Psychology Press
Taylor & Francis Group

The empirical status of data in syntax: A reply to Gibson and Fedorenko

Jon Sprouse[1] and Diogo Almeida[2]

[1]Department of Cognitive Sciences, University of California, Irvine, CA, USA
[2]Department of Psychology, New York University Abu Dhabi, Abu Dhabi, UAE

This is a commentary in response to The need for quantitative methods in syntax and semantics research, by Edward Gibson and Evelina Fedorenko.

Keywords: Acceptability judgments; Syntax; Experimental syntax.

Gibson and Fedorenko (2010, henceforth G&F) claim that the traditional methods of data collection in syntax are invalid. They argue that these methods routinely yield unreliable data, which in turn casts doubt on the validity of the resulting syntactic theories. As a solution to this reliability problem, they propose a formal recipe for data collection that is superficially similar to data collection methods in other domains of experimental psychology. They contend that this recipe will lead to more reliable results, and presumably, better empirical support for syntactic theories. These are fundamentally empirical claims that can be investigated relatively straightforwardly. And as it turns out, the preponderance of available evidence suggests that G&F's claims are empirically false: traditional methods yield remarkably reliable data, and are well-powered with respect to the effect sizes of the phenomena of interest to syntacticians.

IS THE DATA IN SYNTACTIC THEORY UNRELIABLE?

In support of their claims regarding the unreliability of traditional collection methods in syntax, G&F present three examples of pair-based phenomena that were originally reported in the literature as differing in acceptability (from Chomsky, 1986; Gibson, 1991; Kayne, 1983) but failed to show statistically significant differences in the formal experiments reported by G&F. We reprint the examples here for convenience, with the originally reported judgments using the standard diacritic system from the syntax literature, where * means very unacceptable, and ? means mildly unacceptable.

Correspondence should be addressed to Jon Sprouse, 3151 Social Science Plaza A, Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100, USA. E-mail: jsprouse@uci.edu

http://www.psypress.com/lcp http://dx.doi.org/10.1080/01690965.2012.703782

(1) a. *The man that the woman the dog bit likes eats fish. (Gibson, 1991)
b. ?I saw the man that the woman that the dog bit likes.
(2) a. *I'd like to know where who hid it. (Kayne, 1983)
b. ?I'd like to know where who hid what.
(3) a. *I wonder what who saw. (Chomsky, 1986)
b. What do you wonder who saw?

What should we make of these three examples? First and foremost, it should be noted that these three examples were not chosen randomly from the syntax literature. Case (1) is a parsing preference from Gibson (1991), and therefore, arguably not part of the theoretical syntax literature G&F wish to criticize. Case (2) is a frequently cited replication failure that was previously published in the syntax literature (first reported by Clifton, Fanselow, & Frazier, 2006). Case (3) involves two sentence types from Chomsky (1986); however, these two sentence types do not form a minimal pair, and it does not appear as though Chomsky intended them to be a direct contrast [instead he appears to intend a direct contrast between (3b) and his example (107), and cites (3a) as a construction that may share the same violation as his (107)]. Because G&F present a very small, biased sample of phenomena, and because every field of experimental psychology acknowledges that some replication failures will occur (after all, significance is conventionally set at $p < .05$), we argue that little can be concluded from G&F's case studies about the validity of traditional methods in generative syntax and the reliability of the data they generate. A true assessment of the reliability of the data used in theoretical syntax requires testing a large, unbiased sample of data points from the full range of phenomena in the field to estimate a comprehensive replication rate under the formal experimental recipe suggested by G&F. Only then could one draw conclusions about the empirical status of the field.

We have conducted at least two studies that present the type of large-scale, comprehensive assessments of the reliability of data in the syntactic literature that are necessary to answer the questions raised by G&F. In the first (Sprouse & Almeida, in press), we tested the reliability of a large body of data that is used to motivate modern syntactic theories. First, we extracted every unique, English acceptability judgment from a popular syntax textbook (*Core Syntax* by Adger, 2003), which yielded a total of 469 unique sentence types that constitute 365 phenomena and cover 9 broad topic areas in syntactic theory. Then we re-tested those data points using formal magnitude estimation (ME) and yes-no experiments that incorporate all of the best practices suggested by G&F (large samples of naïve participants, multiple items for each condition, etc). If G&F are correct that traditional methods routinely yield unreliable data and that formal experiments will fix this problem, one should expect to find a non-negligible divergence between the results of the traditional experiments (as reported in Adger, 2003) and the results of the formal experiments (as reported in Sprouse & Almeida, in press). Using conservative statistical assumptions (i.e., two-tailed p-values, treating marginal p-values as replication failures), we found that 359 out of the 365 phenomena, or at least 98% of the data, clearly replicated in the formal experiments.

In the second study (Sprouse, Schütze, & Almeida, submitted), we tested the reliability of data points from recent journal articles in order to assess the reliability of data at the cutting edge of syntactic theory. First, we extracted every standard acceptability judgment data point published in *Linguistic Inquiry*, a premiere journal in theoretical linguistics, from 2001 to 2010, which yielded a total of 1743 data points. Next, we randomly sampled 292 sentence types that form 146 two-condition phenomena from that full data set. Finally, we re-tested those 146 phenomena using formal experiments (and following the best practices advocated by G&F) in order to estimate the replication rate of cutting edge journal data. Crucially, the random sampling procedure allows us to use the results of re-testing the sample to estimate the replication rate with a margin of error of ±5% (based on the relative size of the random sample to the population). Again, adopting conservative statistical assumptions, we found that 139 out 146 phenomena, or at least 95% of the sample data, replicated in the formal experiments.

In short, both of the extant large-scale assessments of the reliability of data in syntax, which together cover 511 phenomena found in textbooks and cutting-edge journal articles, have yielded conservative replication rates that are at or above 95%. The wide empirical coverage of these assessments compared to the case studies reported by G&F (511 vs. 3), not to mention the unbiased nature of the selection processes (exhaustive and random sampling), suggest to us that, whatever problems G&F ascribe to traditional methods in syntax, these methods have not generated an epidemic of unreliable data in syntactic theory.

WHY ARE TRADITIONAL METHODS SO RELIABLE?

The reliability revealed by the studies discussed above may be surprising to readers from domains other than syntax. However, there are at least two reasons why traditionally collected acceptability judgment data are as reliable as they are. The first is that acceptability judgments are cheap and easy to systematically replicate (at least in well represented languages like English) because they require no special equipment or procedures. For instance, every presentation of syntactic data, from conference presentations to journal papers, can lead to a round of replication as audiences and readers test the judgments for themselves (see also Marantz, 2005; Phillips, 2009). The second reason for the remarkable reliability is that the phenomena of interest to syntacticians tend to have relatively large effect sizes. Figure 1 presents a histogram (in discrete stacked circles) of standardised effect sizes (Cohen's *d*, the mean difference divided by the mean standard deviation) for the pair-based phenomena tested from Adger (2003) and *Linguistic Inquiry* 2001–2010 that were reported in the previous section.

The distributions in Figure 1 suggest that the preponderance of phenomena (87% in *Linguistic Inquiry*, 97% in Adger, 2003) of interest to syntacticians fall in the range of *d* values that Cohen (1988) described as *medium* (0.5–0.8) and *large* (>0.8). Crucially, Cohen defined *medium* effect sizes as those that a trained researcher could see with the naked eye (i.e., no statistical analysis necessary). To the extent that phenomena in

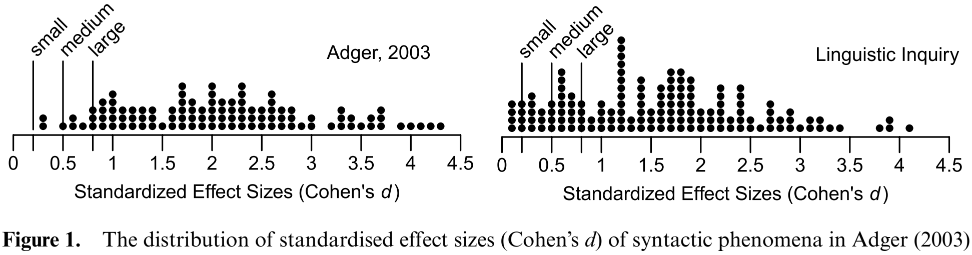

Figure 1. The distribution of standardised effect sizes (Cohen's *d*) of syntactic phenomena in Adger (2003) (left), and Linguistic Inquiry (right). The *x*-axis indicates effect size, which can range from 0 to infinity. The unmarked *y*-axis indicates counts for each effect size: one circle represents one occurrence. The three vertical lines represent Cohen's (1988) suggestion for interpreting effect sizes.

syntax are at least medium in size, it is not surprising that the effects can be reliably observed without the formal methods suggested by G&F.

To see the consequences of these large effect sizes, we (Sprouse & Almeida, submitted) estimated the statistical power of one of the qualitative tasks that are routinely used in traditional syntactic data collection: a two-alternative forced-choice task (2AFC) in which participants are asked to choose which of two sentences is more acceptable (for some evidence of the prevalence of the 2AFC task in traditional methods, see, e.g., Bard, Robertson, & Sorace, 1996, p. 34 and Myers, 2009, p. 414). First, we chose 50 phenomena from *Linguistic Inquiry* (2001–2010) that span the lower half of the effect sizes present in the full sample (Cohen's *d* of 0.15–1.96), as these smaller effect sizes likely provide the most information about statistical power. Next, we collected 2AFC judgments from 144 participants so as to collect a large number of independent judgments for each phenomenon. Then we ran re-sampling simulations to derive empirical power estimates for each phenomenon, and for each possible sample size between 5 and 100 participants. Contrary to G&F's claim that large samples are required for valid data in syntax, we found that a mere 10 judgments (one per subject) was sufficient to detect 70% of the phenomena in *Linguistic Inquiry* (2001–2010) *with 80% power* (a best practice for statistical power suggested by Cohen, 1988), and that the empirical coverage raises to 78% with 15 judgments (one per subject). In other words, the small, qualitative experiments that syntacticians have traditionally employed appear to be well calibrated to the phenomena that syntacticians have traditionally investigated, just as one would expect from a serious scientific field (for fuller details of these results, see Sprouse & Almeida, submitted).

WOULD THE METHODOLOGICAL SUGGESTIONS OF G&F NECESSARILY LEAD TO MORE RELIABLE DATA?

Although the studies above suggest that traditional methods are both reliable and well-suited to the phenomena of interest to syntacticians thus far, one might still wonder whether G&F's methodological suggestions would nonetheless lead to data that are even more reliable, in which case their suggestion would still be helpful (despite their reasons for the suggestion being false). However, G&F offer no empirical evidence in support of their claim that formal experiments of the kind they advocate will yield more reliable or valid results. Instead, they simply assume that the results of their formal experiments are more trustworthy than traditional syntactic methods, which, in turn, leads them to claim that the differences reported in the literature were *false positives*, and that the invariances obtained in their experiments were *true negatives*. However, the reverse is also logically possible: the differences reported in the literature could be *true positives*, and the invariances obtained in their experiments could be *false negatives*, perhaps due to a lack of statistical power. To investigate this possibility, we re-tested the three case studies using the 2AFC task and a large sample size (98 participants). We found that phenomenon (1) actually replicates (62/98, $p = .006$ by sign test) contrary to GF's claim, as does phenomenon (2) (58/98, $p = .04$ by sign test), leaving only phenomenon (3) (arguably a theoretically spurious comparison) as a true replication failure (it trends in the opposite direction: 43/98, $p = .13$ by sign test). In other words, there is reason to believe that two out of three of the null results reported by GF were in fact false negatives that arose because their formal experiments may have been underpowered.

The fact that G&F's case studies (1) and (2) may have been underpowered raises interesting questions about the relative power of the two sets of methods. To investigate this, we also used re-sampling simulations to empirically estimate the statistical power of the different ratings tasks advocated as best practice by critics of traditional methods, such as ME judgment and 7-point Likert-scale tasks, and compared estimates of their statistical power estimates to the estimates of the statistical power of 2AFC judgment tasks for the 50 phenomena from LI discussed previously (Sprouse & Almeida, submitted). The results suggest that ME and 7-point scale tasks are systematically less powerful than 2AFC tasks at detecting differences in acceptability between conditions. To the extent that the 2AFC task can stand as a formalised approximation of (a substantial part of) traditional syntactic methodology, these results further suggest that formal methods cannot simply be assumed to be more powerful than traditional methods without a more detailed discussion of the relative power of the relevant experiments (see also Gigerenzer & Richter 1990 and Gigerenzer, Krauss, & Vitouch, 2004 for a discussion of the comparative merits of forced-choice tasks over simple ratings tasks in other domains of psychology).

SHOULD SYNTACTICIANS WORRY ABOUT COGNITIVE BIAS?

Reliability questions aside, G&F also raise concerns about the use of professional linguists as participants during traditional data collection sessions. G&F argue that linguist-participants may consciously or subconsciously be influenced by their recognition of the experimental manipulation, and their own desires to see a theory supported or falsified. In other words, linguists could be affected by a cognitive bias unlikely to affect naive participants. Although G&F offer no evidence to support this claim, we believe that the studies discussed above can be used to look for evidence of cognitive bias on the part of linguist-participants. If the data reported in the syntax literature were affected by cognitive bias, we might expect two patterns to arise in the data. First, linguists would likely report differences between conditions that were theoretically convenient, but that naive participants did not perceive. In other words, cognitive bias should lead to replication failures when traditionally collected data are tested in formal experiments. As discussed above, there were very few replication failures (13 out of 511 phenomena), likely far fewer than predicted by the hypothesis that traditional methods are severely contaminated by cognitive bias (especially since some of these replication failures could be false negatives that arose due to lack of statistical power). The second prediction would be that linguists would report differences between conditions that went in the opposite direction of the difference reported by naive participants. In other words, cognitive bias should lead to *sign-reversals*. There were only two sign-reversals out of the 13 replication failures, likely far fewer than predicted by the cognitive bias hypothesis.

While cognitive bias is certainly a *potential* problem with the use of linguist-participants, the empirical facts suggest that there is no evidence that it is a *de facto* problem. This distinction should not be surprising, since even G&F implicitly admit that only biases that have been documented to affect experimental results, or that we have good reasons to think might affect experimental results, should be controlled for in an experiment. This is evidenced by the fact that they did not propose double-blind experiments, for instance, which are routinely used to control for experimenter bias in clinical trials, but are routinely ignored in experimental psycholinguistics. We can only conclude that they did not recommend this precaution because experimenter bias is not a confound that has been substantiated in most psycholinguistic experiments. Similarly, there is to our knowledge (and personal experience) no evidence that the kinds of cognitive biases mentioned by G&F negatively impact acceptability judgment experiments (see also Featherston, 2009 for another list of potential confounds that have been previously claimed by critics to unduly influence linguistic judgment tasks, but that in fact do not). There are, in principle, an infinite number of *potential* confounds in any given experiment. Therefore, it is crucial for researchers to be explicit about (1) how the confound in question could potentially impact the interpretation of the results and (2) whether there is any empirical evidence that the confound is affecting the results.

CONCLUSIONS

We, like many syntacticians, share G&F's belief that syntactic theory will be well served by the addition of formal experiments into the syntactician's repertoire. This is because there are some questions, such as those addressed in this special volume, that clearly require formal experiments along the lines of those suggested by G&F (see also Sprouse, Wagers, & Phillips, 2012 for a study that compares acceptability judgments to working memory measures in order to tease apart grammatical and reductionist approaches to island effects). However, we do not share G&F's view that traditional data collection methods used by syntacticians are invalid, nor that syntactic theory is predicated upon unreliable data, because the preponderance of available evidence suggests otherwise. We also do not share their view that all syntacticians should adopt a single data collection recipe to be universally applied to all theoretical questions, because current evidence suggests (1) that traditional methods are, in fact, well calibrated to the phenomena of interest to syntacticians and (2) that blind faith in the reliability or inherent superiority of formal methods can lead to a large number of false negatives, an outcome that would be as problematic as the scenario G&F suggest syntactic theory to be in. Instead, we believe that syntacticians should have the flexibility to decide which methods are best suited for the theoretical question of interest. Fortunately, syntacticians have two great options. Traditional methods are reliable and powerful, and can be quickly deployed with minimal expense (they are often free), but cannot easily provide detailed quantitative information such as gradient judgments or comparisons to other data types (e.g., working memory measures). Formal rating experiments, however, provide additional quantitative information, and can more easily be compared to other data types, but are more costly, and may be less powerful than the 2AFC task at detecting differences between conditions.

REFERENCES

Adger, D. (2003). *Core syntax: A minimalist approach*. Oxford: Oxford University Press.

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

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Gibson, E., & Fedorenko, E. (2010). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*. DOI: 10.1080/01690965.2010.515080

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 D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gigerenzer, G., & Richter, H. (1990). Context effects and their interaction with development: Area judgments. *Cognitive Development*, *5*, 235–264.

Kayne, R. (1983). Connectedness. *Linguistic Inquiry*, *14*, 223–249.

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

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

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

Sprouse, J., & Almeida, D. (in press). Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics*.

Sprouse, J., & Almeida, D. (submitted). Power in acceptability judgment experiments and the reliability of syntactic data. (Manuscript available online: http://ling.auf.net/lingBuzz/001520)

Sprouse, J., Schütze, C. T., & Almeida, D. (submitted). Assessing the reliability of journal data in syntax: Linguistic Inquiry 2001–2010. (Manuscript available online: http://ling.auf.net/lingBuzz/001352)

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working memory capacity and island effects. *Language*, *88*, 82–123.