Assessing the reliability of journal data in syntax: Linguistic Inquiry 2001-2010

Jon Sprouse

Department of Cognitive Sciences

University of California, Irvine

Carson T. Schütze

Department of Linguistics

University of California, Los Angeles

Diogo Almeida

Department of Linguistics and Languages

Michigan State University

Mail correspondence to:

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

Assessing the reliability of journal data in syntax: Linguistic Inquiry 2001-2010

Abstract

There has been a consistent pattern of criticism of the reliability of acceptability judgment data in syntax for at least 50 years (e.g., Hill 1961), culminating in several high-profile criticisms within the past ten years (e.g., Edelman and Christiansen 2003, Ferreira 2005, Wasow and Arnold 2005, Gibson and Fedorenko 2010a, 2010b). The fundamental claim of these critics is that traditional acceptability judgment collection methods, which tend to be relatively informal compared to methods from experimental psychology, lead to an intolerably high number of false positive results. In this paper we empirically assess this claim by formally testing a random sample of 292 sentence types that form 146 two-condition phenomena taken from the most recent ten years of articles in a leading journal of theoretical linguistics (*Linguistic Inquiry* 2001-2010). We report the results of two experiments designed to assess the replication rate of these 146 phenomena under formal experimental methods (Experiment 1 used the magnitude estimation task and 168 participants, Experiment 2 used the forced-choice task and 96 participants). 139 of the 146 phenomena, or 95%, replicated in the formal experiments (with a margin of error of ±5%). This means that even under the (likely unwarranted) assumption that all of the discrepant results are false positives that have found their way into the syntactic literature due to the shortcomings of traditional methods, the maximum proportion of such false positives in LI 2001-2010 is 5% (±5%). We discuss the implications of these results for questions about the reliability of syntactic data, as well as the practical consequences of these results for the methodological options available to syntacticians.

1. Introduction

The reliability of data in syntactic theory has been a contentious issue for at least 50 years (e.g., Hill 1961), with the past 15 years witnessing a particularly forceful set of criticisms (Bard et al. 1996, Schütze 1996, Cowart 1997, Keller 2000, Edelman and Christiansen 2003, Ferreira 2005, Wasow and Arnold 2005, Featherston 2007, Gibson and Fedorenko 2010a, 2010b). The primary claim of these criticisms is that the traditional data collection methods in syntax, which lack some of the properties of the more formal experiments that are familiar from experimental psychology (Marantz 2005), are substantially less reliable than formal acceptability judgment collection methods. However, it would be a mistake to group all of these criticisms together: there are at least two distinct senses of reliability, and each criticism tends to focus on one more than the other. The first type of reliability concerns *false positives*, which occur when there is no difference between conditions, but the experiment falsely indicates that there is. Some critics have claimed that traditional methods are unreliable because they have an intolerably high false positive rate, and that formal experiments are more reliable because they have a substantially lower false positive rate (e.g., Wasow and Arnold 2005, Gibson and Fedorenko 2010b). The second type of reliability concerns *false negatives*, which occur when there really is a difference between conditions, but the experiment falsely indicates that there is none. Other critics have claimed that traditional methods have led to a relatively high false negative rate, and that formal experiments will lead to a lower false negative rate (e.g., Keller 2000, Featherston 2007). Both claims constitute potential challenges to syntactic theory, as the implication is that current theories are either built on spurious differences, or fail to take into account undetected differences, or both. Fortunately, both of these claims are empirical questions. Our goal in this

paper is to systematically investigate both by randomly sampling and formally testing 292 data points (or about 17%) from syntax articles published in *Linguistic Inquiry* from 2001 through 2010.

This particular case study is a natural follow-up to two previous studies that were designed to systematically investigate the two aforementioned reliability claims. First, Sprouse and Almeida (*to appear*) attempted to address the false positive concern by testing all 469 unique, US English data points in a popular syntax textbook (Adger 2003), which formed 365 phenomena covering 9 topic areas in syntax. They found that 98% of the phenomena from Adger (2003) replicated in a formal magnitude estimation experiment, and that most of the non-replications were plausible instances of insufficient statistical power in the experiments. The problem with these results is that the strength of the conclusion depends on how representative one believes that the data in Adger (2003) is of the field in general. For example, whereas Sprouse and Almeida (*to appear*) clearly believe that it is fairly representative, Gibson et al. (2011) have suggested that data from textbooks is likely more reliable than data from journals. This study directly addresses that concern by providing a distinct type of estimate. Relatedly, Sprouse and Almeida (*submitted*) systematically investigated the false negative concern by comparing the false negative rate[1] of experiments that use the magnitude estimation task (a popular choice in formal experiments) with experiments that use the two-alternative forced-choice task (a common choice in traditional, informal experiments). They found that experiments using the forced-choice task have substantially lower false negative rates than experiments using

[1] False negatives can also be thought of as the failure to detect true positives. An ideal experiment would detect true positives 100% of the time. Each failure to detect a true positive (i.e., each false negative) reduces this percentage. So if the false negative rate is 20%, then the detection rate for true positives is 80%. This percentage (100% - the false negative rate) is also known as statistical power in the statistics literature.

the magnitude estimation task for 48 phenomena that span the entire range of effect sizes in Adger (2003). This suggests that, contrary to common assumptions, traditional methods may actually have a lower false negative rates than formal experiments, at least to the extent that traditional methods tend to use the forced-choice task and formal experiments tend to use the magnitude estimation task. The present study follows-up on this general result by using both magnitude estimation and the forced-choice task to test the sample from *Linguistic Inquiry*. This allows us to identify concrete examples of phenomena that were not detectable with a well-designed magnitude estimation experiment, but were detectable with both the traditional methods and a formal forced-choice experiment, in order to highlight the potentially detrimental consequences of uniformly adopting the less powerful methodology that has been advocated by some critics (e.g., Wasow and Arnold 2005, Ferreira 2005, Gibson and Fedorenko 2010b).

2. Linguistic Inquiry 2001-2010

We chose *Linguistic Inquiry* (henceforth LI) for our study because it is a leading theoretical journal among generative syntacticians, and therefore the articles published in LI rely extensively on traditional judgment collection methods. If there is indeed an intolerable false positive rate in (generative) syntax, it seems likely that a comprehensive investigation of LI will reveal it. To be clear, we do not intend the results of this study to be a specific defense or incrimination of LI, but rather a good-faith effort to investigate the reliability of traditional methods using the most appropriate source of data. We chose the most recent ten years of LI (2001-2010) to ensure that the set of data points in our study represent current theoretical debates. There were 308 articles published in LI during those 10 years. Of the 308 articles, 229

were about syntax or sentence-level phenomena, and 79 were about other areas of linguistic theory (e.g., phonology). Of the 229 articles about syntax,114 were predominantly about US English, where predominantly was defined as greater than 80% of the data points. 115 were predominantly about languages other than English. 3 employed formal experimental methods. The 111 articles that were (i) about syntax, (ii) about US English, and (iii) did not report employing formal experimental methods formed the empirical basis for this project. We decided to focus on US-English data points in this project because online participant marketplaces (such as Amazon Mechanical Turk, Sprouse 2011a) are primarily composed of participants who speak US-English (Ipeirotis 2010). Some critics have suggested that the existence of such marketplaces eliminates much of the time cost inherent in formal experiments (e.g., Gibson and Fedorenko 2010b), therefore it seems appropriate to use these marketplaces for this case study. Furthermore, from a theoretical perspective, it seems at least plausible that the distribution and reliability of data points will be similar across languages, as the empirical value of every language is theoretically equal. This is, of course, an empirical question; but here and throughout we will assume that the estimates derived for US English data points can serve as a proxy for data points from other well-studied languages.

The first step in this project was to identify every data point in the 111 remaining articles. We employed several undergraduate research assistants with minimal training in linguistics to help with this stage of the project. They were instructed to identify all numbered examples, ignoring trees, tables, diagrams, definitions, and sentences that were not US English as "non-data-points", and recording all other examples as potential data points. In order to be as comprehensive as possible, we encouraged them to record an example as a data point if they were unsure as to its status. We further asked them to indicate whether each potential data point

included a subscripted pronoun (indicating a coreference judgment) or a greater than/less than sign (indicating a scope-based interpretation judgment), so that we would have a rough cut between those data points that are testable in a standard acceptability judgment experiment, and those that require different methodologies. In addition to the obvious time advantages, these undergraduate assistants eliminated the possibility that our theoretical biases would influence the inclusion or exclusion of potential data points from study.

The undergraduate assistants identified 5396 *potential* data points in the 111 articles. We asked the undergraduate assistants to perform a first-pass categorization procedure, dividing the 5396 potential data points into two groups: group 1 was defined as US-English sentences that could be tested with a standard acceptability judgment experiment, and group 2 was defined as either non-US-English (and therefore untestable on Amazon Mechanical Turk) or untestable in a standard acceptability judgment experiment (e.g., coreference judgments, interpretation judgments, or phenomena based on prosody). This first-pass categorization resulted in 3335 potential data points in group 1 (US-English and testable), and 2061 potential data points in group 2 (non-US-English or untestable). We then randomly sampled 499 items from these two lists (308 items from group 1, and 191 from group 2) to check the accuracy of the first-pass categorization. Based on those samples, we estimate that the total number of US-English data points in LI between 2001 and 2010 is approximately 3635[2] with a margin of error of 4.3-6.9%.[3] This estimate includes all types of US-English data: data points that can be tested with standard

[2] We excluded repeated data points within individual papers; however, we did not exclude repeated data points that spanned two or more papers. Given that journal articles are often responses or extensions of previous articles that include repeats of the previously published data, it is very likely that our estimate of 3635 is an overestimate.

[3] The margin of error is reported as a range because of the bifurcation of the sampling procedure. If we had sampled the 499 from the full 5396, the margin would be 4.3%. However, we took one sample from each of the two sub-populations. The margin of error for a sample of 191 from 2061 is 6.9%; the margin of error for a sample of 308 from 3335 is 5.4%.

acceptability judgment experiments and data points that require special methodologies. We also used the random samples to estimate the categorization of those 3635 data points by methodology as follows:

Table 1: Estimated counts of the number of US-English data points in Linguistic Inquiry from 2001 through 2010. The margin of error for these estimates is 4.3-6.9%.

Type of data point	Estimated Count	Estimated Percentage
Standard acceptability judgments	1743	48%
Coreference judgments	540	15%
Interpretation judgments	854	23%
Judgments about individual lexical items	422	12%
Judgments involving prosody	76	2%
Total	3635	

Because critics of traditional methods have focused exclusively on standard acceptability judgments, and because standard acceptability judgments are the most amenable to online testing (though see Keller 2000 for an example of coreference judgments tested online), we will focus exclusively on standard acceptability judgments in this article. Based on the estimates in Table 1, standard acceptability judgments account for approximately 48% of the (US-English) data in Linguistic Inquiry from 2001 through 2010. The other 52% of the data require different methodologies.

The next step was to create materials to be tested using formal judgment experiments. To do this, we used the random sample from the standard acceptability judgment list created by the research assistants. Because there are some data points in syntax articles that merely demonstrate the existence of a particular construction, and because such existence-based data points are more

amenable to corpus-based studies or yes-no experiments than to the kinds of experiments we conducted, we excluded such examples from the random sample; instead, we only included data points that were accompanied by a judgment diacritc (*, ?, or some combination of the two). The random sample of 308 diacritic-based items yielded a set of 150 data points that were amenable to a standard acceptability judgment experiment (the other 158 required special methodologies). We then looked up each of the 150 data points in their original articles to find a control condition (i.e., a condition without a diacritic), such that we could test 300 conditions that form 150 pairwise phenomena. We found grammatical control conditions for 115 of the phenomena in the original articles; for the remaining 35, we constructed control conditions based upon the theoretical discussion provided by the original authors. After Experiment 1 was run, we discovered that three of the phenomena we included in fact required special methodologies (two required interpretation judgments and one required special prosody), and that the materials for one phenomenon were not complete sentences (they were complex NPs with sentential complements). We excluded these four phenomena from the analysis. Based on the estimate of 1743 US-English data points in LI from 2001 through 2010, any population estimates (e.g., an estimate of the replication rate using formal experiments) derived from our random sample of 292 (analyzable) data points will yield a margin of error of 5.3-5.8%.[4] A full list of the sentence types that were tested, along with example sentences and mean ratings for each, is provided in Appendix A.

[4] The margin of error is a range because there are (at least) two ways to count the 35 additional conditions that we constructed to serve as controls for 35 of the sampled conditions. If we add the 35 constructions to the population count (i.e., treat them as if they were part of the original population), the margin of error would be 5.3%. If instead we subtract them from the sample size (i.e., treat them as if they do not exist in either the sample or the population for purposes of calculating the margin of error), then the margin of error is 5.8%.

3. Experiment 1: Magnitude estimation

As a first estimate of the replication rate for data points from LI, we tested all 292 sentences (146 pairwise phenomena) using the magnitude estimation task. Admittedly, experiments using the magnitude estimation task have been shown to achieve less statistical power than experiments using the more conventional (in syntactic research) two-alternative forced-choice task (Sprouse and Almeida, 2011, *submitted*) and to be based upon cognitive assumptions that do not appear to hold (Sprouse 2011b). However, we believe that the fact that magnitude estimation has been advocated by the proponents of formal judgment experiments (Bard et al. 1996, Keller 2000, Featherston 2005a, 2005b, Alexopoulou and Keller 2007, Sprouse 2008, 2009, Sprouse et al. 2011, Weskott and Fanselow 2011, Sprouse et al. 2012, but cf. Myers 2009, Sprouse 2011b) makes it the logical first choice for estimating the replication rate of traditional judgments under formal conditions. In other words, this experiment gives us some insight on what syntacticians would have observed (and potentially concluded) had they adopted the recommendations made by some critics to forego the use of informal experiments in favor of formal experiments beginning 10 years ago.

Participants

Three groups of 56 participants (N = 168) completed experiment 1. We chose 56 participants (per sub-experiment) using the following logic. Sprouse and Almeida (*submitted*) found that 45 participants, each providing one judgment per condition, would be sufficient to detect 95% of the phenomena in Adger (2003). Given that critics have suggested that the phenomena in journal

articles are less reliable than textbook data, we decided to add 10 participants to the sample sizes (rounded up to 56 to lead to an even number of each survey).

Participants were recruited online using the Amazon Mechanical Turk (AMT) marketplace, and paid $3.00 for their participation (see Sprouse (2011a) for evidence of the reliability of data collected using AMT when compared to data collected in the lab). 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, and consequently there was no incentive to lie. None of the participants answered 'no' to these questions, therefore none were excluded from the analysis.

Materials

Division into 3 sub-experiments. The 300 conditions (recall that 8 were later found to require special methodologies, leaving only 292 in the final analysis) were distributed among 3 separate sub-experiments in order to make the total length of each survey 100 items (plus 6 practice items, see below). The distribution of the conditions among the experiments was random; however, the two conditions that form each phenomenon were always distributed as a pair to the same sub-experiment, such that every phenomenon was tested using a repeated-measures design. Because the 300 conditions form 150 pairwise phenomena consisting of one acceptable and one unacceptable item, the distribution of acceptable and unacceptable items per survey is (by hypothesis) balanced.

Division into eight versions of each experiment. Eight tokens of each condition were constructed such that the structural properties of the condition were maintained but the lexical items used varied. The eight tokens were distributed among eight lists using a Latin Square procedure such that the lists did not contain identical lexicalizations of the two conditions that form each phenomenon. This resulted in 8 lists per sub-experiment (24 total) such that each list contained 100 conditions (50 phenomena) and only one token of each condition. Each list was then pseudorandomized such that related conditions were separated by at least two unrelated conditions. The result was 8 versions each of the 3 sub-experiments, or 24 distinct surveys.

Task

The task was magnitude estimation (Stevens 1957, Bard et al. 1996, Cowart 1997). In the magnitude estimation task, participants are presented with a reference sentence, called the *standard*, which is pre-assigned an acceptability rating, called the *modulus*. Participants are asked to use the standard to estimate the acceptability of the experimental items ("acceptability" was defined following the guidelines suggested in Schütze and Sprouse 2011). 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. The standard sentence was in the middle range of acceptability: *Who said that my brother was kept tabs on by the FBI?* The standard was assigned a modulus of 100 and repeated every seven items to ensure that it was always visible on the screen.

Presentation

In order to familiarize participants with the magnitude estimation task, they 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. After the practice phase, they were told that this procedure can be extended to sentences. No explicit practice phase for sentences was provided; however, six additional "anchoring" items (two each of acceptable, unacceptable, and moderate acceptability) were placed as the first six items of each survey. These items were identical, and presented in the identical order, for every survey. Participants rated these items just like the others; they were not marked as distinct from the rest of the survey in any way. However, these items were not included in the analysis as they served simply to expose each participant to a wide range of acceptability prior to rating the experimental items (a type of unannounced "practice"). This resulted in surveys that were 106 items long. The surveys were advertised on the Amazon Mechanical Turk website (Sprouse 2011a), 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.

The definition of a successful replication

This study is designed around 146 pairwise phenomena (i.e., two conditions per phenomenon). Therefore we will define a successful replication as the detection of a statistically significant difference in the direction reported by the original author. In order to be as informative as possible, we will use both a frequentist statistical test (the *t*-test) and a Bayesian statistical test

(Bayes factor analysis) to look for statistically significant differences. Bayes factors provide an intuitively natural measure of the strength of the evidence for each of the two hypotheses: the experimental hypothesis (H1), which holds that there is a significant difference between the two conditions, and the null hypothesis (H0), which holds that there is no difference between the two conditions. The Bayes factor is an odds ratio of one hypothesis over the other (the directionality of this ratio is up to the experimenter; for our purposes we will use the form H1:H0). 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. In order to be as conservative as possible in the estimation of replication rates, we will use the two-tailed versions of the *t*-test[5], and a non-directional version of the Bayes factor analysis (Rouder et al. 2009). We believe that by (i) employing both frequentist and Bayesian statistical tests, (ii) adopting the more conservative versions of those tests, and (iii) only presenting one token of each condition to each participant, we ensure that the replication rates will be a minimum estimate rather than a maximum estimate. In other words, we are attempting to bias the results against the conclusion that traditionally collected judgments are highly replicable. However, it should be noted that because we employed non-directional versions of the statistical tests (to err on the side of conservativity), we were forced to manually compare the mean ratings of each condition to ensure that they were in the intended direction (mean ratings are reported in Appendix A).

At this point it should also be noted that argumentation within syntactic articles sometimes relies upon information beyond the simple differences between conditions that are reported here. For example, an author may make reference to the size of the difference between two conditions, or could make reference to the relative position of each condition along the

[5] This decision in effect sets the criterion for significance at $p < .025$ because we are actually interested in directional hypotheses, not non-directional hypotheses.

acceptability spectrum (e.g., both are acceptable, but one is more so than the other). It is technically possible to incorporate such additional assumptions into the definition of successful replication; however, we have chosen not to do so for several reasons. First, few (if any) of the original authors explicitly included such criteria, so we would have to create our own criteria and apply it to their hypotheses. Given that these criteria are likely to vary from researcher to researcher, that seems problematic. Second, the fact that acceptability can be (in principle) influenced by several factors (syntax, semantics, pragmatics, and even parsing difficulty; for a recent discussion see Sprouse and Almeida, 2011) means that the specific properties of each sentence need to be taken into account in order to fairly apply extra numerical criteria (otherwise, non-syntactic factors could affect the sentence in a way that leads to a false negative under the additional criteria). Such an analysis is currently impossible because we do not have a comprehensive theory of acceptability judgments. Finally, some critics (e.g., Featherston, 2007, Wasow & Arnold, 2005) have suggested that statistical tests should be routinely adopted for syntactic data analysis. The types of statistics that have been suggested only provide evidence for a simple difference (including additional numeric criteria is possible, but requires a different class of statistical tests, such as randomization tests (Edgington and Onghena, 2007)). Therefore, adopting this definition allows us to more directly engage with the critical literature. Because we are aware that some readers may wish to evaluate these results using additional criteria, we have reported the mean ratings of each condition in Appendix A.

Results

Acceptability judgments from each participant were *z*-score transformed prior to analysis to eliminate some of the forms of scale bias that potentially arise with scaling tasks (see Schütze and Sprouse 2011 for a review). As mentioned above, we ran two types of statistical tests for the pairwise phenomena. First, we calculated two-tailed *p*-values using standard paired *t*-tests. Second, given the growing interest in Bayesian statistics across all domains of cognitive science, we calculated Bayes factors for each comparison (Gallistel 2009, Rouder et al. 2009). Tables 2 and 3 report a summary of the results of these analyses (see Appendix A for more detailed statistical results). The resulting Bayes factors are also categorized using the intuitive classification proposed by Jeffreys (1961). Appendix A reports the *p*-values and Bayes factors for every condition.

Table 2: Summary of the results of Experiment 1 for paired *t*-tests. There were 292 conditions that formed 146 pairwise phenomena. 56 participants per sub-experiment; each participant rated one token of each condition.

Hypothesis	Description	*p*-value	Count	Percentage
H1 in opposite direction	Significant	<.05	1	<1%
	Marginal	<.10	0	--
H0	Non-significant	>.10	14	10%
H1 in predicted direction	Marginal	<.10	2	1%
	Significant	<.05	8	5%
	Significant	<.01	7	5%
	Significant	<.001	6	4%
	Significant	<.0001	108	74%

Table 3: Summary of the results of Experiment 1 for Bayes factor analyses. There were 292 conditions that formed 146 pairwise phenomena. 56 participants per sub-experiment; each participant rated one token of each condition.

Hypothesis	Description	Bayes factor	Count	Percentage
H1 in opposite direction	Extreme evidence	>100	1	<1%
	Very strong evidence	30-100	0	--
	Strong evidence	10-30	0	--
	Substantial evidence	3-10	0	--
	Anecdotal evidence	1-3	0	--
H0	Extreme evidence	<1/100	0	--
	Very strong evidence	1/100-1/30	0	--
	Strong evidence	1/30-1/10	0	--
	Substantial evidence	1/10-1/3	13	9%
	Anecdotal evidence	1/3-1	6	4%
H1 in predicted direction	Anecdotal evidence	1-3	5	3%
	Substantial evidence	3-10	4	3%
	Strong evidence	10-30	3	2%
	Very strong evidence	30-100	5	3%
	Extreme evidence	>100	109	74%

Using the magnitude estimation task with 56 participants rating one token per condition, non-directional versions of the *t*-test and Bayes factor analysis, and including marginal results as equivalent to a replication failure, we can make the following estimates (with a margin of error of 5.3-5.8%): *t*-tests yield a replication rate of 88%, with 1% marginal in the predicted direction, 10% non-significant, and 1% significant in the opposite direction than originally reported; Bayes factor analyses yield a replication rate of 82%, with 3% anecdotal evidence in the predicted direction, 4% anecdotal evidence for no difference, 10% substantial evidence for no difference, and 1% extreme evidence for a difference in the opposite direction than originally reported.

Discussion

The results of Experiment 1 suggest a replication rate for LI 2001-2010 of 88% by *t*-test and 82% by Bayes factor using a magnitude estimation task with 56 participants each judging one token per condition. On the one hand, one could simply assume that all of the replication failures are in fact true negatives (i.e., there really is no difference between the conditions, which means that the results reported in LI were false positives). This would mean that the traditional methods have a false positive rate of 12% by *t*-test or 18% by Bayes factor analysis. However, there is at least one reason to doubt that all of the replication failures are true negatives: Sprouse and Almeida (2011, *submitted*) have independently demonstrated that experiments using the magnitude estimation task are less sensitive than experiments using the forced-choice task. In order to be sure that the replication failures were not due to insufficient statistical power, we should also test the replication failures using the more powerful forced-choice task. In the next section (section 4) we report just such an experiment. We will postpone a discussion of the replication failures from Experiment 1 until after the results of Experiment 2 (i.e., after we assess whether they were false negatives due to the relatively lower power of magnitude estimation experiments). However, it is possible to reconstruct the full list of replication failures from Experiment 1 by either looking through Appendix A for negative results, or by combining Table 6 (the negative results from Experiment 2) and Table 9 (the negative results from Experiment 1 that were positive results in Experiment 2).

4. Experiment 2: Forced choice on the replication failures from Experiment 1

In order to gather more information about the replication failures in Experiment 1, we tested them using the more sensitive forced-choice task (Sprouse and Almeida 2011, *submitted*). Though no experiment can completely eliminate the possibility of spurious results, we believe that by using a task that has been empirically proven to be more sensitive we can provide a more accurate minimum replication rate for the data in LI from 2001 through 2010. There were 26 potential replication failures according to the Bayes factor analyses (of which a proper subset of 18 were replication failures according to the *t*-tests). We combined these 26 potential replication failures with the 24 clearest replications from Experiment 1 in order to create a 100 item survey (50 forced-choice pairs) that was roughly balanced between clear distinctions and less-clear distinctions. One phenomenon was excluded from the analysis because of a problem with the materials (it was also excluded from the results of Experiment 1, see section 3), leaving 25 replication failures (of which 17 were from the *t*-tests).[6]

Participants

96 participants completed Experiment 2. This number was chosen following the results of Sprouse and Almeida (*submitted*) in order to ensure that Experiment 2 had substantially more power than Experiment 1. Participants were recruited online using the Amazon Mechanical Turk (AMT) marketplace, and paid $2.00 for their participation. Participant selection criteria were the same as Experiment 1. No participants were excluded from the experiment.

[6] The excluded phenomenon did replicate in Experiment 2: $p<.0001$, BF>100; however, we excluded it here for consistency.

The forced choice task

(1) a. What do you think that John bought?

b. What do you wonder whether John bought?

In this forced choice task, participants are asked to directly compare two sentence types (for example 2a and 2b) and decide which sentence is more acceptable (using the same definition of "acceptable" from Experiment 1; see also Schütze and Sprouse 2011). The relevant sentence types are designed to be as structurally and lexically similar as possible.

Materials

The materials consisted of a subset of those used in Experiment 1, again with 8 lexicalizations of each condition. Because the forced choice task requires the direct comparison of lexically related 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, two sets of the 8 lists were created to counterbalance the order of presentation within each pair across the lists, and across pairs within each list. The result is 16 pseudorandom lists, such that each list contains one lexicalization of each phenomenon, the numbers of expected *top*-responses and *bottom*-responses are equal, and there are no predictable dependencies between phenomena (e.g., if phenomenon 1 is a *top*-response, phenomenon 2 is necessarily a *bottom*-response). These counterbalancing procedures minimize the effect of response biases on the results (e.g., a strategy of 'always choose the top item'). Finally, the order of the pairs in each list was randomized, resulting in 16 surveys containing 50 randomized and

counterbalanced pairs (100 total sentences). As in Experiment 1, each participant judged one token of each condition. An example of each phenomenon is given in Appendix B, along with the number of responses in the correct direction, *p*-values, and Bayes factors.

Presentation

The surveys were advertised on the Amazon Mechanical Turk website (see Sprouse 2011a), and presented as web-based surveys using an HTML template available on the first author's website. There was no practice for the forced choice experiment, as the task is generally considered intuitively simple. Participants completed the surveys at their own pace.

Results

Responses were coded as either *successes* (if the participant chose the sentence in the pair that is claimed to be *more* acceptable) or *failures* (if the participant chose the sentence that is claimed to be *less* acceptable). Responses were then analyzed in two ways: (i) using the traditional sign-test (with two-tailed *p*-values), and (ii) using the non-directional Bayes factor calculation for binomial responses made available by Jeff Rouder on his website: http://pcl.missouri.edu/bayesfactor. Tables 4 and 5 report a summary of the results of these analyses. Appendix B reports the full list of responses, *p*-values, and Bayes factors.

Table 4: Summary of the results of Experiment 2 for sign-tests. There were 50 target conditions that formed 25 pairwise phenomena. 96 participants each rated one pair of each phenomenon.

Hypothesis	Description	*p*-value	Count
H1 in opposite direction	Significant	<.05	2
	Marginal	<.10	0
H0 (no difference)	Non-significant	>.10	4
H1 in predicted direction	Marginal	<.10	1
	Significant	<.05	3
	Significant	<.01	2
	Significant	<.001	2
	Significant	<.0001	11

Table 5: Summary of the results of Experiment 2 for Bayes factor analyses.

Hypothesis	Description	Bayes factor	Count
H1 in opposite direction	Extreme evidence	>100	2
	Very strong evidence	30-100	0
	Strong evidence	10-30	0
	Substantial evidence	3-10	0
	Anecdotal evidence	1-3	0
H0 (no difference)	Extreme evidence	<1/100	0
	Very strong evidence	1/100-1/30	0
	Strong evidence	1/30-1/10	0
	Substantial evidence	1/10-1/3	4
	Anecdotal evidence	1/3-1	3
H1 in predicted direction	Anecdotal evidence	1-3	2
	Substantial evidence	3-10	1
	Strong evidence	10-30	1
	Very strong evidence	30-100	1
	Extreme evidence	>100	11

The 17 replication failures in Experiment 1 according to the *t*-tests were reduced to only 7 in Experiment 2 (including marginal results as replication failures), of which 2 were significant in the opposite direction than originally reported. The 25 replication failures in Experiment 1 according to Bayesian statistics were reduced to 11 in Experiment 2 (including anecdotal evidence as replication failures), of which 2 were extreme evidence for a difference in the opposite direction than originally reported. This is consistent with the known sensitivity difference between magnitude estimation experiments and forced-choice experiments and the fact that we used a larger sample size in Experiment 2 (Sprouse and Almeida 2011, *submitted*). It should also be noted that all 24 of the previously clear replications from Experiment 1 that were included as filler items also replicated clearly in Experiment 2.

Discussion

The remaining replication failures for Experiment 2 (according to the sign-tests) are presented in Table 6.

Table 6: The clear and marginal replication failures from Experiment 2. Identifier is in the format VOLUME.ISSUE.AUTHOR.EXAMPLE.JUDGMENT, Hits reports the number of responses in the correct direction, *p* reports the *p*-value obtained from the two-tailed sign-test, and BF reports the non-directional Bayes factor.

Identifier	Example	Hits	*p*	BF
34.1.phillips.93b.?*	Wendy stood more buckets in the garage than Peter did in the basement.	10/96	.01	>100
34.1.phillips.92b.g	Wendy stood more buckets than Peter did in the garage.			
35.3.hazout.67c.*	There is likely a man to appear.	25/96	.01	>100
35.3.hazout.67a.g	There is likely to appear a man.			
41.4.bruening.9b.*	What did he prove an account of false?	56/96	.06	0.48
41.4.bruening.9c.g	Who did he give statues of to all the season-ticket holders?			
35.3.richards.17b.*	To whom did you give what?	49/96	.46	0.13
35.3.richards.17a.g	What did you give to whom?			
35.1.bhatt.94a.*	I expect that everyone you do will visit Mary.	42/96	.13	0.27
35.1.bhatt.94b.g	I expect that everyone will visit Mary that you do.			
33.1.fox.69a.*	John wants for everyone you do to have fun.	42/96	.13	0.27
33.1.fox.69b.g	John wants for everyone to have fun that you do.			
32.4.lopez.10a.*	We proclaimed to the public John to be a hero.	42/96	.13	0.27
32.4.lopez.9a.g	We proclaimed John to the public to be a hero.			

The first thing to note about this list is that no author or article appears in this list more than once. To the extent that several items were sampled from each of these articles, this fact makes it unlikely that a particular article has an abnormally high false positive rate. The second thing to note is that there are in fact two effects that go in the opposite direction than the one reported by the original authors. The first is a VP-ellipsis example from Phillips (2003), which was also observed to go in the opposite direction in Experiment 1. The fact that this effect is clearly in the opposite direction even under a forced-choice experiment suggests that it is either a faulty data point, or that ellipsis examples of this type cannot be accurately tested in standard acceptability judgment experiments (perhaps instead requiring specific context or intonation to make the readings intended by Phillips for each condition apparent). The fact that two of the other replication failures also involve VP ellipsis (i.e., Antecedent Contained Deletion from Fox 2002

and Bhatt 2004) may lend some support to the latter possibility, although additional research is clearly necessary. The second reversal is an existential-*there* contrast from Hazout (2004). The Hazout (2004) reversal is notable because it was not a reversal in Experiment 1 (it was significant in the correct direction by *t*-test and marginal by Bayes factor analysis), but became a reversal in Experiment 2. In other words, two methodologies (traditional and magnitude estimation) lead to results in the direction reported by Hazout (2004), and one methodology (forced-choice) leads to a result in the opposite direction. This means that this phenomenon is either a false positive for traditional and magnitude estimation methods, or it is a false positive for the forced-choice method. Therefore, one way or another, this phenomenon is highlighting a problem with one of the major data collection methods; additional research is clearly necessary to determine which result is the correct one, and what exactly went wrong in the other method(s). We will discuss the broader impact of these two reversals for the reliability of syntactic data in the general discussion in section 5.3.

We will not discuss the remaining three replication failures (Bruening (2010), Lopez (2001), and Richards (2004)) in any theoretical detail as they each appear to be testing distinct hypotheses. However, it may be of interest to note that the Bruening (2010) example is very close to significant by non-directional sign-test (p=.06), which suggests that it may be a false negative (and certainly would have been a positive under a directional sign-test). The Lopez (2001) example may also be of interest, as it was attributed by Lopez (2001) to Postal (1974). This suggests that if this phenomenon is indeed a true negative (i.e., a false positive introduced into the literature by traditional methods), it has survived as a false positive through at least two publications (Postal 1974 and Lopez 2001). The final replication failure, from Richards (2004), may also be of interest as it contains two properties that may be likely to prompt erratic

judgments on the part of non-linguist participants: pied-piping and *whom*. Neither of these properties are well-attested among the general population of US-English speakers; they are associated with prescriptive grammars and "correct" writing. It is possible that the presence of these properties in a written judgment experiment may have affected the results.

Leaving the exact cause of the replication failures in Experiment 2 aside as a topic for future research, we can nonetheless use the results of Experiment 2 to derive a more accurate minimum estimate for the replication rate for LI 2001-2010. We can combine the results of the two experiments as follows. First, we will assume that a negative result in Experiment 2 is a true negative. This type of assumption is line with the assumptions of critics, who tend to assume that a negative result in a formal experiment is a true negative. Second, we will assume that a positive result in Experiment 2 is a true positive, even if it was negative in Experiment 1. This assumption is in line with the results of Sprouse and Almeida (2011, *submitted*), who found that experiments that use the forced-choice task are more powerful than experiments that use the magnitude estimation task, which in this case means that negative results in Experiment 1 (magnitude estimation) that become positive in Experiment 2 (forced-choice) were likely false negatives in Experiment 1. Finally, we will assume that a result in Experiment 1 that was positive by both *t*-test and Bayes factor is a true positive (note that these phenomena were only tested once, so the result from Experiment 1 is all we have). Tables 7 and 8 report these new counts:

Table 7: The combined results of Experiment 1 and Experiment 2 for frequentist statistics. Responses were collapsed into three categories because they were obtained using two different statistical tests between Experiment 1 (*t*-tests) and Experiment 2 (sign-tests).

Result	Count	Percentage
Significant in the opposite direction	2	1.4%
Marginal or Non-significant	5	3.4%
Significant in the predicted direction	139	95%

Table 8: The combined results of Experiment 1 and Experiment 2 for Bayes factor analyses. Responses were collapsed into three categories because they were obtained using two different Bayes factor equations between Experiment 1 (pairwise comparisons) and Experiment 2 (binomial responses).

Result	Count	Percentage
Evidence for a difference in the opposite direction	2	1.4%
Anecdotal evidence for a difference or Evidence for no difference	9	6.2%
Evidence for a difference in the predicted direction	135	93%

Using frequentist statistical tests, 139 out of 146 phenomena reached significance in the correct direction in either Experiment 1 or Experiment 2, five phenomena led to negative results in both Experiment 1 and Experiment 2, and two phenomena led to sign reversals in Experiment 2. This suggests a minimum replication rate of 95%, and a reversal rate of just over 1%. For Bayesian statistical tests, the replication rate drops slightly to 135 out of 146, or 93%. The margin of error based on the size of these samples and the estimated number of data points in LI is ±5.3-5.8%.

5. General discussion

5.1 The false positive rate of traditional methods

The first (of two) questions about the reliability of syntactic methods centers around false positives: Do traditional methods lead to an intolerably high false positive rate (and relatedly, do formal experiments lead to a lower one)? We constructed the experiments in this study around phenomena that were originally reported using traditional methods, therefore we can only answer the first half of this question (i.e., we have no information about the false positive rates for formal experiments). Furthermore, it is literally impossible to derive a true false positive rate for either method because there is no independent method of measuring acceptability. The best we can do is compare the results of the two methods to see how well they converge. For the sake of argument we can adopt an assumption espoused by some critics that formal experiments are the more reliable method, and then derive a replication rate for traditional methods relative to formal experiments. We adopted several assumptions that would bias the calculations toward lower replication rates, such as assuming that all of the negative results are true negatives, and using non-directional statistical tests. These assumptions led to a minimum replication rate of 95% ±5.3-5.8% for traditional methods under frequentist statistics, and a minimum replication rate of 93% ±5.3-5.8% for traditional methods under Bayesian statistics. The question then is what can we conclude from these estimates.

The weakest possible conclusion (in the sense that it relies on no additional assumptions) that can be drawn is that there are 139 phenomena that were randomly sampled from LI 2001-2010 that straightforwardly replicate using formal experiments. We can combine this with the

359 phenomena from Adger (2003) reported by Sprouse and Almeida (*to appear*) for a minimum of 498 clearly replicable phenomena that have now been reported in the literature. This means that any critic who wishes to claim that syntactic data is unreliable must also account for these 498 phenomena. Given that these phenomena were selected without bias (exhaustively in the case of Adger (2003), randomly in the case of LI 2001-2010), such an argument should probably take the form of supplying a set of replication failures that increases the replication failure rate to an intolerably high level relative to these 498 phenomena. For example, if a critic assumed that a 10% false positive rate is intolerably high, they would need to provide a set of 50 phenomena that do not replicate in order to make a credible claim that the false positive rate for traditional methods is 10% (and even then, there is still the possibility that some of the replication failures could turn out to be false negatives). We have reported 7 phenomena here, and 6 more in Sprouse and Almeida (*to appear*). Even assuming that the 5 case studies recently provided by Wasow and Arnold (2005) (2) and Gibson and Fedorenko (2010b) (3) are true replication failures (a fact that has been empirically challenged (Culicover and Jackendoff 2010, Sprouse and Almeida 2011, Sprouse and Almeida *in prep*)), that means that 32 more phenomena are necessary to suggest a 10% replication failure (i.e., false positive) rate.

The fact that these results have clarified the empirical requirements for making credible claims about the reliability of syntactic data raises interesting questions about what constitutes tolerable versus intolerable false positive rates. This is clearly a subjective issue; however, there is some degree of consensus in experimental psychology that we can use as a guide. For example, the field of experimental psychology appears to have adopted 5% as an implied maximum tolerable false positive rate for statistical tests over the long run. This is reflected in the widespread convention that differences are considered "significant" if the *p*-value derived

from null hypothesis significance tests is below .05. *P*-values represent the likelihood of the observed results (or results that are more extreme) assuming that in fact there is no difference between the experimental conditions (i.e., assuming the null hypothesis). By setting the criterion for rejecting the null hypothesis at .05, the field is saying that in a world in which the null hypothesis is true 100% of the time, we will tolerate falsely rejecting it (i.e., deriving a false positive) 5% of the time due to statistical uncertainty (Nickerson, 2000). Because statistical uncertainty is only one potential source of false positives (others being experimental and task confounds of various kinds), the 5% tolerance cannot be interpreted as the actual rate of the field, but rather the maximum false positive rate that the field will tolerate from the uncertainty in statistical tests alone. Therefore it is probably safe to assume that most psychologists believe the actual false positive rate may be (at least) somewhat greater than 5% (see also Wetzels et al. 2011 for a comparison of frequentist and Bayesian statistics that suggests that several significant results in the literature would not be considered credible evidence under Bayesian statistics). Although it is impossible to quantify exactly what the true false positive rate is (in any field of cognitive science), it seems reasonable that 10%, or double the implied maximum for statistical tests, would likely be considered high.

To the extent that one is willing to adopt the two rules of thumb discussed above (that a 5% false positive rate due to statistical tests alone is tolerable, and that 10% would likely be considered high), it is possible to evaluate these results in absolute terms. The results of Sprouse and Almeida (*to appear*) suggest that the maximum false positive rate for Adger (2003) is 2%, and the results reported here suggest that the maximum false positive rate for LI 2001-2010 is 5% (±5.3-5.8%). Both of these maximum estimates are in line with the target maximum rate of 5% set by the methodological consensus in experimental psychology. These maximum estimates

are also substantially lower than 10%. Finally, although it is rare to find suggested false positive rates in the published literature, it should be noted that these maximum estimates are also substantially lower than a recent suggestion by Gibson and Fedorenko (2010b) and Gibson et al. (2011), who to their credit were willing to suggest a specific false positive rate that they would consider problematic:

> For example, suppose that 90% of the judgments from an arbitrary paper are correct (which is probably a high estimate). (Gibson and Fedorenko 2010b, p. 10, see also Gibson et al. 2011, p. 511, for a nearly identical quote)

5.2 The false negative rate of formal experiments

The second question about the reliability of syntactic methods centers around false negatives: Do traditional methods lead to an intolerably high false negative rate, and relatedly, do formal experiments lead to a lower one? We constructed the experiments in this study around phenomena that were originally reported as positive results using traditional methods, therefore we have no information about the false negative rate of traditional methods (that would require investigating phenomena that were reported as negatives by traditional methods; see Keller 2000, Featherston 2005b, and Sprouse et al. 2011 for possible examples). However, because we conducted two experiments using two tasks that are known to achieve different levels of statistical power (Sprouse and Almeida *submitted*), we do have some information about the false negative rate of formal experiments. For example, Table 9 reports a list of 11 phenomena that were reported as negatives in Experiment 1 (magnitude estimation) but were reported as positives in Experiment 2 (forced-choice):

Table 9: Phenomena that were reported as negatives in Experiment 1 (magnitude estimation), but reported as positives in Experiment 2 (forced-choice).

Identifier	Example
32.1.martin.28b.??	Sarah convinced Bill that he would have gone to the party by the time he goes to bed this evening.
32.1.martin.27b.g	Sarah convinced Bill that he will have gone to the party by the time he goes to bed this evening.
32.3.culicover.49a.*	Jack asked Sally to be allowed to take care of herself.
32.3.culicover.49a.g	Jack asked Sally to be allowed to take care of himself.
34.4.boskovic.3c.*	They suspected and we believed Peter would visit the hospital.
34.4.boskovic.4c.g	They suspected and we believed that Peter would visit the hospital.
34.4.boskovic.3d.*	Mary believed Peter finished school and Bill Peter got a job.
34.4.boskovic.4d.g	Mary believed that Peter finished school and Bill that Peter got a job.
34.4.boskovic.3e.*	John likes Mary, Jane didn't believe
34.4.boskovic.4e.g	That John likes Mary, Jane didn't believe
34.4.boskovic.7a.??	What did they believe at that time that Peter fixed?
34.4.boskovic.7c.g	At that time, what did they believe that Peter fixed?
35.2.larson.44a.??	A taller man than my father walked in.
35.2.larson.44a.g	A man taller than my father walked in.
38.3.haddican.39.*	Blake said that he would beard his tormentor before the night was up, but the actual doing of so proved rather difficult.
38.3.haddican.39.g	Blake said that he would beard his tormentor before the night was up, but the actual doing of it proved rather difficult.
41.3.landau.10b.*	The game was played shoeless.
41.3.landau.10a.g	The game was played wearing no shoes.
41.3.costantini.2b.??	All the men seem to have all eaten supper.
41.3.costantini.2b.g	The men seem to have all eaten supper.
41.4.haegeman.18a.*	Bill asked if such books John only reads at home.
41.4.haegeman.18a.g	Bill knows that such books John only reads at home.

Because these 11 phenomena were reported by both the traditional method and by the forced-choice task, and because the forced-choice task is known to be more powerful than the magnitude estimation task (Sprouse and Almeida 2011, *submitted*), it is likely that these 11

phenomena are indeed true positives. This means that the negative results obtained for these 11 phenomena in Experiment 1 were false negatives. In other words, there was a minimum false negative rate of about 8% (11 out of 146) in Experiment 1 (magnitude estimation).

The conclusions that we draw based on this false negative rate once again depend on what one is willing to assume. For example, some critics have strongly advocated the universal adoption of formal experiments that use numerical rating tasks along the lines of Experiment 1 (e.g., Featherston 2007, Wasow and Arnold 2005, Gibson and Fedorenko 2010a, 2010b). Although we do not have false negative rates for every type of numerical rating task, one could assume that the overwhelming popularity of magnitude estimation might mean that if syntacticians were to adopt the critics' recommendation, they would do so by conducting magnitude estimation experiments. Given the 8% false negative rate observed for Experiment 1, this would mean that at least 8% of the true positives currently reported in LI 2001-2010 would be erroneously reported as negatives. This suggests to us that the approach advocated by some critics would in fact have been worse for the field than traditional methods, had they been universally adopted. Instead, we believe that syntacticians (and indeed all researchers) must be aware of the relative costs and benefits of each methodology with respect to their research questions, and be allowed to make the decision for themselves. Science cannot be reduced to a simple recipe (see section 5.4).

5.3 Reversals

We have primarily focused on false positives and false negatives in the previous discussions; however, there is a third type of reliability: the direction of the difference. Significant effects in

the wrong direction, often called *sign reversals* (and sometimes called Type III errors), would be devastating to theory construction. The results of Experiment 1 suggest one potential sign reversal (from Phillips 2003). The results of Experiment 2 corroborate the first sign reversal, and add a second potential reversal (from Hazout 2004). In total then we have 2 potential reversals out of 146 phenomena, for a reversal rate of about 1%. Combining this with the Adger (2003) results from Sprouse and Almeida (*to appear*) lowers the estimate even more, as there were no reversals out of 365 phenomena, for a total of 2 out of 511 phenomena. Although we know of no general consensus regarding tolerable and intolerable sign reversal rates, 1% (or lower) suggests to us that sign reversals are not frequent enough to cause a substantial problem for syntactic theory.

It may also be possible to use the low sign reversal rate to assess some of the claims of cognitive bias that have been raised in the critical literature (Dąbrowska 2010, Ferreira 2005, Gibson and Fedorenko 2010a, 2010b), although again this depends upon additional assumptions. For example, Ferreira (2005) and Gibson and Fedorenko (2010a, 2010b) argue that syntacticians should not use other syntacticians (or graduate students) as participants in their experiments because the theoretical knowledge that syntacticians hold could bias their responses to confirm (or disconfirm) their preferred hypothesis. We can call this theory-driven cognitive bias. The question then is what types of replication failures we would expect if theory-driven cognitive bias were influencing the results of traditionally collected acceptability judgments. One possible prediction is that there would be a relatively high rate of sign reversals: theoretically knowledgeable participants could be expected to use their theoretical biases to override the actual direction of the differences. The fact that there are relatively few sign reversals (maximally 2 out of 511 phenomena) would then suggest that there is relatively little theory-

driven cognitive bias. Of course, there are other types of cognitive bias. For example, the 7 negative results that were observed in this study could be examples of a cognitive bias among syntacticians to over-interpret very small acceptability differences as theoretically relevant. However, this type of bias affects all researchers, regardless of whether they use formal experiments. In fact, formal experiments (with naïve participants or not) cannot correct for this sort of bias: formal experiments only allow us to estimate the magnitude of acceptability differences between relevant sentence types; any conclusions about their theoretical (ir)relevance can only be evaluated in light of existing theories.

5.4 The costs and benefits of syntactic methods

The decision about which methodology to use can only be made by weighing the costs and benefits of each methodology relative to the research question at hand. Critics of traditional methods in syntax have suggested that there may be (at least) two heavy costs to the use of traditional methods: a high rate of false positives and a high rate of false negatives. Given that some critics have advocated the nearly universal adoption of formal experiments (e.g., Ferreira 2005, Featherston 2007, Gibson and Fedorenko 2010a, 2010b), we can only conclude that they assume that these costs are high enough to outweigh any benefit that traditional methods may have. However, the results of this study suggest that these concerns have been overstated: traditional methods do seem to have a reasonable false positive rate, according to the standards used in experimental psychology, and may have a lower false negative rate than formal experiments with numerical rating tasks. At best, this suggests that the critics' suggestion for the universal adoption of formal experiments is unwarranted; at worst, the critics' suggestion may be

detrimental to the field, as several phenomena that were detected using traditional methods would not have been detected using formal experiments with numerical rating tasks (see also Sprouse and Almeida 2011, *submitted*).

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

that traditional methods are reliable, researchers in fields that are used to formal experiments may erroneously believe that traditional methods are unreliable because they lack many of the properties of formal experiments (e.g., Ferreira 2005, Gibson and Fedorenko 2010a, 2010b).

The benefits of formal experiments are relatively straightforward as well. Formal experiments are often necessary for the reliable collection of numerical ratings, so they are the best choice for hypotheses about the *size* of the difference between conditions (e.g., Sprouse et al. 2011), hypotheses about the source of gradient acceptability (e.g., Keller 2000, Featherson 2005b), and comparisons between acceptability and other cognitive measures (e.g., Sprouse et al. 2012). As mentioned above, formal experiments also tend to be analyzed using statistical tests, which can provide a type of confidence in the results when replication is difficult or costly. And formal experiments are more likely to be seen as reliable to researchers in fields that rely exclusively on formal experiments (Ferreira 2005, Gibson and Fedorenko 2010a, 2010b). The costs of formal experiments have rarely been discussed in the literature. First and foremost, formal experiments are much more expensive than traditional methods. In the laboratory, participants are routinely paid $5 for the completion of a 100 item magnitude estimation survey; on Amazon Mechanical Turk the same survey would cost $3.30 per participant ($3 to the participant, $.30 to Amazon). A 100 item survey can maximally test 50 two-condition phenomena (one rating per condition per participant), which is probably enough for a medium-length syntax article. Using Sprouse and Almeida (*submitted*) as a guideline, numerical rating tasks of this sort should probably test at least 45 participants, for a cost of $225 in the laboratory and $148.50 on Amazon Mechanical Turk. While this is certainly cheap by experimental psychology standards, it is much more expensive than traditional methods (which are free). It also generally takes more time to recruit participants for formal experiments; however Amazon

Mechanical Turk is neutralizing this cost: 80 participants can be collected per hour on Amazon Mechanical Turk (Sprouse 2011a). Finally, the results of this study (and Sprouse and Almeida 2011, *submitted*) suggest that formal experiments may be less powerful than traditional methods unless special care is taken in selecting an appropriately large sample size.

5.5 The generalizability of these results to data in syntactic theory

Because the phenomena that we tested were randomly sampled from LI articles from 2001 through 2010, these results can be generalized to the full population of standard acceptability data points in LI 2001-2010 within a specific margin of error (±5.3-5.8%). Generalizing these results to other journals or other time periods is less straightforward, as other journals and other time periods likely focused on slightly different topics. Nonetheless, we believe that LI 2001-2010 is a good case study for the field given LI's reputation as a top theoretical journal, and given the general trend in scientific fields to discuss clearer data earlier in their history, and subtler data later. Of course, these results do not generalize to the data types that we specifically excluded from the experiments: coreference judgments, interpretation judgments, judgments about individual lexical items, and judgments involving prosody. We chose to exclude these data types because critics have (to our knowledge) focused exclusively on standard acceptability judgments, and we wanted to address their claims directly. However, this decision means that it is logically possible that there is still a reliability problem in syntactic theory, but that it is in the 52% of data that cannot be tested with standard acceptability judgment experiments, rather than the 48% that can be. We are happy to leave this for future research.

6. Conclusion

In an effort to address concerns about the unreliability of syntactic data, we randomly sampled 292 sentence types forming 146 phenomena from articles published in *Linguistic Inquiry* from 2001 through 2010, and tested them using formal acceptability judgment experiments. We tested them first using a magnitude estimation experiment, and found that that 129 phenomena replicated, for a replication rate of 88%. We then tested the replication failures in a forced-choice experiment (because forced-choice experiments have been independently shown to be more sensitive than magnitude estimation experiments (Sprouse and Almeida 2011, *submitted*)), and found 11 additional replications and 1 additional replication failure in the form of a sign-reversal. Taken together, these two experiments suggest that 139 of the phenomena from LI replicate using standard formal acceptability judgment experiments, for a minimum replication rate of 95%. Because we tested a randomly selected sample of all of the US-English acceptability judgments in LI 2001-2010 that are testable using standard acceptability judgment tasks, this 95% replication rate should in principle hold for all such data points in LI 2001-2010 with a margin of error of ±5.3-5.8%.

These results can be combined with the results of Sprouse and Almeida (*to appear*) to provide a comprehensive estimate of the reliability of standard acceptability judgment data in syntax. Sprouse and Almeida (*to appear*) tested all 365 phenomena from a popular syntax textbook (Adger 2003) and derived a minimum replication rate of 98%. We tested a random sample of 146 phenomena from LI, and derived a minimum replication rate of 95%. To the extent that these results cover all of the standard acceptability judgment data published in syntax (but crucially not the data types that require other methodologies), we can conclude that there is

no evidence of a reliability problem for acceptability judgment data in syntax. Furthermore, we can refine the quantitative criteria for critics who wish to continue to argue that there is a reliability problem: they must first state the false positive rate that they assume is indicative of a reliability problem (e.g., 10% in Gibson and Fedorenko 2010b), and then present a set of false positives that would lead to that rate relative to the 498 phenomena that have been shown to clearly replicate using formal experiments (e.g., 50 phenomena for a 10% false positive rate).

The practical ramifications of these results are relatively straightforward. Some critics have argued for the universal adoption of formal experiments with numerical rating tasks; this would clearly be a mistake, at least in the case of magnitude estimation. Our results (and those of Sprouse and Almeida *submitted*) suggest that magnitude estimation experiments are potentially less powerful than traditional methods. However, it would also be a mistake to universally adopt traditional methods over formal experiments. Several studies, including several by the first author, have clearly demonstrated that formal experiments are a useful tool for some syntactic questions (e.g., Keller 2000, Feathers 2005, Sprouse et al. 2011, Sprouse et al. 2012). Instead, we suggest that syntacticians abandon the idea that there is a single method for every research question (or research environment). Science is not a recipe. Syntacticians need to evaluate each methodology based on its costs and benefits to decide which method is most appropriate for their specific research question.

References

ADGER, DAVID. (2003). *Core Syntax: A Minimalist Approach*. Oxford University Press.

ALEXOPOULOU, THEODORA, and FRANK KELLER. 2007. Locality, Cyclicity and Resumption: At the Interface between the Grammar and the Human Sentence Processor. *Language* 83.110–160.

BALLUERKA, NEKANE; JUANA GOMÉZ; and DOLORES HIDALGO. 2005. Null hypothesis significance testing revisited. *Methodology* 1.55–70 .

BARD, ELLEN GURMAN; DAN ROBERTSON; and ANTONELLA SORACE. 1996. Magnitude estimation of linguistic acceptability. *Language* 72.32–68.

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

CHOMSKY, NOAM. 1986. *Barriers*. Cambridge, MA: The MIT Press.

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

CULICOVER, PETER. W., and RAY JACKENDOFF. 2010. Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences* 14.234–235.

DĄBROWSKA, EWA. (2010). Naïve v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27.1–23.

EDELMAN, SHIMON, and MORTEN CHRISTIANSEN. 2003. How seriously should we take Minimalist syntax? *Trends in Cognitive Sciences* 7.60–61.

EDGINGTON, EUGENE, and PATRICK ONGHENA. 2007. *Randomization tests* (4th ed.). Boca Raton, FL: Chapman and Hall/CRC.

FEATHERSTON, SAM. 2005a. Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua* 115.1525–1550.

FEATHERSTON, SAM. 2005b. Universals and grammaticality: Wh-constraints in German and English. *Linguistics* 43.667–711.

FEATHERSTON, SAM. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33.269–318.

FERREIRA, FERNANDA. 2005. Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review* 22.365–380.

GALLISTEL, RANDY. 2009. The importance of proving the null. *Psychological Review* 116.439–53.

GIBSON, EDWARD. 1991. *A computational theory of human linguistic processing: Memory limitations and processing breakdown.* Pittsburgh, PA: Carnegie Mellon University dissertation.

GIBSON, EDWARD, and EVELINA FEDORENKO. 2010a. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14.233–234.

GIBSON, EDWARD, and EVELINA FEDORENKO. 2010b. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes.*

GIBSON, EDWARD; STEVE PIANTADOSI; and KRISTINA FEDORENKO. 2011. Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass* 5.509–524.

HUBBARD, RAYMOND, and R. MURRAY LINDSAY. 2008. Why *p* values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology* 18.69–88.

IPEIROTIS, PANOS. G. 2010. Demographics of Mechanical Turk. Center for Digital Economy Research Working Papers 10. Available at http://hdl.handle.net/2451/29585

JEFFREYS, HAROLD. 1961. *Theory of Probability*. Oxford University Press.

KAYNE, RICHARD. (1983). Connectedness. *Linguistic Inquiry* 14.223–249.

KELLER, FRANK. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Edinburgh: University of Edinburgh dissertation.

MARANTZ, ALEC. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* 22.429–445.

MYERS, JAMES. 2009) Syntactic judgment experiments. *Language and Linguistics Compass* 3.406–423.

NEWMEYER, FREDERICK J. 2007. Commentary on Sam Featherston, 'Data in generative grammar: The stick and the carrot.' *Theoretical Linguistics* 33.395–399.

NICKERSON, RAYMOND. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* 5.241–301.

ROUDER, JEFFREY N.; PAUL L. SPECKMAN; DONGCHU SUN; RICHARD D. MOREY; and GEOFFREY IVERSON. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* 16.225–237.

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

SCHÜTZE, CARSON T., and JON SPROUSE. 2011. Judgment Data. *Research Methods in Linguistics*, ed. by Devyani Sharma and Rob Podesva. Cambridge University Press.

SPROUSE, JON. 2008. The differential sensitivity of acceptability to processing effects. *Linguistic Inquiry* 39.686–694.

SPROUSE, JON. 2009. Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*. 40.329–341.

SPROUSE, JON. (2011a). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43.155–167.

SPROUSE, JON. 2011b. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87.274–288.

SPROUSE, JON. and DIOGO ALMEIDA. (2011). The role of experimental syntax in an integrated cognitive science of language. *The Cambridge Handbook of Biolinguistics*, ed. by Kleanthes Grohmann and Cedric Boeckx.

SPROUSE, JON. and DIOGO ALMEIDA. (to appear). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*.

SPROUSE, JON. and DIOGO ALMEIDA. (*submitted*). Power in acceptability judgment experiments and the reliability of syntactic data.

SPROUSE, JON. and DIOGO ALMEIDA. (*in prep*). The need for careful discussions of experimental methods in syntax.

SPROUSE, JON; SHIN FUKUDA; HAJIME ONO; and ROBERT KLUENDER. 2011. Grammatical operations, parsing processes, and the nature of wh-dependencies in English and Japanese. *Syntax* 14.179–203.

SPROUSE, JON, MATT WAGERS, and COLIN PHILLIPS, C. 2012. A test of the relation between working memory capacity and island effects. *Language*.

STEVENS, STANLEY SMITH. 1956. The direct estimation of sensory magnitudes: loudness. *The American journal of psychology* 69.1–25.

WASOW, THOMAS, and JENNIFER ARNOLD. (2005). Intuitions in linguistic argumentation. *Lingua* 115.1481–1496.

WESKOTT, THOMAS, and GISBERT FANSELOW. 2011. On the Informativity of Different Measures of Linguistic Acceptability. *Language* 87.249–273.

WETZELS, RUUD; DORA MATZKE; MICHAEL D. LEE; JEFFREY N. ROUDER; GEOFFREY J. IVERSON; and ERIC-JAN WAGENMAKERS. 2011. Statistical evidence in experimental psychology: An empirical comparison using 855 t-tests. *Perspectives on Psychological Science* 6.291–298.

Appendix A: Example materials and results from Experiment 1 (magnitude estimation). Identifier is in the format VOLUME.ISSUE.AUTHOR.EXAMPLE.JUDGMENT, Mean reports the mean rating in z-units, *p* reports the *p*-value obtained from the two-tailed paired *t*-test, and BF reports the non-directional Bayes factor.

Identifier	Example	Mean	*p*	BF
32.1.martin.28b.??	Sarah convinced Bill that he would have gone to the party by the time he goes to bed this evening.	-0.06	**.195**	0.24
32.1.martin.27b.g	Sarah convinced Bill that he will have gone to the party by the time he goes to bed this evening.	0.12		
32.1.martin.79.*	How likely to be a riot is there?	-0.49	.002	12
32.1.martin.77.g	How likely to win the race is John?	-0.15		
32.1.martin.65b.*	John believes without a doubt his team will win.	0.68	.002	13
32.1.martin.65a.g	John believes without a doubt that his team will win.	1.10		
32.1.martin.26b.??	Sarah convinced Bill to have gone to the party.	-0.50	.001	>100
32.1.martin.25b.g	Sarah convinced Bill that he would go to the party.	0.34		
32.1.martin.69b.*	My belief Kim is clever is sincere.	-0.15	.001	>100
32.1.martin.69a.g	My belief that Kim is clever is sincere.	0.67		
32.1.martin.93b.*	John is illegal to park here.	-0.71	.001	>100
32.1.martin.92b.g	John is believed to have parked here.	0.79		
32.1.martin.39a.*	Gino believed Rebecca to win the game.	-0.53	.001	>100
32.1.martin.23a.g	Gino believed Rebecca to be the best.	0.74		
32.1.martin.66b.*	It is illegal one to criticize the government.	-0.65	.001	>100
32.1.martin.66a.g	It is illegal for one to criticize the government.	0.96		
32.1.martin.26a.??	Ginny remembered to have bought the beer.	-0.30	.001	>100
32.1.martin.22a.g	Ginny remembered to bring the beer.	1.27		
32.1.martin.20a.*	He seems to that Kim solved the problem.	-1.10	.001	>100
32.1.martin.20a.g	It seems to him that Kim solved the problem.	0.71		
32.1.martin.2c.*	Sarah saw pictures of.	-1.01	.001	>100
32.1.martin.1a.g	Kerry attempted to study physics.	1.21		
32.2.alexiadou.31a.*	"Don't touch that dial!" suggested abruptly the TV screen.	-0.11	.001	>100
32.2.alexiadou.31b.g	"Don't touch that dial!" suggested the TV screen abruptly.	0.54		
32.2.boeckx.11.*	Debbie ate chocolate, and Kathy milk drank.	-0.66	.001	>100
32.2.boeckx.11.g	Debbie ate chocolate, and Kathy drank milk.	1.06		
32.2.nunes.48b.*	Mary drove Rio and John flew to Sao Paulo.	0.01	.001	>100

32.2.nunes.48b.g	Mary drove to Rio and John flew to Sao Paulo.	0.95		
32.2.nunes.3b.*	Was kissed John.	-1.06	.001	>100
32.2.nunes.3a.g	John was kissed.	1.23		
32.2.nunes.3c.*	John was kissed John.	-1.36	.001	>100
32.2.nunes.3a.g	John was kissed.	1.30		
32.2.stroik.4b.*	Max may have been studying, but Jason may have done so too.	-0.15	.006	4.8
32.2.stroik.4a.g	Max may have been studying, but Jason may have been doing so too.	0.18		
32.2.stroik.13b.*	They all have left and they have done all so deliberately.	-0.20	.001	>100
32.2.stroik.13a.g	They all have left and they have all done so deliberately.	0.35		
32.2.stroik.17a.*	Chris is happy, and Pat does so too.	-0.85	.001	>100
32.2.stroik.17a.g	Chris is happy, and Pat is too.	0.85		
32.3.Culicover.32a.*	John's promise to Susan to take care of herself.	-0.71	**.371**	0.16
32.3.Culicover.32a.g	John's promise to Susan to take care of himself.	-0.81		
32.3.Culicover.49a.*	Jack asked Sally to be allowed to take care of herself.	-0.40	**.121**	0.35
32.3.Culicover.49a.g	Jack asked Sally to be allowed to take care of himself.	-0.18		
32.3.Culicover.22b7.*	John told Sue when to wash himself.	-0.21	.017	1.8
32.3.Culicover.22b7.g	John told Sue when to wash herself.	0.15		
32.3.Culicover.15bii.*	John flattered Mary while insulting herself.	-0.27	.001	77
32.3.Culicover.15bii.g	John flattered Mary while insulting himself.	0.33		
32.3.Culicover.41b.*	Toby said to Sally to take care of himself.	-0.33	.001	>100
32.3.Culicover.41b.g	Toby said to Sally to take care of herself.	0.44		
32.3.Culicover.25d.*	Last night there was an attempt to shoot oneself.	-0.52	.001	>100
32.3.Culicover.25d.g	Last night there was an attempt to shoot me.	0.58		
32.3.Culicover.28c.*	Helen examined Bernie in order for us to vindicate herself.	-0.62	.001	>100
32.3.Culicover.28c.g	Helen examined Bernie in order for us to vindicate ourselves.	0.74		
32.3.Culicover.7b.*	John tried himself to win.	-0.45	.001	>100
32.3.Culicover.7a.g	John tried to win.	1.24		
32.3.fanselow.58d.*	There has been considered a man sick.	-1.01	.001	59
32.3.fanselow.58c.g	There has been a man considered sick.	-0.53		
32.3.fanselow.58b.*	There has been shot a moose in the woods.	-0.22	.001	>100
32.3.fanselow.58a.g	There has been a moose shot in the woods.	0.88		

32.3.fanselow.28b.*	He saw Mary and kissed.	-0.71	.001	>100
32.3.fanselow.28b.g	He saw Mary and kissed her.	0.87		
32.4.lopez.10a.*	We proclaimed to the public John to be a hero.	-0.34	**.154**	0.29
32.4.lopez.9a.g	We proclaimed John to the public to be a hero.	-0.17		
32.4.lopez.14b.*	I expected there three men.	-0.62	.001	>100
32.4.lopez.14b.g	I expected there to be three men.	0.62		
33.1.denDikken.5b.*	I know who the hell would buy that book.	-0.01	.001	>100
33.1.denDikken.5a.g	I know who would buy that book.	0.72		
33.1.denDikken.58a.*	What under no circumstances should he do?	-0.63	.001	>100
33.1.denDikken.58a.g	Under no circumstances should he leave.	0.86		
33.1.denDikken.71a.*	Who is in love with who the hell?	-0.95	.001	>100
33.1.denDikken.67.g	Who the hell is in love with who?	0.37		
33.1.denDikken.62b.*	I don't think that any linguists, I will invite to the party.	-1.04	.001	>100
33.1.denDikken.62a.g	I don't think that I will invite any linguists to the party.	0.67		
33.1.fox.69a.*	John wants for everyone you do to have fun.	-0.85	**.401**	0.15
33.1.fox.69b.g	John wants for everyone to have fun that you do.	-0.93		
33.1.fox.49c.*	I visited a city near the city yesterday that John did.	-0.51	.003	7.5
33.1.fox.49b.g	I visited a city yesterday near the city that John did.	-0.16		
33.1.fox.65b.*	I told you that Bill when we met will come to the party.	-0.52	.001	>100
33.1.fox.65b.g	I told you when we met that Bill will come to the party.	0.62		
33.2.bowers.7b.i.*	The ball perfectly rolled down the hill.	0.57	.013	2.3
33.2.bowers.7b.i.g	The ball rolled perfectly down the hill.	0.94		
33.2.bowers.31c.*	There might mice seem to be in the cupboard.	-0.99	.001	>100
33.2.bowers.31a.g	There might seem to be mice in the cupboard.	-0.04		
33.2.bowers.69b.*	The bureaucrat bribes deliberately.	-0.26	.001	>100
33.2.bowers.69a.g	The bureaucrat was bribed deliberately.	1.00		
33.2.bowers.68b.*	The politician bribes easily to avoid the draft.	-0.27	.001	>100
33.2.bowers.68a.g	The politician was bribed to avoid the draft.	0.93		
33.2.bowers.31b.*	There seem mice to be in the cupboard.	-0.57	.001	>100
33.2.bowers.31a.g	There seem to be mice in the cupboard.	0.99		
33.2.bowers.13a.*	John believes to be sick.	-0.55	.001	>100

33.2.bowers.13a.g	John believes Mary to be sick.	0.88		
33.3.boskovic.48d.*	The was arrested student.	-1.35	.001	>100
33.3.boskovic.48a.g	The student was arrested.	0.97		
33.4.neeleman.100.*	Yesterday seemed that John left.	-0.53	.001	>100
33.4.neeleman.100.g	It seemed that yesterday John left.	0.65		
33.4.neeleman.97b.*	Which book did you sleep before reading?	-0.91	.001	>100
33.4.neeleman.97a.g	Which book did you file before reading?	0.29		
33.4.neeleman.35a.*	What did John wonder what he bought?	-0.71	.001	>100
33.4.neeleman.35a.g	John wondered what he bought.	0.74		
33.4.neeleman.24d.*	Anyone better leave town.	-0.70	.001	>100
33.4.neeleman.24d.g	Someone better leave town.	0.78		
33.4.neeleman.18d.*	Deciding who to see that new movie next makes very happy.	-0.82	.001	>100
33.4.neeleman.18c.g	Deciding which movie to see next makes John very happy.	0.90		
34.1.basilico.96a.??	The children almost all are sleeping.	-0.14	.001	>100
34.1.basilico.96b.??	The children are almost all sleeping.	0.79		
34.1.basilico.62.*	There are linguists tall.	-0.77	.001	>100
34.1.basilico.62.g	There are linguists available.	0.33		
34.1.basilico.44b.*	Who was seen steal the wallet?	-0.86	.001	>100
34.1.basilico.44a.?	Who did you see steal the wallet?	0.53		
34.1.fox.28.*	They said they heard about a Balkan language, but I don't know which Balkan language they did.	0.14	.005	5.2
34.1.fox.27.g	They said they heard about a Balkan language, but I don't know which Balkan language.	0.69		
34.1.fox.26.*	She said that a biography of one of the Marx brothers is going to be published this year, but I don't remember which she did.	-0.15	.001	>100
34.1.fox.23.g	She said that a biography of one of the Marx brothers is going to be published this year, but I don't remember which.	0.84		
34.1.fox.24.*	It appears that a certain senator will resign, but which senator it does is still a secret.	-0.53	.001	>100
34.1.fox.19.g	It appears that a certain senator will resign, but which senator is still a secret.	0.44		
34.1.fox.14.*	What do you worry if the lawyer forgets at the office?	-0.86	.001	>100
34.1.fox.14.g	What do you think that the lawyer forgot at the office?	0.13		

34.1.phillips.96a.*	John intended to give the children something nice to eat, and give the children he did a generous handful of candy.	-0.73	.001	46
34.1.phillips.96a.g	John intended to give the children something nice to eat, and give the children a generous handful of candy he did.	-0.39		
34.1.phillips.93b.?*	Wendy stood more buckets in the garage than Peter did in the basement.	0.48	.001	100
34.1.phillips.92b.g	Wendy stood more buckets than Peter did in the garage.	-0.14		
34.1.phillips.88b.*	John promised Mary to leave, and Sue did to write more poetry.	-0.58	.001	>100
34.1.phillips.88b.g	John promised Mary to leave, and Sue promised to write more poetry.	0.67		
34.1.phillips.67d.*	I gave anything to nobody.	-1.02	.001	>100
34.1.phillips.67c.g	I gave nothing to anybody.	0.06		
34.1.phillips.59b.*	The students were punished and their teachers by their parents.	-0.68	.001	>100
34.1.phillips.59b.g	The students were punished by their parents and their teachers.	0.94		
34.1.phillips.6b.*	Wallace gave at breakfast time his favorite pet beagle an enormous chewy dog-biscuit.	-0.83	.001	>100
34.1.phillips.6b.g	Wallace gave his favorite pet beagle an enormous chewy dog-biscuit at breakfast time.	0.93		
34.1.phillips.3e.*	Each other like Wallace and Greg.	-0.96	.001	>100
34.1.phillips.3d.g	Wallace and Greg like each other.	1.26		
34.2.caponigro.13b.*	The flute was being shiny.	-0.54	.001	>100
34.2.caponigro.13a.g	The flute was being played by the soloist.	1.13		
34.2.panagiotidis.6.*	We students of physics are taller than you of chemistry.	-0.36	.049	0.71
34.2.panagiotidis.6.g	We students of physics are taller than you students of chemistry.	-0.16		
34.3.heycock.82a.*	The dog that I saw's collar was leather.	-0.33	.001	69
34.3.heycock.82a.g	The collar of the dog that I saw was leather.	0.32		
34.3.heycock.37b.??	Knife with the golden blade and fork with the silver handle go on the left.	-0.27	.001	>100
34.3.heycock.37b.g	The knife with the golden blade and the fork with the silver handle go on the left.	0.44		
34.3.heycock.55a.*	Fork is silver-plated and bowl is enameled.	-0.33	.001	>100
34.3.heycock.55a.g	The fork is silver-plated and the bowl is enameled.	0.89		

34.3.heycock.30c.?*	Cat and dog that were fighting all the time had to be separated.	-0.39	.001	>100
34.3.heycock.30c.g	The cat and dog that were fighting all the time had to be separated.	0.50		
34.3.heycock.16.*	He was judge.	-0.22	.001	>100
34.3.heycock.16.g	He was the judge.	1.34		
34.3.heycock.66.*	This is table.	-0.82	.001	>100
34.3.heycock.66.g	This is a table.	1.36		
34.3.landau.39b.*	One interpreter each tried to be assigned to every visiting diplomat.	-0.40	.001	>100
34.3.landau.39a.g	One interpreter tried to be assigned to every visiting diplomat.	0.70		
34.3.landau.32c.*	There expects to be a man in the garden.	-0.72	.001	>100
34.3.landau.32c.g	There seems to be a man in the garden.	1.06		
34.3.takano.10b.*	I bought any books only occasionally.	-0.77	.001	>100
34.3.takano.10b.g	I only occasionally bought any books.	-0.11		
34.3.takano.9e.*	Anything has nobody done.	-1.22	.001	>100
34.3.takano.9e.g	Nobody has done anything.	0.50		
34.4.boskovic.3d.*	Mary believed Peter finished school and Bill Peter got a job.	-0.50	**.720**	0.11
34.4.boskovic.4d.g	Mary believed that Peter finished school and Bill that Peter got a job.	-0.45		
34.4.boskovic.7a.??	What did they believe at that time that Peter fixed?	-0.09	**.494**	0.13
34.4.boskovic.7c.g	At that time, what did they believe that Peter fixed?	-0.01		
34.4.boskovic.3c.*	They suspected and we believed Peter would visit the hospital.	-0.42	**.273**	0.19
34.4.boskovic.4c.g	They suspected and we believed Peter would visit the hospital.	-0.31		
34.4.boskovic.3e.*	John likes Mary, Jane didn't believe	-0.67	**.246**	0.2
34.4.boskovic.4e.g	That John likes Mary, Jane didn't believe	-0.57		
34.4.boskovic.3b.*	What the students believe is they will pass the exam.	-0.19	.038	0.9
34.4.boskovic.4b.g	What the students believe is that they will pass the exam.	0.10		
34.4.boskovic.3a.*	It seemed at that time David had left.	0.25	.001	>100
34.4.boskovic.4a.g	It seemed at that time that David had left.	0.68		
34.4.haegeman.2a.*	This is the man who I think that will buy your house next year.	-0.04	.004	6.7

34.4.haegeman.2a.g	This is the man who I think will buy your house next year.	0.40		
34.4.lasnik.10a.*	Angela wondered how John managed to cook, but it's not clear what food.	-0.28	.001	49
34.4.lasnik.11a.g	Angela wondered how John managed to cook a certain food, but it's not clear what food.	0.18		
35.1.beck.12b.*	Who did you believe a friend of satisfied?	-0.54	.001	>100
35.1.beck.12b.g	I believed a friend of Andy satisfied.	0.25		
35.1.bhatt.94a.*	I expect that everyone you do will visit Mary.	-0.88	**.606**	0.12
35.1.bhatt.94b.g	I expect that everyone will visit Mary that you do.	-0.84		
35.1.bhatt.14a.*	Ralph is more than fit tall.	-0.79	.001	>100
35.1.bhatt.14cf.g	Ralph is more tall than fit.	0.76		
35.1.bhatt.76b.*	I told you that Bill when we met will come to the party.	-0.44	.001	>100
35.1.bhatt.76b.g	I told you when we met that Bill will come to the party.	0.87		
35.1.mcginnis.32b.*	I ran Mary.	-1.00	.001	>100
35.1.mcginnis.32b.g	I ran for Mary.	1.07		
35.2.hazout.6b.*	I find it irritating for usually this street to be closed.	-0.79	.001	>100
35.2.hazout.6a.g	I find it irritating that usually this street is closed.	0.35		
35.2.larson.44a.??	A taller man than my father walked in.	0.11	**.076**	0.5
35.2.larson.44a.g	A man taller than my father walked in.	0.37		
35.2.larson.44c.??	Max talked to as tall a man as his father.	-0.54	.001	>100
35.2.larson.44c.g	Max talked to a man as tall as his father.	0.34		
35.3.embick.13b.*	Mary pounded the apple flattened.	-0.59	.001	>100
35.3.embick.13b.g	Mary pounded the apple flat.	0.82		
35.3.hazout.67c.*	There is likely a man to appear.	-0.32	.016	1.8
35.3.hazout.67a.g	There is likely to appear a man.	-0.60		
35.3.hazout.75a.*	It is unimaginable Mary to arrive on time.	-0.72	.001	>100
35.3.hazout.75a.g	It is unimaginable for Mary to arrive on time.	0.60		
35.3.hazout.63.*	It seems a man to be in the room.	-0.90	.001	>100
35.3.hazout.60b.g	It seems a man is in the room.	0.68		
35.3.richards.17b.*	To whom did you give what?	0.12	**.828**	0.11
35.3.richards.17a.g	What did you give to whom?	0.15		
35.3.sobin.3c.*	Some frogs and a fish is in the pond.	-0.33	.001	>100
35.3.sobin.3c.g	Some frogs and a fish are in the pond.	0.61		
36.4.denDikken.45.*	That much the less you say, the smarter you will	-0.60	.001	>100

	seem.			
36.4.denDikken.45.g	The less you say, the smarter you will seem.	0.98		
37.2.de Vries.39a.*	I talked to with whom you danced yesterday.	-0.94	.001	>100
37.2.de Vries.39b.g	I talked to Mary, with whom you danced yesterday.	0.06		
37.2.Sigurdsson.3d.*	Me would have been elected.	-0.89	.001	>100
37.2.Sigurdsson.2a.g	I would have been elected.	1.14		
37.3.becker.26b.*	I seem eating sushi.	-0.73	.001	>100
37.3.becker.26a.g	I hate eating sushi.	1.36		
37.3.becker.2b.*	There like to be storms at this time of year.	-0.87	.001	>100
37.3.becker.2a.g	There tend to be storms at this time of year.	0.66		
37.3.becker.5b.*	I seem eating sushi.	-0.83	.001	>100
37.3.becker.5a.g	I like eating sushi.	1.26		
37.4.nakajima.20e.*	He existed a dangerous existence.	-0.79	.001	>100
37.4.nakajima.4a.g	The tree grew a century's growth within only ten years.	0.06		
38.2.hornstein.4c.*	Into which room did walk three men?	-0.68	.017	1.8
38.2.hornstein.4b.g	Into which room walked three men?	-0.31		
38.2.hornstein.4e.*	Into which room three men walked?	-0.34	.001	>100
38.2.hornstein.4d.g	Into which room did three men walk?	0.27		
38.2.hornstein.3c.*	How many books there were on the table?	-0.21	.001	>100
38.2.hornstein.3c.g	How many books were there on the table?	0.78		
38.3.haddican.39.*	Blake said that he would beard his tormentor before the night was up, but the actual doing of so proved rather difficult.	-0.29	**.066**	0.56
38.3.haddican.39.g	Blake said that he would beard his tormentor before the night was up, but the actual doing of it proved rather difficult.	-0.04		
38.3.hirose.4a.*	It will take from three five days for him to recover.	0.04	.001	>100
38.3.hirose.3a.g	It will take three to five days for him to recover.	1.30		
38.3.hirose.1b.*	To Mary for Bill I gave a book.	-0.94	.001	>100
38.3.hirose.1a.g	From Alabama to Louisiana John played the banjo.	0.90		
38.3.landau.39a.*	Who did George kick the ball?	-0.41	.001	>100
38.3.landau.38a.g	George kicked the boy the ball.	0.57		
38.3.landau.31b.*	An hour, they slept, and then went to work.	-0.13	.001	>100
38.3.landau.31a.g	They slept an hour and then went to work.	1.18		
38.4.Boskovic.4.*	There seems a man to be in the garden.	-0.01	.001	>100

38.4.Boskovic.17a.g	There seems to be a man in the garden.	1.12		
38.4.kallulli.10b.*	The ship sank deliberately.	-0.05	.001	>100
38.4.kallulli.10a.g	The ship was sunk deliberately.	0.92		
38.4.kallulli.9b.*	The boat sank to collect the insurance.	-0.29	.001	>100
38.4.kallulli.9a.g	The boat was sunk to collect the insurance.	0.97		
38.4.kallulli.4b.*	Eva was killed from John.	-0.24	.001	>100
38.4.kallulli.4b.g	Eva was killed by John.	1.34		
39.1.sobin.20c.*	John broke a cup, and Mary did so with a saucer.	-0.36	.001	>100
39.1.sobin.21c.g	John broke a cup, and Mary did so too.	0.38		
40.1.caponigro.25b.*	Lily will dance who the king chooses.	-0.85	.001	>100
40.1.caponigro.25b.g	Lily will dance with the person the king chooses.	0.82		
40.1.caponigro.23a.*	Jack came the person he is in love with.	-0.95	.001	>100
40.1.caponigro.23cf.g	Jack came with the person he is in love with.	1.41		
40.1.heck.5b.*	Sherry met a man very fond of whom she found herself.	-0.78	.001	>100
40.1.heck.5b.g	Sherry met a man who she found herself very fond of.	0.40		
40.1.stepanov.4b.*	What did who buy?	-0.38	.001	>100
40.1.stepanov.4a.g	Who bought what?	1.03		
40.2.johnson.59b.*	Ice cream gives me in the morning brain-freeze.	-0.72	.001	>100
40.2.johnson.59b.g	Ice cream gives me brain-freeze in the morning.	0.64		
40.4.hicks.23.*	Lloyd Webber musicals are likely to be condemned without anyone even watching	-0.42	.001	>100
40.4.hicks.22.g	Lloyd Webber musicals are easy to condemn without even watching	0.46		
41.1.Muller.14c.*	Who did that Mary was going out with bother you?	-1.00	.001	>100
41.1.Muller.14c.g	That Mary was going out with Luke bothered you.	-0.22		
41.1.Muller.25b.??(*)	Who do you wonder which picture of is on sale?	-0.93	.001	>100
41.1.Muller.25b.g	You wonder which picture of Marge is on sale.	0.46		
41.2.Bruening.36b.*	The man that he gave the creeps last night to is over there.	-0.54	.001	21
41.2.Bruening.36a.g	The man that he gave the creeps to last night is over there.	-0.13		
41.2.Bruening.31a.*	At that battle were given the generals who lost hell.	-0.75	.001	>100
41.2.Bruening.31a.g	At that battle the generals who lost were given hell.	0.32		
41.2.Bruening.3b.*	The count gives the creeps to me.	-0.30	.001	>100

41.2.Bruening.3a.g	The count gives me the creeps.	0.67		
41.2.Bruening.33a.*	At that time were given the tables we inherited from Aunt Selma a good scrubbing.	-0.94	.001	>100
41.2.Bruening.33a.g	The tables we inherited from Aunt Selma were given a good scrubbing at that time.	0.04		
41.3.Costantini.2b.??	All the men seem to have all eaten supper.	0.59	**.479**	0.13
41.3.Costantini.2b.g	The men seem to have all eaten supper.	0.69		
41.3.Landau.10b.*	The game was played shoeless.	-0.80	**.583**	0.12
41.3.Landau.10a.g	The game was played wearing no shoes.	-0.73		
41.3.Landau.25c.*	I told Mr. Smith that I am able to paint the fence together.	-0.24	.045	0.77
41.3.Landau.24c.g	I told Mr. Smith that I wonder when to paint the fence together.	0.06		
41.3.Landau.27b.*	His wife kissed in front of the kids.	-0.41	.001	>100
41.3.Landau.27b.g	He and his wife kissed in front of the kids.	0.83		
41.3.Landau.7b.*	I am now hiring for John to work with.	-0.77	.001	>100
41.3.Landau.7b.g	I am now hiring people for John to work with.	1.01		
41.3.Rezac.3b2.*	There had all hung over the fireplace the portraits by Picasso.	-0.58	.013	2.3
41.3.Rezac.3b1.g	There had hung over the fireplace all of the portraits by Picasso.	-0.35		
41.3.Vicente.4a6.*	Sandy plays the guitar better than Betsy the harmonica.	-0.27	.001	>100
41.3.Vicente.4b.g	She plays the guitar and Betsy the harmonica.	0.70		
41.3.Vicente.5a.*	Amanda went to Santa Cruz, and Bill thinks that Claire to Monterrey.	-0.49	.001	>100
41.3.Vicente.5b.g	Amanda went to Santa Cruz, and Bill thinks that Claire did too.	0.69		
41.3.Vicente.8a.*	Read things, Mike did quickly.	-0.75	.001	>100
41.3.Vicente.8a.g	Mike read things quickly.	0.65		
41.3.Vicente.4a.*	She plays the guitar because Betsy the harmonica.	-0.76	.001	>100
41.3.Vicente.4b.g	Sandy plays the guitar better than Betsy does.	0.70		
41.3.Vicente.8d.*	Want to write, Randy did a novel.	-1.07	.001	>100
41.3.Vicente.8d.g	Randy wanted to write a novel.	1.42		
41.4.Bruening.9b.*	What did he prove an account of false?	-0.60	**.871**	0.11
41.4.Bruening.9c.g	Who did he give statues of to all the season-ticket holders?	-0.61		
41.4.Haegeman.18a.*	Bill asked if such books John only reads at home.	-0.71	**.415**	0.15
41.4.Haegeman.18a.g	Bill knows that such books John only reads at	-0.64		

	home.			
41.4.Haegeman.4a.*	When this column she started to write last year, I thought she would be fine.	-0.79	.001	100
41.4.Haegeman.4c.g	When last year she started to write this column, I thought she would be fine.	-0.20		
41.4.Haegeman.8a.*	We were all much happier when upstairs lived the Browns.	-0.38	.001	>100
41.4.Haegeman.8a.g	Upstairs lived the Browns.	0.46		
41.4.Haegeman.22d.*	If frankly he's unable to cope, we'll have to replace him.	-0.21	.001	>100
41.4.Haegeman.22d.g	If he's unable to cope, we'll have to replace him.	1.15		

Appendix B: Example materials and results from Experiment 2 (forced choice). Identifier is in the format VOLUME.ISSUE.AUTHOR.EXAMPLE.JUDGMENT, Hits reports the number of responses in the correct direction, *p* reports the *p*-value obtained from the two-tailed sign-test, and BF reports the non-directional Bayes factor.

Identifier	Example	Hits	*p*	BF
32.1.martin.28b.??	Sarah convinced Bill that he would have gone to the party by the time he goes to bed this evening.	60	0.01	2.5
32.1.martin.27b.g	Sarah convinced Bill that he will have gone to the party by the time he goes to bed this evening.			
32.1.martin.66b.*	It is illegal one to criticize the government.	92	0.00	>100
32.1.martin.66a.g	It is illegal for one to criticize the government.			
32.1.martin.2c.*	Sarah saw pictures of.	94	0.00	>100
32.1.martin.1a.g	Kerry attempted to study physics.			
32.1.martin.20a.*	He seems to that Kim solved the problem.	95	0.00	>100
32.1.martin.20a.g	It seems to him that Kim solved the problem.			
32.1.martin.26a.??	Ginny remembered to have bought the beer.	96	0.00	>100
32.1.martin.22a.g	Ginny remembered to bring the beer.			
32.2.boeckx.11.*	Debbie ate chocolate, and Kathy milk drank.	95	0.00	>100
32.2.boeckx.11.g	Debbie ate chocolate, and Kathy drank milk.			
32.2.nunes.3b.*	Was kissed John.	95	0.00	>100
32.2.nunes.3a.g	John was kissed.			
32.2.nunes.3c.*	John was kissed John.	96	0.00	>100
32.2.nunes.3a.g	John was kissed.			
32.2.stroik.17a.*	Chris is happy, and Pat does so too.	96	0.00	>100
32.2.stroik.17a.g	Chris is happy, and Pat is too.			
32.3.Culicover.22b7.*	John told Sue when to wash himself.	74	0.00	>100
32.3.Culicover.22b7.g	John told Sue when to wash herself.			
32.3.Culicover.49a.*	Jack asked Sally to be allowed to take care of herself.	79	0.00	>100
32.3.Culicover.49a.g	Jack asked Sally to be allowed to take care of himself.			
32.3.Culicover.32a.*	John's promise to Susan to take care of herself.	82	0.00	>100
32.3.Culicover.32a.g	John's promise to Susan to take care of himself.			
32.4.lopez.10a.*	We proclaimed to the public John to be a hero.	42	**0.13**	0.27
32.4.lopez.9a.g	We proclaimed John to the public to be a hero.			
33.1.denDikken.62b.*	I don↓t think that any linguists, I will invite to the party.	96	0.00	>100
33.1.denDikken.62a.g	I don↓t think that I will invite any linguists to the			

	party.			
33.1.fox.69a.* 33.1.fox.69b.g	John wants for everyone you do to have fun. John wants for everyone to have fun that you do.	42	**0.13**	0.27
33.2.bowers.7b.i.* 33.2.bowers.7b.i.g	The ball perfectly rolled down the hill. The ball rolled perfectly down the hill.	71	0.00	>100
33.3.boskovic.48d.* 33.3.boskovic.48a.g	The was arrested student. The student was arrested.	94	0.00	>100
33.4.neeleman.18d.* 33.4.neeleman.18c.g	Deciding who to see that new movie next makes very happy. Deciding which movie to see next makes John very happy.	96	0.00	>100
34.1.phillips.93b.?* 34.1.phillips.92b.g	Wendy stood more buckets in the garage than Peter did in the basement. Wendy stood more buckets than Peter did in the garage.	10	**0.00**	>100
34.1.phillips.6b.* 34.1.phillips.6b.g	Wallace gave at breakfast time his favorite pet beagle an enormous chewy dog-biscuit. Wallace gave his favorite pet beagle an enormous chewy dog-biscuit at breakfast time.	95	0.00	>100
34.1.phillips.3e.* 34.1.phillips.3d.g	Each other like Wallace and Greg. Wallace and Greg like each other.	96	0.00	>100
34.2.panagiotidis.6.* 34.2.panagiotidis.6.g	We students of physics are taller than you of chemistry. We students of physics are taller than you students of chemistry.	82	0.00	>100
34.3.heycock.66.* 34.3.heycock.66.g	This is table. This is a table.	95	0.00	>100
34.3.landau.32c.* 34.3.landau.32c.g	There expects to be a man in the garden. There seems to be a man in the garden.	95	0.00	>100
34.3.takano.9e.* 34.3.takano.9e.g	Anything has nobody done. Nobody has done anything.	95	0.00	>100
34.4.boskovic.3e.* 34.4.boskovic.4e.g	John likes Mary, Jane didn't believe That John likes Mary, Jane didn't believe	57	0.04	0.68
34.4.boskovic.3d.* 34.4.boskovic.4d.g	Mary believed Peter finished school and Bill Peter got a job. Mary believed that Peter finished school and Bill that Peter got a job.	65	0.00	56
34.4.boskovic.7a.?? 34.4.boskovic.7c.g	What did they believe at that time that Peter fixed? At that time, what did they believe that Peter fixed?	67	0.00	>100

34.4.boskovic.3c.*	They suspected and we believed Peter would visit the hospital.	78	0.00	>100
34.4.boskovic.4c.g	They suspected and we believed Peter would visit the hospital.			
34.4.boskovic.3b.*	What the students believe is they will pass the exam.	78	0.00	>100
34.4.boskovic.4b.g	What the students believe is that they will pass the exam.			
35.1.bhatt.94a.*	I expect that everyone you do will visit Mary.	42	**0.13**	0.27
35.1.bhatt.94b.g	I expect that everyone will visit Mary that you do.			
35.1.mcginnis.32b.*	I ran Mary.	95	0.00	>100
35.1.mcginnis.32b.g	I ran for Mary.			
35.2.larson.44a.??	A taller man than my father walked in.	77	0.00	>100
35.2.larson.44a.g	A man taller than my father walked in.			
35.3.hazout.67c.*	There is likely a man to appear.	25	**0.00**	>100
35.3.hazout.67a.g	There is likely to appear a man.			
35.3.richards.17b.*	To whom did you give what?	49	**0.46**	0.13
35.3.richards.17a.g	What did you give to whom?			
37.2.Sigurdsson.3d.*	Me would have been elected.	96	0.00	>100
37.2.Sigurdsson.2a.g	I would have been elected.			
37.3.becker.2b.*	There like to be storms at this time of year.	95	0.00	>100
37.3.becker.2a.g	There tend to be storms at this time of year.			
37.3.becker.5b.*	I seem eating sushi.	95	0.00	>100
37.3.becker.5a.g	I like eating sushi.			
38.2.hornstein.4c.*	Into which room did walk three men?	78	0.00	>100
38.2.hornstein.4b.g	Into which room walked three men?			
38.3.haddican.39.*	Blake said that he would beard his tormentor before the night was up, but the actual doing of so proved rather difficult.	74	0.00	>100
38.3.haddican.39.g	Blake said that he would beard his tormentor before the night was up, but the actual doing of it proved rather difficult.			
40.1.caponigro.25b.*	Lily will dance who the king chooses.	95	0.00	>100
40.1.caponigro.25b.g	Lily will dance with the person the king chooses.			
40.1.caponigro.23a.*	Jack came the person he is in love with.	95	0.00	>100
40.1.caponigro.23cf.g	Jack came with the person he is in love with.			
41.3.Costantini.2b.??	All the men seem to have all eaten supper.	64	0.00	27
41.3.Costantini.2b.g	The men seem to have all eaten supper.			
41.3.Landau.25c.*	I told Mr. Smith that I am able to paint the fence	57	0.04	0.68

	together.			
41.3.Landau.24c.g	I told Mr. Smith that I wonder when to paint the fence together.			
41.3.Landau.10b.*	The game was played shoeless.	59	0.02	1.6
41.3.Landau.10a.g	The game was played wearing no shoes.			
41.3.Landau.7b.*	I am now hiring for John to work with.	95	0.00	>100
41.3.Landau.7b.g	I am now hiring people for John to work with.			
41.3.Rezac.3b2.*	There had all hung over the fireplace the portraits by Picasso.	74	0.00	>100
41.3.Rezac.3b1.g	There had hung over the fireplace all of the portraits by Picasso.			
41.3.Vicente.8d.*	Want to write, Randy did a novel.	95	0.00	>100
41.3.Vicente.8d.g	Randy wanted to write a novel.			
41.4.Bruening.9b.*	What did he prove an account of false?	56	**0.06**	0.48
41.4.Bruening.9c.g	Who did he give statues of to all the season-ticket holders?			
41.4.Haegeman.18a.*	Bill asked if such books John only reads at home.	61	0.01	4.3
41.4.Haegeman.18a.g	Bill knows that such books John only reads at home.			