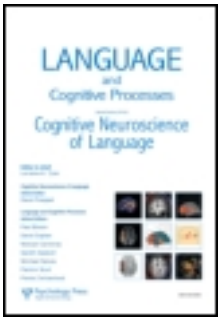
This article was downloaded by: [Edward Gibson]
On: 04 October 2012, At: 01:37
Publisher: Psychology Press
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Language and Cognitive Processes
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/plcp20

Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2012)

Edward Gibson [a b], Steven T. Piantadosi [c] & Evelina Fedorenko [a]

[a] Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, Cambridge, MA, USA

[b] Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA, USA

[c] Brain and Cognitive Sciences Department, University of Rochester, Rochester, NY, USA

Version of record first published: 04 Oct 2012.

To cite this article: Edward Gibson, Steven T. Piantadosi & Evelina Fedorenko (): Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2012), Language and Cognitive Processes, DOI:10.1080/01690965.2012.704385

To link to this article: http://dx.doi.org/10.1080/01690965.2012.704385

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused

arising directly or indirectly in connection with or arising out of the use of this material.

Downloaded by [Edward Gibson] at 01:37 04 October 2012

LANGUAGE AND COGNITIVE PROCESSES, 2012, iFirst, 1–12

Psychology Press
Taylor & Francis Group

Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2012)

Edward Gibson[1,2], Steven T. Piantadosi[3], and Evelina Fedorenko[1]

[1]Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA, USA
[3]Brain and Cognitive Sciences Department, University of Rochester, Rochester, NY, USA

Downloaded by [Edward Gibson] at 01:37 04 October 2012

Sprouse and Almeida (S&A) present quantitative results that suggest that intuitive judgments utilised in syntax research are generally correct in two-condition comparisons: the sentence type that is presented as "good/grammatical" is usually rated better than the sentence type that is presented as "bad/ungrammatical" in controlled experiments. Although these evaluations of intuitive relative judgments are valuable, they do not justify the use of nonquantitative linguistic methods. We argue that objectivity is a universal value in science that should be adopted by linguistics. In addition, the reliability measures that S&A report are not sufficient for developing sophisticated linguistic theories. Furthermore, quantitative methods yield two additional benefits: *consistency of judgments* across many pairs of judgments; and an understanding of the *relative effect sizes* across sets of judgments. We illustrate these points with an experiment that demonstrates five clear levels of acceptability. Finally, we observe that S&A's experiments—where only two authors evaluated 10 years' worth of journal articles and one standard textbook within a few months—further emphasise one of our critical original points: conducting behavioural experiments is in many respects *easy and fast* with the advent of online research tools like Amazon's Mechanical Turk. Given the current ease of performing quantitative experiments (using a platform like Mechanical Turk) and the clear limitations of not doing so, linguistic hypotheses should be evaluated quantitatively whenever it is feasible.

Keywords: Syntax; Semantics; Quantitative methods; Sentence processing.

In our recent papers arguing for the need to gather quantitative evidence to test syntactic and semantic hypotheses, we focused on the danger of cognitive biases in evaluating one's own hypotheses with one's own judgments (Gibson & Fedorenko, 2010a). In response to this concern, Sprouse and Almeida (2012, henceforth S&A)

Correspondence should be addressed to Edward Gibson, 46-3035, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: egibson@mit.edu

We would like to thank Leon Bergen, Peter Graff, Jeremy Hartman, Frank Keller, Melissa Kline, Roger Levy, Kyle Mahowald, Hal Tily, and Tom Wasow for their comments on earlier drafts of this paper.

© 2012 Psychology Press, an imprint of the Taylor & Francis Group, an Informa business

http://www.psypress.com/lcp http://dx.doi.org/10.1080/01690965.2012.704385

present quantitative results that suggest that the intuitive judgments from a recent syntax textbook and recent articles in the journal *Linguistic Inquiry* are generally correct in two-condition comparisons: the sentence type that is presented as "good/grammatical" is usually rated better than the sentence type that is presented as "bad/ungrammatical". Consequently, S&A argue that "syntacticians should have the flexibility to decide which methods are best suited for the theoretical question of interest", including continuing to use traditional nonquantitative methods.[1]

Although we did not explicitly state it in earlier papers, we agree with S&A that nonquantitative methods have been useful in the early stages of the scientific study of language. Much of cognitive science has its roots in the works of philosophers who did not have the tools to evaluate their hypotheses quantitatively and were limited to introspection and "thought experiments". However, in many cases (and certainly in the case of syntactic intuitions), we no longer lack such tools, and thus nothing should stop us from evaluating claims about the human mind with scientific rigor. The ability of S&A to perform a massive experiment—covering 10 years' worth of journal articles and one standard textbook—further emphasises one of our critical original points: conducting behavioural experiments is in many respects *easy and fast* with the advent of online research tools like Amazon's Mechanical Turk.

Here, we respond to S&A's points and present several further arguments in favour of quantitative methods. We focus our discussion on S&A's evaluation of *Linguistic Inquiry* (summarised in S&A from Sprouse, Schutze & Almeida, submitted), because this evaluation is most relevant to the question at hand: what the standards should be for current research. First, we discuss S&A's statistics on the reliability of expert judgments, and argue that one should only consider *theoretically meaningful contrasts* in this calculation, with the consequence that S&A's estimates are inflated. Second, even if we accept the reliability estimates that S&A report, we argue that developing rich linguistic theories requires *even higher reliabilities* only obtainable through quantitative research. Third, we present two important additional reasons to perform quantitative experiments—reasons which hold regardless of the reliability of expert intuitions: quantitative experiments (1) maintain *consistency of judgments* across many pairs of judgments and (2) reveal the *relative effect sizes* across sets of judgments, which can often be used to determine if a factor is likely to be theoretically important. We conclude with a discussion of objectivity: expert intuitions are not considered data in any domain of science, and linguistics should adapt to value objectivity in data collection and analysis.

Downloaded by [Edward Gibson] at 01:37 04 October 2012

THE RELEVANT FALSE POSITIVE RATE IS FOR THEORETICALLY MEANINGFUL CONTRASTS

There are some open questions about how to properly measure the validity of judgments in the current syntax/semantics literature. Following S&A's approach, one would need to establish: (1) how many theoretically critical judgments are there overall

[1] S&A also state the following: "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". We would like to clarify that we are *not* arguing that all syntacticians should adopt a single "data collection recipe": there is no single "best" method, and there is much value in using multiple methods. What we are arguing is that, whichever method a researcher uses, they should apply it quantitatively across multiple naive participants and multiple items, so that statistical significance can be assessed. There is a wide range of methods that satisfy these constraints, from simple offline acceptability rating tasks to functional MRI and electrocorticography.

(the denominator in the fraction that S&A report); and (2) how many of these judgments were correctly determined when using a nonquantitative method (the numerator in the fraction that S&A report). The ratio of these then gives a measure of the reliability of nonquantitative linguistic research, corresponding to the *probability* that the nonquantitative result will be replicated in a rigorous experiment (i.e., will be "real"). Let us first discuss S&A's denominator in this measure.

Sprouse and Almeida's testbed of 146 contrasts from 10 years of *Linguistic Inquiry* articles, out of 1,743 contrasts total in that time period, includes many examples like (1)–(5).[2]

(1) a. * Was kissed John. b. John was kissed.
(2) a. * This is table. b. This is a table.
(3) a. * Me would have been elected. b. I would have been elected.
(4) a. * Sarah saw pictures of. b. Kerry attempted to study physics.
(5) a. * The was arrested student. b. The student was arrested.

All of these contrasts represent well-known phenomena: English subject noun phrases tend to occur to the left of their verb phrases, as in (1a)/(1b); singular count nouns like "table" need a determiner, as in (2a)/(2b); accusatively marked noun phrases like "me" are not good as subjects of finite clauses, as in (3a)/(3b); etc. However, in deciding on the best methods for current research, the relevant examples are not these kinds of sentences: the relevant examples are ones that can *distinguish among current theories*. Although contrasts like these certainly constrain theories, they are not representative of the forefront of syntactic research because all current linguistic theories correctly predict contrasts like (1)–(5). The inclusion of such examples in the denominator (a), therefore, means that S&A's reliability estimate is likely inflated, and not a good estimate for the reliability of cutting-edge syntactic data.

VERY HIGH RELIABILITY IS REQUIRED FOR COMPLEX THEORIES

The second component of a useful measure of the validity of syntactic judgments in the current literature is the number of the theoretically relevant examples that are correct and meaningful. S&A estimate this number as 139 of 146 for their evaluation of materials from *Linguistic Inquiry*, such that there were seven examples where the judgment did not reliably match S&A's quantitative evaluation, for an error rate of about 5%.[3] Their textbook data had an error rate of about 2%. Although S&A concede that there are some true judgment errors in this set, they hypothesise that some others might yet be real, but that their experiments might not have had enough power to detect the relevant differences. As we discuss in detail below, we think that most of these cases are unlikely to be meaningful in distinguishing theories at this point: putative contrasts with tiny effect sizes should be given less weight in theory

[2] Some of the 146 contrasts [e.g., that between (4a) and (4b)] are further not minimal pairs. We asked S&A for the examples that they presented to Mechanical Turk raters, but they provided only the guidelines they used for generating the materials. Consequently, it is difficult to evaluate S&A's study in detail.

[3] S&A's error estimate may be too low. In collaboration with Peter Graff and Jeremy Hartman, we are currently conducting a study of approximately 100 acceptability contrasts from the same articles of *Linguistic Inquiry* that S&A investigated. These contrasts were randomly selected from all the materials that S&A *didn't* investigate. Preliminary results indicate that the error rate in this set of contrasts is larger than the one reported by S&A: in the range of approximately 10%. If the error rate is as large as this, it means that the problems that we discuss here are magnified.

Downloaded by [Edward Gibson] at 01:37 04 October 2012

construction than contrasts associated with larger effect sizes. Furthermore, as discussed above, S&A's reliability estimate is inflated by the inclusion of examples that are not relevant to distinguishing current theories. But even if we accept these estimates, we can consider how reliable linguistic comparisons need in order to build complex linguistic theories: is a 2–5% error rate "good enough"?

Sprouse and Almeida hypothesise that a 2–5% error rate is good enough, because 5% is generally accepted to be the maximum false positive rate for psychology publications ($p < .05$).[4] This, however, is a faulty comparison. What counts as an acceptable error rate for a particular experiment depends on how complex the theories are, which are constructed from the experiment and associated experiments. In most subdomains of linguistics, theories are constructed to get *every* grammaticality rating right in some set, not just most of them. Thus, if we base a theory on a set of grammaticality ratings that has any incorrect data points, the linguistic theory we construct to explain the data will necessarily be the wrong one in some way. There is an interesting question of how close the wrong grammar will be to the correct one, if we assume that the data are "close" to the correct data. Our guess is that this relationship will be somewhat "brittle", such that the simplest grammar for deriving a slightly incorrect set of positive and negative examples may differ substantially from the simplest grammar for the correct set of positive and negative examples.[5] Hence, for our purposes, we will assume that the goal of gathering linguistic data is to get the data set perfectly correct, since even small mistakes in the data to be explained may result in a faulty theory. Ideally, as we discuss later, the theory should predict the entire range of acceptability ratings. But for now we assume that theories only predict binary comparisons, as this will allow a back-of-the-envelope lower-bound estimate of how high a reliability is required.

Given these assumptions, with 95% reliability of linguistic comparisons, the probability of getting all N contrasts right is 0.95^N. We can assume that we want a moderately high probability of getting perfect data, say 80%. In this case, if each data point is correctly measured with 95% probability as in S&A's journal data, we will only be able to base a theory on $\log(0.8)/\log(0.95) = 4.3$ data points (comparisons). This means that when a theory is based on five or more grammatical comparisons, it will have a less than 80% chance of being based on correct data (and thus a less than 80% chance of being correct, assuming it derives the observed data). For S&A's 98% reliability measure on textbook data, we can only build theories based on about 11 comparisons. We can also ask the reverse question: if we had a linguistic theory based on, say, 25 or 50 comparisons, how high a reliability would we need for the individual data points? The answer is $e^{\log(0.8)/25} = 0.9911$ and $e^{\log(0.8)/50} = 0.9955$, respectively, to have an 80% chance of being right on all of them. Note that each of these estimates is higher than the 95% or 98% reliability rates that S&A observed (and much higher than the approximately 90% reliability rate mentioned in footnote 3). Importantly, these reliabilities are not outside the domain of experimental methods, for which, *p*-values can be made arbitrarily low, and statistical power arbitrarily high. Clearly, *very good data are required in order to build rich linguistic theories*, and higher and higher

[4] Though 5% is generally accepted to be the maximum false positive rate for psychology publications ($p < .05$), many if not most reported results have a substantially lower *p*-value. Indeed—assuming an effect is real—the false positive rate can be made arbitrarily low by running more participants. With quantitative methods, one can directly compute this false positive rate and statistical power for any observed set of data. The ability of experimental methods is, therefore, qualitatively unlike that of expert judgments, since expert judgments do not promote reporting of false positive rates for particular contrasts, or have the ability to lower them to acceptable ranges with further effort.

[5] But, of course, this relationship should be formally studied.

Downloaded by [Edward Gibson] at 01:37 04 October 2012

reliability is needed as theories are based on more and more data points. By analogy, although one might be relatively confident driving a car that had a 2–5% chance of breaking down each day, one would certainly not want to drive a car with 50 parts, *each* of which had a 2–5% chance of breaking down each day.

Of course, the 2–5% number is only an estimate of the reliability of linguistic judgments. More pessimistically, what S&A's results actually reveal is that current linguistic theory—even that in textbooks—derives the wrong set of grammatical contrasts, *contra* S&A's assertion that informal consideration of grammaticality effectively weeds out incorrect data. In a certain sense, though, the problem is not *that* there are errors in textbooks or journal articles—this is surely true of all fields engaged in incremental research. And, as S&A point out, experiments too can give the wrong answer since each experiment has imperfect statistical power and a nonzero false positive rate. The problem is that nonquantitative methods have *no hope of recognising or correcting these errors*. It is only once S&A conducted behavioural experiments that they discovered *that* there were errors and, crucially, *which* purported contrasts were erroneous. Experimental methods were required to do this evaluation because experimental methods are the only way to objectively determine which hypothesised contrasts are "real".

In fact, without quantitative methods, there is considerable uncertainty about what the data actually are. It is useful here to repeat a back-of-the-envelope calculation from Gibson and Fedorenko (2010a): if we assume that 5% of linguistic judgments are erroneous, there will be on average about 87 incorrect judgments in a collection of 1,743 expert intuitive ratings. If we rely only on intuitions, we will not know which data points are wrong, giving a total of (1,743 choose 87) $= 5.26 \times 10^{148}$ possibilities for which subset of the data is incorrect. This number is so large as to be unfathomable. Even if we restrict our attention to only a small subset of relevant data for our theorising, say 60 data points, we still end up with (60 choose 3) = 34,220 possibilities for which subset of the data is correct/incorrect. It is clear that this is not a viable situation. Without quantitative methods, we have no way, even in principle, of discovering this vagueness of contemporary linguistic theory, much less choosing among these possibilities.

Finally, S&A argue that the power required to detect typical linguistic contrasts is quite low, and suggest that this justifies the informal (i.e., nonquantitative, nonstatistical) methods. In our view, this is akin to saying "Most t-tests come out, so why run them?" The significance of any *particular* contrast of interest will depend on the *particular* responses that subjects give to it, and so its significance can only be established by statistically analysing the data gathered about *that specific contrast*. Hence, even if it is true that most of the time you will not need many subjects, that does not mean that a good research ethic is to trust findings evaluated informally by a few subjects, especially if the few subjects have not been analysed statistically. Doing so can only decrease the reliability of data and give a false sense of confidence in results. Good scientific inferences—across the entire range of effect sizes—simply *require* quantitative data and statistics.

FURTHER BENEFITS OF QUANTITATIVE METHODS

In these calculations and in interpreting S&A's reliability measures, it is important not to confuse the *correctness* of a judgment with the *meaningfulness* of a judgment to a theory. Doing so critically ignores *the effect size* for the comparison in making this connection. That is, it is more important for a theory to account for contrasts with

Downloaded by [Edward Gibson] at 01:37 04 October 2012

large effect sizes [as in examples (1)–(5)] than it is for a theory to account for small effect sizes [as in (6a) vs. (6b), a putative contrast that both we and S&A have discussed before].

(6) Peter was trying to remember
  a. what who carried.
  b. what who carried when.

Relatedly, one should not confuse statistical significance with meaningful significance (e.g., Cohen, 1994; Gigerenzer, 1994; Nickerson, 2000): even if there is a real difference between two conditions [such as may be the case with respect to (6a) vs. (6b)], the difference may be so small as to be practically meaningless. Under S&A's calculations, 7 of 146 comparisons in their *Linguistic Inquiry* testbed did not reach significance (5%) but a further 13 had small effect sizes (9%), for a total of 14% which were either nonsignificant or had small effects. We probably want our theories to explain the larger effect contrasts first. But without an experimental and statistical evaluation of contrasts like that performed by S&A, there is no way to quantify different effect sizes.[6]

Quantitative methods address this issue and go further, allowing for *consistency of judgments* across many pairs of judgments within and across papers (e.g., Featherston, 2007; Keller, 2001). Typical linguistics papers provide judgments for 50 or more paired contrasts. These judgments need to fit into a global context of linguistic acceptability. Consequently, even if all the relative judgments for pairs of conditions are correct, the overall picture may be oversimplified or wrong.

Consider, for example, a case consisting of two sentence pairs, each of which is proposed to contrast in grammaticality, where one example in each pair is hypothesised to be grammatical as compared to its control condition, which is hypothesised to be ungrammatical. Even if the relative judgments in each pair are correct, the better example from one pair might be worse than the worse example from the other pair, leading to inconsistency. For instance, this situation arises in the examples from Chomsky (1986) that were experimentally investigated in Gibson and Fedorenko (2010a). Chomsky states that (7) (his (105), p. 48) is more grammatical than the appropriate reading of (8) (his (107), p. 49), with "how" interpreted as modifying the embedded verb "fixed" (we provide Chomsky's grammaticality judgments):

(7) What do you wonder who saw?
(8) * How_i do you wonder who fixed the car t_i

Furthermore, in an immediately adjacent section of the same monograph, Chomsky considers multiple-wh extraction ungrammatical in English when the wh-object is fronted, as in (9) (Chomsky's (108), p. 49):

(9) * I wonder what who saw.

[6] We seem to be in disagreement with S&A about the need to pay attention to effect size. S&A state 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". (A false negative is a case where a real difference exists but none was observed in S&A's first quantitative evaluation.) In contrast to S&A, we do not think that these false negatives are problematic to the field. It is useful information when a predicted effect does not come out as predicted in a high-powered experiment. It suggests that the effect is perhaps not worth accounting for initially, and that other, larger, effects are worth explaining first.

Downloaded by [Edward Gibson] at 01:37 04 October 2012

In order to have a consistent set of judgments across these constructions, it follows that (9) must be less acceptable than (7).[7] Gibson and Fedorenko (2010a) tested this contrast in minimal pair contexts that supported each reading, as in (10):

(10) a. The manager tried to figure out what the waiter wondered who had ordered.
b. The manager tried to figure out if the waiter wondered what who had ordered.

If Chomsky's judgments are consistent, then (10b) [corresponding to his ungrammatical (9)] should be rated as less acceptable than (10a) [corresponding to his more grammatical (7)]. In contrast to this judgment, the opposite pattern was observed: materials like (10b) were rated as reliably *more* acceptable than materials like (10a) (see the work of S&A for a similar result in a forced choice paradigm).

Maintaining consistency of judgments is especially problematic in the nonquantitative approach to syntactic acceptability because there are traditionally only two or three values of grammaticality: "grammatical/good" vs. "questionable" vs. "ungrammatical/bad". When performing quantitative experiments, one quickly discovers that syntactic acceptability appears fine-grained and largely continuous. But with only two or three values of syntactic acceptability, it is impossible to achieve global consistency across many levels of acceptability (see also Scontras & Gibson, 2011, who observe that it is difficult to have intuitions about potential interactions among factors in syntactic experiments). We illustrate this problem in an experiment consisting of three subexperiments, using the conditions as in (11), (12), and (13), below.

In (11), we compared the two critical conditions for which Fedorenko and Gibson (2010) and Clifton, Fanselow, and Frazier (2006) found no difference in acceptability. This null effect did not match the intuitions of Bolinger (1978) and Kayne (1983), according to whom sentences like (11b) are more acceptable than sentences like (11a).

In (12) and (13), we compared extractions from dative arguments of double argument verbs like "give" and "offer". In Fillmore (1965), it is stated that sentences like (12a) and (13a) —in which the argument of the prepositional phrase object of the verb is questioned—are grammatical, whereas sentences like (12b) and (13b) — in which the first object of a double object construction is questioned—are ungrammatical.

Following discussions of such examples in Wasow and Arnold (2005) and Langendoen, Kalish-Landon, and Dore (1973), we had probed the acceptability of materials like these in earlier experiments, and found that the acceptability varies a lot depending on the verb: extraction of the first object of a ditranstivie (NP NP) structure is more acceptable for verbs like "offer", "lend", "show" and "give" than for verbs like "toss", "mail", "pass", "lease", "rent", and "throw" (probably due to properties of the usage frequencies of these different verbs, cf. Bresnan & Nikitina, 2009). In the current experiment, we compared conditions involving the extraction of the dative argument across the subcategorisation structures (NP NP or NP PP), for the two sets of verbs. In each of these subexperiments, we also included a completely

[7] S&A argue that the only relevant comparison for Chomsky's (7) is the control that he provides in (8). But surely his theory needs to be internally consistent in its judgments across other comparisons that he provides. This is one reason why we investigated the alternative control in (9) in Gibson and Fedorenko (2010a). Another reason to investigate (9) as a control for (7) is that it is much easier to construct minimal pairs contrasting (7) and (9): (7) and (8) are different on many levels, thus making a comparison between them difficult.

Downloaded by [Edward Gibson] at 01:37 04 October 2012

ungrammatical condition in which the word order was not a possible English word order for a similar meaning as the other conditions. See Appendix 1 for a full list of the experimental materials.

(11) 2WH vs. 3WH extraction
a. Multiple-Wh object-subject
Julius tried to remember what who carried.
b. Multiple-Wh object-subject +third-wh
Julius tried to remember what who carried when.
(12) More acceptable dative extractions (verb-set 1)
a. Extraction from PP, verb-set 1
Madison tried to figure out who Gerald offered a loan to.
b. Extraction of goal NP, verb-set 1
Madison tried to figure out who Gerald offered a loan.
c. Ungrammatical control, verb-set 1
Madison tried to figure out who did offered Gerald a loan to.
(13) More acceptable dative extractions (verb-set 2)
a. Extraction from PP, verb-set 2
Joyce tried to remember who Donovan tossed a ball to.
b. Extraction of goal NP, verb-set 2
Joyce tried to remember who Donovan tossed a ball.
c. Ungrammatical control, verb-set 2
Joyce tried to remember who did tossed Donovan a ball to.

Downloaded by [Edward Gibson] at 01:37 04 October 2012

We posted surveys for 120 workers on Amazon.com's Mechanical Turk using the Turkolizer software from Gibson, Piantadosi, and Fedorenko (2011). Each participant completed a questionnaire consisting of a different randomised order of the 36 items across the three subexperiments, in a Latin Square design for each subexperiment, where each participant saw only one version of each item, and an equal number of items from each condition in each subexperiment. The task was to rate the naturalness of each sentence on a scale from 1 (extremely unnatural) to 7 (extremely natural). In order to ensure that participants read and understood each sentence, they also answered a simple comprehension question about each.

We only analysed participants who (1) self-identified as being native speakers of English; (2) correctly answered at least 75% of comprehension questions following each item; and (3) filled out only one survey. This left 107 participants. Results with other methods of trimming participants revealed similar patterns of results. We then z-transformed the ratings (subtracting a participant's mean across all trials, and dividing that value by the standard deviation across all trials). The means and standard errors for the means by participants for the conditions across the three subexperiments are presented in Figure 1.

Analyses reported here were conducted with the lme4 package (Bates, in press) for the statistical language R (R Core Development Team, 2008). Recent results have shown that including only random intercepts in Linear Mixed Effects regressions can be anti-conservative, so we also include random slopes for participants and items in our model (Barr , Levy, Scheepers, & Tily, submitted). For simplicity, we present pairwise dummy-coded regressions of the relevant contrasts, but note that similar results hold in a more sophisticated and higher power single regression, that uses treatment and Helmert coding. Significance (p) values were estimated from (1) the t-values that were obtained from the lmer function; and (2) conservative estimates of the number of degrees of freedom in the model. The estimates of the number of degrees of freedom in the model

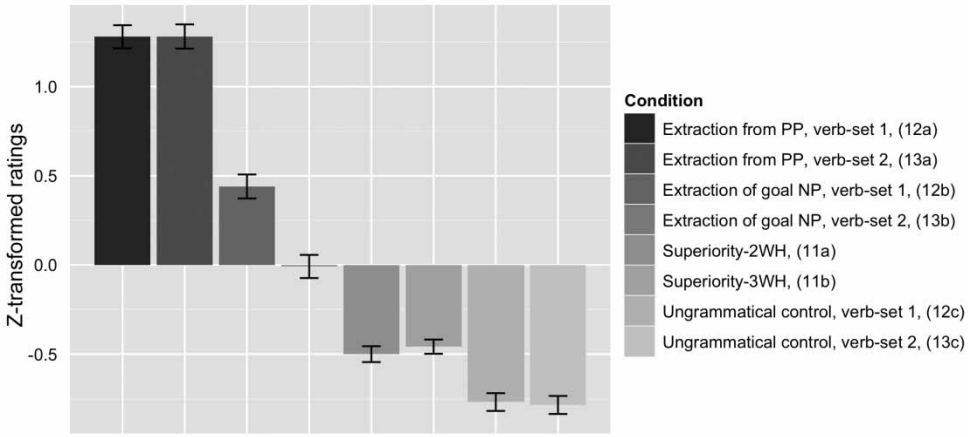


Figure 1. Z-transformed acceptability ratings for the eight conditions across three subexperiments, with standard errors of the mean by participants.

Downloaded by [Edward Gibson] at 01:37 04 October 2012

consisted of the number of observations (1,280) minus the number of intercepts fit in the model (the number of participants + the number of items = 107 + 12 = 119).

As can be seen by inspection of the acceptability ratings, the multiple-wh condition with three wh-phrases ($M = -0.458$ *SDs*) was rated as slightly more acceptable than the multiple-wh condition with two wh-phrases ($M = -0.500$ *SDs*), though not significantly so, despite the large number of participants and items in this experiment ($\beta = .043$, $t = 1.14$, $p = .21$).

In addition, we see that there are four levels of naturalness among the dative extractions. The most acceptable conditions are the extractions from PPs, with means of 1.281 *SDs* and 1.280 across the two verb sets (which did not differ significantly from each other). Extractions of the NP goal argument from the NP-NP structures were rated as less acceptable than the extractions from PPs (verb set 1: $\beta = .841$, $t = 11.94$, $p < .001$; verb set 2: $\beta = 1.290$, $t = 18.58$, $p < .001$). Extractions of the NP goal argument from the NP-NP structures were rated as more acceptable in verb set 1 with verbs like "offer" ($M = 0.440$ *SDs*) than in verb set 2 with verbs like "toss" ($M = -0.009$ *SDs*) ($\beta = .450$, $t = 8.53$, $p < .001$). Finally, the ungrammatical control conditions were rated less acceptable than any other conditions (at a mean rating of -0.785 and -0.769 *SDs*) but not significantly different from each other.

We find, therefore, five different levels of acceptability across these materials, and with more contrasts we likely would have found more (Featherston, 2007; Keller, 2001). Critically, note that although extractions of datives from NP-NP structures are rated as worse than comparable extractions from NP-PP structures, all of these structures are much more acceptable than the multiple-wh conditions in (11a) and (11b). Thus, it is problematic to consider the extractions of datives from NP-NP structures "ungrammatical" as suggested in Fillmore (1965), if one also wants to say that the multiple-wh examples in (11b) are only "questionable". With only three rating categories, it is, therefore, impossible to rate these structures appropriately. Furthermore, note that Gibson and Fedorenko (2010a) showed that Chomsky's (1986) multiple-wh example (7)/(10a) is even less acceptable than the multiple-wh structures that were investigated here. Thus although a structure like (7)/(10a) might be more acceptable than some other control that Chomsky had in mind, it is problematic to consider this kind of structure acceptable, as Chomsky originally indicated.

This experiment also illustrates the point we made above about relative effect sizes: without gathering quantitative evidence, it is impossible to assess the relative sizes of effects across comparisons. Effect size is particularly relevant for the multiple-wh comparison here. As S&A correctly observe, it is difficult to interpret the null effect that we observe here: there could well be a significant difference between these two conditions, as S&A found in their own evaluation of this comparison. However, we can confidently state that the acceptability difference between the two conditions, if there is one, is so small as to be practically meaningless. There are many much larger effects which we probably want our theories to explain first. Once we better understand the larger effects, it is possible that small effects like this one will also be explained by the new theories. But there is no way to assess effect size intuitively: one needs a quantitative experiment.

EXPERT INTUITIONS ARE NOT DATA

We conclude with perhaps the most general statement of our disagreement with S&A and standard linguistic methods: we view expert linguist judgments as essentially *expert predictions*, whereas S&A treat them as *data* (see also Johnson, 2008, for related arguments). Our arguments against treating the subjective grammaticality judgments of the researchers involved in the study were outlined in Gibson and Fedorenko (2010a) and include most notably cognitive biases on the part of the researchers. Indeed, to our knowledge, there is no other field of science where the intuitions of the investigators are treated as admissible data for evaluating theories. Even in fields which—like linguistics—study cognitive phenomena that apply to all humans (i.e., most subfields of psychology/cognitive science), there is a clear divide between the intuitions which motivate experiments or scientific theories, and the data which can falsify those theories. Science, in short, seeks *objectivity*, a value that has a rich history in the development of scientific method (see Daston & Galison, 2007). It is hard to find common ground for debate if objectivity is not a valued aspect of investigation; indeed, objectivity might be viewed as one of the defining characteristics of the scientific enterprise.

To be fair, there is an interesting question of how to treat expert opinions when they are reliably correlated with objective data. S&A argue that because linguists' judgments are strongly correlated with the outcome of behavioural experiments on naive participants, the behavioural experiments are not necessary. However, we believe that this is a serious mistake. One would never argue that because a physicist has been able to predict the outcome of past experiments, more experiments are no longer necessary because we could just rely on the physicist's intuitions. Even in psychology, one would not argue that because an experimenter's, say, moral intuitions have been found to match experimental results, one should no longer conduct the experiments. The experiments are necessary in *all* cases because they provide the only way to objectively measure discrepancies between theory and reality. Theories evaluated only by the intuitions of the investigators involved, are almost necessarily *post hoc*. This is because, lacking quantitative standards, we have little possibility to be wrong and discover discrepancies between theory and data. Such discrepancies drive scientific progress, and while expert intuitions provide a rich source of hypotheses to investigate, reliable *evaluation* of such hypotheses requires more sophisticated quantitative methods. We have argued that quantitative methods allow for the study of absolute and relative differences in grammaticality, and that a linguistic theory should aim to

Downloaded by [Edward Gibson] at 01:37 04 October 2012

explain the full variation in grammaticality judgments, not only pairwise comparisons. However, we also believe that it is a mistake to regard phenomena in the 2–5% tail—which S&A appear content to get wrong—as unworthy of study. Many of the most important advances in science were made by recognising and correcting small discrepancies,[8] and in the science of language such tiny effects can only be studied with quantitative methods. Ideally, science should first aim to explain the largest objectively measured effects with fully formalised theories, and continue to develop methods that encourage richer, more fine-grained quantitative analysis. This kind of scientific progress requires increasingly powerful methods: once the limitations of, say, behavioural grammaticality ratings and magnitude estimation are better understood, the field should continue developing methods in order to gather more and more refined data. At no point—until language is fully understood—should we be content with measurable methodological limitations.

Downloaded by [Edward Gibson] at 01:37 04 October 2012

REFERENCES

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (submitted). Random effects structure in mixed-effect models: Keep it maximal.

Bates, D. M. (in press). *lme4: Mixed-effects modeling with R*. New York: Springer.

Bolinger, D. (1978). Asking more than one thing at a time. In H. Hiz (Ed.), *Questions* (pp. 97–106). Dordrecht: D. Reidel.

Bresnan, J., & Nikitina, T. (2009). The gradience of the dative alternation. In L. Uyechi & L. H. Wee (Eds.), *Reality exploration and discovery: Pattern interaction in language and life* (pp. 161–184). Stanford, CA: CSLI.

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

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–103.

Daston, L., & Galison, P. (2007). *Objectivity*. New York: Zone Books.

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

Fedorenko, E., & Gibson, E. (2010). Adding a third wh-phrase does not increase the acceptability of object-initial multiple-wh-questions. *Syntax*, *13*, 183–195.

Fillmore, C. (1965). *Indirect object constructions in English and the ordering of transformations*. Mouton: The Hague.

Gibson, E., & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Science*, *14*, 233–234.

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

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, *5/8*, 509–524.

Gigerenzer, G. (1994). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Johnson, K. (2008). The need for explicit inferential methods in linguistics. In C. R. Dreyer (Ed.), *Language and Linguistics Emerging Trends* (pp. 1–16). Nova Science Publishers.

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

Keller, F. (2001). Gradience in grammar: Experimental and computational aspects of degrees of grammaticality (Doctoral dissertation, University of Edinburgh).

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

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

[8] A famous example is the perihelion precession of Mercury, in which Newtonian mechanics predicted an orbital trajectory that was off by about 0.01 degrees per *century*, a discrepancy that provided an otherwise puzzling deviation that relativity could explain.

R Core Development Team. (2008). R: A language and environment for statistical computing. Version 2.70. Vienna: R Foundation for Statistical Computing. Online: http://cran.R-project.org.
Scontras, G. & Gibson, E. (2011). A quantitative investigation of the imperative-and-declarative construction in English. *Language*. 87, 817-829.
Sprouse, J., & Almeida, D. (2012). The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes*. doi: 10.1080/01690965.2012.703782.
Sprouse, J., & Almeida, D. (to appear). The role of experimental syntax in an integrated cognitive science of language. In K. Grohmann & C. Boeckx (Eds.), *The Cambridge handbook of biolinguistics*.
Sprouse, J., Schutze, C. T., & Almeida, D. (submitted). Assessing the reliability of journal data in syntax: Linguistic Inquiry 2001–2010. Retrieved from http://ling.auf.net/lingBuzz/001352
Wasow, T., & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua*, *115*, 1481–1496.

APPENDIX 1: EXPERIMENT MATERIALS

Sub-experiment 1: 2WH vs. 3WH extraction

1. Hannah tried to remember what who cooked (when).
2. Dillon tried to remember what who won (when).
3. Mandy tried to remember what who took (when).
4. Julius tried to remember what who carried (when).
5. Carmen tried to remember what who sponsored (when).
6. Hector tried to remember what who brought (when).
7. Jodi tried to figure out what who discarded (when).
8. Daphne tried to figure out what who analyzed (when).
9. Clarence tried to figure out what who broke (when).
10. Darryl tried to figure out what who cherished (when).
11. Angie tried to figure out what who dreaded (when).
12. Sebastian tried to figure out what who appreciated (when).

Sub-experiment 2: Dative extractions, verb-set 1

1. Frank tried to figure out who Jerry mailed a package (to).
2. Glenn tried to remember who Maxwell mailed a letter (to).
3. Susan tried to figure out who Mitchell leased a car (to).
4. Beverly tried to remember who Byron leased a truck (to).
5. Howard tried to figure out who Margaret rented a house (to).
6. Pamela tried to remember who Christy rented an apartment (to).
7. Christopher tried to figure out who Isaac threw a football (to).
8. Marvin tried to remember who Lindsey threw an orange (to).
9. Patrick tried to figure out who Ross tossed a frisbee (to).
10. Joyce tried to remember who Donovan tossed a ball (to).
11. Angela tried to figure out who Laurel passed a newspaper (to).
12. Tyler tried to remember who Renee passed a container (to).

Sub-experiment 3: Dative extractions, verb-set 2

1. Diane tried to figure out who Bryan lent his car (to).
2. Helen tried to figure out who Emerson lent a coat (to).
3. Duncan tried to remember who Valerie lent a cellphone (to).
4. Alice tried to figure out who Monica offered a pencil (to).
5. Madison tried to remember who Gerald offered a loan (to).
6. Norman tried to figure out who Carlton offered a drink (to).
7. Brenda tried to remember who Kyle showed a magazine (to).
8. Rebecca tried to figure out who Ashley showed a painting (to).
9. Neil tried to remember who Chloe showed a sculpture (to).
10. Wesley tried to remember who Christina gave a present (to).
11. Jay tried to remember who Karen gave a book (to).
12. Elaine tried to figure out who Marilyn gave a gift (to).

Downloaded by [Edward Gibson] at 01:37 04 October 2012