Available online at www.sciencedirect.com

SciVerse ScienceDirect

Lingua 134 (2013) 219–248

Lingua

www.elsevier.com/locate/lingua

A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010

Jon Sprouse [a,*], Carson T. Schütze [b], Diogo Almeida [c]

[a] *Department of Linguistics, University of Connecticut, 365 Fairfield Way, Unit 1145, Storrs, CT 06269-1145, USA*
[b] *Department of Linguistics, University of California, Los Angeles, PO Box 951543, Los Angeles, CA 90095-1543, USA*
[c] *Division of Science – Psychology, New York University, Abu Dhabi, PO Box 129188, Abu Dhabi, United Arab Emirates*

Received 17 January 2013; received in revised form 28 June 2013; accepted 1 July 2013
Available online 1 September 2013

Abstract

The goal of the present study is to provide a direct comparison of the results of informal judgment collection methods with the results of formal judgment collection methods, as a first step in understanding the relative merits of each family of methods. Although previous studies have compared small samples of informal and formal results, this article presents the first large-scale comparison based on a random sample of phenomena from a leading theoretical journal (*Linguistic Inquiry*). We tested 296 data points from the approximately 1743 English data points that were published in *Linguistic Inquiry* between 2001 and 2010. We tested this sample with 936 naïve participants using three formal judgment tasks (magnitude estimation, 7-point Likert scale, and two-alternative forced-choice) and report five statistical analyses. The results suggest a convergence rate of 95% between informal and formal methods, with a margin of error of 5.3–5.8%. We discuss the implications of this convergence rate for the ongoing conversation about judgment collection methods, and lay out a set of questions for future research into syntactic methodology.

Keywords: Acceptability judgments; Grammaticality judgments; Experimental syntax; Methodology

1. Introduction

Acceptability judgments provide the primary empirical foundation of many syntactic theories (Chomsky, 1965; Schütze, 1996). The vast majority of the acceptability judgments that have been reported in the syntax literature were collected using methods that appear relatively informal compared to the data collection methods in other domains of cognitive science. However, over the past 15 years or so there has been a shift in data collection practices, with the number of studies employing formal experimental methods, sometimes known as *experimental syntax techniques* following Cowart (1997), increasing every year. This development means that there are two methods for collecting acceptability judgments currently in widespread use in the field of syntax: the relatively informal traditional methods that have largely established the foundation of the field for the past 60 years (henceforth *informal methods*), and the more formal experimental methods that have been gaining popularity over the past 15 years (henceforth *formal methods*). This methodological dichotomy has led a number of researchers to ask which method is empirically superior (e.g., Bard et al., 1996; Keller, 2000; Edelman and Christiansen, 2003; Phillips and Lasnik, 2003; Featherston, 2005a,b, 2007, 2008, 2009; Ferreira, 2005; Sorace and Keller, 2005; Wasow and Arnold, 2005; den Dikken et al., 2007; Alexopoulou and Keller, 2007;

* Corresponding author. Tel.: +1 860 486 4229.
E-mail addresses: jon.sprouse@uconn.edu (J. Sprouse), cschutze@humnet.ucla.edu (C.T. Schütze), diogo@nyu.edu (D. Almeida).

0024-3841/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.lingua.2013.07.002

Bornkessel-Schlesewsky and Schlesewsky, 2007; Fanselow, 2007; Grewendorf, 2007; Haider, 2007; Newmeyer, 2007; Sprouse, 2007; Culbertson and Gross, 2009; Myers, 2009a,b; Phillips, 2010; Bader and Häussler, 2010; Dąbrowska, 2010; Gibson and Fedorenko, 2010; Culicover and Jackendoff, 2010; Gross and Culbertson, 2011; Sprouse, 2011b; Weskott and Fanselow, 2011; Gibson et al., 2011; Sprouse and Almeida, 2012, 2013; Gibson and Fedorenko, 2013). Our goal in this paper is to substantially increase the empirical basis of this line of research by comparing the results of informal and formal methods for a very large and random sample of phenomena from the cutting edge of syntactic theorizing.

The goal of the present study is provide a direct comparison of the results of informal judgment collection methods with the results of formal judgment collection methods. We believe that the first step in understanding the relative merits of each family of methods is to determine to what extent the two methods converge (or diverge). Although there have been several previous studies that have compared results of informal methods with the results of formal methods (e.g., Gordon and Hendrick, 1997; Clifton et al., 2006; Gibson and Fedorenko, 2013), these previous studies cannot in principle be used to estimate a convergence rate between informal and formal methods, for two reasons. First, these studies have investigated a relatively small number of phenomena (e.g., Gibson and Fedorenko, 2013 test seven data points comprising three distinct phenomena in their criticism of informal methods) compared to the sheer number of data points published in the syntax literature. With a small sample size, it is unclear whether the number of divergences is high relative to the number of convergences. Testing this requires a much larger sample of phenomena. Second, the phenomena that have been tested in previous studies have been selected using a biased selection procedure. We cannot know exactly how the phenomena were chosen, but previous studies do not claim to have selected the phenomena randomly from the literature. Only a random selection provides confidence that the divergences and convergences are representative of the population they are drawn from. This confidence is quantified with the margin of error, which can be calculated based on the relative size of the sample compared to the population. For these reasons we decided to randomly select a large sample from the population of phenomena published in *Linguistic Inquiry* (*LI*) from 2001 to 2010. One added benefit of random sampling is that a sufficiently large random sample will likely (although not strictly necessarily) mean that a large number of distinct syntactic phenomena will be investigated, providing a broad empirical base for the comparison of the results of the two methods.

The rest of this article is organized as follows. In Section 2 we present the design of the random sampling study, along with the rationale for each of the design choices that we made. In Section 3 we present the details and results of the acceptability judgment experiments. The results of the three tasks and five statistical analyses suggest convergence rates with the published informal judgments ranging between 86% and 99%, depending on the analysis. In Section 4 we discuss a principled way of selecting a point estimate for the convergence rate, and the potential inferences licensed by that selection. In Section 5 we discuss the information that would be necessary to make additional inferences about syntactic methodology, and the general design of the follow-up experiments that would be necessary to gather that information. Section 6 concludes.

2. The design of the random sampling study

Any random sampling study requires a number of methodological decisions, such as what constitutes the appropriate population of study, what constitutes an appropriate sample, how best to calculate the estimate of interest, etc. In this section we discuss, in detail, the rationale underlying each of the methodological choices that we made in the construction of the present study. For readers under time constraints, a succinct summary of our methodology is as follows: First, we randomly sampled 300 sentence types (forming 150 two-condition, or pairwise, phenomena; see Section 2.2) from the approximately 1743 data points published in *Linguistic Inquiry* 2001–2010 that are (i) unique (i.e., not repeated), (ii) part of US English inter alia, and (iii) based on standard acceptability judgments (as defined in Section 2.2). Next, we constructed a total of 8 items for each of the 300 sentence types (2400 items total). Then we tested the 150 pairwise phenomena in three experiments, one for each of three distinct judgment tasks commonly used in the syntax literature. Finally, we applied five distinct quantitative analyses to the results of the three judgment tasks to derive 15 convergence estimates that span the spectrum of possible tasks and statistical analyses. We discuss each of these methodological choices in detail in the remainder of this section.

2.1. Acceptability versus grammaticality

The first step in deriving a convergence rate between methods is to delineate the type of data that will be the focus of the study. For this study, we are interested in evaluating informal and formal approaches to acceptability judgment tasks, therefore the underlying phenomenon of interest is sentence acceptability. Acceptability judgment tasks are a type of perceptual rating task: they ask participants to provide a report of their perception of the acceptability of a sentence. Acceptability is a property of sentences that speakers have (at least partial) conscious access to. It is often described phenomenologically as ‘‘how good, or acceptable, a sentence sounds.’’ Perceptions of acceptability are often assumed to

arise as an automatic consequence of sentence comprehension, as they cannot be consciously suppressed by native speakers. Acceptability is generally considered a composite property, as several factors appear to affect acceptability judgments. Crucially, one of those factors is (under the assumptions of proponents of both methods) the grammaticality of the sentence, that is, whether the grammar of language generates the sentence in question. It is this relationship between grammaticality and acceptability that has led to the use of acceptability judgments as evidence for the construction of grammatical theories. Although acceptability ratings are used by syntacticians to create grammatical theories, *acceptability* and *grammaticality* are crucially distinct. Similarly, although it is common for some linguists to use the term ''grammaticality judgments'', it is generally assumed that speakers do not have conscious access to the working of the mental grammar, therefore ''grammaticality judgments'' are not possible. In most (if not all) cases, the term ''grammaticality judgment'' appears to be synonymous with the more precise term ''acceptability judgment.'' We use the more precise term here to avoid any confusion: the current study is designed to compare the reports of acceptability that are returned by informal and formal methods.

2.2. The type of judgment task

The second step in deriving a convergence rate between methods is to delineate the different types of acceptability judgments contained in the syntax literature, and decide which of these judgments will be evaluated. We identified at least five judgment types based primarily on the collection method required to elicit judgments:

Standard acceptability judgments: These require only that the participant be presented with a sentence and asked to judge its acceptability on an arbitrary scale or in reference to another sentence.
Coreference judgments: These are primarily used to probe binding relationships. Participants must be presented with a sentence that includes two or more noun phrases that are identified in some way. They are then asked to indicate whether the two noun phrases can or must refer to the same entity.
Interpretation judgments: These are judgments based on the meaning of sentences, such as whether a sentence is ambiguous or unambiguous, or whether one quantifier has scope over another. These may require explicit training of participants to identify multiple potential meanings, and/or explicitly constructed contexts to elicit one or more potential meanings.

Two variants of standard acceptability judgments require additional methodological considerations:

Judgments involving relatively few lexical items: These are acceptability judgments about phenomena that occur with relatively few lexical items, such that the construction of 8 substantially distinct tokens, as was done for the phenomena tested in this study, would likely be impossible. This is not to say that these phenomena cannot be tested in formal experiments, but participants in such experiments may require special instruction to guard against potential repetition confounds.
Judgments involving prosodic manipulations: These are acceptability judgments that are based on specific prosodic properties of the sentence. They require either the presentation of auditory materials or the use of some notational conventions for conveying the critical prosodic properties in writing (e.g., the use of capital letters to indicate emphasis).

The data identification procedure (discussed in Section 2.3) resulted in the (estimated) distribution of data points in articles published in *Linguistic Inquiry* between 2001 and 2010 shown in Table 1.

For the present study we decided to focus exclusively on standard acceptability judgments for three reasons: (i) they are the most straightforward to adapt to formal methods, as they require no special instruction of the participants, and no

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 maximally 6.9% (see Section 2.3 for details).

Type of data point	Estimated count	Estimated percentage
Standard acceptability judgments	1743	48%
Coreference judgments	540	15%
Interpretation judgments	854	23%
Judgments involving relatively few lexical items	422	12%
Judgments involving prosodic manipulations	76	2%
Total number of (unique English) data points in *LI* 2001–2010	3635	100%

special equipment for participants to complete the task; (ii) they form the largest single type of data published in *LI* 2001–2010; and (iii) they are the focus of several recent criticisms of informal methods in the literature (e.g., Ferreira, 2005; Wasow and Arnold, 2005; Gibson and Fedorenko, 2010, 2013). Of course, the fact that the current study is limited to a single data type means that the estimate of convergence derived here applies only to that data type. It is logically possible that the other data types will result in different convergence rates. The same holds for our focus on US-English data points.

2.3. Defining the phenomena to be tested

The third step in deriving a convergence rate is to define the phenomena that will be tested. The goal of any methodology is to measure phenomena that will form evidence for the construction of theories, therefore this step ultimately hinges on the types of phenomena that form the empirical base underlying syntactic theories. Exactly which types of phenomena form the empirical base of syntactic theory, and in what proportion each type is used, is a continually evolving empirical question. For the current study we have decided to focus on what we will call *pairwise phenomena*: two maximally similar sentence types that differ along some dimension that is hypothesized to (i) be relevant for theories of grammar and (ii) lead to a significant difference in acceptability. Although we do not know the exact proportion of pairwise phenomena in the empirical base of current syntactic theories, and have no way of knowing the proportion of pairwise phenomena underlying future iterations of syntactic theories, we believe that they are relatively frequent in the existing literature. For example, we found that 72% of the diacritic-marked data points that we randomly sampled from *LI* were published with explicit control sentences in the article (see also Section 2.3). The other 28% contained discussion in the text surrounding the data point that implied a control condition that any professional syntactician could construct for themselves. Furthermore, the relative frequency of pairwise phenomena has been explicitly recognized in both the experimental syntax literature (e.g., Bard et al., 1996; Myers, 2009a), and the theoretical syntax literature (as Bošković and Lasnik, 2003:527 put it, ''As is standard in the literature, the judgments reported in this article are intended as relative rather than absolute, and most of the data was collected by soliciting relative judgments between pairs of examples.''). Beyond being a relatively frequent source of evidence in syntactic theory, pairwise phenomena are also a relatively useful source of evidence. Pairwise phenomena allow syntacticians to isolate the factors that affect acceptability, and ultimately elucidate the inner workings of the mental grammar. Therefore for this study we randomly selected 150 pairwise phenomena from *LI* 2001 to 2010, consisting of 150 sentences marked with a diacritic indicating unacceptability (*, ?, or some combination thereof), and 150 control sentences. 149 of the control sentences were not marked with any diacritic (indicating acceptability), and one control sentence was marked with a question mark.

There are, of course, other phenomena that could be examined. For example, an anonymous reviewer has asked whether we could move away from the more theory-driven pairwise phenomena, and instead focus on raw acceptability judgments of individual sentences. Although our data could be looked at from this angle, we decided against pursuing this in the main body of the article, in part because raw ratings of individual sentences appear to play a less frequent role in syntactic theorizing than pairwise comparisons in the published literature. However, we understand that some readers may be interested to see such an analysis, so we do present one in Section 5.3. Although the results are roughly in line with the results of the pairwise analysis pursued in Sections 3 and 4, there is at least one potential confound in such an analysis that could only be overcome with a different experimental design. We discuss this in detail in Section 5.3.

2.4. The population of data points to be sampled from

We chose *Linguistic Inquiry* for our study because it is a leading theoretical journal among generative syntacticians, and the articles published in *LI* rely almost exclusively on informal judgment collection methods (only three syntax-focused articles between 2001 and 2010 reported using formal methods). This makes *LI* an ideal candidate for estimating a convergence rate between informal and formal methods. To be clear, we do not intend the results of this study to be a specific defense or incrimination of articles published in *LI*, but rather we intend *LI* to stand as a proxy for the use of informal methods in syntax more generally. We chose a recent ten-year stretch of *LI* (2001–2010) to make it more likely 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. Of the 229 articles about syntax, 114 were predominantly about phenomena that are part of US English inter alia, where predominantly was operationally defined as greater than 80% of the data points. 115 were predominantly about languages other than English(es). Three employed formal experimental methods. We used the remaining 111 articles for this study, as these were (i) about syntax, (ii) about phenomena that hold of US English inter alia, and (iii) did not employ formal experimental methods.

We decided to focus on English data points in this project primarily due to logistical concerns: English is the native language of the first two authors, making materials construction for English data points more manageable than other

languages, and online participant marketplaces (such as Amazon Mechanical Turk) tend to have limited cross-linguistic value at the moment of writing (e.g., Ipeirotis, 2010 reports that the majority of Amazon Mechanical Turk participants come from two countries, the US and India, primarily because US dollars and Indian rupees are the two currencies Amazon makes available). Furthermore, some critics of informal methods have suggested that the existence of such marketplaces reduces the time cost of formal experiments (e.g., Gibson and Fedorenko, 2013; Gibson et al., 2011), therefore it seems appropriate to use these marketplaces for this case study.

We employed several undergraduate research assistants with minimal training in linguistics to conduct a first-pass count of data points in *LI* 2001–2010. This ensured that our theoretical biases would not influence the inclusion or exclusion of potential data points in this study. They were instructed to identify all numbered examples, and then label all trees, tables, diagrams, definitions, and sentences that were not English as ''non-data-points'', while labeling all the remaining numbered examples as potential English 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 subdivide the potential English data points by judgment type: if the example included a subscripted pronoun then label it a coreference judgment, if it included a greater than/less than sign (as used to report scope) or hash-mark (#, as used to report felicity judgments) then label it an interpretation judgment, etc.

This first-pass categorization resulted in 3335 potential English data points and 2061 non-data-points (which includes non-English data points). We then randomly sampled 308 items from the potential English data points, and 191 from the other group, to check the accuracy of the first-pass categorization. Based on those samples, we estimate that the total number of English data points in *LI* between 2001 and 2010 is approximately 3635 with a maximum margin of error of 6.9%,[1] broken down into the data type categories reported in Table 1 above.

2.5. The random sampling procedure

Our goal for this study was to test 150 unacceptable sentence types and 150 (more) acceptable controls (300 sentence types) forming 150 pairwise phenomena. Because we anticipated mistakes in the classification of data points, we knew this would require (i) sampling more than 150 unacceptable sentences from the 3335, and (ii) working through the sample to identify data points of the correct type (unique, English, standard acceptability judgment, and unacceptable). Therefore we used the R statistical computing language (R Core Team, 2012) to randomly sample (without replacement) 355 unacceptable items from the set of potential US English data points (about 10% of that set), and inspected each one sequentially. We had to inspect 308 of the items to find 150 unacceptable sentences that could be used to form the 150 pairwise phenomena. To operationalize ''unacceptable'', we only sampled data points that were published with a judgment diacritic (*, ?, or some combination of the two), which generally indicates that they were judged less acceptable than a minimally contrasting control sentence, according to informal methods. We focused the sampling procedure on unacceptable sentences in order to test claimed contrasts between two phenomena (under the assumption that such contrasts will generally contain at least one diacritically marked sentence). This procedure has the added benefit of reducing the likelihood that our study contained ''example'' sentences that are used simply to illustrate the existence of a specific construction in a language (under the assumption that such sentences would have no diacritic). We found explicit control conditions for 108 of the 150 unacceptable sentences in their original articles; for the remaining 42, we constructed control conditions based upon the theoretical discussion provided by the original authors (see also Section 2.2). Based on the estimated population size of 1743 US English standard acceptability data points in *LI* 2001–2010, the sample of 300 data points allows us to estimate a convergence rate between formal and informal methods for the standard acceptability judgments published in *LI* 2001–2010 with a margin of error of 5.3–5.8%.[2]

In short, we tested a random sample of 300 data points from *LI* 2001 to 2010 that form 150 theoretically meaningful pairwise phenomena. This sample is both randomly selected, thus avoiding any bias in the selection process, and also more than 15 times larger than any previous (biased) comparisons of informal and formal methods. Furthermore, to the extent that *LI* 2001–2010 is representative of the data in the field, the convergence rate will also be representative of the data in the field within the margin of error. A full list of examples of the sentence types that were tested, along with mean ratings for each, is provided in Appendix A.

[1] 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 set of all examples, the margin would be 4.3%. However, we took one sample from each of the two sub-populations. The margin of error for the potential US English data points is 5.4%; the margin of error for the other sub-population is 6.9%; hence the maximal margin of error is 6.9%.

[2] The margin of error is a range because there are (at least) two ways to count the 42 additional conditions that we constructed to serve as controls for 42 of the sampled conditions. If we add the 42 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%.

2.6. The materials to be tested

In the current study we decided to test 8 distinct pairs for each pairwise phenomenon. We also decided to lexically match the two items in each pair, such that most (if not all) of the contribution of lexical items to the acceptability of the sentences was simultaneously distributed across both conditions in each phenomenon (thus limiting the possibility that lexical properties could be driving the effect). For the 108 phenomena with published control sentences, we used the published pair of items as the first of the 8 pairs of items. We then constructed 7 additional, lexically matched, pairs of items ourselves for a total of 8 per phenomenon. For 6 of the 108 phenomena, the originally published pair of items was not lexically matched; however, we decided to keep the unmatched pair in the experiment, and then create 7 additional, lexically matched pairs ourselves, for a total of 8 pairs of items: 1 unmatched, 7 matched. The logic behind this choice is that by testing both the 1 unmatched pair and the 7 matched pairs that we constructed, one could in principle investigate whether the difference reported using informal methods was driven by the unmatched pair, or whether the difference also arises in the 7 matched pairs. In this way, maintaining the 7/1 split potentially provides more information about the source of the effect reported using informal methods; however, we do not present such a follow-up analysis in this article. For the 42 phenomena for which we created the control condition, all 8 pairs were lexically matched. Therefore 144 out of the 150 phenomena consisted of 8 lexically matched pairs of sentences, and 6 phenomena consisted of 7 lexically matched pairs and one (published) non-matched pair. For convenience, the originally published (sometimes unmatched) pairs of each phenomenon are presented in Appendix A as examples of the materials. The full set of materials is available on the first author's website [http://www.sprouse.uconn.edu/], along with the full set of raw results for each of the three formal judgment tasks.

2.7. The experimental methods to be compared

The terms we have been using to describe the two families of methods under investigation in this study, *informal* and *formal*, may give the impression that they differ along a single, potentially categorical, dimension. This is not true. There are a number of dimensions along which acceptability judgment methods can vary, including

- The number of participants
- The number of tokens per condition
- The number of response options available to the participants
- The linguistic training of the participants
- The quality (and quantity) of explicit instruction given to the participants
- The type of statistical analysis performed on the results

and potentially many more. Furthermore, each of these dimensions is multi-valued, rather than dichotomous, in nature. This means that there is no qualitative distinction between an informal method and a formal method. Instead, the two labels refer to general tendencies in the literature. Informal methods tend to involve relatively few participants, relatively few tokens per condition, relatively few response options, relatively expert participants (often professional linguists), relatively little explicit instruction, and relatively little statistical analysis. Formal methods tend to involve substantially more participants, substantially more tokens per condition, substantially more response options, relatively naïve non-linguist participants, substantially more instructions, and substantially more statistical analyses. As such, there is a way in which the two labels can be taken to identify two regions that represent relatively distinct locations in a multi-dimensional space. It would, of course, be ideal to test each of these dimensions independently and in various combinations; however, in order to obtain a first estimate of the convergence between methods, we will limit the current study to two regions in this multi-dimensional space: the clearly informal results reported in *LI* 2001–2010, and a series of three clearly formal experiments that themselves vary only in the specific judgment tasks used in the experiment. We decided to test three of the most common judgment tasks in the formal experimental literature: magnitude estimation (ME), 7-point Likert scale (LS), and two-alternative forced-choice (FC). Each judgment task has a slightly different set of properties that we review here.

In the ME task (Stevens, 1956; Bard et al., 1996), participants are presented with a reference sentence, called the *standard*, which is pre-assigned an acceptability rating, called the *modulus* (which we set at 100). Participants are asked to indicate the acceptability of target sentences as a multiple of the acceptability of the standard by providing a rating that is a multiple of the modulus. One of the proposed benefits of ME is that it asks participants to use the standard as a unit of measure to rate the target sentences, potentially resulting in more accurate ratings than are possible with Likert scale tasks (Stevens, 1956). Recent research suggests that this particular benefit may not hold for acceptability judgments, as participants do not appear to use the standard as a unit of measure (Sprouse, 2011b). A second possible benefit of ME concerns the continuous nature of the response scale (i.e., the positive number line), which could in principle allow

participants to distinguish finer grained differences in acceptability than the fixed response scales of Likert scale tasks. In practice, it appears that the higher degree of freedom permitted by the continuous scale results in slightly more noise in ME responses than LS responses (Weskott and Fanselow, 2011), with no noticeable difference in statistical power between ME and LS (Sprouse and Almeida, submitted for publication).

In the (7-point) LS task, each target sentence is presented with a series of 7 rating options, usually labeled 1–7, with in our case 1 additionally labeled ''least acceptable'' and 7 additionally labeled ''most acceptable''. Participants are asked to use these options to indicate their acceptability judgments. The LS task is a staple of both experimental psychology and the social sciences, as it is very intuitive for most participants. Odd numbered scales such as the one used here allow participants to easily define the most acceptable rating, the least acceptable rating, and a rating that is exactly in the middle. Although the ME task was originally intended by Stevens to supplant the LS task, the fact that participants in acceptability judgment ME tasks do not complete the task in the way envisioned by Stevens (they treat it as an open scale LS task; see Sprouse, 2011b), and the fact that LS and ME yield no difference in statistical power in syntactic experiments (see Sprouse and Almeida, submitted for publication), suggests that the LS task remains a viable alternative to ME.

In the FC task, target sentences are presented in vertically arranged pairs. Participants are asked to indicate which of the two sentences in each vertically arranged pair is more acceptable. In the current FC experiment, the pairs were lexically matched so as to form minimal pairs that varied only by the syntactic property of interest, except for the 6 original pairs that were unmatched (out of 1200 pairs). Unlike the ME and LS tasks, which ask participants to rate sentences in isolation (to later be compared numerically by the experimenter), the FC task is explicitly designed to detect differences between conditions by asking participants to make the comparison themselves. The result is often a dramatic increase in statistical power (Gigerenzer and Richter, 1990; Gigerenzer et al., 2004; Sprouse and Almeida, submitted for publication), but the cost is less information. The FC task reports only indirect information about the size of the difference between conditions in a pair (i.e., one can use the number of selections in each direction as a rough measure of effect size, but it is less sensitive than numerical ratings, cf. Myers, 2009a), and does not allow for comparisons between conditions that were never directly presented as a pair to participants.

Because each of these tasks are viable candidates for use in any given formal acceptability judgment experiment, and because each provides slightly different information that may be of interest to syntacticians, we decided to test the sample of 150 phenomena three distinct times: once each using ME, LS, and FC. For each task we recruited 312 participants on Amazon Mechanical Turk, resulting in three experiments and 936 participants. The full details of the experiments are reported in Section 3.

2.8. The statistical analysis of the formal results

There are several types of quantitative analyses available for any given set of experimental results. The choice of analysis rests upon (i) the type of information that the researcher wishes to extract from the results, and (ii) the researcher's assumptions about the experimental design and the results. For this reason we have decided to present 5 distinct quantitative analyses for each of the 3 experiments (15 analyses in total), each of which is predicated upon a different combination of information and assumptions about the experimental design and the results. Although we will present a principled argument for choosing one specific estimate in Section 4.1, it is our hope that the spectrum of analyses presented here will be useful to readers who may hold assumptions about the results that differ from our own. If the reader is interested in an analysis that is not presented here, the full set of raw results is available on the first author's website [http://www.sprouse.uconn.edu/]. The five analyses are as follows:

Descriptive directionality: In this analysis, we simply ask whether the results are in the direction reported by the informal methods in *LI*. For ME and LS, this means the difference between condition means is in the direction reported originally in *LI*; for FC, this means that the majority (>50%) of responses were in the direction reported originally in *LI*. Descriptive analyses like this do not take into account the possibility that differences between conditions could arise due to chance (e.g., sampling error), therefore this analysis is likely to be simultaneously the most sensitive and the least conservative.

One-tailed null hypothesis tests: Null hypothesis tests (NHTs) take into account the possibility that differences between conditions could arise due to chance, and allow us to make inferences about competing hypotheses. At a logical level, NHTs provide an answer to the following question: Assuming that the null hypothesis were true (i.e., that there really is no difference between conditions), how likely would the observed result (or a result more extreme) be? If the answer to this question is 'extremely unlikely', then one is entitled to conclude with some confidence that the null hypothesis is unlikely to be true. The definition of extremely unlikely is by convention less than 5% ($p < .05$) in most domains of experimental psychology. One-tailed null hypothesis tests assume that the experimental hypothesis is directional (e.g., one condition is predicted to be higher than the other). This means that the rarest 5% of results in one end of the distribution of possible results will be considered *significant* (or, the critical region), but the rarest results in the other end

of the distribution will not. Because all 5% of the critical region is located in one end of the distribution, one-tailed NHTs are more sensitive than two-tailed NHTs. For ME and LS, we ran one-tailed *t*-tests; for FC, we ran one-tailed sign tests; all tests were repeated measures.

Two-tailed null hypothesis tests: Two-tailed NHTs are identical to one-tailed NHTs in both basic logic and calculation, but two-tailed NHTs do not assume a directional experimental hypothesis. Instead, two-tailed NHTs divide the critical region between the two extreme ends of the distribution, such that results in either direction can be considered significant. Because (by convention) only 5% of possible results are considered significant, this means that the two critical regions each contain 2.5% of possible results. In other words, 'extremely unlikely' in two-tailed results is defined as the most extreme 2.5% in each direction. In this way two-tailed NHTs are less sensitive than one-tailed NHTs, but they provide potentially more information in cases where the predicted directionality of results is reversed. For ME and LS, we ran two-tailed *t*-tests; for FC, we ran two-tailed sign tests; all tests were repeated measures.

Mixed effects models: Traditional NHTs assume that participants are randomly sampled from a larger population, and therefore participants must be treated mathematically as a *random factor*. Some researchers have argued that items in language experiments are also randomly sampled from a larger population, and therefore items should also be treated mathematically as a random factor (Clark, 1973). The concern is that if items are indeed randomly chosen from a larger population and not treated as a random factor (instead treated as a *fixed factor*, as they are in traditional NHTs), then there is an increased risk of a false positive result (because the item variation is contributing to the difference between conditions but not being accounted for in the calculation of the test statistic). Modern mixed effects models, unlike the ones used in traditional NHTs, allow one to specify crossed random effects, thus correcting for this potential problem. The result is a lower risk of false positive results. The risk with treating items as random effects is that if the items were not sampled randomly from a population, as other researchers have argued is true for many types of language experiments, treating them as random will result in less statistical power, thereby creating a greater risk of false negative results (Wike and Church, 1976; Cohen, 1976; Keppel, 1976; Smith, 1976; Wickens and Keppel, 1983; Raaijmakers et al., 1999; Raaijmakers, 2003). Although we believe that the current experiments do not require items to be treated as random effects (because the different lexicalizations were not randomly sampled, but rather were instead carefully created to be representative of the conditions of interest, and because the items were lexically matched across conditions), we nonetheless constructed linear mixed effects models treating both participants and items as crossed random effects for the ME and LS experiments, and simulated *p*-values using the languageR package (Baayen, 2007; Baayen et al., 2008). For the FC experiment, we constructed logistic mixed effect models (mixed logit models) treating participants and items as random effects, and report the *p*-values returned by the lme4 package (Bates et al., 2012; see also Jaeger, 2008).

Bayes factor analyses: Whereas NHTs assume that the null hypothesis is true and then ask what the probability is of obtaining the observed result (or a result more extreme), Bayesian approaches to statistical analysis use a logic that is in many ways better aligned with the goals of most scientists: Bayesian approaches assume that the observed results are true of the world, and ask how likely a given hypothesis would be under that assumption (e.g., Gallistel, 2009; Kruschke, 2011; and for accessible reviews of the controversies surrounding NHT, see Shaver, 1993; Cohen, 1994; Nickerson, 2000; Balluerka et al., 2005; Hubbard and Lindsay, 2008). One particularly popular type of Bayesian analysis is to calculate a proportion known as a *Bayes factor*, which simply reports the odds of one hypothesis over another given the experimental results. For example, a Bayes factor of 4 would indicate that the experimental hypothesis is four times more likely than the null hypothesis based on the experimental results. Conversely, a Bayes factor of .25 would indicate that the null hypothesis is four times more likely than the experimental hypothesis. For the ME and LS results, we used the JSZ Bayes factor equation from Rouder et al. (2009), which assumes (i) a non-directional H1 (equivalent to a two-tailed NHT), and (ii) an equal prior probability of the two hypotheses. For the FC results, we used the Bayes factor equation for binomial responses made available by Jeff Rouder on his website: http://pcl.missouri.edu/bayesfactor. Much like mixed effects models that treat items as random factors, Bayes factor analyses tend to return fewer significant results than standard NHT models; it is an empirical question whether this represents a decrease in false positives or an increase in false negatives.

3. The experiments

3.1. Division into nine sub-experiments

As discussed in Section 2, the full test sample consists of 300 conditions that form 150 pairwise phenomena. This means that in order to have a repeated-measures design in which each participant rates each condition once, the three

primary experiments (ME, LS, and FC) would each be 300 sentences long. As a general rule, we prefer to keep the length of acceptability judgment experiments to approximately 100 sentences in order to minimize fatigue-based artifacts. In order to meet this length constraint in a repeated-measures design, we split the 150 phenomena among three sub-experiments: 50 per sub-experiment. The distribution of the phenomena among the sub-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. The same division into three sub-experiments was used for all three primary experiments (ME, LS, and FC), resulting in a total of nine sub-experiments. Because the pairwise phenomena consist of two items, one more acceptable according to informal methods and one less unacceptable according to informal methods, the distribution of acceptable and unacceptable items in every resulting survey was (by hypothesis) balanced.

3.2. Participants

A total of 936 participants were recruited for the present study: 312 per primary experiment (ME, LS, and FC), or 104 per sub-experiment (as per Section 3.1). This means that we collected 104 ratings per condition per task (ME, LS, and FC). Participants were recruited online using the Amazon Mechanical Turk (AMT) marketplace, and paid $2.50 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. 8 participants were removed from the ME results, 8 participants were removed from the LS results, and 5 participants were removed from the FC results for answering ''no'', or failing to answer, one or both of these questions. No response-based outlier removal was performed on the results.

3.3. Materials

For the ME and LS experiments, the 8 items per condition were distributed among eight lists using a Latin Square procedure. Each list was pseudorandomized such that members of the same pairwise phenomenon did not appear sequentially. This resulted in eight surveys per sub-experiment of 100 pseudorandomized items. Six additional ''anchoring'' items (two each of acceptable, unacceptable, and moderate acceptability) were placed as the first six items of each survey. These items were identical, and presented in the identical order, for every survey. Participants rated these items just like the others; they were not marked as distinct from the rest of the survey in any way. However, these items were not included in the analysis as they served simply to expose each participant to a wide range of acceptability prior to rating the experimental items (a type of unannounced ''practice''). This resulted in eight surveys per sub-experiment that were 106 items long. Each survey contained only one token of each condition (i.e., 100 distinct sentence types), meaning that each participant rated each condition in their sub-experiment only once, and that surveys contained the maximal amount of structural and lexical variation possible in a 100 item survey. The lack of repetition of the conditions eliminated any risk of priming effects or response strategies, thus eliminating the need for inflating the length of the surveys with unanalyzed filler items.

For the FC experiment, the 8 pairs of lexically matched items per phenomenon were distributed among the 8 lists as matched pairs, such that both members of a lexically matched pair appeared in the same list. This ensures that the choice within each pair is not influenced by lexically-based variation, thus increasing the likelihood that the choice is predicated upon the structural manipulation of interest. Next, the order of presentation of each pair was counterbalanced across the lists, such that for every pair, four of the lists included one order, and four lists included the other order. This minimized the effect of response biases on the results (e.g., a strategy of 'always choose the first item'). Finally, the order of the pairs in each list was randomized, resulting in 8 surveys containing 50 randomized and counterbalanced pairs (100 total sentences).

3.4. Presentation

For the ME experiment, participants were first asked to complete a practice phase in which they rated the lengths of 6 horizontal lines on the screen prior to the sentence rating task in order to familiarize them with the ME task itself. After this initial practice phase, participants were told that this procedure can be easily extended to sentences. No explicit practice phase for sentences was provided; however, the six unmarked anchor items did serve as a sort of unannounced sentence practice. There was also no explicit practice for the LS and FC experiments, as these tasks are generally considered relatively intuitive. The surveys were advertised on the Amazon Mechanical Turk website, and presented as web-based surveys using an HTML template (including task instructions) available on the first author's website [http://www.sprouse.uconn.edu/]. Participants completed the surveys at their own pace.

3.5. Results

After the experiments were conducted, we discovered that two pairwise phenomena were repeats of others included in the study. These phenomena were excluded from all subsequent analyses, leaving 148 pairwise phenomena (296 sentence types). For the ME and LS experiments, ratings from each participant were *z*-score transformed prior to analysis to eliminate some of the forms of scale bias that potentially arise with rating tasks (see Schütze and Sprouse, 2013 for a review). The *z*-score transformed results were then analyzed using the 5 analyses described in Section 2.8: descriptive directionality, one-tailed *t*-tests, two-tailed *t*-tests, linear mixed effects (LME) models, and Bayes factors. The FC results were converted into *successes* and *failures* at the pair level for the descriptive directionality analysis, one-tailed sign test, two-tailed sign test, and Bayes factor analysis; the FC results were converted into 0 and 1 notation at the item-level for the mixed logit (ML) models. The results of each type of analysis are presented in Tables 2–5. Table 6 contains the crucial convergence rates, and Appendix B presents the results of each individual analysis.

Table 2
Descriptive analysis of the directionality of the responses. For ME and LS, these counts are based on the difference between means for each phenomenon. For FC, these counts are based on the difference between the number of choices in each direction.

Task	Predicted direction	Opposite direction
ME	146	2
LS	143	5
FC	144	4

Table 3
Categorized results of statistical tests for ME. Significant *p*-values are defined at $p < .05$ in each direction; marginal *p*-values are defined at $p \leq .1$ in each direction. Significant Bayes factors are defined at BF > 3 in each direction; marginal Bayes factors are defined at BF > 1 in each direction.

	One-tailed	Two-tailed	LME	Bayes factor
Significant in the opposite direction	–	2	2	2
Marginal in the opposite direction	–	0	0	0
Non-significant in the opposite direction	–	0	0	0
Non-significant in the predicted direction	10	9	16	13
Marginal in the predicted direction	1	1	3	2
Significant in the predicted direction	137	136	127	131

Table 4
Categorized results of statistical tests for LS. Significant *p*-values are defined at $p < .05$ in each direction; marginal *p*-values are defined at $p \leq .1$ in each direction. Significant Bayes factors are defined at BF > 3 in each direction; marginal Bayes factors are defined at BF > 1 in each direction.

	One-tailed	Two-tailed	LME	Bayes factor
Significant in the opposite direction	–	2	2	2
Marginal in the opposite direction	–	0	0	0
Non-significant in the opposite direction	–	3	3	3
Non-significant in the predicted direction	11	6	10	10
Marginal in the predicted direction	0	0	3	1
Significant in the predicted direction	137	137	130	132

Table 5
Categorized results of statistical tests for FC. Significant *p*-values are defined at $p < .05$ in each direction; marginal *p*-values are defined at $p \leq .1$ in each direction. Significant Bayes factors are defined at BF > 3 in each direction; marginal Bayes factors are defined at BF > 1 in each direction.

	One-tailed	Two-tailed	ML	Bayes factor
Significant in the opposite direction	–	3	4	3
Marginal in the opposite direction	–	1	0	0
Non-significant in the opposite direction	–	0	0	1
Non-significant in the predicted direction	7	4	3	5
Marginal in the predicted direction	2	1	1	0
Significant in the predicted direction	139	139	140	139

Table 6
Convergence rates (in percentage) between each analysis and the informal results reported in *Linguistic Inquiry* 2001–2010. In cells with slashes (/) the percentage on the left assumes that marginal results are non-significant; the percentage on the right assumes that marginal results are significant. All rates are estimates based on random sampling, resulting in a margin of error of 5.3–5.8%.

Task	Directionality	One-tailed	Two-tailed	LME/ML	Bayes factor
ME	99	93	92/93	86/88	89/90
LS	97	93	93	88/90	89/90
FC	97	94/95	94/95	95	94

3.6. Two additional experiments

Before moving to the discussion of the results of the three primary experiments, we should mention that we have run two supplementary experiments on these phenomena, described in a previous manuscript that is publicly available (http://ling.auf.net/lingBuzz/001352). The first supplementary experiment used the ME task to test 144 of the 148 phenomena tested here with a sample size of 168 participants, 56 per 100 item survey (about half of the sample size used here). The second supplementary experiment re-tested the divergent results between informal and formal methods from the first supplementary experiment using the more powerful FC task and a larger sample size (96 participants) to derive a convergence rate that is less likely to be contaminated by false negatives. The resulting combined convergence rate was 95%, directly in line with the convergence rate derived using the FC task here. Therefore, the primary results reported here and the supplementary results previously described serve as substantial replications of each other, providing additional confidence in the results. We have chosen to focus on the three experiments reported in this study for space reasons, as they provide slightly more detailed information than the two supplementary experiments (because of larger sample sizes and because the supplementary FC experiment did not test all of the phenomena in the first test set).

4. The convergence rate

The central question at hand is to what extent informal and formal methods yield convergent results, in this case defined over pairwise phenomena. The current study shows that the results of informal and formal methods are not identical, and suggests that the number of divergences is between 1% and 14% (±5.3–5.8%) of the phenomena published in *LI* between 2001 and 2010. The two questions we would like to discuss in this section are (i) whether we can choose a more precise point estimate of the convergence rate in a principled manner, and (ii) whether the resulting convergence rate can be interpreted as relatively high or relatively low in a similarly principled manner. Anticipating the discussion in Section 4.3, we wish to stress that the convergence rates we have observed carry no information concerning which method is superior (if that question even has a general answer).

4.1. Selecting a point-estimate

In the previous section we presented a range of possible convergence estimates based on three distinct judgment tasks and five distinct quantitative analyses: from 86% on the low end to 99% on the high end, with a margin of error of 5.3–5.8% for each estimate. Although we offer these various estimates as a convenience to readers, we do believe that there are principled reasons to prefer certain analyses to others. This is because some of the differences between the convergence rates appear to reflect well-known statistical properties of the tasks and statistical tests themselves. For example, previous work has suggested that the FC task leads to the most statistically powerful experiments in pairwise comparisons, with LS and ME roughly equivalent to each other but less powerful than FC experiments (Sprouse and Almeida, submitted for publication). This appears to be reflected in the convergence rates obtained in this study: FC generally leads to the highest convergence rates, with LS and ME leading to lower convergence rates. Therefore the FC task is probably a more sound choice for this type of study than the ME and LS tasks. Similarly, linear mixed effects/mixed logit models and Bayes factor analyses are known to lead to fewer statistically significant results than traditional frequentist tests. This decrease in significant results could either reflect an increase in accuracy (i.e., the phenomena in question were false positives and are now correctly recognized as true negatives), or a decrease in statistical power (i.e., the phenomena in question were true positives but are now false negatives, as discussed in Section 2.8). Therefore these analyses should only be preferred if there is reason to believe that they represent the former rather than the latter.

The considerations above lead us to believe that the convergence rate estimate yielded by the FC task analyzed using the mixed logit model is likely the most accurate estimate. Our rationale is as follows. First, the logic of the FC task most closely mirrors the logic of the collection of data underlying syntactic theories, as participants are generally asked to

identify a difference between two (or potentially more) maximally similar sentences. Second, previous research has suggested that the FC task is under some circumstances the most sensitive task for the detection of differences between conditions (Gigerenzer and Richter, 1990; Gigerenzer et al., 2004; Sprouse and Almeida, submitted for publication), suggesting that it will result in the fewest false negatives. Finally, the mixed logit models constructed here treat items as random effects, which has been argued by some to lead to a more conservative false positive rate (e.g., Jaeger, 2008; but see the discussion in Section 2.8). The mixed logit models for the FC task yield the same convergence rate as one-tailed sign tests, suggesting that their use incurred no loss of statistical power, while simultaneously maintaining the potential protection against false positives that advocates of mixed effects models have stressed. For these reasons, we believe that the most accurate convergence rate estimate between informal and formal methods for the syntactic phenomena explored in *LI* during the years 2001–2010 that can be derived from the results of this study is 95 ± 5.3–5.8%.

4.2. Evaluating the point-estimate

The inferential value of the convergence rate for broader methodological questions hinges on whether the convergence rate is high or low. If the convergence rate is high, then there is comparatively less at stake in the choice between methods than if the convergence rate is low. Unfortunately, there is no explicit discussion of what would be considered a high (or low) convergence rate in the existing literature; different researchers are likely to reach different conclusions. Nonetheless, we believe it is possible to make a general case for considering the 95% convergence rate high. The field of experimental psychology has, by consensus, signaled a willingness to tolerate a divergence of 5% over the long run between the decision to classify differences as statistically significant and whether there is a real difference between the conditions. This follows from the consensus to set the decision criterion for statistical significance (the alpha level in the Neyman–Pearson framework) at .05. The alpha level represents the maximum long-run frequency of incorrect decisions to reject the null hypothesis when the null hypothesis should not be rejected (also known as Type I errors). This suggests that 5% is considered small, or at least tolerable, as a divergence rate between the results of statistical tests and the true status of the world.

To be clear, we are not suggesting that syntacticians should be satisfied with a 5% divergence rate, either for statistical significance testing or for a comparison of the results of acceptability judgment methods. For example, it is possible to set the decision criterion for statistical significance (the alpha level) lower, perhaps to .01. The cost of such a move is that there is a direct (inverse) relationship between the risk of incorrectly rejecting the null hypothesis (the alpha level) and the risk of incorrectly failing to reject the null hypothesis (the beta level). In other words, being stricter about statistical significance entails sacrificing statistical power, all else equal. The .05 alpha level represents a consensus balance between these two risks. Similarly, syntacticians could decide that a 5% divergence rate between informal and formal methods is too high, and therefore decide to systematically determine which method maximizes the detection of real differences, and minimizes false alarms. The answer to such a question might well be different for different types of linguistic phenomena, and will most likely require a series of specially constructed follow-up studies, which we discuss in principle in Section 5. Our only goal in this section is to note that there is a consensus opinion that 5% is a tolerable divergence rate in statistical significance testing, so this is a reasonable starting point for the current discussion.

4.3. The inferential limits of the convergence rate

Although we believe that estimating a convergence rate between informal and formal methods is a good first step toward understanding the methodological decisions facing the field, and perhaps toward resolving the debate that has been playing out in the literature for more than 40 years, it is also important to be clear about the inferential limits of the convergence rate. The convergence rate allows us to estimate how different the results published in *LI* would be if the field switched wholesale from (nearly) exclusively informal methods to exclusively formal methods. This is useful information, as it means that future iterations of the debate between methods must acknowledge that the empirical scope of the debate is relatively small. However, this does not directly resolve the debate, as one of the driving questions is whether formal methods are (universally) superior to informal methods. The convergence rate provides no information that bears on this question. In fact, no simple comparison of informal and formal results can ever bear on this question. Assuming there are only two types of results (i.e., the two sentences of a pairwise phenomenon are significantly different or not), if the two methods converge, then this tells us that either the results of both methods are correct or the results of both methods are incorrect, but not which of the two scenarios is true. If the two methods diverge then this tells us that one method's results are correct and the other's are incorrect, but not which is which. This is a fundamental limitation of every simple comparison study.

This is not to say that there are no methods for addressing the superiority question. It is just that there is no general method (like a large-scale comparison). Every phenomenon must be investigated separately, with the details of that investigation depending on the specific properties of the phenomenon. In the case of divergent results between methods,

the first step would be to list all of the possible confounds that could have affected each method separately. For example, Ferreira (2005) and Gibson and Fedorenko (2010, 2013) follow earlier literature (e.g., Greenbaum, 1973; Spencer, 1973) in suggesting that knowledge of syntactic theories might influence professional linguists when they provide judgments during informal collection. This then would be a possible confound for informal methods. On the other hand, Newmeyer (1983) has suggested that professional linguists may be better able to distinguish acceptability effects driven by extra-grammatical factors such as plausibility or word frequency from acceptability effects driven by grammatical factors. This then would be a possible confound for formal methods that rely on non-linguist participants. One could then design a series of studies that manipulate the level of syntactic knowledge of the participants (and thus by hypothesis their ability to discriminate between extra-grammatical and grammatical effects on acceptability). If the manipulation leads to a convergent set of results, then that would be evidence that the factor being manipulated was the source of the initial divergence.

It should be clear from this mini-example that follow-up studies of this sort will be resource-intensive: the space of possible confounds is large, and the manipulations involved may require sampling from diverse populations of participants with very specific properties. To our knowledge, no follow-up studies of this sort have been run to test the divergent results that have been reported in the literature. Convergent results present a similar sort of problem. The fact that informal and formal methods converge on the same result either means that both methods yield the correct result, or it means that both methods are affected by confounds that lead them to simultaneously yield an incorrect result. Once again, the general method would be to list potential confounds for each method, and then manipulate those confounds to see if the results could be changed. The difference in the case of convergent methods is that one would be searching for a confound or confounds that change the polarity of the results of both methods, so that they are still convergent. Again, to our knowledge, no such follow-up studies have been run for convergent results in the literature. Instead, it is common to assume that if the two methods converge, the result must be correct, but this is not a logical necessity.

In short, simple convergence estimates cannot bear on the superiority question. One can make assumptions that will convert the convergence rate into an argument for one method over the other, but in all cases, that is simply begging the question. If one assumes that formal methods are always correct, then the null results obtained in the formal experiments reported here (i.e., the results that did not return a significant difference) suggest a lower bound on the false positive rate (Type I error rate) for informal methods. If one assumes that informal methods are always correct, then the null results obtained in the formal experiments reported here suggest a lower bound on the false negative rate (Type II error rate) for formal methods. If one assumes that the 95% of phenomena that converge are correct results, but neither method is 100% correct, then the 5% divergence represents a mixture of false positives and false negatives for the two methods. Crucially, if one makes no assumptions whatsoever, then the false positive and false negative rates for both methods remain completely unknown.

5. How to move the conversation forward

The present study is the first large-scale comparison of informal and formal acceptability judgment collection methods using randomly sampled phenomena from the cutting edge of syntactic theory. The results suggest that the differences between the two methods are relatively small, with a convergence rate of 95 ± 5.3–5.8%. Although this is a substantial new piece of information in its own right, in this section we would like to highlight additional kinds of information about the collection of acceptability judgments that could be useful for moving the methodological conversation forward.

5.1. Other kinds of judgments

One obvious future direction of investigation is to explore the convergence between informal and formal methods for judgment types other than standard acceptability judgments, which make up only approximately 48% of the data points in *LI* 2001–2010. As detailed in Section 2.2, these other types include interpretation judgments, coreference judgments, and judgments involving prosodic manipulations. Clearly these will involve experimental techniques more complex than those used in the current study.

5.2. The consistency of formal tasks for the divergent phenomena

The three experiments in this study have yielded three sets of divergent results: informal vs. ME, informal vs. LS, and informal vs. FC. The status of these phenomena is currently unresolved, as we have no way of measuring acceptability outside of acceptability judgment tasks, and in this case, our two methods yield different results. The ideal scenario is to conduct in-depth follow-up studies to probe the various dimensions of the two methods that could lead to the divergent results. For practical reasons we leave the detailed follow-up studies to future research; however, this does not mean that

Table 7
The eight phenomena that led to divergent results between informal and formal methods based on the forced-choice task and a mixed logit analysis. We report the mixed logit analysis for FC, and the linear mixed effects models and two-tailed *t*-tests for ME and LS, where 0 indicates a null result ($p > .1$), − indicates a sign-reversal ($p < .05$), + indicates a significant result in the same direction as the informal result ($p < .05$), and parentheses indicate a marginal effect ($.05 \leq p \leq .1$) in the direction indicated by the symbol inside the parentheses. For convenience, all divergent analyses are shaded. All statistical tests are two-tailed. A lowercase *g* in the items column indicates that we created the control condition based on the discussion in the text.

Year	(First) Author	Items	FC	ME		LS	
			ML	LME	*t*-Test	LME	*t*-Test
2004	Hazout	67c/67a	−	−	−	−	−
2003	Phillips	93b/92b	−	−	−	(−)	−
2002	Fox	69a/69b	0	0	0	0	0
2004	Richards	17b/17a	0	0	0	0	0
2001	López	10a/9a	−	0	0	0	0
2004	Bhatt	94a/94b	−	0	+	0	0
2010	Haegeman	18a/g	0	0	0	+	+
2003	Bošković	3e/4e	(+)	0	0	0	0

Table 8
Originally published example sentences and judgment diacritics for the eight phenomena that led to divergent results between informal and formal methods based on the forced-choice task and the mixed logit analysis. A lowercase *g* in the item column indicates that we created the control condition based on the discussion in the text.

Year	(First) Author	Item	Example Sentence
2004	Hazout	67c	*There is likely a man to appear.
		67a	There is likely to appear a man.
2003	Phillips	93b	?*Wallace stood more buckets in the garage than Gromit did in the basement.
		92b	Wallace stood more buckets than Gromit did in the garage.
2002	Fox	69a	*John wants for everyone you do to have fun.
		69b	John wants for everyone to have fun that you do.
2004	Richards	17b	*To whom did you give what?
		17a	What did you give to whom?
2001	López	10a	*We proclaimed to the public John to be a hero.
		9a	We proclaimed John to the public to be a hero.
2004	Bhatt	94a	*I expect that everyone you do will visit Mary.
		94b	I expect that everyone will visit Mary that you do.
2010	Haegeman	18a	*Bill asked if such books John only reads at home.
		g	Bill knows that such books John only reads at home.
2003	Bošković	3e	*John likes Mary Jane didn't believe.
		4e	That John likes Mary Jane didn't believe.

there is no information about these phenomena to be gleaned from the current results. One obvious question we can ask is how consistent the set of divergent phenomena returned by the FC task (i.e., the most statistically powerful task) is with the results of the other two formal tasks. Consistency across the formal tasks and statistical analyses would provide strong evidence that these phenomena lead to predictable differences between informal and formal methods, and therefore deserve further scrutiny as potentially critical phenomena for choosing between methods. There were eight phenomena that failed to replicate significantly in the predicted direction in the FC experiment as analyzed using mixed logit models (the analysis that we believe is most appropriate; see Section 4.1). Table 7 reports these eight phenomena along with the directionality of the statistical analyses for each task.

Table 8 presents the originally published example sentences and judgment diacritics. Of these eight phenomena, four show a consistency within the results of the three formal tasks that suggests a strong contrast between informal and formal methods. Of these four, two returned a significant effect in all three formal tasks, but that effect was in the opposite direction from the effect reported using informal methods (this is sometimes known as a sign-reversal). The other two returned a non-significant result (a null result) in all three formal tasks, despite being reported as different using informal methods. (The four remaining phenomena did not show enough consistency within the formal tasks to suggest a strong

contrast between informal and formal methods, although two of them are in line with the claim that FC is simply a more sensitive task.) Although the consistency among the first four phenomena is tantalizing, it should be noted that the cause of this consistency is as yet unknown. The consistency could represent a problem with the informal judgments reported in *LI*, or it could represent a systematic problem with the formal judgments collected (e.g., the non-linguist participants, lacking information about the intended prosody of a sentence, might systematically mis-parse it or fail to parse it at all). Only in-depth follow-up studies like those suggested in Section 4.3 can resolve these questions.

5.3. Controlled comparisons of categorized and continuous judgments for individual sentence types

As previously mentioned in Section 2.3, to the extent that raw ratings of individual sentences are used as data points in the construction of syntactic theories (as opposed to pairwise phenomena), it might be informative to compare the categorized ratings for individual sentence types returned by informal methods with the continuous ratings for individual sentence types returned by formal methods like ME and LS (the LS task is not inherently continuous like ME, but the *z*-score transformation converts the discrete finite scale to an infinite continuous scale). Fig. 1 below provides an idea as to what such a comparison would look like, using the ME and LS results from the present studies.

In Fig. 1, each sentence type is arranged along the *x*-axis in ascending order according to the mean rating obtained from the ME and LS experiments. The informal categorized rating reported for each sentence type in *LI* is encoded by the upper versus lower panel in each graph. We restricted this figure to sentence types with an asterisk and no other marking (unacceptable) or no diacritics (acceptable) in order to simplify the presentation; those with question marks (possibly combined with an asterisk) were left aside (a total of 13 sentence types). We can then attempt to locate a threshold that maximizes the separation of the two sets of sentence types into two categories: acceptable sentences above the threshold and unacceptable sentences below the threshold. Any sentence types that appear on the wrong side of this threshold would be considered divergences. Table 9 presents the results of this analysis by listing the minimum number of sentence types that could be classified as divergences for each task (LS and ME), along with the thresholds (*z*-score ratings) that are necessary to achieve these minima, and the counts for each type of divergence (i.e., the number of

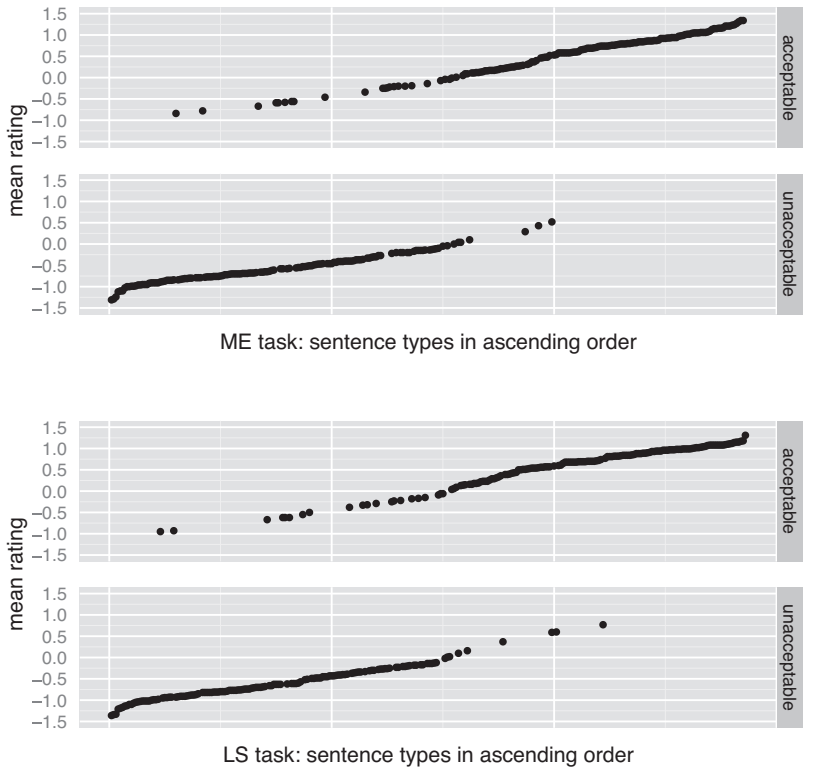

Fig. 1. A comparison of informal categorized ratings published in *LI* 2001–2010 with formal, *z*-score transformed, continuous ratings from the ME and LS tasks. Only sentences with an asterisk and no other marking (labeled ''unacceptable'', lower panels) or no diacritics (labeled ''acceptable'', upper panels) were included. Sentence types are arranged along the *x*-axis in increasing order of acceptability rating.

Table 9
Summary of divergent results in the data from Fig. 1, based on thresholds that minimize divergences.

Task	Minimum number of divergences	Thresholds that achieve the minimum (in mean *z*-scores)	Count of no-diacritic items below threshold	Count of asterisked items above threshold
ME	28	−.08	19	9
		−.04	20	8
		.04	24	4
LS	27	−.10	18	9
		.03	21	6

acceptable sentence types that are errantly below the threshold and the number of unacceptable sentence types that are errantly above the threshold).

The first question we can ask is what the divergence rate would be under this threshold analysis. With 283 sentence types included in the analysis (296 – 13), the ME divergence rate would be 9.8% and the LS divergence rate would be 9.5%. This is roughly in line with the pairwise divergence rates for two-tailed null hypothesis tests (7% or 8% for both LS and ME, depending on how marginal results are counted) and Bayes factor analyses (10% or 11% for both LS and ME, again depending on how marginal results are counted). The second question we can ask is which sentence types were divergent under this threshold analysis. Table 10 lists all of the divergent sentence types for the first threshold listed for each task in Table 9, along with a summary of the results for the pairwise phenomena that each sentence type participated in under the main analysis presented in Section 3.

The first property of Table 10 that may be of interest is how consistent the divergent phenomena were across the ME and LS tasks. There are 25 sentence types that are divergent for both the ME and LS task. There are only 2 divergent sentence types for LS that were not divergent for ME, and only 3 divergent sentence types for ME that were not divergent for LS. This is perhaps unsurprising given the similarities between the two tasks (Weskott and Fanselow, 2011; Sprouse, 2011b). The second property of Table 10 that may be of interest is how consistent the divergences in the threshold analysis are with the divergences in the pairwise analyses. 17 of the 30 divergent sentence types participated in a pairwise phenomenon that was itself divergent under at least one task and statistical analysis (i.e., at least one cell on the right hand side is not a +). 11 of the divergent sentence types participated in a pairwise phenomenon that was itself divergent under 4 or more of the 6 two-tailed NHT analyses performed. Although it is difficult to draw any firm conclusions from this consistency, it at least suggests that detailed follow-up analyses on these phenomena could be enlightening. Note also that all 8 of the divergent pairwise phenomena analyzed in Tables 7 and 8 are represented in Table 10.

The results of the thresholding analysis above are generally in line with the results of the pairwise analysis presented in Section 3. However, we have chosen not to pursue the former as the primary analysis in this paper because the categorized ratings taken from *LI* and the continuous ratings obtained in the current experiments may not be directly comparable. The problem is that different raters tend to use rating scales differently. Some raters might use a wider or narrower range of ratings (scale expansion/compression); some raters might use ratings on one side of the scale or the other (scale bias); and some raters might postulate different boundaries between ratings on the scale. There are ways to minimize such variability through experimental design, e.g., by balancing the distribution of items across the scale, and there are ways to eliminate some of this variability through data analysis, such as the *z*-score transformation applied to the results of the ME and LS experiments reported in Section 3. Although we applied strategies to counteract such differences in calculating the ME and LS means used in the thresholding analysis, it is not possible to apply these strategies to the categorized ratings taken from *LI*. This means that the results from *LI* may contain some amount of uncorrected scale variability, such that the ratings from one author may be misaligned with the ratings from another author with respect to the category boundaries represented by the different diacritics. This variability would surface as sentence types appearing on the wrong side of the threshold, which means that scale variability and divergent results will look identical, potentially misidentifying (and overestimating) the number of divergences.

If one wished to derive a cleaner comparison of the categorized ratings from informal methods with the continuous ratings from formal methods, one would want to introduce the strategies mentioned above for minimizing scale bias (balanced designs and mathematical transformations) to informal methods. For example, one could ask a group of linguists to each rate the same set of sentence types, as is done with non-linguists in formal methods. If the set of sentence types were well-balanced, and if every linguist rated the same set of sentence types, then the full range of variability-minimizing techniques would be available during data analysis. However, that would no longer be an investigation of the informal ratings reported in the literature, but rather of the ratings of linguists under semi-formal circumstances (which the current experiments were not designed to collect).

Table 10
The divergent items from the threshold analysis. On the left hand side items are identified by the code VOLUME.ISSUE.FIRST-AUTHOR.EXAMPLE.JUDGMENT, where "g" stands for grammatical (no diacritics). The right hand side indicates whether the divergent items in the threshold analysis participated in a divergent result in the pairwise analysis presented in Section 3. Items that participated in pairwise phenomena that replicated in the correct direction are marked with a +. Items that participated in a pairwise null result are marked with a 0. Items that participated in a pairwise sign reversal are marked with a −. Items that participated in marginal results in either direction are marked with parentheses. For convenience, all divergent analyses are shaded. All statistical tests are two-tailed.

No-diacritic items below the threshold		Results from pairwise analyses					
ME	LS	ME		LS		FC	
		LME	*t*-Test	LME	*t*-Test	ML	Sign
32.1.Martin.77.g	–	+	+	+	+	+	+
32.3.Fanselow.58c.g	32.3.Fanselow.58c.g	+	+	+	+	+	+
32.4.López.9a.g	32.4.López.9a.g	0	0	0	0	–	–
33.1.Fox.49b.g	33.1.Fox.49b.g	+	+	+	+	+	+
33.1.Fox.69b.g	33.1.Fox.69b.g	0	0	0	0	0	0
34.1.Phillips.96a.g	34.1.Phillips.96a.g	0	+	+	+	+	+
34.1.Phillips.92b.g	–	–	–	(–)	–	–	–
34.4.Bošković.4c.g	34.4.Bošković.4c.g	0	+	0	+	+	+
34.4.Bošković.4d.g	34.4.Bošković.4d.g	0	0	0	0	+	+
34.4.Bošković.4e.g	34.4.Bošković.4e.g	0	0	0	0	(+)	0
35.1.Bhatt.94b.g	35.1.Bhatt.94b.g	0	+	0	0	(–)	0
35.3.Hazout.67a.g	35.3.Hazout.67a.g	–	–	–	–	–	–
38.2.Hornstein.4b.g	38.2.Hornstein.4b.g	+	+	+	+	+	+
–	38.3.Haddican.39.g	+	+	0	(+)	+	+
41.1.Müller.14c.g	41.1.Müller.14c.g	+	+	+	+	+	+
41.3.Landau.10a.g	41.3.Landau.10a.g	0	0	0	0	+	+
41.3.Rezac.3b1.g	41.3.Rezac.3b1.g	+	+	+	+	+	+
41.4.Bruening.9c.g	41.4.Bruening.9c.g	0	(+)	0	0	+	+
41.4.Haegeman.18a.g	41.4.Haegeman.18a.g	0	0	+	+	0	0
41.4.Haegeman.4c.g	41.4.Haegeman.4c.g	+	+	+	+	+	+

Asterisked items above the threshold		Results from pairwise analyses					
ME	LS	ME		LS		FC	
		LME	*t*-Test	LME	*t*-Test	ML	Sign
32.1.Martin.65b.*	32.1.Martin.65b.*	+	+	+	+	+	+
32.2.Stroik.4b.*	32.2.Stroik.4b.*	+	+	0	+	0	+
33.1.den Dikken.5b.*	33.1.den Dikken.5b.*	+	+	+	+	+	+
33.2.Bowers.7b.*	33.2.Bowers.7b.*	+	+	0	+	+	+
34.1.Fox.26.*	–	+	+	+	+	+	+
34.4.Bošković.3a.*	34.4.Bošković.3a.*	+	+	0	0	+	+
34.4.Haegeman.2a.*	34.4.Haegeman.2a.*	+	+	+	+	+	+
35.3.Richards.17b.*	35.3.Richards.17b.*	0	0	0	0	0	0
–	38.4.Kallulli.9b.*	+	+	+	+	+	+
38.4.Kallulli.10b.*	38.4.Kallulli.10b.*	+	+	+	+	+	+

5.4. Finer-grained comparisons of dimensions along which informal and formal methods differ

As briefly mentioned in Section 2.7, the labels *informal* and *formal* are simply convenient idealizations of data collection tendencies in the syntax literature. There are a number of multi-valued dimensions along which acceptability judgment methods can vary. A comprehensive investigation of judgment methods will require a large-scale effort to vary each of these dimensions independently, across the range of their potential values. We hope that the results of the present experiments will provide a useful starting point for this investigation. We would also suggest that a useful starting point might be to design survey studies to determine how much variability there is in the methods routinely deployed by professional linguists. Because informal methods, by their very nature, do not involve any reports of the data collection technique, it is difficult to engage in discussions of the "typical" data collection method. It is not uncommon for critics of informal methods to claim that linguists consult only one participant for judgments (the linguist herself), and use only one item per sentence type (the examples published in the journal article) (e.g., Gibson and Fedorenko, 2013). In our

experience, linguists tend to use many more participants and many more items than these reports suggest; however, it is an empirical question exactly how much variation there is in informal methods.

5.5. Finer-grained classifications of data points

There are finer-grained distinctions between data types that may be relevant for a comprehensive picture of judgment methodologies. One such distinction is the evidential value of each data point. In all sciences, some data points have more evidential value than others. These evidential differences can arise for any number of reasons, from the fact that different theories might have overlapping empirical coverage (thus increasing the value of data points that are captured by only one theory), to the fact that the empirical domain of each theory is determined by the scientist. Now that the general convergence of the two methods has been established (through random sampling), it might be interesting to use evidential value as a finer-grained distinction in future studies to better quantify the effect that method choice could have on syntactic theories. For example, Colin Phillips (personal communication) notes that the central analysis of Phillips (2003) could be bolstered by the sign-reversal obtained in these experiments for his examples 92b/93b. This is because the phenomenon in question (an additional restriction on verb phrase ellipsis that is not present for right node raising) raises a potential problem for his central analysis. The discussion surrounding this phenomenon in the original article is intended to modify the central analysis to account for this potential problem. If the sign-reversal turns out to be the true result (i.e., the sign-reversal is not due to a confound in the present formal experiments), and if the other data points presented in the relevant section of Phillips (2003) also turn out to be incorrect, then the potential problem would disappear—the evidential ''value'' of these data points was to complicate the analysis, not to support it. This example illustrates the subtleties involved in examining not only the empirical status of any given phenomenon, but also the consequences of that phenomenon for a particular syntactic analysis.

Another finer-grained distinction one could consider is the ''age'' of the data points, that is, whether they have been reported in previous scholarly publications, such that their report in *LI* 2001–2010 is a re-report rather than a new empirical claim. The general idea behind taking age into account is to determine to what extent the convergence between informal and formal methods is dependent upon the number of times a phenomenon has been tested using informal methods. It could be that the two methods eventually arrive at the same results, but on different time scales. Now that the convergence rate has been established, future studies can ask the finer-grained question of what time-scale is required, although there may be some difficulty in determining whether each re-report involved re-testing or not.

5.6. The source of acceptability differences

Finally, it is important to note that the present study provides no information about how to interpret the acceptability judgments obtained by any of the methods discussed. As mentioned in Section 2.1, acceptability contrasts can be driven by any number of factors, with grammatical mechanisms being only one possibility. Syntacticians are, of course, free to assume that a contrast is driven by grammatical mechanisms, and then explore the theoretical consequences of that assumption. However, if syntacticians are interested in providing empirical justification for the assumption, then some sort of experimental manipulation will be necessary to determine to what extent the acceptability contrast could be caused by non-grammatical factors (e.g., Sprouse et al., 2012). The exact nature of those manipulations will vary with each phenomenon of interest, depending on the possible non-grammatical factors that could be driving the effect.

6. Conclusion

We have conducted the first large-scale comparison of informal and formal methods based on a random sample of phenomena from the cutting-edge of syntactic theory, and obtained a convergence rate of 95% with a margin of error of 5.3–5.8%. As we move forward as a field in the conversation about judgment methods, these results suggest that we can no longer assume or assert that the choice of methods would affect a large proportion of the empirical base of syntax, at least with respect to standard acceptability judgments. Of course, this holds only for the most statistically powerful of the formal tasks, the forced-choice task, as both magnitude estimation and Likert scale tasks yielded lower convergence rates, suggesting that statistical power should play a role in future methodological conversations. We have also identified a series of additional studies that might be relevant to the conversation, such as investigating other types of judgment data in the literature, investigating each of the dimensions along which informal and formal methods vary, investigating the divergent results of this study in more detail, and investigating finer-grained distinctions among the data points reported in the literature. Although this conversation is far from over, we hope that these results contribute to bringing the field closer to a consensus about data collection in syntax.

Acknowledgments

We would like to thank audiences at the following universities for helpful comments on earlier stages of this project: Harvard University, Johns Hopkins University, Michigan State University, Pomona College, Princeton University, University of Connecticut, University of Michigan, and the attendees of TEAL 7 at Hiroshima University. We would also like to thank Colin Phillips and one anonymous reviewer for helpful comments on an earlier draft. This work was supported in part by NSF grant BCS-0843896 to JS.

Appendix A

Example materials and descriptive results from all three formal experiments (magnitude estimation, Likert scale, and two-alternative forced-choice). Identifier is in the format VOLUME.ISSUE.FIRST-AUTHOR.EXAMPLE.JUDGMENT, where "g" stands for grammatical (no diacritics). The example sentences are the originally published sentences, so not all will be lexically matched. Control sentences that we constructed are shaded. Unshaded acceptable sentences with the same identifier as their unacceptable counterparts indicate that the authors put both versions in the same numbered example, e.g. by using slashes. The ratings for ME and LS are mean *z*-scores; the ratings for FC are choices/trials.

Identifier	Example	ME	LS	FC
32.1.Martin.2c.*	Sarah saw pictures of.	−0.95	−0.91	2/100
32.1.Martin.1a.g	Kerry attempted to study physics.[a]	1.17	1.15	98/100
32.1.Martin.20a.*	He seems to that Kim solved the problem.	−1.10	−1.00	6/104
32.1.Martin.20a.g	It seems to him that Kim solved the problem.	0.84	0.97	98/104
32.1.Martin.26a.??	Ginny remembered to have bought the beer.	−0.35	−0.35	1/103
32.1.Martin.22a.g	Ginny remembered to bring the beer.	1.31	1.11	102/103
32.1.Martin.26b.??	Sarah convinced Bill to have gone to the party.	−0.62	−0.51	11/104
32.1.Martin.25b.g	Sarah convinced Bill that he would go to the party.	0.25	0.44	93/104
32.1.Martin.28b.??	Sarah convinced Bill that he would have gone to the party by the time he goes to bed this evening.	0.02	0.03	31/103
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.	0.05	0.26	72/103
32.1.Martin.39a.*	Gino believed Rebecca to win the game.	−0.48	−0.49	3/104
32.1.Martin.23a.g	Gino believed Rebecca to be the best.	0.82	0.84	101/104
32.1.Martin.65b.*	John believes without a doubt his team will win.	0.52	0.59	15/103
32.1.Martin.65a.g	John believes without a doubt that his team will win.	0.99	1.02	88/103
32.1.Martin.66b.*	It is illegal one to criticize the government.	−0.70	−0.75	2/104
32.1.Martin.66a.g	It is illegal for one to criticize the government.	1.05	1.10	102/104
32.1.Martin.69b.*	My belief Kim is clever is sincere.	−0.20	−0.14	11/100
32.1.Martin.69a.g	My belief that Kim is clever is sincere.	0.74	0.88	89/100
32.1.Martin.79.*	How likely to be a riot is there?	−0.40	−0.36	36/104
32.1.Martin.77.g	How likely to win the race is John?	−0.14	−0.09	68/104
32.1.Martin.93b.*	John is illegal to park here.	−0.79	−0.74	4/100
32.1.Martin.92b.g	John is believed to have parked here.	0.93	0.84	96/100
32.2.Alexiadou.31a.*	"Don't touch that dial!" suggested abruptly the TV screen.	−0.15	−0.14	2/103
32.2.Alexiadou.31b.g	"Don't touch that dial!" suggested the TV screen abruptly.	0.79	0.61	101/103
32.2.Boeckx.11.*	Debbie ate chocolate, and Kathy milk drank.	−0.69	−0.72	1/104
32.2.Boeckx.11.g	Debbie ate chocolate, and Kathy drank milk.	1.15	1.08	103/104
32.2.Nunes.3b.*	Was kissed John.	−1.10	−1.14	2/100
32.2.Nunes.3a.g	John was kissed.	1.02	1.08	98/100
32.2.Nunes.3c.*	John was kissed John.	−1.29	−1.36	0/103
32.2.Nunes.3a.g	John was kissed.	1.06	1.01	103/103
32.2.Nunes.48b.*	Mary drove Rio and John flew to Sao Paulo.	−0.10	−0.13	3/104
32.2.Nunes.48b.g	Mary drove to Rio and John flew to Sao Paulo.	0.92	0.98	101/104
32.2.Stroik.4b.*	Max may have been studying, but Jason may have done so too.	−0.05	0.10	42/100

Appendix A (*Continued*)

Identifier	Example	ME	LS	FC
32.2.Stroik.4a.g	Max may have been studying, but Jason may have been doing so too.	0.18	0.29	58/100
32.2.Stroik.13b.*	They all have left and they have done all so deliberately.	−0.12	−0.17	13/104
32.2.Stroik.13a.g	They all have left and they have all done so deliberately.	0.27	0.33	91/104
32.2.Stroik.17a.*	Chris is happy, and Pat does so too.	−0.95	−0.93	1/100
32.2.Stroik.17a.g	Chris is happy, and Pat is too.	0.87	0.99	99/100
32.3.Culicover.7b.*	John tried himself to win.	−0.70	−0.66	0/103
32.3.Culicover.7a.g	John tried to win.	1.23	1.08	103/103
32.3.Culicover.15bii.*	John flattered Mary while insulting herself.	−0.20	−0.23	6/103
32.3.Culicover.15bii.g	John flattered Mary while insulting himself.	0.33	0.50	97/103
32.3.Culicover.22b.*	John told Sue when to wash himself.	−0.39	−0.21	20/104
32.3.Culicover.22b.g	John told Sue when to wash herself.	0.17	0.19	84/104
32.3.Culicover.25d.*	Last night there was an attempt to shoot oneself.	−0.53	−0.41	7/104
32.3.Culicover.25d.g	Last night there was an attempt to shoot me.	0.70	0.70	97/104
32.3.Culicover.28c.*	Helen examined Bernie in order for us to vindicate herself.	−0.43	−0.49	9/100
32.3.Culicover.28c.g	Helen examined Bernie in order for us to vindicate ourselves.	0.61	0.54	91/100
32.3.Culicover.32a.*	John's promise to Susan to take care of herself.	−0.37	−0.27	10/103
32.3.Culicover.32a.g	John's promise to Susan to take care of himself.	0.14	0.31	93/103
32.3.Culicover.41b.*	Toby said to Sally to take care of himself.	−0.37	−0.19	14/104
32.3.Culicover.41b.g	Toby said to Sally to take care of herself.	0.46	0.71	90/104
32.3.Culicover.49a.*	Jack asked Sally to be allowed to take care of herself.	−0.15	−0.37	25/100
32.3.Culicover.49a.g	Jack asked Sally to be allowed to take care of himself.	−0.04	0.06	75/100
32.3.Fanselow.28b.*	He saw Mary and kissed.	−0.80	−0.77	2/103
32.3.Fanselow.28b.g	He saw Mary and kissed her.	0.69	0.84	101/103
32.3.Fanselow.58b.*	There has been shot a moose in the woods.	−0.20	−0.34	7/100
32.3.Fanselow.58a.g	There has been a moose shot in the woods.	0.84	0.96	93/100
32.3.Fanselow.58d.*	There has been considered a man sick.	−0.96	−1.07	1/104
32.3.Fanselow.58c.g	There has been a man considered sick.	−0.59	−0.33	103/104
32.3.Fanselow.59b.*	He gave a book Mary.	−0.58	−0.74	5/100
32.3.Fanselow.59a.g	He gave Mary a book.	1.13	1.05	95/100
32.4.López.10a.*	We proclaimed to the public John to be a hero.	−0.27	−0.26	65/100
32.4.López.9a.g	We proclaimed John to the public to be a hero.	−0.25	−0.17	35/100
32.4.López.14b.*	I expected there three men.	−0.81	−0.91	1/103
32.4.López.14b.g	I expected there to be three men.	0.59	0.68	102/103
33.1.den Dikken.5b.*	I know who the hell would buy that book.	−0.04	0.02	3/103
33.1.den Dikken.5a.g	I know who would buy that book.	0.84	0.84	100/103
33.1.den Dikken.58a.*	What under no circumstances should he do?	−0.78	−0.76	4/100
33.1.den Dikken.58a.g	Under no circumstances should he leave.	0.79	0.69	96/100
33.1.den Dikken.62b.*	I don't think that any linguists, I will invite to the party.	−1.00	−1.10	2/103
33.1.den Dikken.62a.g	I don't think that I will invite any linguists to the party.	0.79	0.90	101/103
33.1.den Dikken.71a.*	Who is in love with who the hell?	−0.84	−0.95	2/104
33.1.den Dikken.67.g	Who the hell is in love with who?	0.53	0.51	102/104
33.1.den Dikken.72b.*	John didn't give every charity a red cent.	−0.58	−0.40	6/103
33.1.den Dikken.72a.g	John didn't give Mary a red cent.	0.52	0.57	97/103
33.1.Fox.49c.*	I visited a city near the city yesterday that John did.	−0.61	−0.77	27/104
33.1.Fox.49b.g	I visited a city yesterday near the city that John did.	−0.20	−0.29	77/104
33.1.Fox.65b.*	I told you that Bill when we met will come to the party.	−0.52	−0.61	1/100
33.1.Fox.65b.g	I told you when we met that Bill will come to the party.	0.74	0.81	99/100
33.1.Fox.69a.*	John wants for everyone you do to have fun.	−0.91	−0.90	49/104
33.1.Fox.69b.g	John wants for everyone to have fun that you do.	−0.84	−0.95	55/104
33.2.Bowers.7b.*	The ball perfectly rolled down the hill.	0.43	0.77	23/103

Appendix A (*Continued*)

Identifier	Example	ME	LS	FC
33.2.Bowers.7b.g	The ball rolled perfectly down the hill.	0.94	0.94	80/103
33.2.Bowers.13a.*	John believes to be sick.	−0.51	−0.41	4/100
33.2.Bowers.13a.g	John believes Mary to be sick.	0.69	0.70	96/100
33.2.Bowers.31b1.*	There seem mice to be in the cupboard.	−0.66	−0.69	5/104
33.2.Bowers.31a1.g	There seem to be mice in the cupboard.	−0.07	0.23	99/104
33.2.Bowers.31c2.*	There might mice seem to be in the cupboard.	−1.00	−1.19	2/103
33.2.Bowers.31a2.g	There might seem to be mice in the cupboard.	0.85	0.99	101/103
33.2.Bowers.68b.*	The politician bribes easily to avoid the draft.	−0.46	−0.45	4/103
33.2.Bowers.68a.g	The politician was bribed to avoid the draft.	0.94	0.98	99/103
33.2.Bowers.69b.*	The bureaucrat bribes deliberately.	−0.41	−0.31	3/100
33.2.Bowers.69a.g	The bureaucrat was bribed deliberately.	1.06	0.96	97/100
33.3.Bošković.48d.*	The was arrested student.	−1.31	−1.34	1/103
33.3.Bošković.48a.g	The student was arrested.	0.94	0.93	102/103
33.4.Neeleman.18d.*	Deciding who to see that new movie next makes very happy.	−0.70	−0.77	2/100
33.4.Neeleman.18c.g	Deciding which movie to see next makes John very happy.	0.74	0.81	98/100
33.4.Neeleman.24d.*	Anyone better leave town.	−0.69	−0.70	4/104
33.4.Neeleman.24d.g	Someone better leave town.	0.87	0.88	100/104
33.4.Neeleman.35a.*	What did John wonder what he bought?	−0.83	−0.82	2/104
33.4.Neeleman.35a.g	John wondered what he bought.	0.77	0.82	102/104
33.4.Neeleman.97b.*	Which book did you sleep before reading?	−0.91	−1.05	4/103
33.4.Neeleman.97a.g	Which book did you file before reading?	0.28	0.39	99/103
33.4.Neeleman.100.*	Yesterday seemed that John left.	−0.62	−0.61	2/100
33.4.Neeleman.100.g	It seemed that yesterday John left.	0.58	0.60	98/100
34.1.Basilico.44b.*	Who was seen steal the wallet?	−0.81	−0.88	4/100
34.1.Basilico.44a.?	Who did you see steal the wallet?	0.50	0.68	96/100
34.1.Basilico.62.*	There are linguists tall.	−0.70	−1.02	9/100
34.1.Basilico.62.g	There are linguists available.	0.21	0.50	91/100
34.1.Basilico.96a.??	The children almost all are sleeping.	0.02	−0.12	8/100
34.1.Basilico.96b.g	The children are almost all sleeping.	0.78	0.81	92/100
34.1.Fox.14.*	What do you worry if the lawyer forgets at the office?	−0.87	−0.88	12/104
34.1.Fox.14.g	What do you think that the lawyer forgot at the office?	0.16	0.13	92/104
34.1.Fox.24.*	It appears that a certain senator will resign, but which senator it does is still a secret.	−0.53	−0.63	8/103
34.1.Fox.19.g	It appears that a certain senator will resign, but which senator is still a secret.	0.54	0.59	95/103
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.	0.04	−0.23	6/100
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.	0.76	0.86	94/100
34.1.Fox.28.*	They said they heard about a Balkan language, but I don't know which Balkan language they did.	−0.15	−0.14	4/104
34.1.Fox.27.g	They said they heard about a Balkan language, but I don't know which Balkan language.	0.73	0.82	100/104
34.1.Phillips.3e.*	Each other like Wallace and Greg.	−0.96	−0.99	0/100
34.1.Phillips.3d.g	Wallace and Greg like each other.	1.27	1.18	100/100
34.1.Phillips.6b.*	Wallace gave at breakfast time his favorite pet beagle an enormous chewy dog-biscuit.	−0.65	−0.70	7/103
34.1.Phillips.6b.g	Wallace gave his favorite pet beagle an enormous chewy dog-biscuit at breakfast time.	0.81	0.72	96/103
34.1.Phillips.59b.*	The students were punished and their teachers by their parents.	−0.77	−0.91	4/104
34.1.Phillips.59b.g	The students were punished by their parents and their teachers.	1.02	0.98	100/104
34.1.Phillips.67d.*	I gave anything to nobody.	−0.98	−1.21	2/103

Appendix A (*Continued*)

Identifier	Example	ME	LS	FC
34.1.Phillips.67c.g	I gave nothing to anybody.	−0.01	0.22	101/103
34.1.Phillips.88b.*	John promised Mary to leave, and Sue did to write more poetry.	−0.77	−0.80	0/103
34.1.Phillips.88b.g	John promised Mary to leave, and Sue promised to write more poetry.	0.58	0.68	103/103
34.1.Phillips.93b.?*	Wendy stood more buckets in the garage than Peter did in the basement.	0.35	0.36	88/104
34.1.Phillips.92b.g	Wendy stood more buckets than Peter did in the garage.	−0.24	−0.06	16/104
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.	−0.46	−0.63	15/100
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.	−0.25	−0.38	85/100
34.2.Caponigro.13b.*	The flute was being shiny.	−0.58	−0.62	0/100
34.2.Caponigro.13a.g	The flute was being played by the soloist.	1.16	1.08	100/100
34.2.Panagiotidis.6.*	We students of physics are taller than you of chemistry.	−0.40	−0.34	15/103
34.2.Panagiotidis.6.g	We students of physics are taller than you students of chemistry.	0.11	0.16	88/103
34.3.Heycock.16.*	He was judge.	−0.46	−0.39	3/100
34.3.Heycock.16.g	He was the judge.	1.16	1.17	97/100
34.3.Heycock.30c.?*	Cat and dog that were fighting all the time had to be separated.	−0.30	−0.45	5/104
34.3.Heycock.30c.g	The cat and dog that were fighting all the time had to be separated.	0.40	0.55	99/104
34.3.Heycock.37b.??	Knife with the golden blade and fork with the silver handle go on the left.	−0.38	−0.44	7/103
34.3.Heycock.37b.g	The knife with the golden blade and the fork with the silver handle go on the left.	0.60	0.69	96/103
34.3.Heycock.55a.*	Fork is silver-plated and bowl is enameled.	−0.51	−0.43	3/103
34.3.Heycock.55a.g	The fork is silver-plated and the bowl is enameled.	0.88	0.97	100/103
34.3.Heycock.66.*	This is table.	−0.57	−0.93	2/104
34.3.Heycock.66.g	This is a table.	1.57	1.31	102/104
34.3.Heycock.82a.*	The dog that I saw's collar was leather.	−0.40	−0.43	28/103
34.3.Heycock.82a.g	The collar of the dog that I saw was leather.	0.09	0.29	75/103
34.3.Landau.32c.*	There expects to be a man in the garden.	−0.70	−0.70	2/104
34.3.Landau.32c.g	There seems to be a man in the garden.	1.09	1.07	102/104
34.3.Landau.39b.*	One interpreter each tried to be assigned to every visiting diplomat.	−0.40	−0.30	17/104
34.3.Landau.39a.g	One interpreter tried to be assigned to every visiting diplomat.	0.80	0.74	87/104
34.3.Takano.9e.*	Anything has nobody done.	−1.24	−1.17	0/100
34.3.Takano.9e.g	Nobody has done anything.	0.21	0.69	100/100
34.3.Takano.10b.*	I bought any books only occasionally.	−0.69	−0.71	4/104
34.3.Takano.10b.g	I only occasionally bought any books.	0.11	0.14	100/104
34.4.Bošković.3a.*	It seemed at that time David had left.	0.29	0.60	33/100
34.4.Bošković.4a.g	It seemed at that time that David had left.	0.71	0.70	67/100
34.4.Bošković.3b.*	What the students believe is they will pass the exam.	−0.13	−0.18	19/103
34.4.Bošković.4b.g	What the students believe is that they will pass the exam.	0.12	0.18	84/103
34.4.Bošković.3c.*	They suspected and we believed Peter would visit the hospital.	−0.36	−0.43	22/103
34.4.Bošković.4c.g	They suspected and we believed that Peter would visit the hospital.	−0.22	−0.22	81/103
34.4.Bošković.3d.*	Mary believed Peter finished school and Bill Peter got a job.	−0.56	−0.48	29/103
34.4.Bošković.4d.g	Mary believed that Peter finished school and Bill that Peter got a job.	−0.46	−0.55	74/103
34.4.Bošković.3e.*	John likes Mary, Jane didn't believe	−0.68	−0.66	43/100
34.4.Bošković.4e.g	That John likes Mary, Jane didn't believe	−0.58	−0.62	57/100
34.4.Bošković.7a.??	What did they believe at that time that Peter fixed?	−0.23	−0.23	31/104
34.4.Bošković.7c.g	At that time, what did they believe that Peter fixed?	0.09	0.04	73/104
34.4.Haegeman.2a.*	This is the man who I think that will buy your house next year.	0.00	0.01	4/100
34.4.Haegeman.2a.g	This is the man who I think will buy your house next year.	0.48	0.56	96/100
34.4.Lasnik.10a.*	Angela wondered how John managed to cook, but it's not clear what food.	−0.31	−0.35	25/104

Appendix A (*Continued*)

Identifier	Example	ME	LS	FC
34.4.Lasnik.11a.g	Angela wondered how John managed to cook a certain food, but it's not clear what food.	0.12	0.09	79/104
35.1.Beck.12b.*	Who did you believe a friend of satisfied?	−0.79	−0.82	8/104
35.1.Beck.12b.g	I believed a friend of Andy satisfied.	0.17	0.23	96/104
35.1.Bhatt.14a.*	Ralph is more than fit tall.	−0.82	−0.76	2/103
35.1.Bhatt.14cf.g	Ralph is more tall than fit.	0.92	0.89	101/103
35.1.Bhatt.76b.*	I told you that Bill when we met will come to the party.	−0.64	−0.51	2/104
35.1.Bhatt.76b.g	I told you when we met that Bill will come to the party.	0.93	0.96	102/104
35.1.Bhatt.94a.*	I expect that everyone you do will visit Mary.	−0.88	−0.82	57/100
35.1.Bhatt.94b.g	I expect that everyone will visit Mary that you do.	−0.78	−0.93	43/100
35.1.McGinnis.32b.*	I ran Mary.	−0.95	−1.04	1/100
35.1.McGinnis.32b.g	I ran for Mary.	1.06	0.99	99/100
35.1.McGinnis.63b.*	The article angered Bill at the government.	−0.55	−0.61	7/100
35.1.McGinnis.63a.g	The article angered Bill.	1.21	1.00	93/100
35.2.Hazout.6b.*	I find it irritating for usually this street to be closed.	−0.75	−0.89	3/104
35.2.Hazout.6a.g	I find it irritating that usually this street is closed.	0.29	0.51	101/104
35.2.Larson.44a.??	A taller man than my father walked in.	0.28	0.09	24/103
35.2.Larson.44a.g	A man taller than my father walked in.	0.47	0.57	79/103
35.2.Larson.44c.??	Max talked to as tall a man as his father.	−0.46	−0.51	4/100
35.2.Larson.44c.g	Max talked to a man as tall as his father.	0.48	0.53	96/100
35.3.Embick.13b.*	Mary pounded the apple flattened.	−0.49	−0.45	4/104
35.3.Embick.13b.g	Mary pounded the apple flat.	1.01	0.77	100/104
35.3.Hazout.63.*	It seems a man to be in the room.	−0.99	−1.03	1/103
35.3.Hazout.60b.g	It seems a man is in the room.	0.58	0.68	102/103
35.3.Hazout.67c.*	There is likely a man to appear.	−0.27	−0.21	65/100
35.3.Hazout.67a.g	There is likely to appear a man.	−0.56	−0.50	35/100
35.3.Hazout.75.*	It is unimaginable Mary to arrive on time.	−0.68	−0.81	2/104
35.3.Hazout.75.g	It is unimaginable for Mary to arrive on time.	0.58	0.65	102/104
35.3.Richards.17b.*	To whom did you give what?	0.10	0.37	46/100
35.3.Richards.17a.g	What did you give to whom?	0.16	0.40	54/100
35.3.Sobin.3c.*	Some frogs and a fish is in the pond.	−0.30	−0.26	4/104
35.3.Sobin.3c.g	Some frogs and a fish are in the pond.	0.69	0.91	100/104
36.4.den Dikken.45.*	That much the less you say, the smarter you will seem.	−0.67	−0.79	3/103
36.4.den Dikken.45.g	The less you say, the smarter you will seem.	1.15	1.08	100/103
37.1.Boeckx.5.*	Sue asked what who bought.	−0.74	−0.77	3/100
37.1.Boeckx.8.g	Sue asked me who bought what.	0.25	0.56	97/100
37.2.de Vries.39a.*	I talked to with whom you danced yesterday.	−0.91	−1.02	4/104
37.2.de Vries.39b.g	I talked to Mary, with whom you danced yesterday.	0.13	0.16	100/104
37.2.Sigurðsson.3d.*	Me would have been elected.	−0.89	−1.10	2/100
37.2.Sigurðsson.2a.g	I would have been elected.	0.76	1.09	98/100
37.3.Becker.2b.*	There like to be storms at this time of year.	−0.85	−1.02	2/103
37.3.Becker.2a.g	There tend to be storms at this time of year.	0.60	0.57	101/103
37.3.Becker.5b.*	I seem eating sushi.	−0.72	−0.74	3/100
37.3.Becker.5a.g	I like eating sushi.	1.04	1.04	97/100
37.3.Becker.26b.*	I seem eating sushi.	−0.79	−0.84	1/104
37.3.Becker.26a.g	I hate eating sushi.	1.37	1.33	103/104
37.4.Nakajima.20e.*	He existed a dangerous existence.	−0.80	−0.80	4/103
37.4.Nakajima.4a.g	The tree grew a century's growth within only ten years.	0.19	0.15	99/103
38.2.Hornstein.3c.*	How many books there were on the table?	−0.11	−0.28	5/104
38.2.Hornstein.3a.g	How many books were there on the table?	0.82	0.69	99/104

Appendix A (*Continued*)

Identifier	Example	ME	LS	FC
38.2.Hornstein.4c.*	Into which room did walk three men?	−0.76	−0.94	11/103
38.2.Hornstein.4b.g	Into which room walked three men?	−0.34	−0.32	92/103
38.2.Hornstein.4e.*	Into which room three men walked?	−0.41	−0.42	12/103
38.2.Hornstein.4d.g	Into which room did three men walk?	0.37	0.42	91/103
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.	−0.33	−0.33	19/103
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.	0.01	−0.15	84/103
38.3.Hirose.1b.*	To Mary for Bill I gave a book.	−0.76	−0.92	1/104
38.3.Hirose.1a.g	From Alabama to Louisiana John played the banjo.	0.86	0.89	103/104
38.3.Hirose.4a.*	It will take from three five days for him to recover.	−0.22	−0.20	7/100
38.3.Hirose.3a.g	It will take three to five days for him to recover.	1.21	1.08	93/100
38.3.Landau.31b.*	An hour, they slept, and then went to work.	−0.20	−0.25	4/104
38.3.Landau.31a.g	They slept an hour and then went to work.	1.24	1.02	100/104
38.3.Landau.39a.*	Who did George kick the ball?	−0.41	−0.40	4/103
38.3.Landau.38a.g	George kicked the boy the ball.	0.74	0.68	99/103
38.4.Bošković.4.*	There seems a man to be in the garden.	−0.15	−0.18	2/104
38.4.Bošković.17a.g	There seems to be a man in the garden.	1.34	1.12	102/104
38.4.Kallulli.4b.*	Eva was killed from John.	−0.46	−0.48	0/103
38.4.Kallulli.4b.g	Eva was killed by John.	1.21	1.14	103/103
38.4.Kallulli.9b.*	The boat sank to collect the insurance.	−0.14	−0.02	7/104
38.4.Kallulli.9a.g	The boat was sunk to collect the insurance.	0.92	1.03	97/104
38.4.Kallulli.10b.*	The ship sank deliberately.	0.04	0.16	9/100
38.4.Kallulli.10a.g	The ship was sunk deliberately.	1.05	0.88	91/100
39.1.Sobin.20c.*	John broke a cup, and Mary did so with a saucer.	−0.30	−0.30	10/103
39.1.Sobin.21c.g	John broke a cup, and Mary did so too.	0.31	0.54	93/103
40.1.Caponigro.23a.*	Jack came the person he is in love with.	−0.84	−0.98	0/100
40.1.Caponigro.23cf.g	Jack came with the person he is in love with.	0.97	1.08	100/100
40.1.Caponigro.25b.*	Lily will dance who the king chooses.	−0.71	−0.85	3/104
40.1.Caponigro.25b.g	Lily will dance with the person the king chooses.	0.86	0.93	101/104
40.1.Heck.5b.*	Sherry met a man very fond of whom she found herself.	−0.92	−0.94	3/104
40.1.Heck.5b.g	Sherry met a man who she found herself very fond of.	0.27	0.52	101/104
40.1.Stepanov.4b.*	What did who buy?	−0.37	−0.45	5/100
40.1.Stepanov.4a.g	Who bought what?	0.66	0.85	95/100
40.2.Johnson.59b.*	Ice cream gives me in the morning brain-freeze.	−0.85	−0.81	3/103
40.2.Johnson.59b.g	Ice cream gives me brain-freeze in the morning.	0.75	0.68	100/103
40.4.Hicks.23.*	Lloyd Webber musicals are likely to be condemned without anyone even watching	−0.55	−0.57	25/103
40.4.Hicks.22.g	Lloyd Webber musicals are easy to condemn without even watching	0.58	0.54	78/103
41.1.Müller.14c.*	Who did that Mary was going out with bother you?	−1.03	−1.13	14/100
41.1.Müller.14c.g	That Mary was going out with Luke bothered you.	−0.21	−0.23	86/100
41.1.Müller.25b.??(*)	Who do you wonder which picture of is on sale?	−0.99	−1.13	5/104
41.1.Müller.25b.g	You wonder which picture of Marge is on sale.	0.23	0.43	99/104
41.2.Bruening.3b.*	The count gives the creeps to me.	−0.47	−0.52	8/104
41.2.Bruening.3a.g	The count gives me the creeps.	0.80	0.70	96/104
41.2.Bruening.31a.*	At that battle were given the generals who lost hell.	−0.99	−0.93	3/104
41.2.Bruening.31a.g	At that battle the generals who lost were given hell.	0.37	0.36	101/104
41.2.Bruening.33a.*	At that time were given the tables we inherited from Aunt Selma a good scrubbing.	−0.85	−0.99	4/100

Appendix A (*Continued*)

Identifier	Example	ME	LS	FC
41.2.Bruening.33a.g	The tables we inherited from Aunt Selma were given a good scrubbing at that time.	0.24	0.18	96/100
41.2.Bruening.36b.*	The man that he gave the creeps last night to is over there.	−0.33	−0.63	18/103
41.2.Bruening.36a.g	The man that he gave the creeps to last night is over there.	−0.04	−0.06	85/103
41.3.Costantini.2a.??	All the men seem to have all eaten supper.	0.47	0.38	21/104
41.3.Costantini.2a.g	The men seem to have all eaten supper.	0.67	0.82	83/104
41.3.Landau.7b.*	I am now hiring for John to work with.	−0.76	−0.87	3/103
41.3.Landau.7b.g	I am now hiring people for John to work with.	0.98	0.94	100/103
41.3.Landau.10b.*	The game was played shoeless.	−0.68	−0.68	34/103
41.3.Landau.10a.g	The game was played wearing no shoes.	−0.67	−0.62	69/103
41.3.Landau.25c.*	I told Mr. Smith that I am able to paint the fence together.	−0.20	−0.17	38/104
41.3.Landau.24c.g	I told Mr. Smith that I wonder when to paint the fence together.	0.22	0.23	66/104
41.3.Landau.27b.*	His wife kissed in front of the kids.	−0.65	−0.60	7/103
41.3.Landau.27b.g	He and his wife kissed in front of the kids.	0.86	0.94	96/103
41.3.Rezac.3b2.*	There had all hung over the fireplace the portraits by Picasso.	−0.47	−0.63	15/100
41.3.Rezac.3b1.g	There had hung over the fireplace all of the portraits by Picasso.	−0.20	−0.18	85/100
41.3.Vicente.4a3.*	Sandy plays the guitar because Betsy the harmonica.	−0.91	−1.02	5/100
41.3.Vicente.4a1.g	Sandy plays the guitar and Betsy the harmonica.	0.17	0.39	95/100
41.3.Vicente.4a6.*	Sandy plays the guitar better than Betsy the harmonica.	−0.14	−0.27	5/100
41.3.Vicente.4b6.g	Sandy plays the guitar better than Betsy does.	1.05	0.97	95/100
41.3.Vicente.5a.*	Amanda went to Santa Cruz, and Bill thinks that Claire to Monterrey.	−0.43	−0.48	3/100
41.3.Vicente.5b.g	Amanda went to Santa Cruz, and Bill thinks that Claire did too.	0.65	0.83	97/100
41.3.Vicente.8a.*	Read things, Mike did quickly.	−0.79	−0.80	0/103
41.3.Vicente.8a.g	Mike read things quickly.	0.58	0.68	103/103
41.3.Vicente.8d.*	Want to write, Randy did a novel.	−1.12	−1.33	0/103
41.3.Vicente.8d.g	Randy wanted to write a novel.	1.34	1.15	103/103
41.4.Bruening.9b.*	What did he prove an account of false?	−0.77	−0.81	39/104
41.4.Bruening.9c.g	Who did he give statues of to all the season-ticket holders?	−0.56	−0.67	65/104
41.4.Haegeman.4a.*	When this column she started to write last year, I thought she would be fine.	−0.72	−0.82	17/100
41.4.Haegeman.4c.g	When last year she started to write this column, I thought she would be fine.	−0.19	−0.25	83/100
41.4.Haegeman.18a.*	Bill asked if such books John only reads at home.	−0.66	−0.82	45/100
41.4.Haegeman.18a.g	Bill knows that such books John only reads at home.	−0.59	−0.62	55/100
41.4.Haegeman.22a*	If frankly he's unable to cope, we'll have to replace him.	−0.17	−0.12	5/103
41.4.Haegeman.22a.g	If he's unable to cope, we'll have to replace him.	1.07	0.99	98/103

[a] These two items stand out as less structurally matched than the other pairwise phenomena in the experiment. Martin used these two items to illustrate the standard claim that PRO is distributionally restricted to the subject position of nonfinite clauses. He presented them with ''PRO'' in the relevant positions, which we obviously could not. This pair highlights the subjectivity of the notion ''maximally similar'' in the definition of pairwise phenomena given in section 2.3. Whether a more similar pair could be constructed to make the same distributional claim is an open question, as is the broader question of how acceptability judgments can bear on claims about theory-internal constructs. For the current experiment we decided to be true to the published pair, as we do not believe that our general conclusions hinge on how these questions may be resolved.

Appendix B

Results of the statistical tests for each phenomenon and each formal judgment task. Identifier is in the format VOLUME.ISSUE.FIRST-AUTHOR.EXAMPLE.JUDGMENT. For space reasons only the *p*-values and Bayes factors are reported. All *p*-values have been rounded to two decimal places for ease of presentation. Any *p*-values below .01 have been rounded up to .01. Only the two-tailed *p*-values are reported; one-tailed *p*-values can be calculated by dividing by two. Bayes factors are reported in scientific notation (e.g., 8.E+34 = 8×10^{34}) and also rounded to two exponential digits. Shaded cells indicate results significantly or marginally in the opposite direction from the direction reported in the original article. The raw results are available on the first author's website [www.sprouse.uconn.edu] for further analysis.

Item ID	Magnitude estimation			Likert scale			Forced choice		
	LME	*t*-test	Bayes	LME	*t*-test	Bayes	ML	Sign test	Bayes
35.3.Hazout.67c.*	.01	.01	9.E+01	.05	.01	2.E+01	.03	.01	1.E+01
34.1.Phillips.93b.?*	.02	.01	8.E+01	.07	.01	8.E+00	.01	.01	7.E+10
33.1.Fox.69a.*	.46	.38	1.E-01	.65	.51	1.E-01	.54	.62	1.E-01
35.3.Richards.17b.*	.58	.47	1.E-01	.81	.73	8.E-02	.26	.48	2.E-01
32.4.López.10a.*	.93	.85	8.E-02	.71	.49	1.E-01	.01	.01	1.E+01
35.1.Bhatt.94a.*	.12	.04	7.E-01	.22	.16	2.E-01	.05	.10	3.E-01
41.4.Haegeman.18a.*	.41	.28	1.E-01	.02	.01	3.E+00	.20	.37	2.E-01
34.4.Bošković.3e.*	.29	.18	2.E-01	.74	.67	9.E-02	.08	.19	3.E-01
32.1.Martin.20a.*	.01	.01	8.E+34	.01	.01	2.E+32	.01	.01	1.E+20
32.1.Martin.26a.??	.01	.01	1.E+24	.01	.01	2.E+25	.01	.01	9.E+26
32.1.Martin.26b.??	.01	.01	6.E+17	.01	.01	5.E+10	.01	.01	9.E+14
32.1.Martin.28b.??	.78	.74	8.E-02	.18	.03	8.E-01	.01	.01	5.E+02
32.1.Martin.2c.*	.01	.01	4.E+40	.01	.01	4.E+37	.01	.01	3.E+24
32.1.Martin.39a.*	.01	.01	8.E+22	.01	.01	2.E+24	.01	.01	1.E+24
32.1.Martin.65b.*	.01	.01	9.E+03	.01	.01	8.E+04	.01	.01	2.E+11
32.1.Martin.66b.*	.01	.01	3.E+29	.01	.01	1.E+35	.01	.01	4.E+25
32.1.Martin.69b.*	.01	.01	4.E+12	.01	.01	2.E+12	.01	.01	9.E+13
32.1.Martin.79.*	.01	.01	9.E+00	.05	.01	3.E+00	.01	.01	2.E+01
32.1.Martin.93b.*	.01	.01	5.E+36	.01	.01	9.E+22	.01	.01	3.E+21
32.2.Alexiadou.31a.*	.01	.01	4.E+11	.01	.01	1.E+09	.01	.01	2.E+25
32.2.Boeckx.11.*	.01	.01	2.E+35	.01	.01	5.E+31	.01	.01	2.E+27
32.2.Nunes.3b.*	.01	.01	3.E+36	.01	.01	4.E+37	.01	.01	3.E+24
32.2.Nunes.3c.*	.01	.01	4.E+33	.01	.01	1.E+49	.01	.01	1.E+29
32.2.Nunes.48b.*	.01	.01	2.E+10	.01	.01	1.E+14	.01	.01	1.E+24
32.2.Stroik.13b.*	.02	.01	2.E+02	.01	.01	2.E+03	.01	.01	2.E+13
32.2.Stroik.17a.*	.01	.01	4.E+36	.01	.01	5.E+36	.01	.01	1.E+26
32.2.Stroik.4b.*	.01	.01	3.E+01	.10	.03	8.E-01	.04	.07	4.E-01
32.3.Culicover.15bii.*	.01	.01	7.E+02	.01	.01	3.E+05	.01	.01	7.E+19
32.3.Culicover.22b.*	.01	.01	9.E+04	.06	.01	3.E+01	.01	.01	2.E+08
32.3.Culicover.25d.*	.01	.01	5.E+20	.01	.01	1.E+20	.01	.01	9.E+18
32.3.Culicover.28c.*	.01	.01	1.E+14	.01	.01	5.E+11	.01	.01	7.E+15
32.3.Culicover.32a.*	.01	.01	1.E+04	.01	.01	2.E+03	.01	.01	4.E+15
32.3.Culicover.41b.*	.01	.01	2.E+11	.01	.01	5.E+12	.01	.01	2.E+12
32.3.Culicover.49a.*	.35	.20	2.E-01	.01	.01	1.E+03	.01	.01	5.E+04
32.3.Culicover.7b.*	.01	.01	8.E+37	.01	.01	6.E+31	.01	.01	1.E+29
32.3.Fanselow.28b.*	.01	.01	3.E+25	.01	.01	3.E+33	.01	.01	2.E+25
32.3.Fanselow.58b.*	.01	.01	4.E+12	.01	.01	6.E+19	.01	.01	8.E+17
32.3.Fanselow.58d.*	.01	.01	5.E+04	.01	.01	6.E+08	.01	.01	2.E+27
32.3.Fanselow.59b.*	.01	.01	3.E+26	.01	.01	5.E+26	.01	.01	2.E+20
32.4.López.14b.*	.01	.01	2.E+24	.01	.01	7.E+28	.01	.01	9.E+26
33.1.den Dikken.58a.*	.01	.01	3.E+31	.01	.01	5.E+21	.01	.01	3.E+21
33.1.den Dikken.5b.*	.01	.01	7.E+09	.01	.01	3.E+13	.01	.01	6.E+23
33.1.den Dikken.62b.*	.01	.01	1.E+32	.01	.01	1.E+45	.01	.01	2.E+25
33.1.den Dikken.71a.*	.01	.01	3.E+23	.01	.01	1.E+25	.01	.01	4.E+25
33.1.den Dikken.72b.*	.01	.01	4.E+14	.01	.01	2.E+15	.01	.01	7.E+19
33.1.Fox.49c.*	.01	.01	2.E+03	.01	.01	5.E+04	.01	.01	3.E+04
33.1.Fox.65b.*	.01	.01	4.E+22	.01	.01	3.E+22	.01	.01	1.E+26
33.2.Bowers.13a.*	.01	.01	5.E+22	.01	.01	4.E+16	.01	.01	3.E+21
33.2.Bowers.31b1.*	.01	.01	1.E+22	.01	.01	9.E+29	.01	.01	2.E+21
33.2.Bowers.31c2.*	.01	.01	5.E+13	.01	.01	7.E+24	.01	.01	2.E+25

Appendix B (*Continued*)

Item ID	Magnitude estimation			Likert scale			Forced choice		
	LME	*t*-test	Bayes	LME	*t*-test	Bayes	ML	Sign test	Bayes
33.2.Bowers.68b.*	.01	.01	1.E+20	.01	.01	3.E+25	.01	.01	2.E+22
33.2.Bowers.69b.*	.01	.01	1.E+24	.01	.01	3.E+20	.01	.01	8.E+22
33.2.Bowers.7b.*	.02	.01	2.E+03	.28	.05	6.E-01	.01	.01	2.E+06
33.3.Bošković.48d.*	.01	.01	4.E+28	.01	.01	2.E+36	.01	.01	9.E+26
33.4.Neeleman.100.*	.01	.01	1.E+21	.01	.01	1.E+21	.01	.01	3.E+24
33.4.Neeleman.18d.*	.01	.01	8.E+23	.01	.01	1.E+31	.01	.01	3.E+24
33.4.Neeleman.24d.*	.01	.01	5.E+30	.01	.01	5.E+33	.01	.01	4.E+22
33.4.Neeleman.35a.*	.01	.01	1.E+34	.01	.01	5.E+35	.01	.01	4.E+25
33.4.Neeleman.97b.*	.01	.01	2.E+18	.01	.01	8.E+23	.01	.01	2.E+22
34.1.Basilico.44b.*	.01	.01	7.E+23	.01	.01	4.E+24	.01	.01	3.E+21
34.1.Basilico.62.*	.01	.01	4.E+07	.01	.01	3.E+28	.01	.01	7.E+15
34.1.Basilico.96a.??	.01	.01	7.E+08	.01	.01	8.E+13	.01	.01	7.E+16
34.1.Fox.14.*	.01	.01	2.E+17	.01	.01	1.E+09	.01	.01	1.E+14
34.1.Fox.24.*	.01	.01	1.E+16	.01	.01	2.E+18	.01	.01	4.E+17
34.1.Fox.26.*	.01	.01	2.E+08	.01	.01	6.E+16	.01	.01	1.E+19
34.1.Fox.28.*	.01	.01	3.E+11	.01	.01	6.E+16	.01	.01	4.E+22
34.1.Phillips.3e.*	.01	.01	3.E+38	.01	.01	5.E+47	.01	.01	1.E+28
34.1.Phillips.59b.*	.01	.01	1.E+31	.01	.01	3.E+37	.01	.01	4.E+22
34.1.Phillips.67d.*	.01	.01	1.E+12	.01	.01	3.E+23	.01	.01	2.E+25
34.1.Phillips.6b.*	.01	.01	3.E+22	.01	.01	3.E+23	.01	.01	5.E+18
34.1.Phillips.88b.*	.01	.01	1.E+24	.01	.01	6.E+24	.01	.01	1.E+29
34.1.Phillips.96a.*	.10	.01	1.E+02	.02	.01	6.E+00	.01	.01	5.E+10
34.2.Caponigro.13b.*	.01	.01	3.E+31	.01	.01	1.E+33	.01	.01	1.E+28
34.2.Panadiotidis.6.*	.01	.01	3.E+05	.01	.01	3.E+05	.01	.01	2.E+11
34.3.Heycock.16.*	.01	.01	5.E+22	.01	.01	8.E+24	.01	.01	8.E+22
34.3.Heycock.30c.?*	.01	.01	1.E+07	.01	.01	3.E+13	.01	.01	2.E+21
34.3.Heycock.37b.??	.01	.01	3.E+11	.01	.01	2.E+18	.01	.01	5.E+18
34.3.Heycock.55a.*	.01	.01	2.E+22	.01	.01	4.E+23	.01	.01	6.E+23
34.3.Heycock.66.*	.01	.01	1.E+28	.01	.01	6.E+44	.01	.01	4.E+25
34.3.Heycock.82a.*	.01	.01	4.E+02	.01	.01	2.E+05	.01	.01	7.E+03
34.3.Landau.32c.*	.01	.01	9.E+27	.01	.01	5.E+31	.01	.01	4.E+25
34.3.Landau.39b.*	.01	.01	1.E+18	.01	.01	7.E+11	.01	.01	1.E+10
34.3.Takano.10b.*	.01	.01	4.E+10	.01	.01	2.E+09	.01	.01	4.E+22
34.3.Takano.9e.*	.01	.01	1.E+26	.01	.01	1.E+32	.01	.01	1.E+28
34.4.Bošković.3a.*	.01	.01	4.E+01	.32	.20	2.E-01	.01	.01	4.E+01
34.4.Bošković.3b.*	.17	.01	2.E+00	.04	.01	1.E+02	.01	.01	4.E+08
34.4.Bošković.3c.*	.24	.03	9.E-01	.23	.02	1.E+00	.01	.01	6.E+06
34.4.Bošković.3d.*	.46	.20	2.E-01	.53	.40	1.E-01	.01	.01	3.E+03
34.4.Bošković.7a.??	.06	.01	7.E+00	.09	.01	3.E+00	.01	.01	7.E+02
34.4.Haegeman.2a.*	.01	.01	3.E+03	.01	.01	1.E+05	.01	.01	3.E+21
34.4.Lasnik.10a.*	.01	.01	3.E+03	.01	.01	2.E+04	.01	.01	3.E+05
35.1.Beck.12b.*	.01	.01	6.E+12	.01	.01	3.E+12	.01	.01	7.E+17
35.1.Bhatt.14a.*	.01	.01	1.E+26	.01	.01	2.E+24	.01	.01	2.E+25
35.1.Bhatt.76b.*	.01	.01	2.E+30	.01	.01	4.E+27	.01	.01	4.E+25
35.1.McGinnis.32b.*	.01	.01	8.E+33	.01	.01	7.E+37	.01	.01	1.E+26
35.1.McGinnis.63b.*	.01	.01	2.E+26	.01	.01	2.E+25	.01	.01	8.E+17
35.2.Hazout.6b*	.01	.01	4.E+17	.01	.01	2.E+26	.01	.01	1.E+24
35.2.Larson.44a.??	.24	.06	5.E-01	.01	.01	2.E+04	.01	.01	5.E+05
35.2.Larson.44c.??	.01	.01	1.E+12	.01	.01	2.E+15	.01	.01	3.E+21
35.3.Embick.13b.*	.01	.01	4.E+24	.01	.01	4.E+13	.01	.01	4.E+22
35.3.Hazout.63.*	.01	.01	3.E+25	.01	.01	5.E+32	.01	.01	9.E+26
35.3.Hazout.75.*	.01	.01	3.E+20	.01	.01	1.E+27	.01	.01	4.E+25
35.3.Sobin.3c.*	.01	.01	9.E+14	.01	.01	6.E+20	.01	.01	4.E+22
36.4.den Dikken.45.*	.01	.01	9.E+27	.01	.01	7.E+38	.01	.01	6.E+23
37.1.Boeckx.5.*	.01	.01	3.E+17	.01	.01	6.E+22	.01	.01	8.E+22
37.2.de Vries.39a.*	.01	.01	7.E+14	.01	.01	1.E+19	.01	.01	4.E+22
37.2.Sigurðsson.3d.*	.01	.01	9.E+22	.01	.01	4.E+39	.01	.01	3.E+24
37.3.Becker.26b.*	.01	.01	1.E+38	.01	.01	8.E+43	.01	.01	2.E+27
37.3.Becker.2b.*	.01	.01	1.E+28	.01	.01	7.E+29	.01	.01	2.E+25
37.3.Becker.5b.*	.01	.01	6.E+27	.01	.01	6.E+31	.01	.01	8.E+22
37.4.Nakajima.20e.*	.01	.01	3.E+12	.01	.01	2.E+13	.01	.01	2.E+22

Appendix B (*Continued*)

Item ID	Magnitude estimation			Likert scale			Forced choice		
	LME	*t*-test	Bayes	LME	*t*-test	Bayes	ML	Sign test	Bayes
38.2.Hornstein.3c.*	.01	.01	8.E+07	.01	.01	2.E+08	.01	.01	2.E+21
38.2.Hornstein.4c.*	.01	.01	8.E+04	.01	.01	1.E+09	.01	.01	5.E+14
38.2.Hornstein.4e.*	.01	.01	2.E+11	.01	.01	5.E+12	.01	.01	6.E+13
38.3.Haddican.39.*	.03	.01	1.E+01	.20	.06	4.E-01	.01	.01	4.E+08
38.3.Hirose.1b.*	.01	.01	6.E+25	.01	.01	4.E+41	.01	.01	2.E+27
38.3.Hirose.4a.*	.01	.01	1.E+19	.01	.01	1.E+19	.01	.01	8.E+17
38.3.Landau.31b.*	.01	.01	6.E+23	.01	.01	5.E+19	.01	.01	4.E+22
38.3.Landau.39a.*	.01	.01	3.E+13	.01	.01	2.E+09	.01	.01	2.E+22
38.4.Bošković.4.*	.01	.01	1.E+14	.01	.01	7.E+15	.01	.01	4.E+25
38.4.Kallulli.10b.*	.01	.01	1.E+09	.01	.01	5.E+07	.01	.01	7.E+15
38.4.Kallulli.4b.*	.01	.01	3.E+25	.01	.01	7.E+31	.01	.01	1.E+29
38.4.Kallulli.9b.*	.01	.01	2.E+14	.01	.01	2.E+13	.01	.01	9.E+18
39.1.Sobin.20c.*	.01	.01	3.E+08	.01	.01	6.E+13	.01	.01	4.E+15
40.1.Caponigro.23a.*	.01	.01	4.E+27	.01	.01	1.E+41	.01	.01	1.E+28
40.1.Caponigro.25b.*	.01	.01	2.E+29	.01	.01	2.E+37	.01	.01	1.E+24
40.1.Heck.5b.*	.01	.01	5.E+24	.01	.01	5.E+27	.01	.01	1.E+24
40.1.Stepanov.4b.*	.01	.01	2.E+13	.01	.01	2.E+20	.01	.01	2.E+20
40.2.Johnson.59b.*	.01	.01	2.E+21	.01	.01	6.E+21	.01	.01	6.E+23
40.4.Hicks.23.*	.01	.01	3.E+14	.01	.01	1.E+16	.01	.01	2.E+05
41.1.Müller.14c.*	.01	.01	4.E+13	.01	.01	2.E+13	.01	.01	3.E+11
41.1.Müller.25b.??(*)	.01	.01	2.E+20	.01	.01	5.E+29	.01	.01	2.E+21
41.2.Bruening.31a.*	.01	.01	2.E+18	.01	.01	7.E+18	.01	.01	1.E+24
41.2.Bruening.33a.*	.01	.01	7.E+17	.01	.01	2.E+16	.01	.01	3.E+21
41.2.Bruening.36b.*	.06	.01	7.E+00	.01	.01	3.E+05	.01	.01	2.E+09
41.2.Bruening.3b.*	.01	.01	4.E+14	.01	.01	2.E+13	.01	.01	7.E+17
41.3.Costantini.2a.??	.27	.02	1.E+00	.01	.01	3.E+04	.01	.01	4.E+07
41.3.Landau.10b.*	.96	.98	8.E-02	.84	.63	9.E-02	.03	.01	5.E+01
41.3.Landau.25c.*	.07	.01	3.E+01	.07	.01	5.E+00	.04	.01	5.E+00
41.3.Landau.27b.*	.01	.01	2.E+24	.01	.01	2.E+24	.01	.01	5.E+18
41.3.Landau.7b.*	.01	.01	2.E+27	.01	.01	3.E+34	.01	.01	6.E+23
41.3.Rezac.3b2.*	.02	.01	6.E+00	.01	.01	9.E+02	.01	.01	5.E+10
41.3.Vicente.4a3.*	.01	.01	9.E+33	.01	.01	9.E+39	.01	.01	2.E+20
41.3.Vicente.4a6.*	.01	.01	8.E+13	.01	.01	3.E+21	.01	.01	2.E+20
41.3.Vicente.5a.*	.01	.01	1.E+21	.01	.01	3.E+23	.01	.01	8.E+22
41.3.Vicente.8a.*	.01	.01	4.E+24	.01	.01	2.E+32	.01	.01	1.E+29
41.3.Vicente.8d.*	.01	.01	4.E+36	.01	.01	2.E+59	.01	.01	1.E+29
41.4.Bruening.9b.*	.12	.05	5.E-01	.40	.21	2.E-01	.01	.01	3.E+00
41.4.Haegeman.22a.*	.01	.01	4.E+19	.01	.01	9.E+17	.01	.01	1.E+21
41.4.Haegeman.4a.*	.01	.01	6.E+05	.01	.01	3.E+06	.01	.01	2.E+09

References

Alexopoulou, T., Keller, F., 2007. Locality, cyclicity and resumption: at the interface between the grammar and the human sentence processor. Language 83, 110–160.

Baayen, R.H., 2007. Analyzing Linguistic Data: A Practical Introduction to Statistics using R. Cambridge University Press, New York.

Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59, 390–412.

Bader, M., Häussler, J., 2010. Toward a model of grammaticality judgments. Journal of Linguistics 46, 273–330.

Balluerka, N., Goméz, J., Hidalgo, M.D., 2005. Null hypothesis significance testing revisited. Methodology 1, 55–70.

Bard, E.G., Robertson, D., Sorace, A., 1996. Magnitude estimation of linguistic acceptability. Language 72, 32–68.

Bates, D.M., Maechler, M., Bolker, B., 2012. lme4: Linear Mixed-effects Models using S4 Classes. R Package Version 0.999999-0. http://CRAN.R-project.org/package=lme4.

Bornkessel-Schlesewsky, I., Schlesewsky, M., 2007. The wolf in sheep's clothing: against a new judgment-driven imperialism. Theoretical Linguistics 33, 319–333.

Bošković, Ž., Lasnik, H., 2003. On the distribution of null complementizers. Linguistic Inquiry 34, 527–546.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

Clark, H.H., 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior 12, 335–359.

Clifton, C., Jr., Fanselow, G., Frazier, L., 2006. Amnestying superiority violations: processing multiple questions. Linguistic Inquiry 27, 51–68.

Cohen, J., 1976. Random means random. Journal of Verbal Learning and Verbal Behavior 15, 261–262.
Cohen, J., 1994. The Earth is round ($p < .05$). American Psychologist 49, 997–1003.
Cowart, W., 1997. Experimental Syntax: Applying Objective Methods to Sentence Judgments. Sage, Thousand Oaks, CA.
Culbertson, J., Gross, S., 2009. Are linguists better subjects? British Journal for the Philosophy of Science 60, 721–736.
Culicover, P.W., Jackendoff, R., 2010. Quantitative methods alone are not enough: response to Gibson and Fedorenko. Trends in Cognitive Sciences 14, 234–235.
Dąbrowska, E., 2010. Naïve v. expert intuitions: an empirical study of acceptability judgments. Linguistic Review 27, 1–23.
den Dikken, M., et al., 2007. Data and grammar: means and individuals. Theoretical Linguistics 33, 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, 353–367.
Featherston, S., 2005a. Magnitude estimation and what it can do for your syntax: some *wh*-constraints in German. Lingua 115, 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, 269–318.
Featherston, S., 2008. Thermometer judgments as linguistic evidence. In: Riehl, C.M., Rothe, A. (Eds.), Was ist Linguistische Evidenz? Shaker Verlag, Aachen.
Featherston, S., 2009. Relax, lean back, and be a linguist. Zeitschrift für Sprachwissenschaft 28, 127–132.
Ferreira, F., 2005. Psycholinguistics, formal grammars, and cognitive science. Linguistic Review 22, 365–380.
Gallistel, R., 2009. The importance of proving the null. Psychological Review 116, 439–453.
Gibson, E., Fedorenko, E., 2010. Weak quantitative standards in linguistics research. Trends in Cognitive Sciences 14, 233–234.
Gibson, E., Fedorenko, E., 2013. The need for quantitative methods in syntax and semantics research. Language and Cognitive Processes 28, 88–124.
Gibson, E., Piantadosi, S., Fedorenko, K., 2011. Using Mechanical Turk to obtain and analyze English acceptability judgments. Language and Linguistics Compass 5, 509–524.
Gigerenzer, G., Richter, H., 1990. Context effects and their interaction with development: area judgments. Cognitive Development 5, 235–264.
Gigerenzer, G., Krauss, S., Vitouch, O., 2004. The null ritual: what you always wanted to know about significance testing but were afraid to ask. In: Kaplan, D. (Ed.), The Sage Handbook of Quantitative Methodology for the Social Sciences. Sage, Thousand Oaks, CA.
Gordon, P., Hendrick, R., 1997. Intuitive knowledge of linguistic co-reference. Cognition 62, 325–370.
Greenbaum, S., 1973. Informant elicitation of data on syntactic variation. Lingua 31, 201–212.
Grewendorf, G., 2007. Empirical evidence and theoretical reasoning in generative grammar. Theoretical Linguistics 33, 369–381.
Gross, S., Culbertson, J., 2011. Revisited linguistic intuitions. British Journal for the Philosophy of Science 62, 639–656.
Haider, H., 2007. As a matter of facts – comments on Featherston's sticks and carrots. Theoretical Linguistics 33, 381–395.
Hubbard, R., Lindsay, R.M., 2008. Why *p* values are not a useful measure of evidence in statistical significance testing. Theory and Psychology 18, 69–88.
Ipeirotis, P.G., 2010. Demographics of Mechanical Turk. Center for Digital Economy Research Working Papers 10, Available at http://hdl.handle.net/2451/29585.
Jaeger, T.F., 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language 59, 434–446.
Keller, F., 2000. Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality. University of Edinburgh, (Ph.D. Dissertation).
Keppel, G., 1976. Words as random variables. Journal of Verbal Learning and Verbal Behavior 15, 263–265.
Kruschke, J.A., 2011. Doing Bayesian Data Analysis: A Tutorial with R and BUGS. Academic Press, New York.
Myers, J., 2009a. The design and analysis of small-scale syntactic judgment experiments. Lingua 119, 425–444.
Myers, J., 2009b. Syntactic judgment experiments. Language and Linguistics Compass 3, 406–423.
Newmeyer, F.J., 1983. Grammatical Theory: Its Limits and its Possibilities. University of Chicago Press, Chicago.
Newmeyer, F.J., 2007. Commentary on Sam Featherston, 'Data in generative grammar: the stick and the carrot'. Theoretical Linguistics 33, 395–399.
Nickerson, R.S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods 5, 241–301.
Phillips, C., 2003. Linear order and constituency. Linguistic Inquiry 34, 37–90.
Phillips, C., 2010. Should we impeach armchair linguists? In: Iwasaki, S., Hoji, H., Clancy, P., Sohn, S.-O. (Eds.), Japanese-Korean Linguistics, vol. 17. CSLI Publications, Stanford, CA, pp. 49–64.
Phillips, C., Lasnik, H., 2003. Linguistics and empirical evidence: reply to Edelman and Christiansen. Trends in Cognitive Sciences 7, 61–62.
R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0. http://www.R-project.org/.
Raaijmakers, J.G., 2003. A further look at the ''Language-as-fixed-effect fallacy''. Canadian Journal of Experimental Psychology 57, 141–151.
Raaijmakers, J.G., Schrijnemakers, J.M.C., Gremmen, F., 1999. How to deal with the ''Language-as-fixed-effect fallacy'': common misconceptions and alternative solutions. Journal of Memory and Language 41, 416–426.
Rouder, J.N., et al., 2009. Bayesian *t* tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin and Review 16, 225–237.
Schütze, C.T., 1996. The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology. University of Chicago Press, Chicago.
Schütze, C.T., Sprouse, J., 2013. Judgement data. In: Podesva, R.J., Sharma, D. (Eds.), Research Methods in Linguistics. Cambridge University Press, (in press).
Shaver, J.P., 1993. What statistical significance testing is, and what it is not. Journal of Experimental Education 61, 293–316.
Smith, J.E.K., 1976. The assuming-will-make-it-so fallacy. Journal of Verbal Learning and Verbal Behavior 15, 262–263.
Sorace, A., Keller, F., 2005. Gradience in linguistic data. Lingua 115, 1497–1524.
Spencer, N.J., 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. Journal of Psycholinguistic Research 2, 83–98.

Sprouse, J., 2007. A Program for Experimental Syntax. University of Maryland, (Ph.D. Dissertation).
Sprouse, J., 2011a. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. Behavior Research Methods 43, 155–167.
Sprouse, J., 2011b. A test of the cognitive assumptions of magnitude estimation: commutativity does not hold for acceptability judgments. Language 87, 274–288.
Sprouse, J., Almeida, D., 2012. Assessing the reliability of textbook data in syntax: Adger's *Core Syntax*. Journal of Linguistics 48, 609–652.
Sprouse, J., Almeida, D., 2013. The role of experimental syntax in an integrated cognitive science of language. In: Grohmann, K., Boeckx, C. (Eds.), The Cambridge Handbook of Biolinguistics. Cambridge University Press, New York, pp. 181–202.
Sprouse, J., Almeida, D. Power in acceptability judgment experiments and the reliability of data in syntax, submitted for publication.
Sprouse, J., Wagers, M., Phillips, C., 2012. A test of the relation between working memory capacity and island effects. Language 88, 82–123.
Stevens, S.S., 1956. The direct estimation of sensory magnitudes: loudness. American Journal of Psychology 69, 1–25.
Wasow, T., Arnold, J., 2005. Intuitions in linguistic argumentation. Lingua 115, 1481–1496.
Weskott, T., Fanselow, G., 2011. On the informativity of different measures of linguistic acceptability. Language 87, 249–273.
Wickens, T.D., Keppel, G., 1983. On the choice of design and of test statistic in the analysis of experiments with sampled materials. Journal of Verbal Learning and Verbal Behavior 22, 296–309.
Wike, E.L., Church, J.D., 1976. Comments on Clark's ''The language-as-fixed-effect fallacy''. Journal of Verbal Learning and Verbal Behavior 15, 249–255.