Power in acceptability judgment experiments

Jon Sprouse, Department of Cognitive Sciences, University of California, Irvine
Diogo Almeida, Department of Linguistics and Languages, Michigan State University

Abstract

The literature on acceptability judgment methodology has been recently dominated by two trends: criticisms of traditional informal judgment experiments as coarse and unreliable, and endorsement of larger, formal judgment experiments as more sensitive and reliable. In order to empirically investigate these claims, we present a systematic comparison of the statistical power of two types of judgment experiments: the Forced-Choice task, which is traditionally used in informal syntactic experiments, and the Magnitude Estimation task, which is traditionally used in formal syntactic experiments. We tested 48 pairwise phenomena spanning the full range of effect sizes found in a recent large-scale empirical survey of the core phenomena of syntactic theory (Sprouse & Almeida, *submitted*), deriving estimates (via resampling simulations) of statistical power for each phenomena for sample sizes from 5 to 100 participants. The results show that (i) contrary to recent criticisms, Forced-Choice experiments are generally more powerful than formal Magnitude Estimation experiments at detecting differences between sentence types, and that (ii) even under the most conservative assumptions, Forced-Choice experiments with small sample sizes achieve the “best practice” guideline of 80% statistical power (established for experimental psychology and the social sciences) for 95% of the phenomena in syntactic theory. We also compare the standardized effect sizes of the syntactic phenomena with phenomena in other domains of experimental psychology, and show that the former are, on average, four times larger than the latter. These results suggest that well-constructed, small-scale, informal syntactic experiments may in fact be among the most powerful experiments in experimental psychology.

1. Introduction

Acceptability judgments form a significant portion of the empirical base of generative syntactic theories (Chomsky 1965, Schütze 1996). The methodological issues surrounding acceptability judgment collection have been a steady topic of interest since the earliest days of generative grammar (Hill 1965, Spencer 1973); however, the past fifteen years in particular have seen a dramatic increase in the number of discussions of the methodology of judgment collection (Bard et al. 1996, Keller 2000, 2003, Edelman and Christiansen 2003, Phillips and Lasnik 2003, Featherston,2005a, 2005b, 2007, 2008, 2009, Ferreira 2005, Sorace and Keller 2005, Wasow and Arnold 2005, den Dikken et al. 2007, Alexopoulou and Keller 2007, Fanselow 2007, Newmeyer 2007, Culbertson and Gross 2009, Myers 2009a, 2009b, Phillips 2009, Bader and Häussler 2010, Dabrowska, 2010, Gibson and Fedorenko 2010a, 2010b, Culicover and Jackendoff 2010, Gross and Culbertson 2011, Weskott and Fanselow 2011). The issue at hand is that the majority of judgments reported in the generative syntactic literature were collected relatively informally: under typical circumstances (Marantz 2005), a syntactician constructs 5 or fewer tokens of each condition (type) of interest, asks 10 or fewer friends, students, or colleagues to directly compare the relevant examples (i.e, a forced-choice task), and then reports the results in an article as a single diacritic attached to a single example sentence. This informal approach to data collection can be contrasted with the more formal approach to data collection that is typical in other domains of experimental psychology: for example, a formal acceptability judgment experiment would consist of 8 or more tokens of each example, 20 or more naïve participants (e.g. undergraduate students), a survey-based rating task, and some sort of inferential statistical analysis (either frequentist or Bayesian). The question facing syntacticians is which (if any) of these methods is most appropriate for the advancement of syntactic theory.

Comparisons of traditional informal experiments and formal experiments have generally proceeded on two fronts. On the first front, critics of informal experiments have suggested that the data collected using informal experiments may be biased and unreliable. Such criticisms have potentially far-reaching consequences for syntactic theories: If the data collection methods used by generative syntacticians are inadequate, the data they produce might be riddled with unexpected problems, which would in turn cast serious doubts on the soundness of the theories built upon them (Edelman & Christiansen 2003, Ferreira 2005, Wasow & Arnold 2005, Dabrowska 2010, Gibson & Fedorenko 2010b). The suggested conclusion from this critical literature is that the data collection issues and all the problems they engender for generative theory can only be fully remedied by abandoning the informal data collection protocols in favor of formal acceptability experiments. On the second front, some researchers have argued that formal experiments, especially those involving quantitative tasks such as Likert scales and magnitude estimation, can yield valuable new data that may be impossible to collect with traditional informal methods (e.g., Keller 2000, Featherston 2005a, 2005b, 2007, 2008, Alexopoulou and Keller 2007, Sprouse 2008, Sprouse et al. 2011, Sprouse et al. 2012). Though the nuanced differences between these two strands of research has generally made it difficult to address both simultaneously, in this paper we will attempt to address what we see as a common thread between both: the concern that informal experiments are coarser and less sensitive than formal experiments in ways that will affect syntactic theory. In other words, both strands of methodological research appear to be concerned about the *power* of informal experiments relative to the power of formal experiments. In this paper, we will present the first (to our knowledge) systematic study of power for both experiments using the Forced-Choice task, which is traditionally used in informal syntactic experiments, and experiments using the Magnitude Estimation task, which is traditionally used in formal syntactic experiments. Our experiments test 96 sentence types that form 48 pairwise phenomena from syntactic theory that were chosen to span much of the range of effect sizes and many of the topics covered in modern syntactic theories.

To see the immediate relevance of power in discussions of syntactic methodology, we should first turn to the (arguably) most pressing issue: the integrity of the existing body of data upon which generative theories are constructed. On this topic, there seems to be (i) some consensus amongst experimentally inclined syntacticians that the bases of current syntactic theories are by and large well-established (Featherston 2009, Marantz 2005, Phillips 2009), and (ii) reports that formal experiments generally replicate the results of informal experiments (Featherston 2009, Phillips 2009). Complementing these meta-analytic reports, Sprouse and Almeida (*submitted*) formally tested (using the Magnitude Estimation (ME) task) all 469 English language data points used in a popular introductory syntax textbook (Adger 2003). The data reported in that book was presumably obtained via informal acceptability judgment experiments, according to the traditional procedures in the field of generative syntax. Sprouse and Almeida (*submitted*) report a conservative estimate of 98% agreement between the results of the informal acceptability judgment experiments presented in the book and the results of their ME experiments. If it is indeed true that the informal data collection methods in syntax are inherently unreliable and that formal acceptability judgment experiments provide the appropriate alternative to them, these results should be encouraging for generative syntacticians: Despite the potential vicissitudes associated with traditional methods, they at least have not significantly derailed the construction of syntactic theories.

Encouraging though they may be, results showing a high degree of agreement between the outcomes of formal and informal acceptability judgment experiments are still *in principle* compatible with the main thrust of the critical literature: that informal experiments are suboptimal experimental procedures that should be substituted by formal experiments whenever possible. It is a logical possibility, for instance, that generative theories have been developed based primarily on sentence types that elicit relatively robust acceptability contrasts, which are detectable by both formal and informal methodologies. However, as the theories develop and progress to more fine-grained predictions, the less robust contrasts in acceptability ratings between the relevant test cases might lead to problems (Ferreira 2005, Gibson & Fedorenko 2010b). Therefore, even with the encouraging results from recent methodological studies (e.g., Sprouse and Almeida *submitted*), it is still possible that further progress in syntactic theory will crucially depend on the adoption of formal, putatively more sensitive experiments at the expense of the traditional, putatively less sensitive informal experiments.

Two things should be clear at this point. First, the concern regarding power that is hinted at by the critical literature is similar to the concern regarding power underlying recent claims that formal experiments may provide new, valuable data to syntactic theory (e.g., Keller 2000, Featherston 2005a, 2005b, Sprouse et al 2011). The difference between these two claims (with respect to power) is simply a matter of degree. Second, these arguments rest upon the assumption that formal experiments are more powerful than informal experiments. Indeed, we would argue that this is a common assumption among all language researchers; however, to our knowledge, there have been no systematic comparisons of the power of the two methodologies. As a first step toward addressing the two empirical questions surrounding the use of informal versus formal syntactic experiments (i.e., whether informal experiments are somehow suboptimal or unreliable, and whether formal experiments necessarily provide more useful information to the syntactician), this paper presents two large-scale experiments (96 sentence types each, 148 and 140 participants) and a series of resampling simulations designed to illustrate the relationship between effect size and sample size for both the Forced-Choice (FC) task normally used in traditional syntactic experiments and the Magnitude Estimation (ME) task used in formal syntactic experiments.

The rest of this paper is organized as follows: Section 2 presents a brief introduction to statistical power, with discussions covering both the assumptions underlying power criteria and the calculations proposed to assess them. Section 2 may be particularly interesting to syntacticians, as the issue of statistical power is rarely (if ever) discussed in the syntactic literature, and from our perspective, the relevant assumptions underlying current recommendations for adequate statistical power adopted in the field of experimental psychology may not necessarily be shared by syntacticians. As such, we hope that this paper may spark conversations in the field about what is an appropriate level of power for syntactic experiments. Section 3 reports the details of the two experiments, as well as three possible experimental predictions that can be derived from the existing methodological literature. Section 4 describes the resampling simulations used to assess the power of acceptability judgment experiments. Section 5 reports the results of the analyses in relation to the three predictions from the critical literature, which as we shall see shortly suggest that informal experiments are appropriately powered under typical circumstances. Section 6 presents a discussion of the power of acceptability judgments to detect the phenomena in syntactic theory relative to the power of formal experiments in other branches of experimental psychology. As we shall see, this analysis suggests that informal experiments in syntactic theory are in fact very well-powered compared to the experiments in other branches of experimental psychology, contrary to the claims of some critics. Section 7 presents a deeper discussion of the issues surrounding the choice of ideal power levels in syntactic theory, such as the role of invariances in syntactic theorizing, and the benefits of Bayesian statistics. Section 7 also presents a set of power analyses using Bayesian statistics to demonstrate that the results discussed in sections 5 and 6 remain virtually the same regardless of the choice between frequentist and Bayesians statistics. Section 8 concludes with a general discussion of the implication of these results for syntactic methodology moving forward.

2. Power

Power (also known as statistical power) is a measure of the ability of a given experiment to detect a difference between (two or more) conditions when there is in fact a true difference between the conditions. Power is often stated as a percentage: for example, an experiment with 80% power indicates that the experimenter has an 80% probability of detecting a difference when a difference exists, and conversely, a 20% chance of failing to detect a difference when an effect exists (also known as a false negative or a Type II error). Though this percentage is very intuitive, it does not represent a single aspect of an experiment, but rather a derived property of all of the features of a given experiment and the statistical analysis that was performed on the data. With respect to acceptability judgment experiments, the critical properties affecting power are: the experimental task (FC, ME, Likert scales, Yes-No, etc.), the size of the experimental sample (larger samples generally yield greater power), the size of the difference that one wishes to detect (larger differences generally yield greater power), and the rate of false positives that one is willing to tolerate (also known as the Type I error rate, conventionally set at 5% in experimental psychology and the social sciences).

Because power is expressed as a numerical value between 0 and 100%, the criterion at which an experiment may be considered “well-powered” may vary from field to field, or even from researcher to researcher. As a matter of convention, many fields of experimental psychology and the social sciences have adopted the suggestion by Cohen (1962, 1988, 1992) that the target power rate should be at or above 80%. This suggestion is based on the following logic: (i) most experimenters conventionally tolerate a false-positive (type I error) rate of 5%, (ii) false positives are approximately 4 times more troublesome than false negatives (type II errors), (iii) power is mathematically equal to 1-β, where β is the false negative rate (type II error rate), therefore (iv) β should be set at 20%, and (v) power should be set at 80% (Cohen, 1992). For the purposes of addressing concerns that informal experiments are underpowered, a criterion of 80% power seems appropriate given its widespread recommendation in experimental psychology and the social sciences, as it allows a direct comparison between informal acceptability judgment experiments and the explicit “best practice” guidelines that behavioral experiments in experimental psychology (and psycholinguistics) strive to achieve. To the extent that informal experiments reach 80% power under routine circumstances, the widespread assumption that informal experiments are “under-powered”, coarser, or less sensitive than formal experiments will need to be revised. Therefore one of our analyses of the present studies will adopt the conventions of experimental psychology: a 5% false positive rate, and a target of a minimum 80% power (i.e., a 20% false negative rate) in order to address concerns about the relative lack of sensitivity of informal experiments directly.

Though the 80% power recommendation may be appropriate for comparing the results of informal experiments with other experiments in experimental psychology and the social sciences, it may not ultimately be the most appropriate for syntactic theory. Recall that Cohen arrived at this figure by assuming that false positives (type I errors) are 4 times more troublesome than false negatives (type II errors). It is an open question whether this assumption is shared by syntacticians. For example, many syntactic theories seek to capture both the differences between conditions and the invariances between conditions. False negatives, which fail to indicate a difference when there is in fact one, may therefore have more of an effect on the theorizing of syntacticians than it would on other experimental psychologists, since this detection failure could be interpreted as indicating that no difference exists between the relevant conditions[1], a result that could have important theoretical implications. Arriving at a consensus regarding the correct ratio of false positives to false negatives is a complex problem. On the one hand, it would be easy to simply assume that both types of errors are equally problematic and therefore should be equally minimized. As one concrete example, one could assume a 5% false positive rate (p<.05) and a 5% false negative rate (95% power). However, this would in a very real sense be holding syntax to a higher statistical standard than other fields of experimental psychology and the social sciences. As a field, syntacticians may agree that this is a warranted step in pursuit of the goals of syntactic theory, but it would substantially alter the nature of the methodological debates that have occurred to date. Another option would be for syntacticians to assume that both error rates should be equal, but “less strict” than the 5% rate. For example, syntacticians could decide that a false positive rate of 10% and a false negative rate of 10% are tolerable given the reliability of judgment data suggested by previous studies (e.g., Sprouse and Almeida *submitted*) and the ease of replication of informal syntactic experiments, which severely minimizes the impact of false positives in the long run. In such a case, syntacticians would be holding themselves to a higher power criterion (90%), but a less strict significance criterion (p<.1). We will not attempt to advocate for a specific power criterion here, as we prefer to leave this to individual researchers and the field as a whole to determine. However, in order to advance this discussion empirically, we will provide all three analyses in this paper: we will present power curves at 5% - 80% (Cohen’s level), 5% - 95% (the strict and equal level) and at 10% - 90% (the less strict and equal level).

3. The experiments

In order to illustrate the relationship between effect size, sample size, and task we need data sets that (i) span a wide range of effect sizes, (ii) span a wide range of sample sizes, and (iii) span multiple tasks. In this section, we describe the details of two experiments designed to meet these criteria for the magnitude estimation and forced-choice tasks.

[1] Although at this point it should be noted that standard null hypothesis significance testing is inappropriate for establishing invariances, therefore this approach to invariances would require more than a shift in attitudes toward the false negative rate. It would probably also require a shift to Bayesian statistics, which allow direct testing of the null hypothesis (for a review see Gallistel 2009). We address this issue in more detail in Section 7.

The magnitude estimation task

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

The forced choice task

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

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

Selection of phenomena

In order to identify a set of phenomena that span a wide range of effect sizes that arise in syntactic experiments, we drew on Sprouse and Almeida's (*submitted*) Magnitude Estimation results. Sprouse and Almeida (*submitted*) formally tested 208 sentence types from a popular minimalist textbook (Adger 2003) that form 104 pairwise phenomena (i.e., a direct comparison of the ratings of two sentence types). The 104 pairwise phenomena span 9 fundamental topics in syntactic theory (e.g., phrase structure, theta-theory, functional structure, wh-movement), and arguably represent the core empirical coverage of minimalist theories. Sprouse and Almeida (*submitted*) collected ratings from 40 participants for each sentence type. We used these results as a guide for choosing the phenomena for the present studies by using the previously obtained effect sizes as a rough approximation of the effect sizes that we would obtain in the present studies. Crucially, Sprouse and Almeida used the z-score transformation to standardize the ratings in their ME experiments such that 1 z-unit equals one standard deviation for each participant. Therefore we can use the difference between the mean ratings of the pairs in each phenomenon as a direct measure of effect size (i.e., the mean difference between the two sentence types). The range of effect sizes for the 104 pairwise phenomena in Sprouse and Almeida (*submitted*) is 0.22 z-units to 2.33 z-units. We chose 48 of these pairwise phenomena for the current experiments using the following procedure. First, we coded the phenomena numerically such that our knowledge of syntactic theory could not influence our selection of the phenomena for inclusion in these experiments. Second, because at a practical level the most useful information for syntacticians will be the behavior of smaller effect sizes, since those would be the ones most likely to escape detection, we selected all of the phenomena from Adger (2003) that were in the range 0.22 to 0.86 (a total of 20 phenomena), which were the smallest observed differences in the Sprouse and Almeida (*submitted*) study. Finally, we selected two phenomena at each 0.1 increment between 0.9 and 2.3 (28 phenomena). It should be noted that the previous effect sizes were not used in any of the analyses of the present experiments; only the effect sizes from the present experiments were used in the analyses. The previous effect sizes were only used to select phenomena in a systematic way in order to maximize the information provided by the experiments.

To make the idea of effect sizes more tangible, here we present six example phenomena at a range of effect sizes for the reader to judge for herself. The full list of phenomena selected, along with relevant statistics, are available as Table 3 in the appendix.

(3) I'd planned to have finished, and finished I have. (.26 z-units)
*I'd planned to have finished, and have finished I did.

(4) What she thought was that the poison was neutralized. (.52 z-units)
*What she thought that was the poison was neutralized.

(5) I believed she might be pregnant. (.73 z-units)
*I believed she may be pregnant.

(6) I worry if the lawyer forgets his briefcase at the office. (1.02 z-units)
*What do you worry if the lawyer forgets at the office?

(7) He has known him. (1.50 z-units)
*Him has he known.

(8) I shaved myself. (1.99 z-units)
*Myself shaved me.

Materials

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

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

For the forced choice task, which requires the direct comparison of pairs of sentences, a slightly different distribution process was used. First, the 8 lexicalizations were distributed among 8 lists by pairs, such that each pair of related lexicalizations appeared in the same list. Next, the order of presentation of each pair was counterbalanced across the lists, such that for every pair, four of the lists included one order, and four lists included the other order. This minimized the effect of response biases on the results (e.g., a strategy of 'always choose the first item'). Finally, the order of the pairs in each list were randomized, resulting in eight surveys containing 48 randomized and counterbalanced pairs (96 total sentences).

Presentation

For the magnitude estimation 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 nine unmarked anchor items (which were not included in the analysis) did serves as a sort of unannounced sentence practice. There was also no practice for the forced choice experiment, as the task is generally considered intuitively simple. The surveys were advertised on the Amazon Mechanical Turk website (see Sprouse 2011a, for evidence of the equivalence of data collected using AMT when compared to data collected in the lab), and presented as web-based surveys using an HTML template available on the first author's website. Participants completed the surveys at their own pace.

Participants

152 participants completed the magnitude estimation experiment, and 152 participants completed the forced choice experiment. Participants were recruited online using the Amazon Mechanical Turk (AMT) marketplace, and paid $2.00 for their participation. 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. 12 participants were excluded from the magnitude estimation 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). The remaining 140 participants were included in the resampling simulations. 4 participants were excluded from the forced choice experiment for answering 'no' to one of the language history questions. It was not possible to establish an objective criterion for identifying obvious

attempt to cheat for the forced choice experiment, so the remaining 148 participants were included in the resampling simulations.

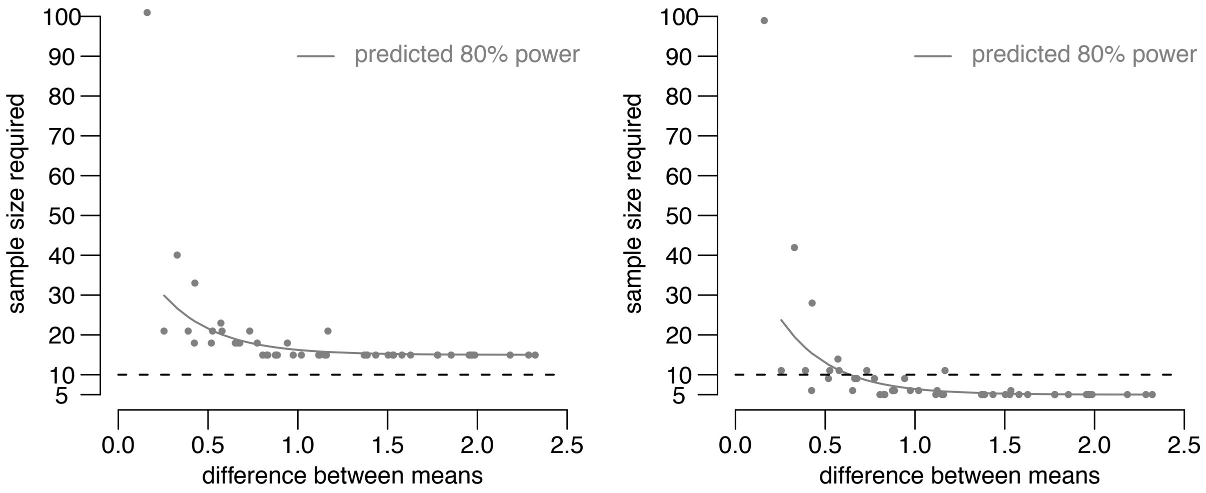
The number of judgments per condition

Each participant rated only one token of each condition. From the perspective of both informal experiments and formal experiments, this number is quite low. We chose to only test one token of each condition per participant for several reasons. First, this is the lower limit of possible experimental designs. This means that the power curves that we derive will provide a lower bound for such experiments. By simply increasing the number of tokens per condition to 2 or 4, syntacticians can easily increase the power at any given sample size. Second, only including one token per condition allowed us to test all 48 phenomena (96 sentences) in a single survey without risking fatigue on the part of the participants. This means that the z-score transformation is based upon the same sentence types for every participant. Finally, some critics (e.g., Gibson & Fedorenko 2010b) have suggested that informal methods are predicated upon a single judgment per condition. While in our experience this is false (see also Marantz 2005), incorporating that claim into our design allows us to address the concerns of critics of the informal syntactic experiments directly.

The experimental predictions

To our knowledge none of the literature that is critical of the reliability of informal experiments has made explicit predictions about the power of informal experiments relative to the power of formal experiments. However, it is possible to articulate a good-faith formulation of the prediction that is implied by the claims in the critical literature. First, this literature claims that informal experiments are unreliable under standard circumstances. In terms of our experimental design, this would predict that the FC task will not reach 80% power (the recommended minimum level in experimental psychology and social sciences) with sample sizes that are less than or equal to 10. We can test this prediction directly by plotting the sample size required to reach 80% power for each of the 48 phenomena and looking to see if these sample sizes must be above or below 10 participants (see predictions schematized in Figure 1).

Figure 1: Schematized prediction of the literature that is critical of informal acceptability judgment experiments. The sample size criterion of 10 is indicated with a black dashed line. The hypothesized points and fitted line are drawn in grey to emphasize that this is a prediction, not real data.



a. FC is unreliable at small sample sizes

b. FC is reliable at small sample sizes

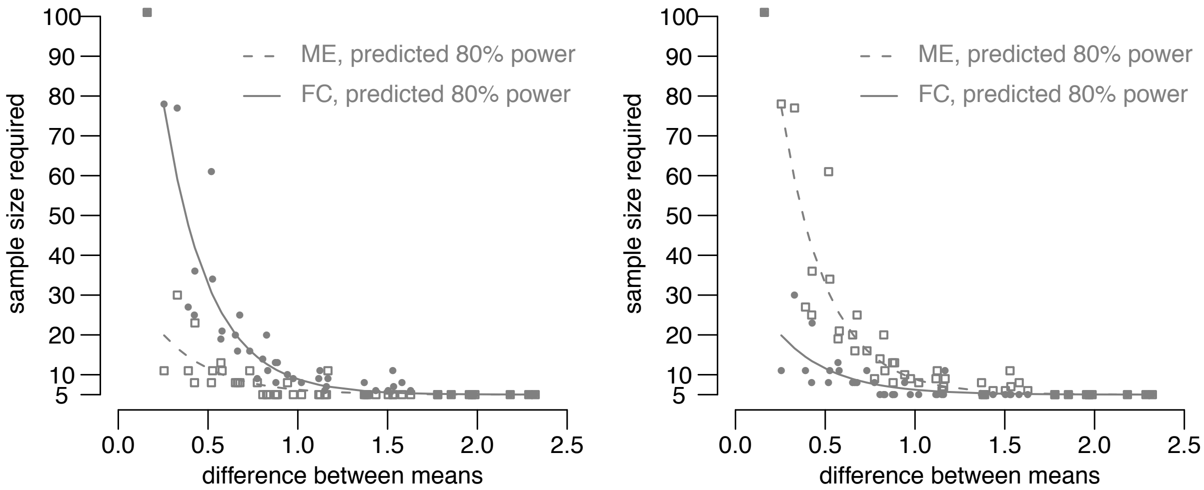
It should be noted that some critics have suggested that traditional informal experiments are conducted with absurdly small sample sizes (e.g., a sample size of one, see Gibson & Fedorenko, 2010b); however, in our experience, careful syntacticians always ask a number of friends, students, and colleagues for judgments before publishing a given datum. Therefore we believe a sample size criterion of 10 is an appropriate starting point for testing the critical predictions. Furthermore, the experiments presented here only involve one judgment per participant, which is the absolute minimum power that a sample could yield. In other words, though we have adopted the sample size that we believe is more reflective of the actual methodology of syntacticians, we have also biased our results against a favorable outcome for syntacticians by observing only one judgment per condition per participant. Of course, if the reader disagrees with these assumptions, there is an easy remedy. All of the information about the sample sizes required for 80% power will be contained in the graph (as well as 5%-95% and 10%-90% power); the sample size of 10 will simply be chosen as a point of discussion (see Figure 4 and Table 1 for observed results). The interested reader can choose any sample size criterion they want and look in Figure 4 and Table 1 to see how many phenomena reach the desired power at that criterion.

The second possible prediction of the critical literature is also shared by the current experimental syntax literature: in terms of our experiments, the one using the ME task, which is typically used in formal experiments, will be more powerful than the one using the FC task, which is typically used in informal experiments. The question then is what counts as “more powerful”. The most straightforward interpretation would be whichever experiment requires fewer participants to reach the target power level. This can be easily tested by comparing the power curves of the experiments using the FC and ME tasks for all 48 participants to see which one reaches the target power level with the smallest samples (see the predictions schematized in Figure 2).

Figure 2: Schematized prediction of the literature that is critical of informal acceptability judgment experiments. The hypothesized points and fitted line are drawn in grey to emphasize that this is a prediction, not real data. ME points are empty squares and fitted with the dashed line. FC points are solid circles and fitted with the solid line.

a. ME reaches target power faster

b. FC reaches target power faster



It is also the case that a more nuanced prediction that establishes a "typical" sample size for each kind of experiment could be tested. For example, one could assume that a typical FC experiment sample size is 10, whereas a typical ME experiment sample size is 20. The question then is which type of experiment achieves higher power at the typical sample size. If ME experiments have higher power at 20 participants than FC experiments have at 10 participants, then in one sense, ME experiments would be the more powerful type of experiment. We will operationalize this question in Section 5 (see Figure 6) by comparing the power of each experiment at its typical sample size for each phenomenon to see which type of experiment achieves higher power more often. This direct phenomenon-by-phenomenon comparison could also allow syntacticians to see if specific areas of syntactic theory are better detected with ME than FC, though for space reasons we will not attempt such a meta-analysis here.

4. The resampling simulations

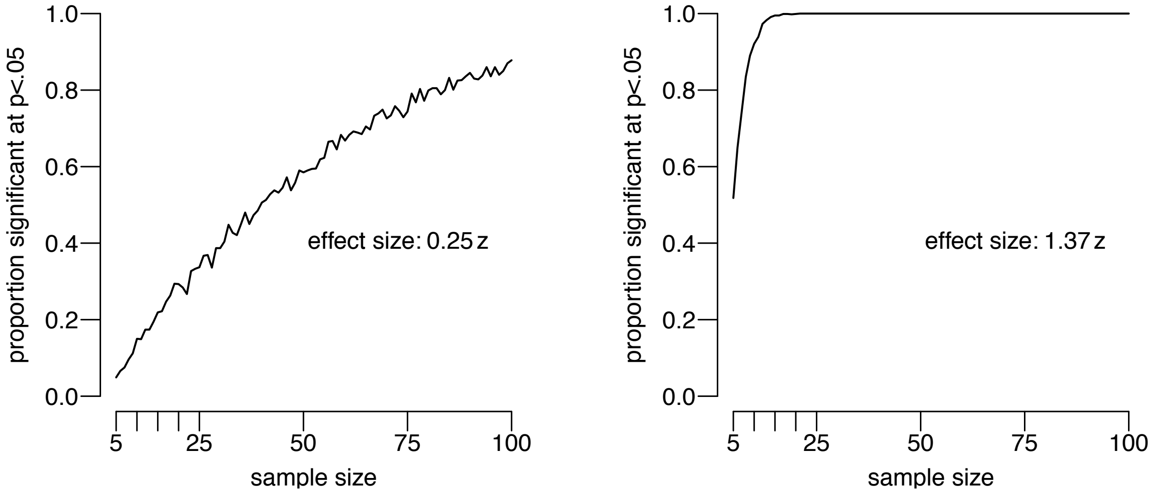
In order to assess the role of sample size in statistical power, we performed resampling simulations on each sample. In essence, these resampling simulations treated our large samples (N=140, N=148) as full populations, and sampled from them to determine the statistical power (operationalized here as a *rate of detection*) at each sample size that is possible with the population (1 to N). For example, imagine that we are interested in the *whether* island effect illustrated by the pair of sentences in (2). To establish a rate of detection for a sample size of 5, we could perform the following procedure:

1. Draw a random sample of 5 participants (allowing participants to be potentially drawn more than once; this is called *sampling with replacement*)

2. Run a statistical test on the sample (for the magnitude estimation experiment, we used paired t-tests and Bayes factor analyses for paired t-tests; for the forced choice experiment, we used sign tests and Bayes factor analyses for sign tests).
3. Repeat steps 1 and 2 1000 times to simulate 1000 experiments with a sample size of 5.
4. Calculate the proportion of simulations (out of the 1000) that resulted in a significant result (i.e., a *p*-value that is less than .05, or a Bayes factor that is above 3).

This procedure would tell us the rate of detection of the *whether* island effect for samples of size 5. We can then repeat this procedure for samples of size 6, 7, 8… N to derive a complete relationship between sample size and detectability for *whether* islands. Finally, we can repeat this procedure for all 48 phenomena to derive power relationships (operationalized here as empirically derived rates of detection) for the full range of effect sizes in Adger (2003). Figure 1 presents results of these resampling simulations for two of the phenomena tested in the magnitude estimation:

Figure 3: The results of the resampling simulation for two phenomena from the magnitude estimation experiment. The x-axis represents sample sizes from 0 to 100. The y-axis represents the proportion of simulations that resulted in a significant t-test at *p* < .05. The effect size (difference between means) in z-units is labeled within each graph. Participants only judged one token per condition (e.g., a sample size of 5 represents only 5 judgments).



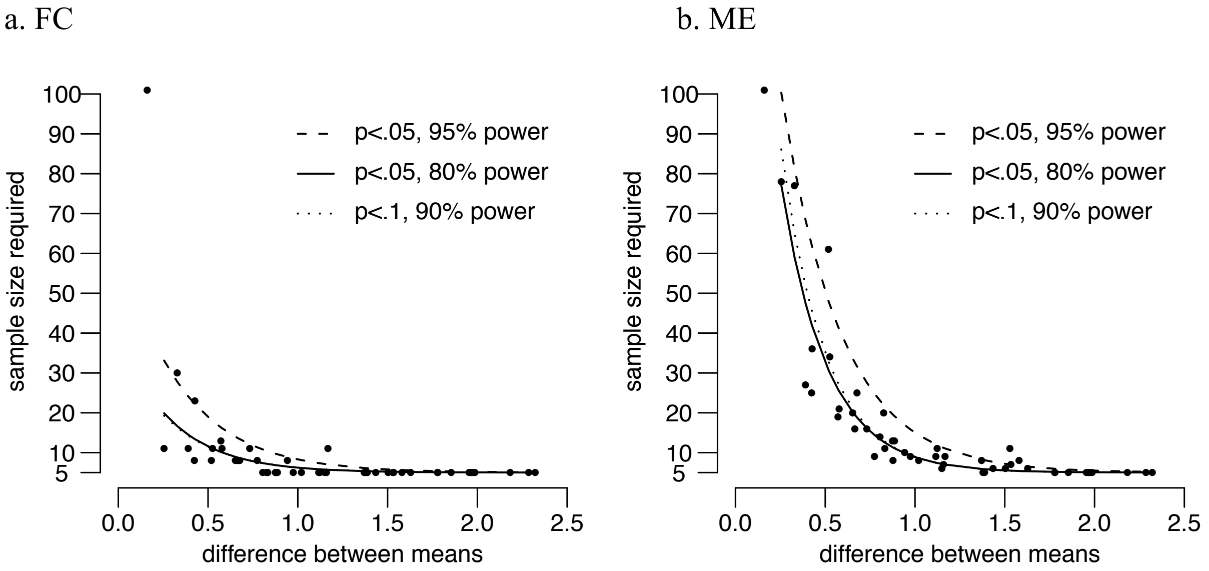
For the magnitude estimation experiment, we used two-tailed paired *t*-tests with *p*-values set at a criterion of *p*<.05 for the primary analysis. We chose two-tailed tests to make the power curves as conservative as possible (i.e., biased against the conclusion that informal experiments are well-powered). We also ran a second set of simulations with two-tailed tests with *p*-values set at a criterion of *p*<.1 to provide a point of discussion for syntacticians interested in exploring the various combinations of false positive and false negative error rates (see sections 2 and 7 for additional discussion). For the forced choice experiment, the samples were analyzed using the traditional two-tailed sign-test (with a significance criterion of *p*<.05 and *p*<.1). For all of the resampling simulations, the smallest sample size used in the resampling simulations was 5, as this is the minimum number of participants required for t-tests and sign tests to converge reliably, and the largest sample size in the simulations was 100, as this seems like a liberal upper-bound for potential sample sizes in acceptability judgment experiments. This resulted in the graphs in Figure 4 in the next section.

5. Assessing the power of informal and formal experiments

Prediction 1:Informal (FC) experiments are in general underpowered

The first potential prediction derived from the critical literature is that experiments using the FC task will be underpowered under typical circumstances. For the present discussion we have chosen to adopt 80% power as the criterion for a well-powered experiment (following the widely adopted recommendations for experimental psychology and the social sciences) and a sample size of 10 participants each making only 1 judgment per condition as "typical circumstances" for informal acceptability experiments (see section 3 for justification). Figure 4 graphically represents the sample sizes required to reach 80% power for the FC experiment with solid black dots and solid black line of best fit, as well as other power combinations with dashed and dotted lines (note that there are no points plotted for the other power combinations). Table 1 reports the number of phenomena captured at each sample size and power combination.

Figure 4: Power curves for FC and ME tasks. The x-axis is the difference between means for each phenomenon as determined by the results of the ME experiment; the y-axis is the sample size required for the indicated power level. The points represent actual values required to achieve 80% power when p < .05 according to the resampling simulations. To maximize clarity of exposition, actual values are not graphed for the other power levels, only their estimated curves. Power curves were estimated using the NLS function (for a non-linear least-squares line of best fit) in R based on phenomena 2 through 48. Phenomenon 1 was removed from the curve because its inclusion results in a sub-optimal fit.


a. FC

b. ME

Table 1: The total number of phenomena that reach a specified power level at each sample size up to 30 for each task. The results for three power levels are reported: the recommended criterion for well-powered experiments in experimental psychology and the social sciences (5% false positives - 80% power), a higher minimum power standard that linguists might be more interested in (5% - 95%), and a balance between the two (10% - 90%), with higher tolerance for false positives.

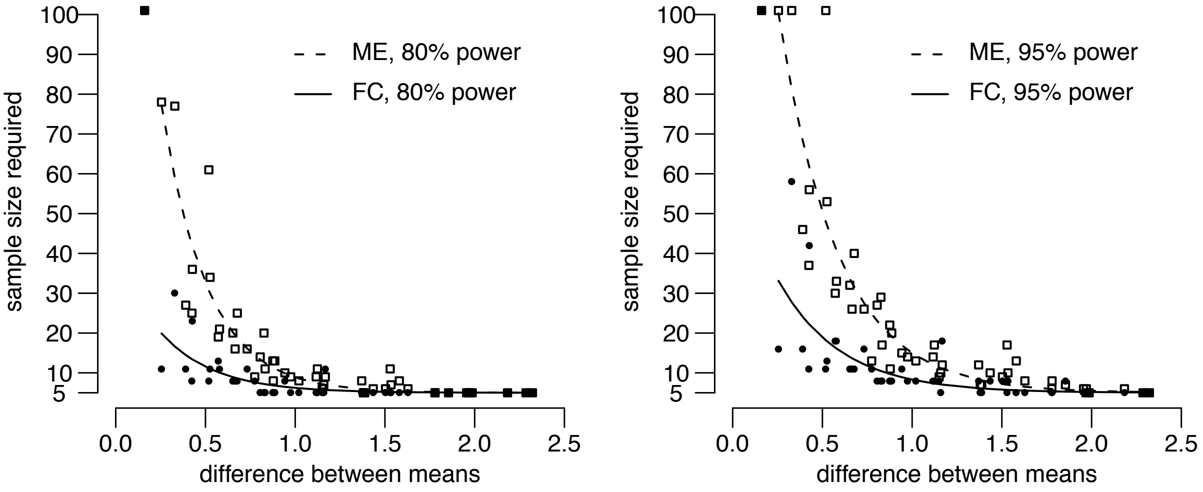
	Forced Choice			Magnitude Estimation		
	# of phenomena that reach target power			# of phenomena that reach target power		
sample size	5% - 80%	5% - 95%	10% - 90%	5% - 80%	5% - 95%	10% - 90%
5	31	15	24	12	4	8
6	31	15	24	16	8	11
7	31	15	24	18	11	14
8	38	31	33	22	13	17
9	38	31	33	26	15	19
10	38	31	33	27	18	23
11	44	38	38	30	19	26
12	44	38	38	30	21	27
13	45	39	44	32	24	27
14	45	39	44	33	26	30
15	45	39	44	33	27	30
16	45	42	45	35	27	31
17	45	42	45	35	30	32
18	45	45	45	35	30	32
19	45	45	45	36	30	33
20	45	45	45	38	31	33
21	45	45	45	39	31	35
22	45	45	45	39	32	35
23	46	45	45	39	32	35
24	46	45	45	39	32	35
25	46	45	45	41	32	38
26	46	45	45	41	34	39
27	46	45	45	42	35	39
28	46	45	45	42	35	39
29	46	45	45	42	36	39
30	47	45	46	42	37	39

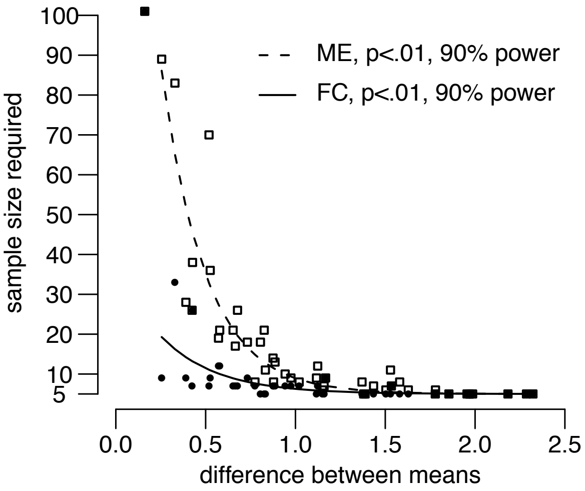
With a sample size of 10, 38 out of the 48 of the phenomena reach a minimum 80% power; increasing the sample size by a single extra participant to 11 participants captures 44 phenomena at a minimum 80% power. Given that we tested more smaller effect sizes than larger effect sizes in order to maximize the information obtained about smaller effect sizes, the large number of phenomena (44) that reach 80% power strongly suggests that informal (FC) experiments are not underpowered under typical circumstances. While there are still 4 phenomena that do not reach a minimum of 80% power, this simply suggests that there are a small percentage of phenomena for which larger samples (or more judgments per condition) are required (cf. Phillips 2009, Culicover and Jackendoff 2010, Sprouse & Almeida *submitted*). To the extent that these 48 phenomena do span the range of effect sizes of the phenomena in syntactic theory, as suggested by the results of Sprouse and Almeida (*submitted*), these results indicate that informal (FC) experiments with 10 participants should be considered an appropriately powered task for investigating most syntactic phenomena according to the "best practice" guidelines that are adopted in experimental psychology and the social sciences. Please see section 6 for a discussion of the distribution of effect sizes in syntactic theory, and how this distribution compares to the distribution of effect sizes in other branches of experimental psychology.

Prediction 2: Formal (ME) experiments yield higher statistical power than informal (FC) experiments

The power curves plotted for the ME experiment in Figure 4b above, as well as the table entries for the ME experiment in Table 1 above, already suggest that the potential prediction that ME experiments are better powered than FC experiments is false. To make this clearer, we can re-plot the FC and ME points and power curves together in the same figure:

Figure 5: A direct comparison of the power curves for FC and ME experiments. The x-axis is the difference between means for each phenomenon; the y-axis is the sample size required for the indicated power level. The solid circles represent actual values required to achieve the target power when p < .05 for the FC task. The solid line was fitted using the NLS function in R for phenomena 2-48 of the FC task. The empty squares and dashed line represent the same for the ME task.



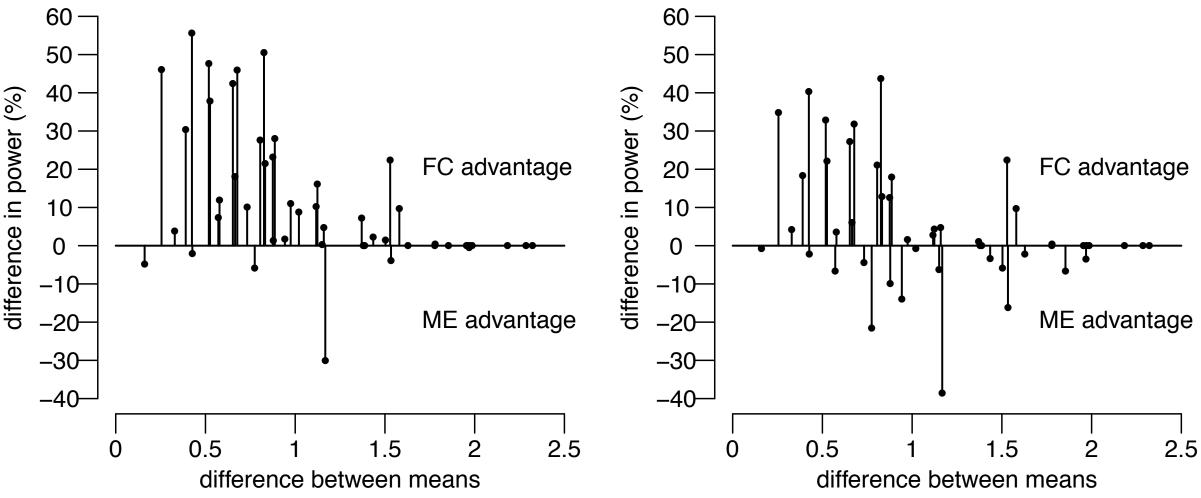


It is clear from Figure 5 that the ME experiment is less powerful than the FC experiment at all three of the power criteria investigated here when it comes to detecting differences between two conditions, as the ME experiment requires larger samples than the FC experiment to reach the target power for almost every phenomenon. While it may be tempting to interpret this result as evidence that FC experiments are universally "better" than ME experiments, we must caution against such over-interpretation. The criterion for success in these analyses is the simple detection of a difference between two conditions. The FC task is ideally suited for such a purpose: participants are asked to directly compare two conditions to each other and choose the more acceptable condition. In contrast, the ME task asks participants to rate conditions relatively to a third sentence (the standard). As such, the only comparison that can be made between two conditions is mediated by their numerical rating relative to the standard. The data presented in Figure 5 clearly demonstrates that the indirect method of comparison in ME is less efficient than the direct comparison that happens in the FC task. Therefore all that can be concluded is that FC experiments are more powerful at detecting differences between two conditions, which is consistent with the design of the FC task. Crucially, the ME task (and other numerical rating tasks like the Likert scale) provide more information than the FC task: they provide information about the size of the differences between conditions. In contrast, the FC task can only indirectly provide information about the size of the difference between conditions by comparing the total number of responses for each condition (see also Myers 2009a). Based on these facts, we would suggest the following: if the question of interest only hinges on detecting a difference between two conditions, then an FC experiment is likely to be the most powerful test (requiring the smallest sample size) that one could use; if the question of interest hinges on the size of the differences between conditions, then a numerical rating task like ME or Likert scales will be the more appropriate task.

Prediction 3: Formal (ME) experiments are better powered than informal (FC) experiments at their respective typical sample sizes

The last potential prediction of the critical literature is slightly more nuanced: Formal (ME) experiments may be more powerful than informal (FC) experiments at *typical sample sizes relative to each kind of experiment*. To assess this possibility, we performed the following analysis. First, we calculated the power of the FC experiment at a sample size of 10 (what we deem as a good approximation of the typical size in informal experiments) for each phenomenon under investigation. Next, we calculated the power of the ME experiment at a sample size of 20 (again, an approximation of the typical size in formal experiments) for each phenomenon. Then we calculated the difference between the FC and ME experimental power for each phenomenon. Finally, we plotted these difference scores relative to the effect size of each phenomenon in order to provide a visual representation of any advantage that one task may have over the other. We then repeated the process for an FC experiment with a sample size of 5 (and ME sample size of 20) to see what would happen with the absolute smallest possible FC sample size (and thus better engage the critics who have claimed that informal experiments are performed on absurdly small samples). The results are in Figure 6.

Figure 6: The difference in power between the FC experiment and the ME experiment for each phenomenon at sample sizes that are typical for each type of experiment. The left panel is the difference between FC with 10 participants and ME with 20 participants; the right panel is the difference between FC with 5 participants and ME with 20 participants. The x-axis reports the effect size of each phenomenon; the y-axis reports the difference in percent power. Differences favoring FC are positive and differences favoring ME are negative (differences of 0 favor neither task).



a. FC: N=10, ME: N=20

b. FC: N=5, ME: N=20

It is quite easy to see in Figure 6 that the majority of phenomena receive a power advantage in the FC experiment. When the FC sample size is 10 and the ME sample size is 20 (Figure 6a), only 6 phenomena show a power advantage for the ME experiment, 10 are equal, and the remaining 32 show a power advantage for the FC experiment. When the FC sample size is reduced to 5 (Figure 6b), 16 phenomena show an advantage for the ME experiment, 8 are equal, and the remaining 24 show a power advantage for the FC experiment. These results suggest that ME experiments do not have a power advantage over the FC experiments (with respect to detecting differences) even when the ME experiment is given a sample size that is four times larger than the FC sample size. In fact, at both sample sizes the FC experiment appears to be better equipped for detecting the smaller effect sizes, which are precisely the ones that critics of informal experiments worry about. If the ME experiment were better at detecting the smaller effect sizes, we would expect all of the ME advantage phenomena to be on the left side of the figure (where the x-axis is smallest); however, the FC experiment appears to have the advantage for the majority of the phenomena at the low end of the effect size scale. In other words, when it comes to detecting differences between conditions, the FC experiment with 5 participants is generally more powerful than the ME experiment with 20 participants.

Are informal experiments appropriately powered?

A question that is at the heart of both the literature that is critical of informal experiments and the literature that extolls the benefits of formal experiments is whether informal acceptability judgment experiments are appropriately sensitive to the comparisons of sentence types of interest to the practicing syntactician. In this section we have seen three different analyses that all point to the same conclusion: when it comes to detecting differences between conditions, informal experiments, which typically rely on the FC task, routinely achieve a minimum of 80% power at a sample size of 10 participants *with only 1 judgment per participant*, and thus qualify as well-powered. Furthermore, there is no clear power-based benefit to adopting the ME task (and its required larger sample size) when it comes to detecting differences between conditions. Of course, as we previously mentioned, there are other reasons to adopt numerical rating tasks like ME, such as the ability to quantify the size of the differences between conditions. Nonetheless, to the extent that syntactic theory is predicated upon differences between conditions (an open question that we do not attempt to address here), the informal experiments relying on the FC task appear not only to be completely appropriate, but also to qualify as surprisingly well-powered at the sample sizes that are commonly used in the field of syntax.

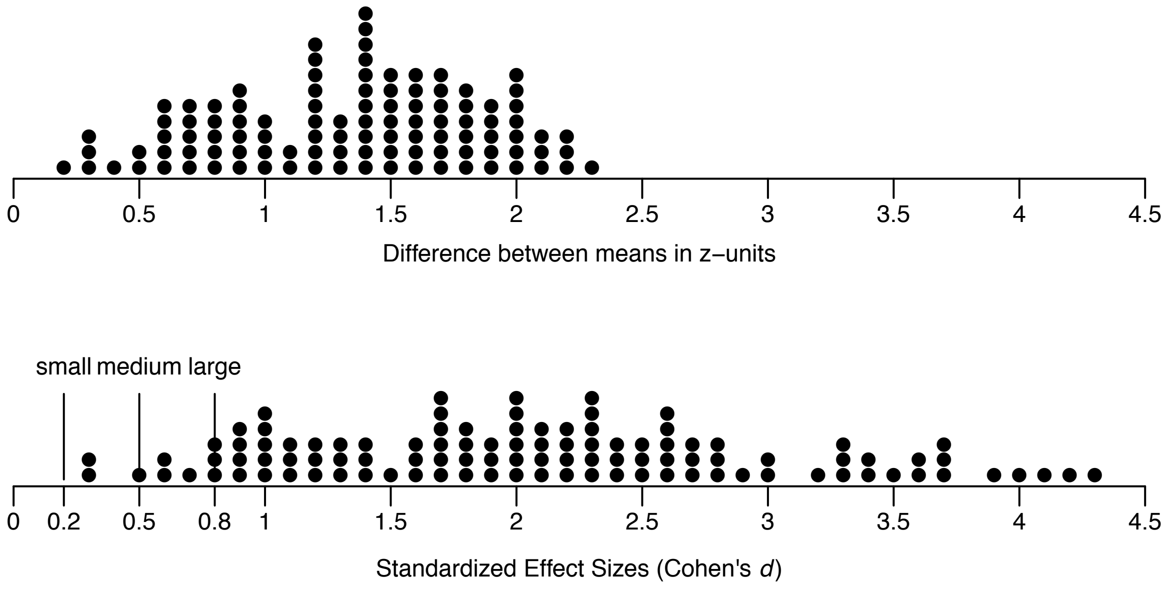
6. Effect sizes in syntactic theory and other branches of experimental psychology

The results in section 5 go a long way toward disproving the potential predictions of the current methodological literature, and establishing that informal experiments are indeed well-powered. Along the way, we have seen that a large portion of the phenomena (e.g. 44/48) in these experiments reach normative recommendations of power levels with the FC task and relatively small sample sizes (e.g., 11). However, the set of phenomena tested in these experiments were not selected randomly: smaller effect sizes were chosen more often in order to maximize the information about smaller effect sizes. To really address the question of whether informal experiments are an appropriate choice for syntactic theory moving forward, we need an estimate of how many phenomena in syntactic theory (as opposed to our sample) will reach recommended power levels. In this section, we will use the distribution of effect sizes from the 104 pairwise phenomena in Adger (2003), reported by Sprouse and Almeida (*submitted*), as a representative sample from syntactic theory (which, as a textbook, covers 9 different topics in syntactic theory in an attempt to be theoretically comprehensive). After estimating the proportion of phenomena in syntactic theory that will be well-powered with informal experiments, we will also compare the distribution of effect sizes in syntactic theory with other branches of experimental psychology. This comparison will require the introduction of *standardized* effect sizes (for our purposes, we will use Cohen's *d*) as opposed to the intuitive effect sizes that we've been using thus far (z-units); however, we believe this is worthwhile, as one of the primary question in the methodological literature is whether effects in syntactic theory are generally larger than effects in other areas of psychology (so-called *blowout* effects, e.g. Schütze 1996, Gibson and Fedorenko 2010b). Such a comparison will provide an answer to that question once and for all.

What is the distribution of effect sizes in syntactic theory?

Sprouse and Almeida (*submitted*) identified 104 unique, English-language, pairwise phenomena in Adger (2003). Using the intuitive definition of effect size as a difference between mean ratings, the distribution of effect sizes for the 104 phenomena in Adger (2003) can be plotted as in Figure 7a.

Figure 7: The frequency of effect sizes for the 104 unique, English-language, pairwise phenomena in Adger (2003). The x-axis is effect size. There is no y-axis as each circle represents one phenomena: the number of circles in each stack represents the number of phenomena at that effect size. Figure 7a uses difference between means as the effect size. Figure 7b uses a standardized effect size known as Cohen's *d*, which can be computed for any type of pairwise comparison. The vertical lines in 7b indicate the intuitive interpretation that Cohen (1988, 1992) suggested for effect sizes. The scale of the x-axis was kept identical to illustrate the effect of the Cohen's *d* calculation: Cohen's *d* is the difference between means divided by the mean of the standard deviations of the two conditions.



For the FC task we have seen that a sample size of 11 with only a single judgment per participant will yield a minimum 80% power for effect sizes over .5 z-units. The question then is how many of the phenomena in syntactic theory have effect sizes over .5 z-units. Using Figure 7a, we can see that 99 out of the 104 phenomena in Adger (2003) have an effect size above .5 z-units. This suggests that an FC experiment with a sample size of 11 *when only 1 judgment per participant* is collected will be adequately powered to detect 95% of the phenomena in syntactic theory. This is a lower-bound estimate: increasing the number of judgments per condition (or the number of participants in the sample) will likely increase the coverage. For the ME task we have already seen that larger sample sizes are required. In the current experiment, a sample size of 21 will yield a minimum 80% power for effect sizes greater than .55 z-units. In Adger (2003), 98 of the 104 phenomena (94%) have effect sizes greater than .55 z-units. This suggests that an ME experiment with a sample size of 21 *when only 1 judgment per participant* is collected will be well-powered to detect 94% of the phenomena in syntactic theory. Again, this is a lower-bound estimate: increasing the number of judgments per condition (or the number of participants in the sample) will likely increase the coverage.

How do effects sizes in syntactic theory compare with effect sizes in other branches of experimental psychology?

On one hand, the successful performance of the FC task may be surprising to some readers given the prevailing claims in the methodological literature that informal experiments lead to less reliable (or even faulty) data. One the other hand, some readers may be thoroughly unsurprised by this result, as some linguists have claimed that the effect sizes of phenomena studied in syntactic theory are substantially larger than the effect sizes studied in the branches of psychology that primarily rely on formal experimental methodologies (the existence of this claim is noted in Schütze 1996 and Gibson and Fedorenko 2010b, but it is difficult to track down these claims in print). We can use the phenomenon from Adger (2003) as tested by Sprouse and Almeida (*submitted*) to evaluate this claim as well by using a measure of effect size known as Cohen's *d*. Up to this point, we've been using the intuitive notion of effect size captured by the difference between means. While this is an appropriate effect size when dealing with a single type of data (e.g., acceptability judgments), it is impossible to compare this measure across data types because the scale of different effects will be different (e.g., reaction times are on a different scale than acceptability judgments). Therefore Cohen's *d* attempts to standardize across data types by setting the scale of the difference between means on a common scale that is meaningful to all data types. This is achieved by dividing the difference between means by the standard deviation of the conditions. Because standard deviations can vary between conditions even within the same data type, we have used the following calculation for Cohen's *d*: *d* is equal to the difference between means divided by the mean of the standard deviations of the conditions.

Because Cohen's *d* involves the division of a positive number (the difference between means) by a positive number (the mean of the standard deviations), it can take any value between 0 and positive infinity. Smaller values indicate smaller effect sizes. Cohen (1988, 1992) suggested the following criteria for the intuitive interpretation of *d* values: a *d* greater than 0.2 and less than 0.5 is considered a "small" effect, a *d* greater than 0.5 but less than 0.8 is considered a "medium" effect, and a *d* greater than 0.8 is considered a "large" effect. Here is what Cohen (1992) said about the intent behind these criteria:

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

Cohen's *d* values for the 104 phenomena in Adger (2003) are reported in Figure 7b along with vertical lines representing each of Cohen's interpretation criteria. Out of the 104 phenomena in Adger (2003), 2 would be considered small by Cohen's criteria, 4 would be considered medium, and 98 would be considered large. In fact, the median Cohen's *d* for the 104 phenomena in Adger (2003) is 2.0, which is more than double Cohen's criterion for "large" effects. In other words, all but 2 phenomena should be "apparent to the naked eye of a careful observer". These results strongly suggest that syntactic theory does indeed deal with large ("blowout") effects.

Moving beyond Cohen's criteria for effect size interpretation, the direct comparison of the distribution of effect sizes in syntactic theory with other areas of psychology becomes a bit tricky because most meta-analyses in experimental psychology are concerned with average power across studies rather than the actual distribution of effect sizes for a single data type. However, we can make the following comparisons. First, recall that we found that 95% of phenomena in syntactic theory would reach 80% power with 11 or fewer participants in an FC experiment. To put this striking result in perspective, Cohen (1962) demonstrated that in the 1960 volume of the *Journal of Abnormal and Social Psychology* the median power of the experiments was 46% for the average effect size of the phenomena under investigation (i.e., not much different than a coin toss). A follow-up study by Sedlmeier and Gigenrenzer (1989) for the 1984 issue of same journal found virtually identical results (44% median power for the average effect size of the phenomena of interest). In a review of the 1993 and 1994 volumes of the *British Journal of Psychology*, Clark-Carter (1997) reported a slightly larger average 59% power for the average phenomena of interest. Finally, Bezeau and Graves (2001) reported a mean of 50% power for "medium" effect sizes (*d* between 0.5 and 0.8) in their review of three neuropsychology journals.[2] Taken as whole, these results suggest that the typical informal experiments of syntactic theory, with 95% of phenomena reaching a minimum 80% power, would in fact be considered extremely well-powered experiments by the standards of the rest of experimental psychology contrary to the claims of some critics of informal experiments.

7. Power levels and Bayesian statistics

As discussed in section 2, the 5% false positive rate and 80% power level that has been recommended in experimental psychology and the social sciences may not be optimal guidelines for syntactic theory. On the one hand, syntactic theory attempts to capture both differences and invariances between conditions, which means that a 20% false negative rate may not be tolerable for some syntacticians, since it would too often introduce confusion between unreliable differences and reliable invariances. On the other hand, given the empirical findings pointing to an extremely low likelihood of actual type I errors (Sprouse & Almeida, *submitted*), and the relative low impact these would have in the field to begin with, given how easy and cheap direct

[2] It must be noted that Bezeau & Graves (2001) also reported that this low power to detect medium effect sizes was somewhat offset by the fact that the average effect size in their review of three neuropsychology journals would qualify as "large" according to the criteria proposed by Cohen (1988), which strongly suggests that neuropsychology seems to be interested in phenomena yielding larger effect sizes that other areas of psychology, and thus can make progress with their typically smaller sample sizes – an argument that seems to mirror the situation suggested by our data for syntactic theory.

replication of informal syntactic experiments is, a false positive rate of 5% might be considered too stringent, especially if the experimenter must pay for it in terms of statistical power. At any rate, the most conservative solution would be to adopt a 5% rate for both false positives and false negatives, which is equivalent to a 95% power level. Based on these results (see Figure 4 and Table 1), achieving a 95% power level with a similar level of empirical coverage would require a sample size of 18 for an FC experiment and a sample size of 35 for an ME experiment. Both of these sample sizes are certainly reasonable, especially with the availability of participants on Amazon Mechanical Turk (Sprouse 2011a). However, before endorsing this as the preferred solution, it is important to note that the null hypothesis significance tests (NHST) that were used to detect the effects in these experiments (sign test and *t*-test) are not capable of establishing invariances (i.e, establishing that there are no differences between the conditions of interest). In other words, a negative result in the context of a 5% false negative rate is not equivalent to establishing that there is no difference between conditions. If the importance of establishing invariances (no difference between conditions) is the driving factor behind adopting higher power levels, then the field should also consider adopting a different set of statistical tests so as to be able to truly establish the existence of invariances. The current best candidate for establishing invariances are Bayesian statistics.

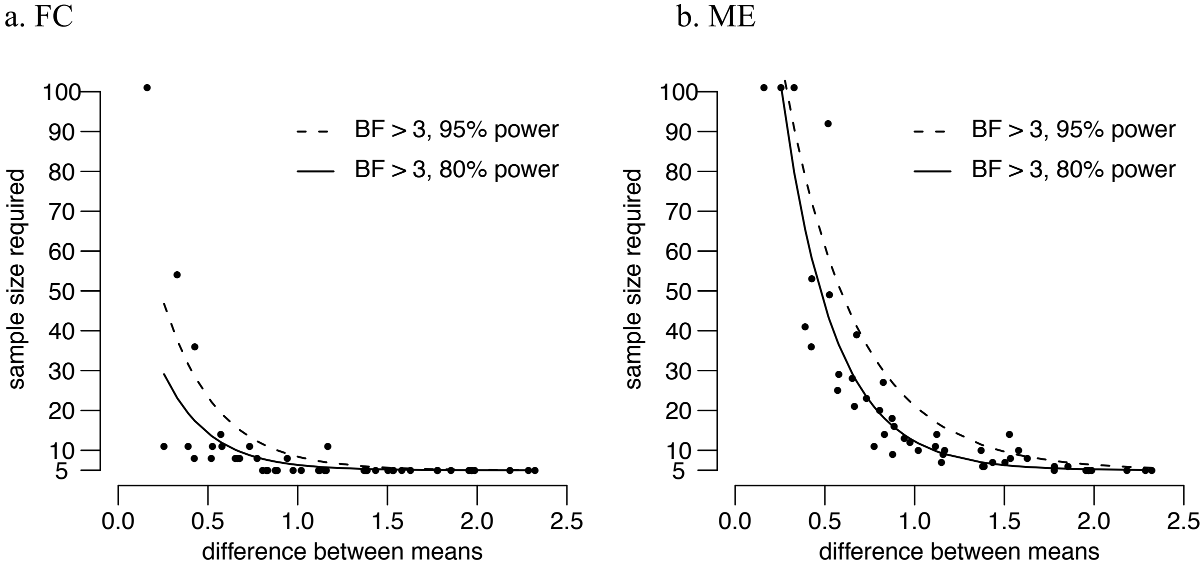
To see the benefit of Bayesian statistics for establishing invariances, it is perhaps best to start with a discussion of classic null hypothesis significance testing, also known as *frequentist statistics*, which have been a mainstay of experimental psychology and the social sciences for nearly 100 years. Crucially, the logic of frequentist statistics is not amenable to establishing invariances. This is because of two factors: (i) invariances are the *null hypothesis* in frequentist statistics, and (ii) frequentist statistics are designed to *disprove* the null hypothesis, not prove it. These two facts can be seen when one schematizes the logic underlying frequentist tests. First, assume that the null hypothesis is true (i.e., there is no differences between conditions). Second, generate the hypothetical set of data that would arise if one tested all possible samples of the relevant size in a world in which *the null hypothesis is true*. Finally, compare the actual data that the experimenter obtained to the hypothetical set of all possible data to see how likely the actual data is. The result is the famous *p*-value: it is a measure of how likely the data would be *if the null hypothesis were true*. So what can we conclude from the *p*-value? Well, if the *p*-value is sufficiently small, we can conclude that this data would be very *unlikely* if the null hypothesis were true. In such a case, we can *reject* the null hypothesis as a poor assumption for this data. But what can we conclude if the *p*-value is higher? While it is tempting to conclude that the null hypothesis is true, we can't. This is because we assumed that the null hypothesis is true when we generated the hypothetical data that led to the *p*-value. Therein lies the problem: NHST can only be used to reject invariances; if syntacticians want to establish invariances, then a different set of statistical tests must be used.

Bayesian statists offer one possible solution to the problem of invariances. This is because the logic of Bayesian statistics is in some ways the reverse of frequentist statistics. Whereas frequentist statistics assume a theory (the null hypothesis) and then ask how likely the experimenter's data is given that theory, Bayesian statistics assume that the experimenter's data is an accurate reflection of the world, and then ask how likely each theory is given that data. This logic is in many ways much closer to the logic of science itself: as experimenters we collect data because we believe it will be evidence for or against a particular theory. Crucially, every possible theory is equally testable in Bayesian statistics. Because the null hypothesis (an invariance) is just one theory among many, it can be tested just as easily as the others (see Gallistel 2009 for an accessible review). In addition to establishing the probability of each hypothesis (often called *posteriors*), Bayesian statistics also lead to a measure known as a Bayes factor. Bayes factors provide an intuitively natural measure of the strength of the evidence for each of the two standard hypotheses (the experimental hypothesis that there is a difference between conditions, and the null hypothesis that there is no differences) in the form of an odds ratio. For example, a Bayes factor of 4 indicates that the data favors the experimental hypothesis (H1) over the null hypothesis (H0) in a ratio of 4:1.

The foundation of Bayesian statistics, Bayes' rule, has been around for over 200 years; however, it has only been in the last decade or two that Bayesian statistics have gained in popularity in experimental psychology and the social sciences, despite the benefits mentioned above. This is because there are several obstacles that have prevented a more widespread adoption of Bayesian statistics. First, the calculation of Bayesian models is computationally complex. Prior to the introduction of computers, many Bayesian models were simply impossible to evaluate. Second, the construction of Bayesian models is often itself a complex task that requires specialized mathematical knowledge. Unlike the classic frequentist NHST tests, for which several end-user accessible pieces of software are available, Bayesian models generally require knowledge of both the mathematics underlying the model and the specific scientific questions being investigated. These two sets of knowledge rarely co-occur in the same individual. Third, Bayesian statistics require one to explicitly specify a set of starting assumptions (called *priors*) about the likelihood of each theory. Though most scientists have such assumptions in mind when investigating a specific question, it can sometimes be difficult to arrive at a set of starting assumptions that all researchers will agree upon. In such cases, it is sometimes necessary to run several Bayesian models using a range of starting assumptions. For an accessible introduction to Bayesian statistics for experimentalists, as well as citations for most of the claims in the previous two paragraphs, we suggest the textbook *Doing Bayesian Data Analysis* (Kruschke 2010).

Given that Bayesian statistics appear to be the next logical step for the analysis of syntactic data, we decided to re-run our previous analyses using Bayesian statistics instead of frequentist statistics to determine whether any of the conclusions would change. Specifically, we calculated Bayes factors using the equation provided in Rouder et al. (2009) instead of calculating *p*-values from sign tests and *t*-tests. As mentioned above, Bayes factors provide an intuitively natural measure of the strength of the evidence for each of the two hypotheses in the form of an odds ratio. For the ME data, we used the JSZ Bayes factor equation from Rouder et al. (2009), which assumes (i) a non-directional H1 (equivalent to a two-tailed *t*-test), and (ii) an equal prior probability of the two hypotheses. We chose the non-directional form of the Bayes factor calculation over directional versions (such as the Savage-Dickey test proposed by Wetzels et al. (2009)) because the non-directionality and equal prior assumptions result in more conservative Bayes factors. We used a Bayes factor of 3 (or above) as the criterion of significance, following the classification from Jeffreys (1961) that a Bayes factor of 3 and above is "substantial evidence" for the experimental hypothesis. For the FC data, we used the Bayes factor calculation for binomial responses (with a criterion of >3) made available by Jeff Rouder on his website (http://pcl.missouri.edu/bayesfactor). Figure 8 presents power curves (similar to those in Figure 4) for 80% and 95% power (with a BF > 3), and Table 2 presents counts of the number of phenomena that reach the target power (similar to Table 1).

Figure 8: Bayesian versions of the power graphs. Sample sizes required to reach a minimum of 80% power are plotted with solid circles, and the power curve is plotted with a solid black line. The power curve for 95% power is plotted with a dashed line. No points for 95% power are plotted.



a. FC

b. ME

Table 2: Bayesian versions of the total number of phenomena that reach a specified power level at each sample size for each task.

	Forced Choice		Magnitude Estimation	
	# of phenomena that reach target power		# of phenomena that reach target power	
sample size	5% - 80%	5% - 95%	5% - 80%	5% - 95%
5	31	15	8	3
6	31	15	12	6
7	31	15	15	8
8	38	31	17	11
9	38	31	19	12
10	38	31	23	15
11	44	38	25	16
12	44	38	26	17
13	44	38	27	19
14	45	38	30	19
15	45	38	30	21
16	45	38	31	24
17	45	41	31	26
18	45	41	32	26
19	45	41	32	27
20	45	44	33	28
21	45	44	34	29

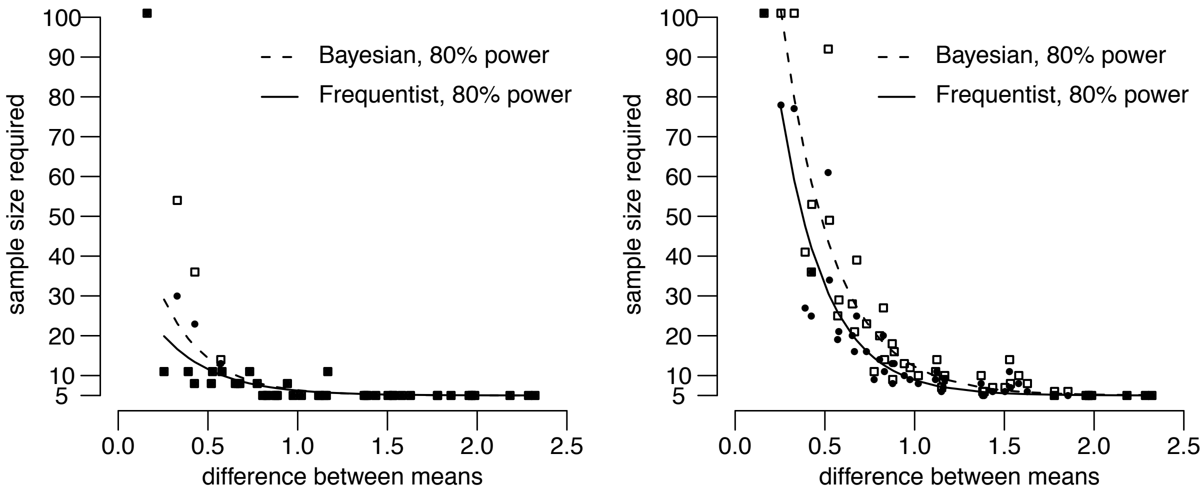
22	45	44	34	30
23	45	45	35	30
24	45	45	35	30
25	45	45	36	31
26	45	45	36	32
27	45	45	37	32
28	45	45	38	32
29	45	45	39	32
30	45	45	39	32

The first thing to note is that for the FC task, not much has changed: about 95% of phenomena achieve a minimum of 80% power at a sample size of 11, and 95% of phenomena achieve 95% power at a sample size of 20. For the ME task, the Bayesian statistics appear to be a bit more conservative: reaching 80% power for more than 40 of the phenomena requires a sample size of 35 or more, and reaching 95% power for more than 40 of the phenomena requires a sample size of 50 or more. This increase in sample sizes required for the ME task (but not for the FC task) can be seen by plotting the frequentist and Bayesian power curves together as in Figure 9.

Figure 9: A direct comparison of frequentist and Bayesian power curves for a criterion of 80% power. Frequentist results are indicated with solid circles and solid lines; Bayesian results are indicated with empty squares and dashed lines.

a. FC

b. ME



These results suggest that the switch to Bayesian statistics will not change the fact that informal experiments relying on FC tasks are well-powered at typical sample sizes (11 participants when there is only 1 judgment per participant) for 95% of the phenomena in syntactic theory. However, the switch to Bayesian statistics may require an increase in the sample sizes used for ME experiments, with ideal samples being at least 35 participants when there is only 1 judgment per participant. Of course, as with all of these estimates, an increase in the number of judgments per participant will likely decrease the required sample sizes.

8. Conclusion

In this paper we set out to directly investigate an understudied property of acceptability judgment experiments that is at the heart of both the criticism that has been levied against traditional informal judgment experiments, and the support that some linguists have given to formal judgment experiments: statistical power. Combined with current evidence suggesting that informal experiments have not led to problematic data in the past (e.g., Sprouse and Almeida *submitted*), this investigation of the power of informal judgment experiments helps determine whether informal experiments will lead to unreliable data in the future. We presented two large scale experiments designed specifically to establish a reliable estimate of the power of informal and formal experiments. The experiments and analyses were intentionally conservative: the participants in the experiments only made one judgment per condition, the frequentist statistical analyses used two-tailed tests, and the Bayes factor analyses used equal priors and bi-directional experimental hypotheses (equivalent to two-tailed *p*-values). The results of these experiments strongly suggest that informal judgment experiments are well-powered with typical sample sizes (~10 participants) *even when only 1 judgment per conditions is collected*. Furthermore, these results strongly suggest that there is no inherent power advantage for formal experiments, even when they are granted much larger sample sizes than informal experiments. Using the distribution of the 104 phenomena in Adger (2003) as an approximation of the core phenomena in syntactic theory, we estimated that forced-choice experiments with a sample size of 11 participants and one judgment per condition per participant would be sufficient to detect 95% of the phenomena in syntactic theory with a minimum of 80% power. This result is striking when compared to the statistical power in other branches of experimental psychology, as other branches often only reach 50% power at average effect sizes (see section 6). In short, our results suggests that typical informal experiments in syntactic theory are in fact very well-powered experiments by experimental psychology standards, contrary to critical claims in the literature on acceptability judgment methodology.

Of course, there are some caveats when it comes to interpreting our results. First, the definition of success in informal experiments is the simple detection of a difference between conditions. While we believe that much of syntactic theory is predicated upon differences between conditions, some syntacticians may require more information such as the relative location of the judgments on the scale (e.g., an "unacceptable zone"), or the size of the differences between conditions (e.g., only "large" differences would count as relevant for the theory). In such cases, the FC task that is typical of informal experiments will not be appropriate, as it is explicitly designed to detect the existence of differences, not quantify their magnitudes. Numerical rating tasks such as the ME task are better equipped to provide that sort of information, and as such the power advantage of experiments relying on the FC task will be less relevant (see also Sprouse and Almeida 2011). The second caveat is that the recommendations regarding statistical power normally adopted in experimental psychology and the social sciences may not appropriate for syntactic theory. Cohen (1962, 1988, 1992) suggested 80% power as a reasonable criterion based on the belief that false negatives are 4x more tolerable than false positives. This logic may not prove attractive for syntactic theory, where false negatives may obscure true invariances that need to be captured by the theory. To account for this possibility, and to help spur future investigations, we have included two additional analyses: power analyses that assume a 5% false negative rate, and Bayesian analyses that allow for explicit identification of invariances. Crucially, we found that the switch to Bayesian statistics would not affect the reliability of informal experiments at typical sample sizes. The final caveat is that sample sizes are rarely, if ever, reported in the informal experimental literature. This means that readers are not able to judge the power of the experiments for themselves without re-running the informal experiment. In well-studied languages, such as English, this is not really an issue, but it could prove problematic for less well-studied languages, for which access to native speakers is not a trivial matter. As many researchers have recently suggested (Featherston 2009, Phillips 2009) it may be time for all syntacticians to report the minimum sample that they consulted prior to publication. As discussed in section 6, readers could then use the results from the experiments reported here to make an informed decision about whether the reported informal experiment was adequately powered relative to the phenomenon of interest. This might eliminate critics' concern about the reliability of informal experiments, at least for the 95% of phenomena that yield normatively well-powered informal experiments at typical sample sizes.

What this pattern of results suggests is a much more nuanced role for formal experiments in the development of syntactic theories than the one advocated by in the literature critical of informal syntactic experiments. In lieu of adopting some of the radical recommendations from the latter – for instance, that informal experiment should be completely abandoned in favor of the more formal data collection methods – our findings suggest instead that a more parsimonious strategy should include the complementary use of the two methodologies: If the theorist is simply interested in establishing whether a difference exists between two relevant sentence types without any regard for the magnitude of the difference, the traditional methods seem more than appropriate. However, if the theorist is interested in comparing the magnitude of perceived differences between sets of sentence types, the traditional methods offer little to no help, whereas the ME task is particularly well suited for it. Therefore, critical claims that a syntactician must disavow informal experiments in order to be supportive of the use of formal experiments are presenting the field with a false dichotomy. It appears to be the case that both claims are true: informal experiments are reliable *and* formal experiments can provide additional information that may be relevant for specific research questions (for example, see the explicit endorsements of formal experiments by the first author: Sprouse 2007a, 2007b, 2008, 2009, 2011, Sprouse and Almeida 2011, Sprouse et al. 2011, Sprouse et al. 2012, Fukuda and Sprouse *submitted*). This seems to us to be the best of all possible worlds: we can all be confident in previous (informal) results; syntacticians who wish to continue to use traditional, well-constructed, informal experiments may continue to do so; and syntacticians who wish to pursue research questions that require formal experiments can be excited to see the new (formal) results that are waiting to be observed.

References

Adger, D. (2003). *Core Syntax: A Minimalist Approach*. Oxford University Press.

Bader, M. & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics*, *46*, 273-330.

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

Bezeau, S. & Graves, R. (2001). Statistical Power and Effect Sizes of Clinical Neuropsychology Research. *Journal of Clinical and Experimental Neuropsychology, 23*(3), 399-406.

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

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

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

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

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98-101.

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

Culbertson, J., & Gross, J. (2009). Are linguists better subjects? *British Journal for the Philosophy of Science, 60*, 721-736.

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

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

den Dikken, M., Bernstein, J., Tortora, C., & Zanuttini, R. (2007). Data and grammar: means and individuals. *Theoretical Linguistics, 33*(3), 335-352.

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

Fanselow, G. (2007). Carrots – perfect as vegetables, but please not as a main dish. *Theoretical Linguistics, 33*(3), 353-367.

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

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

Featherston, S. (2007). Data in generative grammar: The stick and the carrot. *Theoretical Linguistics, 33*(3), 269-318.

Featherston, S. (2008). Thermometer judgments as linguistic evidence. In C. M. Riehl & A. Rothe (ed.) *Was ist linguistische evidenz?*, Aachen: Shaker Verlag.

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

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

Gallistel, R. (2009). The importance of proving the null. *Psychological Review, 116*(2):439-53.

Grewendorf, G. (2007). Empirical evidence and theoretical reasoning in generative grammar. *Theoretical Linguistics, 33*(3), 369-381.

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

Gibson, E. & Fedorenko, E. (2010b). The need for quantitative methods in syntax. *Language and Cognitive Processes.*

Gross, J. & Culbertson, J. (2011). Revisited linguistic intuitions. *British Journal for the Philosophy of Science.*

Haider, H. (2007). As a matter of facts – comments on Featherston's sticks and carrots. *Theoretical Linguistics, 33*(3), 381-395.

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

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.

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

Kruschke, J. (2010). Doing Bayesian Data Analysis: A tutorial with R and BUGS. Academic Press.

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

Myers, J. (2009a). The design and analysis of small-scale syntactic judgment experiments. Lingua, 119, 425-444.

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

Newmeyer, F. J. (2007). Commentary on Sam Featherston, 'Data in generative grammar: The stick and the carrot.' *Theoretical Linguistics, 33*(3), 395-399.

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

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237.

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

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

Sprouse, J. (2007a). A program for experimental syntax. Doctoral dissertation, University of Maryland.

Sprouse, J. (2007b). Continuous Acceptability, Categorical Grammaticality, and Experimental Syntax. *Biolinguistics, 1*, 118-129.

Sprouse, J. (2008). The differential sensitivity of acceptability to processing effects. *Linguistic Inquiry, 39*(4), 686-694.

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

Sprouse, J. (2011a). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*.

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

Sprouse, J. & Almeida, D. (2011). The role of experimental syntax in an integrated cognitive science of language. In C. Boeckx and K. Grohmann (Eds.), *The handbook of Biolinguistics*, Cambridge University Press.

Sprouse, J. & Almeida, D. (submitted). A formal experimental investigation of the empirical foundation of generative syntactic theory.

Sprouse, J. & Fukuda, S. (submitted). Experimental evidence in the unaccusativity debate: Variable behavior verbs in Japanese suggest distinct underlying syntactic structures.

Sprouse, J., Fukuda, S., Ono, H. & Kluender, R. (2011). Reverse island effects and the backward search for a licensor in multiple wh-questions. *Syntax*.

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working memory capacity and syntactic island effects. *Language*.

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

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

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

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

Weskott, T. & Fanselow, G. 2011. On the informativity of different measures of linguistic acceptability. *Language, 87*, 249-273.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E. J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t–test. *Psychonomic Bulletin & Review, 16*, 752–760.

Appendix

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

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

Table 3 (continued)

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

Table 3 (continued)

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

Table 3 (continued)

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

Table 3 (continued)

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