the results of the current study to provide ranges of possible power rates depending on the assumptions that one makes about the properties of informal linguistic experiments. For example, the median effect size in the random sample of 150 phenomena from

Linguistic Inquiry 2001–2010 tested by Sprouse et al. (2013) was a *d* of 1.61. If we take this as the "average effect size" similar to the psychology studies, we can ask what the statistical power of this effect size would be under different tasks and different sample sizes. Figure 6 and Table 3 provide this information for the specific phenomenon that has this effect size. The results suggest that 6 participants already provide better power than what has been reported for average psychology experiments in leading journals, for FC, LS, and ME. Ten participants provide over 80% power for LS and ME, and close to 80% power for FC and YN. If these are likely sample sizes for informal experiments (which we believe they are), syntactic data compares favorably to the rest of psychology in terms of how much statistical power their experiments have, presumably because effect sizes in syntax tend to be large (see also Figure 2).

5.2 The design and evaluation of judgment experiments

For readers wishing to evaluate published experiments, the first and most obvious method for evaluating power is to analytically calculate a post-hoc power estimate. For readers who do not wish to do that work, the graphs and tables from this project can provide a rough estimate of the likely power of the experiments. Readers simply need to extract the specific design components from the study: task, philosophical approach to hypothesis testing, statistical test, significance criterion, and sample size. Readers can then look up the estimated power in the graphs and tables. For syntacticians planning to conduct their own studies, the graphs and tables from this project can be used to provide a rough estimate of the sample size required to reach a specific level of power depending on the hypothesized effect size of interest. Again, the process involves specifying the components of the experiment (task, philosophy, statistical test, significance criterion, and effect size), and using those components to locate the sample size required for the desired level of power. While this process is relatively straightforward in principle, a few words are in order about what this entails in practice.

The first component is the choice of task. The choice of task should follow from the type of information that is required by the hypotheses. FC exhibits a clear advantage in sensitivity for detecting pairwise contrasts under a relative approach to acceptability.

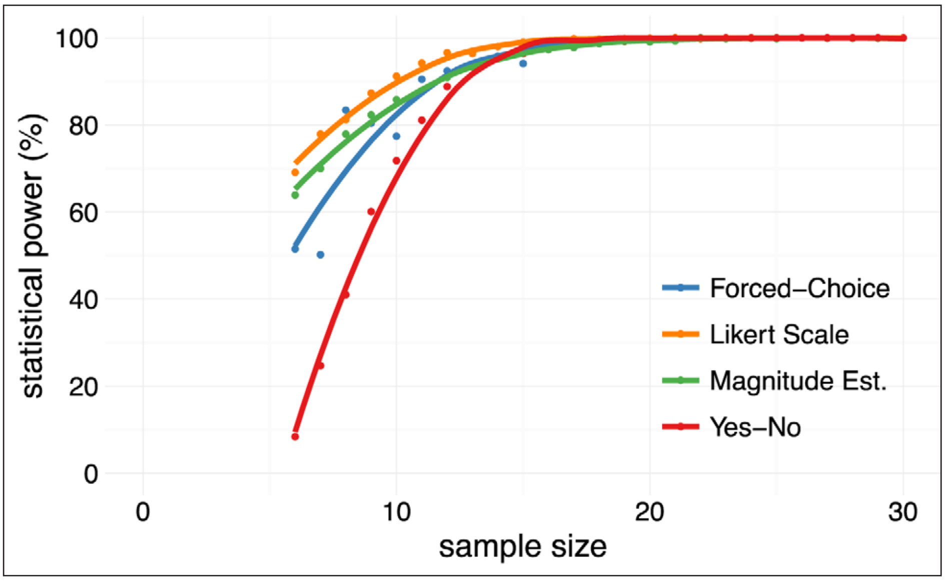

Figure 6: The statistical power (y-axis) for the median effect size (Cohen's *d* of 1.61) reported in Linguistic Inquiry (2001–2010) that would be detectable by each task (lines) as a function of the sample size (x-axis) assuming only 1 judgment per participant per condition.

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%

Table 3: Statistical power for a phenomenon with a Cohen's *d* of 1.61 for several sample sizes, assuming only 1 judgment per participant per condition.

However, FC will be ill-suited for hypotheses that require more than two conditions. FC will also be ill-suited to post-hoc exploration, as the only comparisons that are possible are those that are directly built into the experiment as pairs. FC will also be less well-suited for hypotheses that hinge on differences in effect sizes between pairs (e.g., factorial designs; though not impossible). LS and ME are well-suited for both relative and absolute approaches to acceptability. They are also well-suited for comparisons involving multiple conditions and differences between effect sizes (e.g., factorial designs), and they both allow for the comparison of every condition to every other condition. YN is perhaps the best choice for absolute approaches to acceptability, as it explicitly asks participants to divide sentences into acceptability categories with meaningful labels (e.g., "acceptable" and "unacceptable"). YN shares FC's disadvantage when it comes to factorial designs and hypotheses predicated upon effect sizes, but it shares LS and ME's advantage when it comes to allowing comparisons between every condition in the experiment. Furthermore, the choice between FC and YN is theoretically-laden. FC is appropriate for hypotheses that rely on a relative notion of acceptability, i.e., whether condition X is more/less acceptable than condition Y. However, FC is inappropriate for hypotheses that rely on a more absolute, or categorical, notion of acceptability, i.e., whether condition X is high on the rating scale (or considered categorically "acceptable"). The converse is true for YN. YN is at a disadvantage for relative acceptability, particularly when the two conditions both fall in the same category; however, YN (and other categorical tasks) are perfectly suited for absolute acceptability (though the number of categories should match the theory).

The second component is the choice of philosophical approach to hypothesis testing. This should follow from the type of information that the syntactician would like to use to make inferences. Null hypothesis approaches provide the probability of obtaining the observed data (or data more extreme) under the null hypothesis: p(data | H_0). Full Bayesian modeling approaches can provide the probability of a specific hypothesis given the observed data, p(H | data), but require the full specification of a prior probability and a likelihood. Bayes Factor approaches provide an odds ratio of the probability of obtaining the data under two hypotheses (typically H_A and H_0). Relatedly, the choice of statistical test should follow from the interaction of the type of data being collected and the philosophical approach that has been chosen. FC and YN require binomial statistical tests (e.g., sign-tests, logistic regression), whereas LS and ME require continuous statistical tests (e.g., *t*-tests, linear regression). As discussed in section 2, the type I and type II error rates are set explicitly in NPHT by α and β (where power is $1 - \beta$), and implicitly in FHT and BHT by the choice of significance criterion and power level.

The final component is the effect size of the phenomenon of interest. This, of course, follows from the properties of the phenomenon itself. The (minimum) effect size must be specified before the experiment is conducted, in order to calculate the intended power and target sample size; however, in many cases we simply will not have a reasonable estimate for the effect size of a syntactic phenomenon before the experiment (because syntactic theories do not currently make effect size predictions). One possibility is to use the results of this experiment (or the other large scale replications) to estimate the effect size. Native speakers can search for phenomena in the appendix that, to their judgments, have roughly the same effect size as the phenomenon of interest (i.e., the two sentences in the appendix feel like they are the same distance from each other as the two sentences in the phenomenon of interest). Syntacticians can then use the effect size reported in the appendix as an estimate of the effect size for the new phenomenon of interest (with the caveat that it is only as good as the judgments used in the estimation procedure). This is obviously not ideal – it would be much better for syntactic theories to predict effect sizes. But in the absence of that information, large scale projects such as this one provide a possible solution.

Finally, it is important to note that the prospective use of the estimates of statistical rate of detection provided here should be used as an *absolute lower bound* because only one item per experimental condition was used here, making them conservative estimates if the experimental design under consideration uses multiple items per condition.

5.3 The role of effect sizes

As we have seen, effect sizes play a critical role in any discussion about statistical power. Even though effect sizes are rarely discussed in the syntax literature, we'd like to end this section by pointing out that one take-home message of this project is that effect sizes have been lurking in the background of syntactic theory since the beginning, and that it may be time to start discussing them explicitly as part of syntactic theories. Effect sizes are in the background of every hypothesis test, whether formal or informal, that is conducted in syntax. For example, we know before the experiment is run that there is a difference between the conditions in our experiment because we designed the conditions to have a difference. This difference may be large or exceedingly small, but it is there, because we put it there. The question that we are answering with our hypothesis test is not whether there is an actual difference between sentence types (because we know there is), but whether the structural difference we introduced leads to a large enough difference in acceptability to be detected in our experiment (i.e., is the difference larger than the measurement noise?). In other words, when we interpret a statistically significant result as revealing "a difference", what we are really saying is that the difference that we built into the conditions is larger than the minimum effect size that could be detected by our experiment. And when we interpret a null result as revealing "no difference", what we are really saying is that the difference that we built into the conditions is smaller than the minimum effect size that could be detected by our experiment. Claims of statistical (in)significance are effect size claims, even if the effect size is never mentioned explicitly.

This suggests that we should start discussing effect sizes explicitly, both in the design of our experiments, and in our syntactic theories themselves. Effect sizes already play a role in the development of many theories. Keller (2000) develops a grammatical theory (Linear Optimality Theory) that predicts gradient effect sizes using gradient constraint weights. Featherston (2005) develops a grammatical theory that predicts gradient effect sizes using gradient probabilities. Chomsky (1986) famously uses effect sizes to infer the number and type of violations. Hoji (2015) develops a heuristic that assumes that grammatical effects are likely to have larger effect sizes than extra-grammatical effects. There are many others that we do not have space to mention here. Though these proposals differ dramatically in their details, what they share is the idea that we need to avoid the mistake of treating statistical hypothesis testing as a discovery procedure for "differences" vs "no differences", and instead focus on the real information that experiments provide: estimates of the effect size.

6 Conclusion

Our two goals in this paper were (i) to provide a fuller picture of the status of acceptability judgment data in syntax (i.e., a complement to the validity experiments in Sprouse et al. 2013 and Sprouse & Almeida 2012), and (ii) to provide detailed information that syntacticians can use to design and evaluate the sensitivity of acceptability judgment experiments. To that end, we conducted a set of experiments and simulations to cover a wide range of possible experimental designs, fully crossing four acceptability judgment tasks (yes-no, two-alternative forced-choice, Likert scale, and magnitude estimation), a set of 50 real phenomena that span a large portion of effect sizes in the literature, sample sizes from 5 to 100 participants (obtained through resampling simulations out of a sample of 144 per task), and two approaches to hypothesis testing (null hypothesis testing and Bayesian hypothesis testing). The result is a database of information regarding the rate of statistical detection (our proxy measure of statistical power) that covers a substantial portion of possible experimental designs in syntax, and that is freely available to syntacticians on the first author's website for use in the design and analysis of judgment experiments.

The results of our experiments and resampling simulations revealed several notable trends in the statistical power of acceptability judgment experiments. First, the forced-choice task is generally the most sensitive task (of the four we tested) for detecting differences between two conditions (which makes sense given the fact that the FC task is the only one to explicitly contrast two conditions). Second, the Likert scale and magnitude estimation tasks have roughly the same sensitivity across effect sizes and sample sizes. This accords well with previous research suggesting that participants cannot actually perform the magnitude estimation task as imagined by psychophysicists for linguistic material (Sprouse 2011b; Weskott & Fanselow 2011), and suggests that participants might instead treat the magnitude estimation task as a modified Likert scale task. Third, the yes-no task is generally the least sensitive task. Again, this makes sense given that the task is predicated upon a category boundary that may not separate all of the pairs of sentences in the phenomena that we tested. Finally, and perhaps most importantly, these results suggest that syntax compares favorably with other domains of cognitive science in terms of statistical power. Of course, it is difficult to assess the statistical power of informal experiments; nonetheless, these results suggest that acceptability judgment experiments with relatively small sample sizes are actually relatively well powered to detect many of the phenomena currently in the syntax literature. This is compatible with previous investigations that demonstrated low type I error rates in the field of syntax, as far as well-studied languages such as English are concerned (Sprouse & Almeida 2012; Sprouse et al. 2013) and, in combination with these, the results presented here provide a more complete picture of the quantitative properties of acceptability judgment experiments, and provide a foundation of baseline data that syntacticians can use to plan and evaluate their own acceptability judgment studies.

Abbreviations

Alpha (α) = upper limit for type I error rate in NPHT, Beta (β) = upper limit for the type II error rate in NPHT, AMT = Amazon Mechanical Turk, BF = Bayes factor, BHT = Bayesian hypothesis testing, FC = Forced-choice task, FHT = Fisher hypothesis testing, H_0 = Null Hypothesis, H_A = Alternative Hypothesis, LS = Likert scale task, ME = Magnitude estimation task, NPHT = Neyman-Pearson hypothesis testing, NPV = negative predictive value, PPV = positive predictive value, YN = Yes-or-no task

Additional Files

The additional files for this article can be found as follows:

- **Appendix.** https://doi.org/10.5334/gjgl.236.s1

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant no. BCS-0843896 to JS. We would like to thank two anonymous reviewers for helpful comments on an earlier draft. We would also like to thank audiences at the École Normale Supérieure, Michigan State University, the University of Michigan, Princeton University, Harvard University, and Johns Hopkins University. We would also like to thank Carson T. Schütze for assistance in creating the test materials used in this study. All errors remain our own.

Competing Interests

The authors have no competing interests to declare.

References

Baayen, Harald, Doug Davidson & Douglas Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59. 390–412. DOI: https://doi.org/10.1016/j.jml.2007.12.005

Bader, Marcus & Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46. 273–330. DOI: https://doi.org/10.1017/S0022226709990260

Bard, Ellen, Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72. 32–68. DOI: https://doi.org/10.2307/416793

Bayes, Thomas. 1764. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* 53. 370–418. DOI: https://doi.org/10.1098/rstl.1763.0053

Bezeau, Scott & Roger Graves. 2001. Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology* 23. 399–406. DOI: https://doi.org/10.1076/jcen.23.3.399.1181

Button, Katherine, John Ioannidis, Claire Mokrysz, Brian Nosek, Jonathan Flint, Emma Robinson & Marcus Munafò. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14. 365–376. DOI: https://doi.org/10.1038/nrn3475

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

Chomsky, Noam. 1986. *Barriers*. Cambridge, MA: MIT Press.

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. DOI: https://doi.org/10.1111/j.2044-8295.1997.tb02621.x

Clark, Herbert. 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. DOI: https://doi.org/10.1016/S0022-5371(73)80014-3

Cohen, Jacob. 1962. The statistical power of abnormal social psychological research: A review. *Journal of Abnormal and Social Psychology* 65. 145–153. DOI: https://doi.org/10.1037/h0045186

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 edition; Hillsdale, NJ: Erlbaum

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

Featherstone, Sam. 2005. Universals and grammaticality: Wh-constraints in German and English. *Linguistics*, *43*(4): 667–711. DOI: https://doi.org/10.1515/ling.2005.43.4.667

Fisher, Ronald. 1925. *Statistical methods for research workers*. Oliver and Boyd.

Fisher, Ronald. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B. Statistical Methodology* 17. 69–78.

Gibson, Edward & Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28(1–2). 88–124. DOI: https://doi.org/10.1080/01690965.2010.515080

Gigerenzer, Gerd. 2004. Mindless statistics. *Journal of Socio-Economics* 33. 587–606. DOI: https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, Gerd & Hans Richter. 1990. Context effects and their interaction with development: Area judgments. *Cognitive Development*. 235–264. DOI: https://doi.org/10.1016/0885-2014(90)90017-N

Hoji, Hajime. 2015. *Language faculty science*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781107110731

Hubbard, Raymond. 2004. Alphabet soup: Blurring the distinctions between p's and α's in psychological research. *Theory & Psychology* 14. 295–327. DOI: https://doi.org/10.1177/0959354304043638

Ioannidis, John. 2005. Why most published research findings are false. *PLoS Medicine* 2(8). e124. DOI: https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, John. 2008. Why most discovered true associations are inflated. *Epidemiology* 19. 640–648. DOI: https://doi.org/10.1097/EDE.0b013e31818131e7

Jeffreys, Harold. 1939/1961. *Theory of probability*. Oxford: Oxford University Press.

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

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

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

Morey, Richard & Jeffrey Rouder. 2015. BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.12–2. http://bayesfactorpcl.r-forge.r-project.org/.

Mulder, Joris & Eric-Jan Wagenmakers. 2016. Editors' introduction to the special issue 'Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments'. *Journal of Mathematical Psychology* 72. 1–5. DOI: https://doi.org/10.1016/j.jmp.2016.01.002

Neyman Jerzy & Egon Pearson. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* 20A. 175–240.

Neyman Jerzy & Egon Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A* 231. 289–337. DOI: https://doi.org/10.1098/rsta.1933.0009

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6151). aac4716.

Raaijmakers, Jeroen. 2003. A further look at the "language-as-fixed-effect fallacy". *Canadian Journal of Experimental Psychology* 57. 141–151. DOI: https://doi.org/10.1037/h0087421

Rouder, Jeffrey, Paul Speckman, Dongchu Sun, Richard Morey & Geoffrey Iverson. 2009. Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* 16. 225–237. DOI: https://doi.org/10.3758/PBR.16.2.225

Rouder, Jeffrey, Richard Morey, Josine Verhagen, April Swagman & Eric-Jan Wagenmakers. in press. Bayesian analysis of factorial designs. *Psychological Methods*. DOI: https://doi.org/10.1037/met0000057

Schütze, Carson. 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. DOI: https://doi.org/10.1037/0033-2909.105.2.309

Simmons, Joseph, Lief, Nelson & Uri, Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22. 1359–1366. DOI: https://doi.org/10.1177/0956797611417632

Smith, Keith. 1976. The assuming-will-make-it-so fallacy. *Journal of Verbal Learning and Verbal Behavior* 15. 262–263.

Sprouse, Jon. 2011a. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43. 155–167. DOI: https://doi.org/10.3758/s13428-010-0039-7

Sprouse, Jon. 2011b. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87. 274–288. DOI: https://doi.org/10.1353/lan.2011.0028

Sprouse, Jon. 2013. Acceptability judgments. In Mark Aronoff (ed.), *Oxford Bibliographis Online: Linguistics*. http://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0097.xml.

Sprouse, Jon, Carson Schütze & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134. 219–248. DOI: https://doi.org/10.1016/j.lingua.2013.07.002

Sprouse, Jon & Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48. 609–652. DOI: https://doi.org/10.1017/S0022226712000011

Wagenmakers, Eric-Jan, Richard Morey & Michael Lee. 2016. Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science 25*. 169–176. DOI: https://doi.org/10.1177/0963721416643289

Weskott, Thomas & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87. 249–273. DOI: https://doi.org/10.1353/lan.2011.0041

Wickens, Thomas & 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. DOI: https://doi.org/10.1016/S0022-5371(83)90208-6

Wike, Edward & James Church. 1976. Comments on Clark's "The language-as-fixed effect fallacy". *Journal of Verbal Learning and Verbal Behavior* 15. 249–255. DOI: https://doi.org/10.1016/0022-5371(76)90023-2

How to cite this article: Sprouse, Jon and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: a journal of general linguistics* 2(1): 14. 1–32, DOI: https://doi.org/10.5334/gjgl.236

Published: 28 February 2017

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS