The 469 data points that form the empirical foundation of generative syntactic theory are at least 98% replicable using formal experiments.

Jon Sprouse[a]
Diogo Almeida[b]

[a]Department of Cognitive Sciences, University of California, Irvine
[b]Department of Linguistics and Languages, Michigan State University

Abstract

Acceptability judgment collection in the field of generative syntax has generally proceeded informally, that is, without the formal methods familiar from experimental psychology. Two types of arguments have been proposed for the adoption of formal experimental techniques in generative syntax: (i) that formal experiments provide a potentially more sensitive measurement tool, and (ii) that informal techniques are in fact an unreliable measurement tool. While the first is relatively widely accepted, the second has become a matter of considerable debate because it suggests that the data that was used to construct current versions of generative theories are in fact faulty. In order to empirically test this claim, we tested all 469 data points in a popular generative syntax textbook (Adger, 2003) using 440 naïve participants, the magnitude estimation and yes-no tasks, and three different types of statistical analyses (traditional frequentist tests, linear mixed effects models, and Bayes factor analyses). This study suggests that the maximum replication failure rate for the informally reported results is 2%, or put another way, that the empirical foundation of generative syntactic theory is at least 98% replicable with formal experiments. These results suggest that (i) the extensive use of informally collected judgments in generative syntax has not led to theories constructed upon faulty data, and (ii) though there are several reasons for generative syntacticians to adopt formal experimental methods for data collection, the putative inadequacy of the empirical foundation of generative syntactic theories is not one of them.

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

1. Introduction

There are two undisputable facts concerning data collection in the field of generative syntax: (1) Acceptability judgments form a substantial component of the empirical foundation of generative syntax (Chomsky, 1965; Schütze, 1996), and (2) the vast majority of the acceptability judgments that have been reported in the generative syntax literature over the past 50 years were collected informally, that is, without the use of formal collection and statistical analysis techniques that are familiar from experimental psychology. Therefore, it is no surprise that criticisms of the data in generative syntax have been around nearly as long as the field itself (e.g., Hill, 1961; Spencer, 1973). Though there has been a steady stream of methodological discussions over the years, the past fifteen years in particular have seen a dramatic increase in the number of papers devoted to

the topic of acceptability judgment collection (Bard et al., 1996; Keller, 2000, 2003; Edelman and Christiansen, 2003; Phillips and Lasnik, 2003, Featherston, 2005a, 2005b, 2007, 2008, 2009; Ferreira, 2005; Sorace and Keller, 2005; Wasow and Arnold, 2005; den Dikken et al., 2007; Alexopoulou and Keller, 2007; Fanselow, 2007; Newmeyer, 2007; Culbertson and Gross, 2009; Myers, 2009; Phillips 2009; Bader and Häussler, 2010; Dabrowska, 2010; Gibson and Fedorenko 2010a, 2010b; Fedorenko and Gibson, 2010, Culicover and Jackendoff, 2010; Weskott and Fanselow *to appear*). If there is a consensus to be drawn from this literature, it is that everyone (including the authors; see Sprouse, 2007a, 2007b, 2008, 2009, 2011, *submitted*; Sprouse et al., 2011; Sprouse et al., *submitted*; Fukuda and Sprouse, *submitted*) agrees that it would be a positive development if generative syntacticians started using formal experimental techniques when collecting acceptability judgments. The consequences of this consensus can be seen in the fact that graduate programs in generative syntax have begun to incorporate formal experimental methods for judgment collection as part of their curricula, and that several recent conferences have devoted special sessions to experimental issues in generative syntax. The popularity of formal experiments in the field has even led to the adoption of a concise term for the relevant methods – *experimental syntax* (Cowart 1997), which we will adopt throughout this article.

Given that the widespread adoption of experimental syntax appears to be well underway, we will not focus on the question of whether or not generative syntacticians should adopt experimental syntax methods, but rather the reasons that have been offered in the literature for their adoption. Though the specific reasons that have been offered are numerous, it seems that they all can be reducible to two basic types:

1. Claims that experimental syntax yields superior data to the classic informal methods. (Bard et al., 1996; Keller, 2000, 2003; Featherston, 2005a, 2005b, 2008, 2009; Sorace and Keller, 2005; Myers, 2009; Sprouse, 2007a, 2007b, 2008, 2009, 2011, *submitted*; Sprouse et al., 2011; Sprouse et al., *submitted*; Fukuda and Sprouse, *submitted*)

2. Claims that informal methods are inherently unreliable and might routinely yield flawed data. (Edelman and Christiansen, 2003; Ferreira, 2005; Wasow and Arnold, 2005; Gibson and Fedorenko 2010a, 2010b)

While these two claims appear interrelated, they are in fact logically separable, and have dramatically different consequences for existing generative syntactic theories. This becomes clear if the claims are recast in terms of false negatives (the failure to detect a true effect) and false positives (the incorrect conclusion that there is an effect when none exists). Claims of type 1 suggest that the use of informal methods has generated false negatives, and that experimental syntax may reveal the unobserved true positives that the false negatives have obscured, thereby allowing researchers to explore previously unavailable data. In essence, claims of type 1 hold that experimental syntax offers a potentially more sensitive measurement device than informal methods and that experimental syntax may enrich the empirical bases of syntactic theories. Claims of type 2 suggest instead that the use of informal methods has resulted in false positives, and that experimental syntax is necessary to correct the record. In essence, claims of type 2 hold that informal methods are an unreliable measurement device, and its extensive adoption has resulted in unreliable syntactic theories. Framed in this way, it is easy to see that the two claims have drastically different consequences for the status of the existing syntactic theories that were constructed using informally collected judgments.

Claims of type 1 are relatively widely accepted (see especially Featherston, 2005a, 2005b; Myers, 2009; and Sprouse et al. *submitted*), though there is some mild concern that some of the techniques may have been oversold (Grewendorf, 2007; Haider, 2007; Newmeyer, 2007; Phillips, 2009; Bader and Häussler 2010; Weskott and Fanselow, *to appear*; Sprouse *submitted*). Claims of type 2, on the other hand, remain controversial. To date, the critics of informal methods that have endorsed claims of type 2 have relied on proof-of-concept arguments, demonstrating that there are at least a few informal judgments that have not (yet) been replicated using formal experiments. The question raised by such results is how representative these replication failures are of the broader field. Though difficult to answer, this is a straightforward empirical question: What is the replication failure rate for informal judgments in generative syntax? The existing proof-of-concept studies unfortunately cannot answer this question, since the examples of replication failures that have been previously published were admittedly not chosen randomly from the entire population of data points in generative syntax. Gibson and Fedorenko (2010b) is perhaps the most recent example: of the 7 sentence types reported in the paper, two were previously reported (Clifton et al., 2006), two were selected from the first author's dissertation (Gibson, 1991) and three were statistically analyzed as a unitary phenomenon despite not being presented as a unitary phenomenon by the original author (Chomsky, 1986). Because the selection of data points in these types of proof-of-concept studies was admittedly biased, the results of such studies are unlikely to convince generative syntacticians that the data supporting existing theories is faulty. Of course, the converse is also true: a biased selection of 7 sentence types from the existing literature that can be conclusively replicated in formal experiments would be equally unlikely to convince critics that existing syntactic theories are built on solid empirical foundations. In short, it seems clear that this debate requires a large, systematic investigation to overcome the prior beliefs held by the two sides.

As a first step toward resolving this issue, we decided to test all 469 sentence types that are presented as data in the popular introductory syntax textbook: *Core Syntax: A Minimalist Approach* (2003) by David Adger (published by Oxford University Press). There were three reasons for choosing the data from Adger (2003). First, the number of data points in the textbook (469) is two orders of magnitude larger than the number of data points tested in previous experimental syntax studies. It is our hope that this number is large enough to render extrapolation from the results of the experiment a less contentious matter. Second, exhaustively testing the data presented in an introductory textbook is a solid, yet tractable, step toward overcoming the selection bias of previous studies. The ideal solution to the selection bias problem would be to randomly select data points from the complete population of data points presented in the generative syntax literature; however, the collection of an exhaustive list of every data point presented during the past 50 years of generative syntax research is beyond our resources. The exhaustive testing of 469 data points that span the wide range of topics covered by an introductory syntax textbook offers a good compromise, as the only possible selection bias is in the choice of the population of data points (the data presented in Adger, 2003), not the sample. Finally, Adger (2003) is a popular textbook among adherents of the research program within generative syntactic theory known as the Minimalist Program. It stands to reason that minimalist syntacticians will agree that the data points in Adger (2003) accurately represent the empirical foundation of the Minimalist Program (at least as much as any other cohesive set of data points). Since a substantial number of criticisms have been levied at informal judgments and the Minimalist Program simultaneously (e.g., Edelman and Christiansen, 2003; Ferreira, 2005), it

seems appropriate to test the empirical foundation of the Minimalist Program rather than another type of generative syntactic theory.

2. Data Identification Procedure

The procedure for identifying the data points in Adger (2003) was as follows: First, all examples that were obviously not data points (syntactic trees, terminological definitions, etc) were excluded. This yielded 873 data-like examples, which were sorted into the following categories:

Pattern: These are sentences that are reported as part of a group of two or more sentence types that form a pattern of acceptability as is standard in generative syntax. A pattern always included at least one starred example[1] and one un-starred example.

Existence: These are sentences that were used to demonstrate the existence or inexistence of a given construction in English.

Repeats: As a textbook, some sentences are repeated for expository or pedagogical reasons.

Not English: These are examples that are a non-English. We included non-US and non-standard dialects of English in this category because the participant population spoke US English.

Untestable: These are sentences that required a task different from acceptability judgment. For example, some data in syntax is based on the availability or unavailability of specific interpretations or readings of a potentially ambiguous string of words. These can't be tested using a standard judgment survey.

The distribution of data-like examples in Adger (2003) is given in Table 1. Though we attempted to apply the above criteria consistently throughout the entire textbook, it is possible that some readers may disagree with a few of the classifications. Nonetheless, we believe that the classification in Table 1 is by and large correct.

Table 1: The distribution of data-like examples in Adger (2003) according to the above categories.

	Tokens	
Pattern	261	(29.9%)
Existence	250	(28.6%)
Repeats	124	(14.2%)
Not English	144	(16.5%)
Untestable	94	(10.8%)
Total	873	

[1] Stars (or asterisks) are the traditional linguistic notation to convey ungrammaticality, a theoretical construct that should not be confused with acceptability, which is a phenomenal judgment made by informants (Chomsky, 1964; Schütze, 1996).

The 261 tokens that were categorized as pattern examples represented 198 distinct sentence types. 21 of those sentence types were presented without an explicit control condition, though the intended control condition was described in the text (e.g., by discussing the grammatical operation that led to the unacceptability). Therefore we constructed 21 additional sentence types to serve as the grammatical control conditions for these sentence types. The resulting 219 sentence types served as the conditions of the magnitude experiments (ME) described in section 3. We chose magnitude estimation to test the pattern sentence types because the empirical claim in Adger (2003) is that there is a relative difference (or pattern) in acceptability among these sentence types, and magnitude estimation has been proposed as a good method for assessing relative differences in acceptability (Bard et al., 1996; Cowart, 1997; Keller, 2000; Featherston, 2005; for more cautious endorsements, see Bader and Häussler 2010; Weskott and Fanselow, *to appear*; Sprouse *submitted*). The 250 existence tokens became the target materials for the yes-no experiments, which are described in section 4. We chose the yes-no task to test the existence tokens because the empirical claim in Adger (2003) is that these constructions are possible in US English (a fact that could also be tested using a representative corpus of US English). All but three of the existence tokens were un-starred in the text, and presented without any discussion of potential comparison conditions. Three starred examples were also included in the existence category because there was no discussion of an explicit acceptable comparison condition or a grammatical operation that could be used to construct an appropriate comparison condition. The 219 conditions from the ME experiments combined with the 250 existence tokens from the yes-no experiments to form the 469 data points referenced in the introduction. The repeat, non English, and untestable tokens were not tested in the experiments.

3. Magnitude estimation experiments (Experiments 1-6)

Participants

240 participants (40 in each of 6 experiments) completed the magnitude estimation experiments. Participants were recruited online using the Amazon Mechanical Turk (AMT) marketplace, and paid $3.00 for their participation (see Sprouse (2011) 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, consequently there was no incentive to lie. 5 participants answered ‘no’ to one or both of these questions and were therefore excluded from the analysis.

Materials

Division into six experiments. The 219 conditions were distributed among 6 separate experiments in order to keep the total length of each survey under 100 items. The distribution of the conditions among the experiments was pseudorandom according to the following 2 constraints: (i) conditions that were related (i.e., formed a pattern) were placed in the same

experiment so that the resulting statistical analyses were always repeated measures, (ii) the balance of by-hypothesis acceptable and unacceptable conditions was approximately balanced across all 6 experiments. The percentage of by-hypothesis acceptable items was 51% in three of the experiments, and 54% in the other three experiments.

Division into four 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 4 lists using a Latin Square procedure such that each list contained 2 tokens of each condition, and such that the lists did not contain identical lexicalizations of structurally related conditions. Each list was combined with 2 tokens each of 8 filler conditions, which were chosen from a previous, large-scale study (Sprouse, 2011). Two additional acceptable items were added to three of the lists to yield 90 items per list. Finally, each list was pseudorandomized such that related conditions never appeared consecutively. The result was 4 versions each of the 6 experiments.

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. 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, nine additional "anchoring" items (three each of acceptable, unacceptable, and moderate acceptability) were placed as the first nine items of each survey. These items were identical, and presented in the identical order, for every survey. Participants rated these items just like the others; they were not marked as distinct from the rest of the survey in any way. However, these items were not included in the analysis as they served simply to expose each participant to a wide range of acceptability prior to rating the experimental items (a type of unannounced "practice"). This resulted in surveys that were 99 items long. The surveys were advertised on the Amazon Mechanical Turk website (see Sprouse, 2011, for evidence of the reliability of data collected using AMT), 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.

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 also Featherston 2005). We used the discussion in Adger (2003) to identify the appropriate analyses for each pattern. This resulted in 104 pairwise comparisons, 7 one-way analyses (each of three conditions), and 4 two-way (2x2) factorial analyses.

Pairwise comparisons. We ran three types of statistical tests for the pairwise comparisons. First, we calculated two-tailed *p*-values using standard paired *t*-tests. Second, because there has been growing interest in the use of linear mixed-effects models (LMEMs) for the analysis of language data, we ran linear mixed-effect models with one (two-level) fixed factor and two random factors: participants and items. We used the `pvals` function from the `languageR` package to estimate *p*-values (Baayen, 2007; Baayen et al., 2008). Finally, 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). Bayes factors provide an intuitively natural measure of the strength of the evidence for each of the two hypotheses in the form of an odds ratio. For 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. We used the JSZ Bayes factor equation from Rouder et al. (2009), which assumes (i) a non-directional H1 (equivalent to a two-tailed *t*-test), and (ii) an equal prior probability of the two hypotheses. We chose this form of the Bayes factor calculation over directional versions (such as the Savage-Dickey test proposed by Wetzels et al. (2009)) because the non-directionality and equal prior assumptions result in more conservative Bayes factors. For ease of exposition, the resulting Bayes factors in Table 2 are categorized using the classification proposed by Jeffreys (1961).

Table 2: Results of the pairwise analyses for three types of statistical tests: *t*-tests, linear mixed effects models, and Bayes factor analysis. There were 104 pairwise comparisons.

Paired *t*-Tests		LMEM		Bayes Factors	
n.s	0	n.s	1	Evidence for H0 (<1)	0
p<.10	0	p<.10	0	No evidence (1)	0
p<.05	2	p<.05	4	Anecdotal evidence for H1 (1-3)	2
p<.01	1	p<.01	4	Substantial evidence for H1 (3-10)	1
p<.001	3	p<.001	5	Strong evidence for H1 (10-30)	0
p<.0001	98	p<.0001	90	Very strong evidence for H1 (30-100)	2
				Extreme evidence for H1 (>100)	99

One-way and two-way analyses. For the one-way and two-way analyses we ran both repeated-measures ANOVAs and LMEMs with participant and item as random factors. For the 2 x 2 factorial ANOVAs, we report the results for the predicted interaction. We do not report Bayes factors for the one-way and two-way analyses because there are currently no straightforward methods for calculating Bayes factors for these designs.[2]

[2] Jeff Rouder (p.c.) has developed methods for calculating Bayes factors for both one-way and factorial ANOVA; however, he has not yet made those methods public. If these methods become public prior to publication, we will include Bayes factors for these 11 analyses. However, we feel that the current lack of BFs for 11 out of 365 analyses is unlikely to impact the conclusions of this paper.

Table 3: Results of the one-way and two-way analyses for two types of statistical tests: repeated measures ANOVA and linear mixed effects models. There were 7 one-way analyses and 4 two-way analyses. The results of the two-way analyses concern the predicted interaction.

	One-way analyses		Two-way analyses	
	ANOVA	LMEM	ANOVA	LMEM
n.s	0	0	2	2
p<.10	1	0	0	0
p<.05	0	0	1	1
p<.01	0	1	0	0
p<.001	0	0	0	0
p<.0001	6	6	1	1

Though we have chosen to report only summaries of the statistical tests in this section, a complete list of the results of the statistical analyses is available online as Appendix A.

Discussion

Traditional statistical tests (i.e., paired *t*-tests, one-way repeated measures ANOVA and two-way factorial 2x2 repeated measures ANOVA) and linear mixed-effects models yielded nearly identical results. Traditional statistical tests (with two-tailed *p*-values) resulted in 112 replications, 1 marginal replication (p=.09), and 2 replication failures (p =.36, p =.14), while linear mixed-effects models resulted in 112 replications and 3 replication failures (p =.11, p =.38, p =.11). In order to derive a maximum replication failure rate, we equated marginal replications with replications failures, resulting in a maximum replication failure rate for both traditional statistical analysis and linear mixed-effect models of 2.6% (3/115). Bayes factor analyses were only conducted for the pairwise comparisons. Out of the 104 pairwise comparisons, 102 yielded "substantial" to "extreme" evidence for H1, and 2 yielded only "anecdotal" evidence for H1. In order to derive a maximum replication failure rate for Bayes factor analysis, we counted "anecdotal" evidence as a replication failure, resulting in a maximum replication failure rate of 1.9%. Taken together, the maximum replication failure rate for the 115 analyses tested in the ME experiments is in the range of 1.9%-2.6%. Table 4 summarizes these counts.

Table 4: Counts of the replications and failures for the ME experiments. The failure rate includes marginal results as replication failures to derive a maximum failure rate.

	Significant	Marginal	Non-significant	Replication failure rate
Traditional statistical tests	112	1	2	2.6%
LMEM	112	0	3	2.6%
Bayes factors	102	2	0	1.9%

Though it is beyond the scope of this article, some readers may be interested in precisely which sentence types failed to replicate under formal data collection methods, and which aspects of the theory they are intended to justify. A list of the sentences that yielded replication failures,

along with their descriptive and inferential statistics, and their theoretical relevance is available online as Appendix B.

4. Yes-No experiments (Experiments 7-11)

Participants

200 participants completed the yes-no experiments (40 participants in each of 5 experiments). Participants were once again recruited online using the Amazon Mechanical Turk (AMT) marketplace (see Sprouse, 2011), and paid $2.00 for their participation. Participant selection criteria were identical to those of the ME experiments. 3 participants answered 'no' to one or both of the language history questions and were therefore excluded from the analysis.

Materials

Adger (2003) follows a practice that is common in linguistics textbooks: several of the example tokens were obviously constructed to maintain the attention of undergraduate and graduate students, rather than present semantically and pragmatically neutral examples of the syntactic structures in question. As such, we made minor changes to 107 of the 250 existence tokens prior to running them in the yes-no experiments. We changed proper names (usually Greek mythological figures) in 68 sentences to common US proper names; we changed the lexical items in 66 sentences to eliminate references to violent, fictional, or otherwise implausible items (e.g., executioners, gorgons); finally, we added antecedent clauses to 9 sentences to make certain pragmatically restricted constructions, such as ellipsis and topicalization, more plausible in a single sentence. In all 107 cases, the structural properties of the sentences were maintained.

The 250 existence tokens were distributed into 5 separate lists. Four lists contained 50 acceptable target items and 50 unacceptable filler items. The fifth list contained 47 acceptable target items, 3 unacceptable target items, 47 unacceptable filler items, and 3 acceptable filler items. Each list was 100 items long, with a ratio of acceptable items to unacceptable items of 1:1, and a ratio of target items to filler items of 1:1. Four versions of each list were created to counterbalance the order of presentation: original order, reversed order, transposition of the first and second half, and reversed order of the transposed halves.

Task and Presentation

The task was a standard two-choice yes-no task. Participants were asked to click radio buttons that were labeled YES or NO. The surveys were advertised on the Amazon Mechanical Turk website (see Sprouse, 2011), and presented as web-based surveys using an HTML template that is available on the first author's website. Participants completed the surveys at their own pace.

Results

Responses were analyzed in two ways: (i) using the traditional sign-test (with two-tailed *p*-values), and (ii) using the Bayes factor calculation for binomial responses made available by Jeff Rouder on his website: http://pcl.missouri.edu/bayesfactor.

Table 5: Results of the yes-no experiments for two statistical tests: sign-tests and binomial Bayes factor analyses. There were 250 items tested.

Sign tests		Bayes Factors	
n.s	1	Strong evidence for H0 (1/30-1/10)	1
p<.10	2	Anecdotal evidence for H0 (1/3-1)	2
p<.05	2	No evidence (1)	0
p<.01	5	Anecdotal evidence for H1 (1-3)	2
p<.001	7	Substantial evidence for H1 (3-10)	2
p<.0001	233	Strong evidence for H1 (10-30)	2
		Very strong evidence for H1 (30-100)	7
		Extreme evidence for H1 (>100)	234

Discussion

The sign-tests yielded 247 replications, 2 marginal replications (p=.054, p=.077), and 1 replication failure (p=.44). Again, in order to derive a maximum replication failure rate, we equated marginal replications as replications failures, resulting in a maximum replication failure rate of 1.2% (3/250). From the perspective of Bayes factor analysis, 245 replications yielded "substantial" to "extreme" evidence for H1, 2 yielded only "anecdotal" evidence for H1, 2 yielded "anecdotal" evidence for H0, and 1 yielded "strong" evidence for H0. Again, in order to derive a maximum replication failure rate for Bayes factor analysis, we counted "anecdotal" evidence for H1 as a replication failure, as well as any evidence for H0, resulting in a maximum replication failure rate under Bayes factor analysis of 2% (5/250). Taken together, the replication failure rate for the yes-no experiments was in the range of 1.2%-2%. Table 6 summarizes these counts. A list of the failed sentences, along with their descriptive and inferential statistics, and their theoretical relevance is presented in Appendix B.

Table 6: Counts of the replications and failures for the yes-no experiments. The failure rate includes marginal results as replication failures to derive a maximum failure rate.

	Significant	Marginal	Non-significant	Replication failure rate
Sign tests	247	2	1	1.2%
Bayes factors	245	2	3	2.0%

5. General Discussion

In this study, we compared the rate of agreement between the results of the informal experiments reported in the popular textbook by Adger (2003) and the results elicited by the same material under formal experimental conditions. A small number of disagreements were identified, but these were all null results (i.e., a replication failure); no results in opposing directions were identified. To estimate the maximum number of replication failures in Adger (2003) we propose the following algorithm: (i) select the statistical tests that resulted in the most replication failures for each design, and (ii) include marginal results as replication failures. Using this algorithm, the

maximum number of replication failures is 7 out of 365 statistical tests. This suggests a replication failure rate of 1.9%; which for reasons of resolution we will round to 2%.

Before discussing the implications of this number, it is important to be clear about what it actually represents. The discussion of the 'replication failure rate' thus far has assumed a specific directionality. We have assumed that the formal results represent a true description of the world (i.e., true negatives and true positives), and that the informal judgments represent an inaccurate description of the world (i.e., false positives). This is an assumption shared by some critics of informal acceptability judgment methods (Wasow and Arnold, 2005; Dabrowksa, 2010; Gibson and Fedorenko 2010a, 2010b). However, there is no *a priori* reason to assume one directionality over the other, as it is a completely empirical question as to which set of results better reflects reality (see also Grewendorf, 2007; Haider, 2007). In other words, though we think of the 2% discrepancy between the informal and formal judgments as a failure of the informal methods, it could also be a failure of the formal methods. For example, it may be the case that the marginal results reported above are not true negatives, but instead were caused by insufficient statistical power in the experiments. Ultimately, the question of which results are closer to the truth is an empirical question that is beyond the scope of this article. For the sake of argument, we will follow the critical literature and assume that the 2% discrepancy represents a set of false positives caused by deficiencies in the informal methods used to collect them, though we note that this discrepancy alone does not tell us which result is more accurate. This assumption allows us to interpret the 2% discrepancy as a maximum replication failure rate for the empirical foundation of the Minimalist Program, and ask the question: what can be concluded from it?

One of the primary implications of the critical literature is that generative syntactic theories, such as the Minimalist Program, rest on faulty empirical foundations:

> The result has been the construction of elaborate theoretical edifices supported by disturbingly shaky empirical evidence. (Wasow & Arnold, 2005, p. 1482)

> The lack of validity of the standard linguistic methodology has led to many cases in the literature where questionable judgments have led to incorrect generalizations and unsound theorizing. (Gibson & Fedorenko, 2010a, p. 233)

The question then is whether the maximum replication failure rate of 2% is considered a tolerable replication failure rate for a syntactic theory, or whether it constitutes a serious problem for the empirical foundation of generative syntax. A 2% rate of false positives is substantially lower than the 5% false positive rate (type 1 error rate) tolerated by the null hypothesis significance testing (NHST) that is standard practice in cognitive science. This suggests that informal methods in fact provide a tolerably low rate of false positives. To the extent that the 469 data points in Adger (2003) form the empirical foundation of the Minimalist Program, it seems safe to conclude that this empirical foundation is sound, and that the results of traditionally informal methods can, at least in some circumstances, reach formal replication rates of 98%.

This leads to a second important question: Is the replication failure rate for the Adger (2003) textbook (2%) indicative of the replication failure rate for the entire generative syntax literature? Prior to this study, estimates of the replication failure rate tended to be speculative:

> 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)

> My guess is that the percentage [of replication failures] is quite low, though that is an empirical issue, difficult as it might be to resolve adequately. (Newmeyer, 2007, p. 396)

While it is likely that generative syntacticians will be willing to accept the replication failure for Adger (2003) as indicative of the entire field, it may be the case that stalwart critics of informal judgments will wish to argue that the failure to replicate rate for the rest of the field (i.e., the cutting edge theories discussed in modern journal articles) may be higher. However, if that were so, it would raise an interesting logical dilemma for critics of informal judgments, since such a situation would require them to assume that there is a qualitative difference between the "core" data points in Adger (2003), which have a low replication failure rate, and the "non-core" data points that appear in journal articles, which have a higher replication failure rate (see also Phillips 2009). This distinction is not without consequences. As Phillips (2009) observes, not all data points are treated equally during theory construction. It seems to be a fact of all scientific theories that there is a core set of phenomena that are used to initially construct the basic architecture of the theory. Crucially, these core data points also tend to be weighted more than non-core data points when theoretical changes are proposed. This core/non-core distinction has been explicitly rejected by some critics (e.g., Gibson and Fedorenko, 2010b) as a subjective distinction that cannot be grounded in scientific criteria. If critics wish to argue that the 2% replication failure of Adger (2003) is not representative of the rest of the literature (despite encompassing 469 data points across a wide range of topics), then either their claim is subjective (and can therefore be dismissed as suggested by Gibson and Fedorenko, 2010b), or their claim is based on a true scientific distinction between core/non-core phenomena. If there is indeed a legitimate scientific basis for the core/non-core distinction, then the data points in Adger (2003) would certainly be considered core, and would therefore carry more weight than non-core data points in discussions about the problem of faulty data in generative syntax. Thus, there seems to be no easy way to dismiss the 2% replication failure rate as unrepresentative of the data that is used to construct generative syntactic theories.

This discussion leads to the third interesting question raised by the experimental syntax literature: Will the broad adoption of formal experimental syntax techniques lead to drastic changes in generative syntactic theory? If it were true that the empirical foundation of generative syntax were faulty, then the answer would be simple – yes. However, the fact that the empirical foundations appear to be sound suggests that the answer will be more nuanced. It is possible that over time the (potentially) increased sensitivity of experimental syntax will lead to the collection of a body of data that will motivate dramatically different theoretical architectures (as proposed by Keller, 2000; Featherston, 2005a, 2005b, 2009). However, it is also possible that acceptability judgments will continue to be only one small part of the puzzle. It is a widely accepted fact in the syntactic literature that the difference among syntactic theories is rarely based on debates about the acceptability of specific sentences. Instead, syntactic theories tend to explain the same set of (acceptability) data using different assumptions about the types of grammatical operations made available by the human brain (see also Phillips, 2009; Sprouse and Lau, 2011). Because there is significant overlap in the empirical coverage of all syntactic theories, the formal experimental validation of the data points in Adger (2003) is also a validation of the data points that underlie Head-driven Phrase Structure Grammar (HPSG), Lexical-Functional Grammar (LFG), Role and Reference Grammar, Construction Grammar, Categorial Grammar, and any number of other syntactic theories (Phillips, 2009).

6. Conclusion

Two logically separable types of arguments have been proposed in the literature for the adoption of formal experimental techniques for the collection of acceptability judgments. One claims that formal experiments provide a potentially more sensitive measurement tool than informal techniques; the other claims that informal techniques are in fact an unreliable measurement tool. These are logically distinct, since one can simultaneously believe that the adoption of experimental syntax is a positive development for the field of generative syntax, without believing that the previous 50 years of research was substantially flawed. While this nuanced position has been generally adopted by the practitioners of experimental syntax (e.g., Keller, 2000; Featherston, 2005; Sprouse 2007), there is a growing critical literature that explicitly claims that informal judgment collection has yielded faulty data and led to unsound syntactic theories. In this paper, we proposed a methodology for assessing the empirical validity of this claim. We formally tested all 469 data points from a popular generative syntax textbook (Adger, 2003) on 440 naïve participants. Using three different statistical analysis approaches (traditional statistical tests, linear mixed-effects models, and Bayes factor analysis), and adopting the assumption of critics that formal results are more 'true' than informal judgments, we estimated a maximum replication failure rate of 2% for the 469 data points in Adger (2003).

Such a low replication failure rate suggests to us that the data that has been used to construct generative syntactic theories is sound. In reaching this conclusion, we have attempted to avoid any biased assumptions. First, we adopted the position of most critics that the results of formal experiments should be believed over informal results (though this is technically an open empirical question). Second, we adopted the conservative versions of three different types of statistical tests, and chose the tests that led to the highest replication failure rates. Third, we attempted to ensure the generalizability of the results, and avoid any selection bias, by exhaustively testing all 469 data points presented in a popular generative syntax textbook (Adger, 2003). By choosing a textbook, we ensured that (i) the number of data points tested would be two orders of magnitude larger than previous studies, (ii) the data points tested would span a wide range of phenomena, (iii) the data points tested would represent the empirical foundation of generative syntactic theory (in particular, the Minimalist Program). These assumptions and design choices yielded a maximum false positive rate for informal methods of 2%, which is smaller than the 5% false-positive rate that is widely tolerated in cognitive science. The only open question is whether critics will argue that there is a qualitative distinction between the core phenomena in Adger (2003) and the cutting edge phenomena that appear in syntax journals. Though this is an interesting question, it seems clear that such a claim will not lead to foundational problems for generative syntactic theories. If there is indeed a qualitative difference between the data points in Adger (2003) and the data points in journal articles, then surely the former should be weighted more in debates about the empirical foundations of generative syntactic theories. In either case, a replication rate of 98% for the 469 data points in Adger (2003) suggests that informal data collection did not lead to generative syntactic theories being constructed upon faulty data.

Acknowledgments

This research was supported in part by National Science Foundation grant BCS-0843896 to Jon Sprouse. We would like to thank Andrew Angeles, Melody Chen, and Kevin Proff for their assistance constructing materials.

References

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

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

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

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

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Clifton, C., Jr., Fanselow, G., & Frazier, L. (2006). Amnestying superiority violations: Processing multiple questions. *Linguistic Inquiry, 37*, 51-68.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown.* Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

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

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

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

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

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

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

Langendoen, D.T., Kalish-Landon, N., & Dore, J. (1973). Dative questions: a study in the relation of acceptability to grammaticality of an English sentence type. *Cognition, 2*, 451-477.

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

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

Newmeyer, F. J. (1983). *Grammatical theory: Its limits and its possibilities*. Chicago: University of Chicago Press.

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

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

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

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

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

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

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

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

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

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory.

Sprouse, J. (under review). Evaluating the assumptions of magnitude estimation of linguistic acceptability.

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

Sprouse, J. & Lau, E. (2011). Syntax and the brain. In M. den Dikken (ed) *The handbook of generative syntax*, Cambridge University Press.

Sprouse, J., Wagers, M., & Phillips, C. (under review). A test of the relation between working memory capacity and island effects.

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

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

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

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

Weskott, T. & Fanselow, G. (*to appear*). *Language*.

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