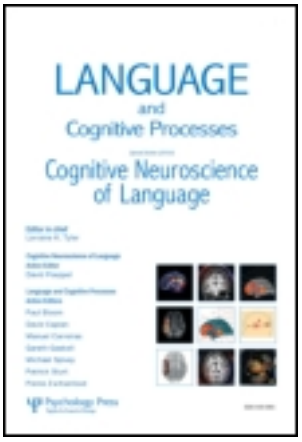

Language and Cognitive Processes
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/plcp20

The need for quantitative methods in syntax and semantics research

Edward Gibson [a b] & 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

Version of record first published: 27 Oct 2010.

To cite this article: Edward Gibson & Evelina Fedorenko (2013): The need for quantitative methods in syntax and semantics research, Language and Cognitive Processes, 28:1-2, 88-124

To link to this article: http://dx.doi.org/10.1080/01690965.2010.515080

LANGUAGE AND COGNITIVE PROCESSES, 2013
Vol. 28, Nos. 1/2, 88–124, http://dx.doi.org/10.1080/01690965.2010.515080

The need for quantitative methods in syntax and semantics research

Edward Gibson[1,2] 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

The prevalent method in syntax and semantics research involves obtaining a judgement of the acceptability of a sentence/meaning pair, typically by just the author of the paper, sometimes with feedback from colleagues. This methodology does not allow proper testing of scientific hypotheses because of (a) the small number of experimental participants (typically one); (b) the small number of experimental stimuli (typically one); (c) cognitive biases on the part of the researcher and participants; and (d) the effect of the preceding context (e.g., other constructions the researcher may have been recently considering). In the current paper we respond to some arguments that have been given in support of continuing to use the traditional nonquantitative method in syntax/semantics research. One recent defence of the traditional method comes from Phillips (2009), who argues that no harm has come from the nonquantitative approach

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

This research conducted here was supported by the National Science Foundation under Grant No. 0844472, "Bayesian Cue Integration in Probability-Sensitive Language Processing". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We would like to thank Diogo Almeida, Leon Bergen, Joan Bresnan, David Caplan, Nick Chater, Morten Christiansen, Mike Frank, Adele Goldberg, Helen Goodluck, Greg Hickok, Ray Jackendoff, Nancy Kanwisher, Roger Levy, Maryellen MacDonald, James Myers, Colin Phillips, Steve Piantadosi, Steve Pinker, David Poeppel, Omer Preminger, Ian Roberts, Greg Scontras, Jon Sprouse, Carson Schütze, Mike Tanenhaus, Vince Walsh, Duane Watson, Eytan Zweig, members of TedLab, and three anonymous reviewers for their comments on an earlier draft of this paper. We would also like to thank Kristina Fedorenko for her help in constructing the materials for the experiment in Case Study 3.

in syntax research thus far. Phillips argues that there are no cases in the literature where an incorrect intuitive judgement has become the basis for a widely accepted generalisation or an important theoretical claim. He therefore concludes that there is no reason to adopt more rigorous data collection standards. We challenge Philips' conclusion by presenting three cases from the literature where a faulty intuition has led to incorrect generalisations and mistaken theorising, plausibly due to cognitive biases on the part of the researchers. Furthermore, we present additional arguments for rigorous data collection standards. For example, allowing lax data collection standards has the undesirable effect that the results and claims will often be ignored by researchers with stronger methodological standards. Finally, we observe that behavioural experiments are easier to conduct in English than ever before, with the advent of Amazon.com's Mechanical Turk, a marketplace interface that can be used for collecting behavioural data over the internet.

Keywords: Syntax; Semantics; Sentence processing; Confirmation bias.

1. INTRODUCTION

The prevalent method in syntax and semantics research over the past 45 years has involved obtaining a judgement of the acceptability of a sentence/meaning pair, typically by just the author of the paper, sometimes with feedback from colleagues. As has often been noted in recent years (Cowart, 1997; Edelman & Christiansen, 2003; Featherston, 2007; Ferreira, 2005; Gibson & Fedorenko, 2010a; Marantz, 2005; Myers, 2009; Schütze, 1996; Wasow & Arnold, 2005), the results obtained using this method are not necessarily generalisable because of (a) the small number of experimental participants (typically one); (b) the small number of experimental stimuli (typically one); (c) cognitive biases on the part of the researcher and participants; and (d) the effect of the preceding context (e.g., other constructions the researcher may have been recently considering). In order to address this methodological weakness, the authors above have argued that future syntax and semantics research should adopt the following standards, common in all behavioural sciences, when using acceptability judgements as evidence: (a) include many experimental participants; (b) include many experimental materials, in order to rule out effects due to idiosyncratic properties of individual experimental items (e.g., particular lexical items used in the critical sentences); (c) use naïve participants, all of whom are blind with respect to the research questions and hypotheses, and include distractor materials so that the experimental manipulations are not obvious to the participants; and (d) present counterbalanced lists of items to different experimental participants (in different random orders) to avoid effects of presentation order.

Despite the fact that these observations have been made many times, research in syntax and semantics continues in much the same way as before, as can be seen by examining recent issues of syntax and semantics journals. The point of the current paper is to respond to some arguments that have

been given in support of continuing to use the traditional nonquantitative method, including the recent defence by Phillips (2009). The long-term goal is to bring a change in the fields of syntax and semantics so that research questions in these fields are evaluated using quantitative methods.

The paper is organised as follows. We first briefly discuss how syntax and semantics questions are typically studied by language researchers in general and by researchers in the field of linguistics in particular (Section 2). We then summarise and respond to several arguments that have been provided in the literature in support of continuing to use the traditional nonquantitative method (Section 3). One recent claim, made by Phillips (2009), is that there are no cases in the literature where an incorrect intuitive judgement has become the basis for a widely accepted generalisation or an important theoretical claim. Contrary to Phillips' claim, we present three cases from the literature where a faulty intuition has led to incorrect generalisations and influential but mistaken theorising (Sections 4–6). We finish with some concluding remarks in Section 7.

2. BACKGROUND: EMPIRICAL INVESTIGATIONS OF SYNTAX AND SEMANTICS

There are two general ways to gather data in any scientific domain: through naturalistic observation or through controlled experiments. In language research, the naturalistic observation method consists of gathering corpora of naturally produced (written or spoken) language and analysing the distributions of particular linguistic elements (e.g., syntactic constructions) in order to make inferences about the architecture of the language system. Several researchers have argued that corpus analyses constitute the optimal way to investigate language (Bresnan, 2007; Bresnan & Nikitina, 2009; Labov, 1975, 1996). However, in language, as in any scientific domain, controlled experiments are useful for (a) dissociating theoretically interesting dimensions of linguistic stimuli that may be correlated in naturalistic productions; and (b) exploring uncommon phenomena that may be important for distinguishing among theories. Regarding the latter, it is a mistake to equate the rarity of a particular construction in a corpus with the syntactic and/or semantic ill-formedness of that construction, because production frequency is affected by many factors other than the syntactic and semantic complexity of an utterance, such as working memory demands (e.g., constructions with longer-distance dependencies may be dispreferred), or contextual factors (e.g., particular constructions may be licensed only by very specific contexts that may be very infrequent). Thus controlled experiments are critical in investigating syntactic and semantic phenomena.

Experiments that are designed to investigate syntactic or semantic research questions typically compare sentence materials in one condition to sentence materials in one or more other conditions, using some dependent measure (e.g., reading times, ratings of intuitive complexity/naturalness, event-related potentials, etc.). Because dependent measures in all experimental paradigms that involve the comprehension or production of sentences reflect the influences of many representational and processing factors (including but not limited to the syntax and semantics of the sentences under investigation), the interpretation of the results as being relevant to a particular syntactic or semantic question relies on a good experimental design, where factors other than the critical one(s) are not also different across conditions. Factors that must be controlled for in experiments investigating syntactic/semantic questions include: properties of the lexical items in the sentence (e.g., their frequencies and their lexico-semantic complexity); the pragmatics of the sentence; the prosody of the sentence; the plausibility of the event(s) described in the sentence according to general world knowledge; the local discourse context; the working memory resources that are needed to process the sentence; and any other properties of the materials that have been shown to affect language comprehension/production (e.g., temporary ambiguities that may lead to processing difficulty; Gibson & Pearlmutter, 1998).

Language researchers have used a wide variety of paradigms/dependent measures to investigate syntactic and semantic research questions, including self-paced reading or listening, eye-tracking during reading, elicited production (e.g., sentence completions or picture descriptions), eye-tracking in the visual domain while listening to/producing sentences, and brain imaging (PET, fMRI, EEG, MEG, NIRS) during reading or listening. A commonly used method in syntax and semantics research is the *acceptability judgment task* where participants are asked to judge the acceptability/naturalness of a sentence/meaning pair. The dependent measure is a judgement on some scale[1] where the ends of the scale correspond to "acceptable"/

[1] In a nonquantitative (single-participant/single-item) version of the acceptability judegment task a 2- or 3-point scale is typically used, usually consisting of "good"/"natural"/"grammatical"/"acceptable" vs. "bad"/"unnatural"/"ungrammatical"/"unacceptable" (usually annotated with an asterisk "*" in the papers reporting such judgements), and sometimes including a judgement of "in between"/"questionable" (usually annotated with a question mark "?"). In a quantitative version of the acceptability judgement task (with multiple participants and items) a fixed scale (a "Likert scale") with five or seven points is typically used. Alternatively, a geometric scale is used where the acceptability of each target sentence is compared to a reference sentence. This latter method is known as *magnitude estimation* (Bard, Robertson, & Sorace, 1996). Although some researchers have hypothesised that magnitude estimation allows detecting more fine-grained distinctions than Likert scales (Bard et al., 1996; Featherston, 2005, 2007; Keller, 2000), controlled experiments using both Likert scales and magnitude estimation suggest that the two methods are equally sensitive (Fedorenko & Gibson, 2010a; Fukuda, Michel, Beecher, & Goodall, 2010; Weskott & Fanselow, 2008, 2009).

"not acceptable", "natural"/"unnatural", "grammatical"/"ungrammatical", "good"/"bad", etc. Although the acceptability judgement task is a good paradigm for investigating questions in syntax and semantics,[2] most researchers do not use the acceptability judgement task in a quantitative way. In particular, they do not gather multiple observations in each condition, across experimental participants and items. A major advantage of quantitative methods is that they enable the use of inferential statistics to evaluate the likelihood of particular hypotheses.[3]

Before proceeding to discuss some arguments that have been made in defence of the traditional nonquantitative method in syntax/semantics research, we would like to clarify two issues that sometimes confuse discussions of methods in language research. First, we have encountered arguments that the acceptability judgement task is somehow optimal for probing syntactic/semantic knowledge, compared to other tasks or dependent measures, like reading times. The claim is that in this paradigm it may be easier to focus on the syntactic/semantic properties of the stimulus. However, to the best of our knowledge, no evidence for such a claim exists. In a judgement of a sentence's acceptability, all factors that affect language comprehension/production will be reflected. In order to isolate the effect of interest to syntax or semantics, it is necessary to use sentence materials that differ only in the critical syntactic/semantic property and are matched for all the other factors discussed above (and this is true regardless of what dependent measure is used: acceptability ratings; reading times; eye movements in the visual-world paradigm; etc.).

Second, we have encountered a claim that the reason for different kinds of methods being used across the different fields of language study (i.e., in linguistics vs. psycho-/neuro-linguistics) is that the *research questions* are different across these fields, and some methods may be better suited to ask some questions than others. Although the latter is likely true, the premise—that the research questions are different across the fields—is false. The

[2] It should be noted that some researchers have criticised the acceptability judgement method because it requires participants to be aware of language as an object of evaluation, rather than simply as a means of communication (Edelman & Christiansen, 2003). Whereas this concern is worth considering with respect to the research questions that are being evaluated, one should also consider the strengths of the acceptability judgement method: (1) it is an extremely simple and efficient task; and (2) results from acceptability judgement experiments are highly systematic across speakers and correlate with other dependent measures, presumably because the same factors affect participants' responses across different measures (Schütze, 1996; Sorace & Keller, 2005).

[3] In analysing quantitative data, it is important to examine the distributions of individual responses in order to determine whether further analyses may be necessary, in cases where the population is not sufficiently homogeneous with respect to the phenomena in question. A wide range of analysis techniques are available for not only characterising the population as a whole, but also for detecting stable sub-populations within the larger population or for characterising stable individual differences (e.g., Gibson et al., 2009).

typical claim is that researchers in the field of linguistics are investigating linguistic *representations*, and researchers in the fields of psycho-/neuro-linguistics are investigating the *computations* that take place as language is understood or produced. However, many researchers in the fields of psycho-/neuro-linguistics are also interested in the nature of the linguistic representations (at all levels; e.g., phonological representations, lexical representations, syntactic representations, etc.).[4] By the same token, many researchers in the field of linguistics are interested in the computations that take place in the course of online comprehension or production. However, inferences—drawn from *any* dependent measure—about either the linguistic representations or computations are always indirect. And these inferences are no more indirect in reading times or event-related potentials, etc., than in acceptability judgements: across all dependent measures we take some observable (e.g., a participant's rating on an acceptability judgement task or the time it took a participant to read a sentence) and we try to infer something about the underlying cognitive representations/processes. More generally, methods in cognitive science are often used to jointly learn about representations and computations, because inferences about representations can inform questions about the computations, and vice versa. For example, certain data structures can make a computation more or less difficult to perform, and certain representations may require assumptions about the algorithms being used.

In our opinion then, the distinction between the fields of linguistics and psycho-/neuro-linguistics is purely along the lines of the kind of data that are used as evidence for or against theoretical hypotheses: typically nonquantitative data in linguistics vs. typically quantitative data in psycho-/neuro-linguistics. Given the superficial nature of this distinction, we think that there should be one field of language study where a wide range of dependent measures is used to investigate linguistic representations and computations.

3. ARGUMENTS IN FAVOUR OF USING THE TRADITIONAL NONQUANTITATIVE METHOD

In this section, we summarise seven arguments that we have encountered in favour of continuing to use the traditional single-sentence/single-participant nonquantitative method that is prevalent in the syntax/semantics literature.[5]

[4] In fact, some methods in cognitive science and cognitive neuroscience were specifically developed to get at representational questions (e.g., lexical/syntactic priming methods, neural adaptation or multi-voxel pattern analyses in functional MRI).

[5] See http://www.talkingbrains.org/2010/06/weak-quantitative-standards-in.html for a recent presentation and discussion of some of these arguments.

3.1.a. A valid argument: Gathering quantitative data can sometimes be difficult

The first argument in favour of using the traditional nonquantitative method is a valid one: that, in circumstances where gathering data is difficult, it is better to gather as much data as possible, even if no statistical evaluation of hypotheses is possible (without a sufficient number of participants and/or items). For example, a speaker of a dying language may only be available for testing during a brief period of time. Of course, the conclusions that can be drawn from these data will be weaker and more speculative in nature than the conclusions based on quantitative data.

3.1.b. An invalid argument: Gathering quantitative data is generally difficult

Circumstances like those described in Section 3.1.1 do occur. However, the argument about *general* difficulty in gathering quantitative data is not valid. In particular, although there certainly are circumstances in language research where constraints exist on the amount of data that can be collected, in most of these circumstances, there is a constraint *only* on the number of participants or *only* on the number of trials. For example, consider a case of a speaker with a particular cognitive/neurological/genetic disorder. Such cases are often of great theoretical interest to language researchers because they enable inferences about the properties of the human cognitive system that are difficult or not possible to make by studying only healthy normally functioning individuals. However, it is often not possible to find many speakers with a similar cognitive/neurological/genetic profile, or—in some cases—averaging data from multiple speakers is not meaningful. For example, Caramazza (1986) advocates the single-patient case study approach in neuropsychological investigations of the language system because of high levels of variability in the cognitive deficits of brain-damaged patients, resulting from variability in the location, size, and severity of brain lesions. This variability makes inter-subject averaging meaningless: patients aren't similar enough to one another, and, as a result, valuable insights from cognitive profiles of single patients are likely to get lost in the averaging process. In circumstances where only a single individual or a small group of individuals are being investigated, *many trials per participants per condition* are necessary. This ensures that the observed phenomena are stable within the individual and are not due to idiosyncratic properties of particular experimental materials.

An example where multiple participants are relatively easily available but only one or a few trials are possible due to a short attention span and/or other factors is a case of human infants/young children or participants with

cognitive deficits affecting the individual's ability to tolerate prolonged testing sessions (e.g., medium-/low-functioning individuals with autism). A solution in cases where only a small number of data points can be gathered from any given individual is to test *large numbers of participants* to ensure that the observed phenomena are stable in the population under investigation and are not due to idiosyncratic properties of particular individuals.

Critically, in the case of languages like English and in studying neurologically healthy adults (which is the case for most of research in syntax and semantics), the constraints above (on the number of participants or on the number of trials) do not apply. First, multiple speakers are easily available, and it is generally assumed by language researchers that people's language systems are similar across native speakers of a language (the fact that hundreds of language phenomena, including syntactic/semantic phenomena, have been documented to hold across speakers of a language validates this assumption). Note also that this assumption—about the cognitive systems of healthy individuals being similar to one another—is standard across all fields of psychology, and averaging responses across participants is therefore a standard practice. And second, neurologically healthy adults are capable of undergoing testing sessions lasting extended periods of time, which means many trials per participant per condition can be gathered.

3.2. An invalid argument: Analyses of a small number of experimental participants are acceptable in other domains of cognitive research

A second argument for the continued use of the traditional nonquantitative method makes a parallel to fields of cognitive psychology like psychophysics, where data are gathered from a small number of participants, often fewer than five. If it is valid to investigate the responses of a small number of participants in those studies, then it should be similarly valid in linguistic research. We have two responses to this argument. First, psychophysics experiments include many (often hundreds of) trials per participant per condition (similar to cases where only a few participants are available, as discussed in Section 3.1.2), unlike syntax/semantics research, where there is typically a single experimental trial. The use of many trials per condition allows the use of inferential statistics. Second, unlike tasks involving higher-level cognitive processes (such as syntactic/semantic processing), there is plausibly less variability within and between participants in tasks typically investigated in psychophysics experiments, such as low-level visual or auditory perception tasks. Thus it is easier to generalise the findings from a small number of participants to the population at large. Such general-

isation is more difficult for a high-level cognitive process such as syntactic/semantic interpretation.

3.3. An invalid argument: Quantitative evidence is not always/ever *sufficient* for conclusively answering research questions, so why bother?

A third argument that we have encountered for using the traditional nonquantitative method is that even if one gathers quantitative evidence using the methods that we advocate here, this will not guarantee the lack of confounding factors in the materials or the accurate interpretation of data by the experimenter. Although the premises that there may be confounds in experimental materials and/or that the researcher(s) may make incorrect inferences about the meaning of the experimental results are valid, it does not follow that quantitative methods shouldn't be adopted. Indeed, if we apply this argument to all of science, we would conclude that because a single experiment is not enough to test a hypothesis adequately, we should never do any experiments, which is clearly absurd.

By adopting the methods advocated here, the experimenter will help minimise the presentation of data that are spurious, driven by one or more of the four problems with the traditional nonquantitative method presented above: (a)–(b) nongeneraliseability due to small numbers of participants or items; (c) cognitive biases on the part of the researcher or participants; and (d) context effects. Issues such as (1) how theoretically important and clearly formulated the research question is; (2) how well designed the critical contrast is; and (3) whether reasonable alternative hypotheses are considered are orthogonal to whether the data that are collected are quantitative or nonquantitative in nature.

3.4. An invalid argument: Gathering quantitative evidence is inefficient and will slow down progress in linguistic research

A fourth argument that has sometimes been given to us for the continued use of the traditional nonquantitative method is that it would be too inefficient to evaluate every syntactic/semantic hypothesis or phenomenon quantitatively. For example, Culicover and Jackendoff (2010) make this argument explicitly in their response to a letter we published in *Trends in Cognitive Sciences* summarising the key problems with nonquantitative methods in linguistic research (Gibson & Fedorenko, 2010a): "It would cripple linguistic investigation if it were required that all judgements of ambiguity and grammaticality be subject to statistically rigorous experiments on naive subjects, especially when investigating languages whose speakers are hard to access" (Culicover & Jackendoff, 2010, p. 234). Whereas we agree that in circumstances where gathering data is difficult, some evidence is better than

no evidence (as discussed in Section 3.1 above), we do not agree that research would be slowed with respect to languages where experimental participants are easy to access, such as English. In contrast, we think that the opposite is true: the field's progress is probably slowed by *not* doing quantitative research.

Suppose that a typical syntax/semantics paper that lacks quantitative evidence includes judgements for 50 or more sentences/meaning pairs, corresponding to 50 or more empirical claims. Even if most of the judgements from such a paper are correct or are on the right track, the problem is in knowing *which* judgements are correct. For example, suppose that 90% of the judgements from an arbitrary paper are correct (which is probably a high estimate).[6] This means that in a paper with 50 empirical claims 45/50 are correct. But which 45? There are 2,118,760 ways to choose 45 items from 50. That's over 2 million different theories. Indeed, the number of possible theories is even larger, because this result assumes that an experiment will either support or reject a hypothesis, whereas in many cases, the results may be in between, offering partial support or partial rejection of a hypothesis. By quantitatively evaluating the empirical claims, we reduce the uncertainty a great deal. To make progress, it is better to have theoretical claims supported by solid quantitative evidence, so that even if the interpretation of the data changes over time as new evidence becomes available—as is often the case in any field of science—the empirical pattern can be used as a basis for further theorising.

Furthermore, as will be discussed in more depth in the presentation of Case study 3 below, it is no longer expensive to run behavioural experiments, at least in English and other widely spoken languages. There now exists a marketplace interface—Amazon.com's Mechanical Turk—which can be used for collecting behavioural data over the internet quickly and inexpensively. The cost of using an interface like this is minimal, and the time that it takes for the results to be returned is short. For example, currently on Mechanical Turk, a survey of approximately 50 items will be answered by 50 or more participants within a couple of hours, at a cost of approximately US$1 per participant. Thus a survey can be completed within a day, at a cost of less than US$50. (The hard work of designing the experiment, and constructing controlled materials remains of course.)

[6] Colin Phillips and some of his former students/postdocs have commented to us that, in their experience, quantitative acceptability judgement studies almost always validate the claim(s) in the literature. This is not our experience, however. Most experiments that we have run which attempt to test some syntactic/semantic hypothesis in the literature end up providing us with a pattern of data that had not been known before the experiment (e.g., Breen, Fedorenko, Wagner, & Gibson, 2010; Fedorenko & Gibson, 2010a; Patel et al., 2009; Scontras & Gibson, 2010).

Note that there is possibly some validity to not evaluating *every* empirical claim quantitatively. A problem then arises, however, in deciding when it is ok *not* to gather quantitative evidence. For any new interesting claim, there has to be *some* quantitative evidence in support of this claim (which is not the case in the current literature). It is possible that some cases are so similar to others that an experiment is not needed to evaluate those cases. However, we recommend gathering quantitative evidence for all empirical claims, because conducting experiments is so easy to do. As we discuss in Case study 2 below, for example, we ran the study in the discussion of Case study 2 in a single day, including writing the items (which were admittedly similar to a previously existing set of items), posting the survey on Mechanical Turk, obtaining and analysing the results.

3.5. An invalid argument: Linguists make better experimental participants because they can ignore nonrelevant factors in evaluating linguistic materials

A fifth argument that has sometimes been presented to us as a reason to use the traditional nonquantitative methodology in linguistics research takes the form of a claim about the kind of participants that are most useful in evaluating theoretical hypotheses in linguistics. According to this claim, it is necessary to have detailed theoretical knowledge of the critical hypotheses in order to have useful judgements on the relevant examples.[7] Naïve participants would therefore not be useful in evaluating linguistic hypotheses: only syntactic theoreticians or semanticists who understand the relevant features of the critical examples would be capable of providing useful judgements on theoretically important contrasts. If this claim were true, then the study of language would be unlike the study of any other domain of cognition, where behavioural effects are present not only in theoretically aware individuals, but also those who are theoretically naïve. Indeed, in contrast to the present claim about the most useful experimental participants, naïve participants are actually *more useful* for evaluating theoretical hypotheses than participants who are aware of the theoretical hypotheses, because of the presence of *cognitive biases* in theoretically aware participants. The existence of such cognitive biases makes interpreting the results of experiments with theoretically aware participants difficult, perhaps even meaningless. We summarise some of these cognitive biases below. Because humans are subject to the

[7] Note that there is a tension between this claim and the Chomsky an idea that there is a universal language faculty possessed by all native speakers of a language, including naïve subjects. Moreover, as discussed above, the need to ignore irrelevant features of examples should be eliminated by good experimental design: the experimenter reduces the possibility of confounding influences by controlling theoretically irrelevant variables in the materials to be compared.

cognitive biases discussed below, linguistic experiments should be carried out on naïve experimental participants, just like in every other area of science investigating human behaviour.

3.5.1. Cognitive biases

There exist at least two types of cognitive biases related to the issue of being naïve vs. non-naïve with respect to research questions/hypotheses that can negatively affect the results of single-subject/single-item experiments (see Dabrowska, 2010; Ferreira, 2005; Schütze, 1996; Wasow & Arnold, 2005; for similar observations):

3.5.1.1. Confirmation bias on the part of the researcher. The researcher often suffers from a confirmation bias that favours the outcome that is predicted by his/her hypothesis, resulting in a tendency to justify ignoring data from participants/materials that do not fit the predicted pattern. That is, the researcher will tend to pay attention to the data that are consistent with the expected pattern of results, but will tend to ignore the data that are inconsistent with the hypothesis and would thus falsify it (e.g., Wason, 1960), treating such data as noise or problematic in some way (e.g., from a speaker who is not quite native, or from a speaker who may speak a different dialect).

3.5.1.2. Confirmation bias on the part of the participants. People that are asked by the researcher to provide a judgement for some construction(s)—including the researcher him/herself—may be biased due to their understanding of the theoretical hypotheses. Faced with complex materials (e.g., a multi-clause sentence or a sentence involving ambiguity), they may then rely on this knowledge of the theoretical hypotheses to arrive at the judgement. This type of cognitive bias is known as belief or confirmation bias, such as the bias demonstrated in experiments by Evans, Barston, and Pollard (1983) (cf. other kinds of confirmation bias, such as first demonstrated by Wason, 1960; see Nickerson, 1998, for an overview of many similar kinds of cognitive biases). In Evans et al.'s experiments, people were asked to judge the acceptability of a logical argument. Although the experimental participants were sometimes able to use logical operations in judging the acceptability of the arguments that were presented, they were most affected by their knowledge of the plausibility of the consequents of the arguments in the real world, independent of the soundness of the argumentation. They thus often made judgement errors, because they were unconsciously biased by the real-world likelihood of events.

More generally, when people are asked to judge the acceptability of a linguistic example or argument, they seem unable to ignore potentially relevant information sources, such as world knowledge, or theoretical

hypotheses. Hence the participant population should be naïve to the theoretical hypotheses, and factors like world knowledge should be controlled in the materials, as discussed in Section 2.

3.6. An invalid argument: Presenting research at public forums is similar to testing multiple experimental participants

A sixth argument for continuing to use the traditional nonquantitative methodology in linguistic research is that although the initial observations are only evaluated by a single individual (the researcher him/herself), through the process of sharing these observations with colleagues, as well as presenting them at talks and conferences (and eventually in published works), the judgements are subject to scrutiny from many different individuals and would only "survive" if they are shared by most native speakers of the relevant language. Aside from the fact that most of these individuals will not be naïve with respect to the research questions/hypotheses and therefore vulnerable to cognitive biases discussed in Section 3.5, research in social psychology has demonstrated that individuals are highly sensitive to social factors which have a strong effect on behaviour in interactions with other individuals. Effects of social factors can be observed in both one-on-one interactions (e.g., a researcher sharing some observation with a colleague and eliciting judgements informally), and in group interactions (e.g., a researcher presenting his/her observations at a talk).

For example, colleagues that are asked by the researcher to provide a judgement may be biased due to their subconscious desire to please the researcher in some way, and, as a result, taking into account the researcher's subtle positive or negative reactions in arriving at judgements. Such "observer-expectancy" effects have been reported in the social psychology literature for decades, and are sometimes referred to as the "clever Hans" effect (named after a horse that appeared to be able to count, but was instead responding to subtle subconscious cues from his trainer). This kind of an effect is also related to the Hawthorne effect, named after a factory where workers appeared to work harder under slightly different lighting conditions, but were instead responding to the presence of the experimenter.

In larger groups—for example, during a talk—if you are an audience member and most people in the audience are agreeing about something, it is difficult to trust—and continue to defend—your own judgements when your judgements disagree with those of the majority, even if you are confident in them (e.g., Asch, 1951; Bond & Smith, 1996). This effect of social conformity is stronger in cases where the "majority" consists of members of your "in group" (e.g., Allen, 1965), which is always the case in the situations in question. Furthermore, social conformity is higher when individuals that are in disagreement with you are in a position of authority (e.g., Milgram, 1963),

which is relevant in cases where, for example, the talk is given by someone who is well respected in the field.

As a result, taking positive feedback (agreement on some judgements) from your colleagues or from audiences at talks as evidence of a particular judgement being correct is not justified. Instead, naïve experimental participants who are performing the task based solely on their own intuitions and are not affected by cues from the experimenter and/or other individuals should be used.

3.7. An invalid argument: No harm has come from the nonquantitative approach in syntax/semantics research thus far

A seventh argument for continuing to use the traditional nonquantitative methodology in linguistic research is provided by Phillips (2009). Phillips (2009) argues that no harm has come from the nonquantitative approach in syntax research thus far, and therefore he doesn't see a reason for the field to adopt stricter data collection and reporting standards. In particular, Phillips argues that the use of the traditional nonquantitative methodology in syntax has not caused a "crisis" in the field of syntax:

> In order for there to be a crisis, however, it would need to be the case that (i) Intuitive judgments have led to generalizations that are widely accepted yet bogus. (ii) Misleading judgments form the basis of important theoretical claims or debates. (iii) Carefully controlled judgment studies would solve these problems. (Phillips, 2009, p. 3)

Phillips argues that the existence of incorrect intuitive judgements by itself is not a serious problem for research in syntax, because other researchers would evaluate the intuitive judgements carefully themselves before allowing those judgements to become the basis for either widely accepted generalisations or important theoretical claims (cf. Section 3.6 above). Thus, unless there are cases in the literature where an incorrect intuitive judgement has become the basis for a widely accepted generalisation or an important theoretical claim, there is no evidence of any damage to the field of syntax ("a crisis") by allowing intuitive judgements obtained from single often non-naïve subjects as an accepted method for evaluating theories.

Phillips may be right that researchers will often ignore claims from papers with nongeneraliseable intuitive judgements. But many researchers study languages that they don't speak, and therefore they won't be able to decide which judgements are reasonable and which ones are not. Thus linguistics is a field in which getting the empirical judgements correct in print is especially important. Moreover, allowing papers to be published in conference proceedings and journals that don't adhere to high methodological standards

has the undesirable effect that the results presented in these papers will be ignored by researchers with stronger methodological standards because these results are based on an invalid methodology. If the standards are strengthened, then researchers with higher methodological standards are more likely to pay attention, and the field will progress faster. Relatedly, researchers from neighbouring fields such as psychology, cognitive neuroscience, and computer science are more likely to pay attention to the claims and the results if they are based on data gathered in a rigorous way (as pointed out in Section 2, many language researchers are interested in the same questions; besides, researchers in related domains—e.g., social cognition, reasoning—are also interested in the questions of linguistic representations and computations). Thus, whether or not there are any instances of faulty intuitions from the literature that have resulted in either widely accepted generalisations or important theoretical claims, it is undesirable to accept weak methodological practices in the field.

Indeed, we would argue that the use of the traditional nonquantitative methodology has in fact created a crisis in the field of syntax. Evidence of this crisis is that researchers in other fields of cognitive science no longer follow many developments in syntactic research, as has been observed by several researchers recently (e.g., Edelman & Christiansen, 2003; Ferreira, 2005; Roberts, 2008). Although data gathering and reporting may not be the only source of this crisis, improving data gathering and reporting can only help the situation. However, even setting this point aside, there are many examples from the literature of the kind that Phillips claims do not exist. We review a few that have been discussed in previous papers (Section 4) and discuss three additional cases (Sections 5–6).

4. FAULTY JUDGEMENTS FROM THE LITERATURE

There are several examples of questionable judgements from the literature that have been discussed by others as potentially leading to incorrect generalisations and unsound theorising. Wasow & Arnold (2005) discuss three cases: the first having to do with extraction from datives; the second having to do with the syntactic flexibility of idioms; and the third having to do with acceptability of certain heavy-noun phrase(NP) shift items that were discussed in Chomsky (1955/1975). Phillips himself points out an example discussed by Schütze (1996): a disagreement in the literature between Aoun, Hornstein, Lightfoot, and Weinberg (1987) and Lasnik and Saito (1984) about whether a complementiser can block the long-distance interpretation of *Why do you think that he left?* Phillips observes that few, if any, theoretical proposals attempt to explain these phenomena anymore, so that this is not an example that contradicts his claim.

Bresnan and Nikitina (2009) document several cases involving the dative alternation in English where incorrect judgements from the literature have led to incorrect theorising. For example, according to judgements from Pinker (1989), Levin (1993), and Krifka (2001) manner-of-speaking verbs are grammatical with prepositional phrase syntax (e.g., *Susan whispered/yelled/mumbled/barked/muttered ... the news to Rachel*.) but ungrammatical with dative NP syntax (e.g., **Susan whispered/yelled/mumbled/barked/muttered ... Rachel the news*). Bresnan and Nikitina (2009) observe, however, that people naturally generate dative NP syntax with these verbs, as in the following examples: *Shooting the Urasian a surprised look, she* ***muttered him a hurried apology*** *as well before skirting down the hall*, or *''Hi baby.'' Wade says as he stretches. You just* ***mumble him an answer****. You were comfy on that soft leather couch. Besides* ... On the basis of the existence of examples like these from naturalistic corpora, Bresnan and Nikitina (2009) argue that intuitively judging self-generated materials is a problematic method.

Featherston (2007) discusses two cases from the German syntax literature where the judgements of theoreticians disagree with the results of experiments that Featherston himself conducted. These are interesting cases, but, as noted by Fanselow (2007) and Grewendorf (2007), the results from Featherston's experiments are not clear-cut, because other experiments have provided results that differ from Featherston's and that are instead more consistent with the original claims from the theoretical literature. The inconsistencies between different sets of experimental findings suggest that there may be factors affecting the target judgements that are not well understood yet and that differed between the sets of experimental materials used in different investigations (e.g., animacy; Fanselow, Schlesewksy, Vogel, & Weskott, 2010). The way to proceed in cases where inconsistent results are obtained in different experiments/by different research groups is to gather additional quantitative data, using better controlled materials and more similar methods across studies. Understanding the influences of factors that lead to such discrepancies may provide important insights into the nature of the representations/processes in question.

Because Phillips referenced the articles discussed above when he stated that there have never been cases that adversely affected the progress of the field of syntax (except for the studies by Bresnan and colleagues), he presumably would argue that these cases were not influential in leading to either widely accepted false generalisations or important theoretical claims. We disagree. This is where definitions of what constitutes a ''widely accepted'' generalisation or an ''important'' theoretical claim become subjective. Rather than argue about these definitions, we will provide three additional examples, which we think make similar points to the others in the literature. We hope that the increasing number of such examples will

cause others to question Phillips' claim that such damaging examples are not present in the literature.

5. CASE STUDY 1: SUBJECT- VS. OBJECT-MODIFYING RELATIVE CLAUSES

The first new example of a questionable judgement from the literature that we will present here pertains to research on nesting level and sentence acceptability. Many researchers (including an author of this paper (Edward Gibson)) have argued that comparing the complexity of nested and non-nested syntactic structures is informative with regard to how working memory constrains sentence production and comprehension. This research topic does not concern syntactic representations, but it is closely related. In the first author's Ph.D. thesis it is argued that doubly nested relative clause structures are more complicated to understand when they modify a subject (1) than when they modify an object (2) (Gibson, 1991, examples (342b) and (351b) from pp. 145–147):

(1) The man that the woman that the dog bit likes eats fish.
(2) I saw the man that the woman that the dog bit likes.

That is, it was claimed that (1) is harder to understand than (2). Judgements to this effect were given in the thesis, but no quantitative data were presented. A theory of nesting is presented such that (1) is correspondingly more complex than (2). In particular, it was proposed, following Yngve (1960) and Chomsky and Miller (1963), that sentence structures that contain a greater number of unresolved dependency relations are more complex than those that contain fewer. Sentence (1) has a maximum of five open dependency relations, at the point of processing the most embedded subject "the dog": three open subject–verb relations corresponding to the three initial subjects "the man", "the woman", and "the dog"; and two open object–verb relation for the two occurrences of the relative pronoun "that". In contrast, sentence (2) has only at most four open dependency relations during its processing, occurring at the same point in this sentence, at the most embedded subject "the dog". The open relations are all the same as for (1), with the exception that the NP "the man" has already received a dependency relation in the syntactic/semantic structure at this point, as the object of the verb "saw". Thus (2) was argued to be less complex than (1).

A few years later, the first author attempted to "confirm" this intuition in collaboration with a student, James Thomas, in a series of acceptability rating experiments, each with several conditions, many items per condition, and many participants (Gibson & Thomas, 1996, 1997). Despite several

attempts, Gibson & Thomas found no significant acceptability contrast between these two kinds of structures: people uniformly rated them as very unacceptable—less acceptable than several control structures—but not differentially so.

Later, Gibson, Desmet, Grodner, Watson, and Ko (2005) ran an on-line reading time experiment investigating simpler versions of (1) and (2), with only one relative clause instead of two:

(3) The reporter who the senator attacked on Tuesday ignored the president.
(4) The president ignored the reporter who the senator attacked on Tuesday.

Contrary to the prediction of the theory proposed by Yngve (1960), Chomsky and Miller (1963), and Gibson (1991) among numerous others, Gibson et al. (2005) observed that relative clauses modifying a subject were actually read *faster* than relative clauses modifying an object. That is, the relative clause "who the senator attacked" was read faster in (3) than in (4), and in several other similar comparisons (cf. Hakes, Evans, & Brannon, 1976; Holmes, 1973; who found related results using other methods).

Although these behavioural observations alone do not falsify the incomplete-dependency complexity hypothesis—in fact, Gibson et al. (2005) still argue that there is much evidence in support of such a hypothesis—these observations minimally suggest that there are other factors at play in the judgement of acceptability of examples like (1)–(4). Intuitive judgements alone, like those provided in Gibson (1991), will often lead the theorist astray.

6. CASE STUDIES 2 AND 3: MULTIPLE-WH-EXTRACTION EFFECTS

We will now discuss two further examples from the syntactic literature. For *Case study 2* some quantitative data are already available in the literature, and for *Case study 3* we will present new quantitative data.

6.1. Case study 2

Kuno and Robinson (1972) and Chomsky (1973) observed a subject/object asymmetry in multiple-wh-questions and embedded mutiple-wh-clauses in English. Specifically, in clauses in which both the subject and object NPs are questioned, the items in which the wh-subject (e.g., *who*) is clause-initial such

as (5a) and (6a) are more acceptable than items in which the wh-object (e.g., *what*) is clause-initial such as (5b) and (6b):

(5) a. Who bought what?
    b. * What did who buy?

(6) a. Mary wondered who bought what.
    b. * Mary wondered what who bought.

The unacceptability of sentences like (5b) and (6b) was proposed by Chomsky to be due to *the Superiority Condition*, which is a structural constraint on the movement of elements such as wh-phrases. According to the Superiority Condition, an element X cannot move to a structural position above another element Y in cases where Y is superior to X, where "superior" was defined in terms of a c-/m-command relationship between X and Y. The Superiority Condition therefore required NPs to appear in the order of their "superiority", with more superior items preceding less superior ones (subject NPs preceding direct object NPs, direct object NPs preceding indirect object NPs, etc.). What is relevant for the first case study is the further claim due to Bolinger (1978), and Kayne (1983) that sentences like (5b) and (6b) become more acceptable with a third wh-phrase added at the end:

(7) a. ?* What did who buy where?
    b. ?* Mary wondered what who bought where.

This empirical generalisation resulted in several theoretical attempts to explain it, by postulating additional mechanisms, or making additional assumptions (e.g., Kayne, 1983; Pesetsky, 1987, 2000; Richards, 2001). It turns out, however, that when this intuition is evaluated using quantitative experimental methods, it does not hold. In particular, Clifton, Fanselow, and Frazier (2006) evaluated the relative acceptability of examples like (7b), (6a), and (6b), and whereas they found a reliable difference between examples like (6a) on the one hand and (6b) and (7b) on the other, they found no difference between examples like (6b) and (7b). More recently, Fedorenko and Gibson (in press) evaluated the same kinds of materials in supportive contexts, and replicated Clifton et al., critically finding no difference between examples like (6b) and (7b). In particular, Fedorenko and Gibson (in press) evaluated materials like the following in supportive contexts:

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

c. what who carried.
d. what who carried when.

Whereas a large complexity difference was observed between (8a,b) on the one hand and (8c,d) on the other, no difference was observed between examples like (8c) and (8d), contrary to the judgements in the literature. This is therefore a further example of the kind that Phillips had claimed did not exist: an intuitive contrast from the literature that has led to a widely accepted incorrect generalisation, and important theorising that assumes the contrast.

Recently, in response to Fedorenko and Gibson's (in press) evidence, Culicover and Jackendoff (2010) state that some further controls are needed in order to interpret the lack of a difference between examples like (8c) and (8d) in Fedorenko and Gibson's experiment. In particular, Culicover and Jackendoff state that the pair of control structures in (9) would better elucidate the theoretical landscape:

(9) Peter was trying to remember
a. who carried what last week.
b. what who carried last week.

Culicover and Jackendoff hypothesise that (9a) is as good as (8a,b), but (9b) is worse than (8c,d). If so, then *some* violations with two wh-phrases *are* worse than counterparts with three. In order to test this claim quantitatively, Gibson and Fedorenko (2010b) evaluated all six sentence types in (8) and (9) in an off-line survey on Amazon.com's Mechanical Turk, a marketplace interface that can be used for collecting behavioural data over the internet. The results from Fedorenko and Gibson (in press) were replicated. Furthermore, it was observed that the ratings for (9a) and (9b) were statistically identical to those for (8a) and (8c), respectively, contrary to Culicover and Jackendoff's intuition.

More generally and as discussed in Section 3.3, it is always going to be the case that there are factors that are uncontrolled in any experimental comparison. The main point of this paper is that such comparisons should be evaluated quantitatively. For a language like English, this is very easy to do.

6.2. Case study 3

Another case study that involves materials with multiple-wh-extractions is due to Chomsky (1986). Chomsky (1986, p. 48, example (105)) presented example (10) with the claim that it is more acceptable than sentences that violate the Superiority Condition like (11) (example (108) from Chomsky, 1986, p. 49), due to a process of "vacuous movement":

(10) What do you wonder who saw?
(11) * I wonder what who saw.

Although Chomsky acknowledges later in the text that (10) may not be fully acceptable, he presents this example without an asterisk, whereas (11) is marked with an asterisk, making it clear that Chomsky believes that (10) is more acceptable than (11).[8] He then makes theoretical claims that rely on this and other relative acceptability judgements. In particular, Chomsky proposes the Vacuous Movement Hypothesis, whereby "vacuous movement is not obligatory at S-Structure" (Chomsky, 1986, pp. 49–50). According to this hypothesis, in a derivation from D-structure to S-Structure, movement of wh-elements to a position in the specifier of CP (Spec-CP) is not obligatory when such a position is linearly adjacent to the Spec-CP (e.g., in the subject position of the embedded clause), thereby allowing the wh-phrase "what" in (10) to first move from the object position of "saw" to the Spec-CP in the embedded clause, and then to the top-level Spec-CP position in this derivation. In the derivation from S-Structure to logical form (LF), wh-phrases must move to Spec-CP positions in order to get their interpretations. At this point, the wh-phrase "who" in the embedded subject position must move to the embedded Spec-CP position, and the derivation is grammatical. Consider now a sentence like (11). As in the derivation of a sentence like (10), the Vacuous Movement Hypothesis also applies in (11), allowing the wh-phrase "what" to move from the object position of "saw" to the Spec-CP in the embedded clause in a derivation from D-structure to S-Structure. But in the ensuing derivation from S-Structure to LF, the wh-phrase "who" must also move to a Spec-CP position, and there is no unfilled Spec-CP position for it to move to. Hence the derivation is disallowed, rendering (11) ungrammatical.[9]

[8] These two sentences are of course not a minimal pair, because of several uncontrolled differences between the items, including (a) the wh-question in (11) is a matrix wh-question, while the wh-question in (12) is an embedded wh-question; and (b) the lexical items in the two sentences aren't the same ("I" vs. "you"). These differences were controlled in the experimental comparison reported below.

[9] An anonymous reviewer has noted that "the importance of vacuous movement has plummeted greatly since Barriers days, and thus that the judgements in question really don't matter all that much". Although this may be true, this is certainly a case that, in the words of Phillips (2009) "adversely affected the progress of the field of syntax". In particular, Chomsky's writings have a much greater impact than other syntacticians' writings, so any errors in his work are exacerbated in the field for years to come. To give a specific example, the first author of this paper (Edward Gibson) began to work in the field of syntax in the late 1980s, but was so disenchanted by several of the judgements in Chomsky (1986) that he shifted his research focus to a different topic within the area of language research.

In order to evaluate the relative acceptability of the structures in (10) and (11), we conducted an acceptability rating experiment using Amazon.com's Mechanical Turk. In addition to evaluating Chomsky's empirical predictions about multiple-wh-extractions, we wanted to evaluate the viability of using the participant pool in Amazon.com's Mechanical Turk for testing linguistic hypotheses more generally. Consequently, we also evaluated the materials from an established contrast in the literature (a control experiment), which manipulated the degree of centre-embedding of relative clauses in a target sentence. Replicating effects observed in a laboratory setting for these materials would help to validate the method.

The critical experiment had three conditions: the key condition of interest (modelled after example (10) above), along with two control conditions: an example violating the Superiority Condition modelled after (11), and a control example that did not violate the Superiority Condition. We presented the three conditions in supportive contexts, as in (12). (For the complete set of experimental materials, see the Appendix.)

(12) Example of a critical item

Context: The waiter was trying to keep track of the orders for all of the tables that he was serving. He suddenly became very tired and confused. He then collapsed on the floor. The manager of the restaurant came quickly to try to understand what the waiter had been confused about.

a. *Vacuous movement (Chomsky, 1986)*
The manager tried to figure out what the waiter wondered who had ordered.

b. *Superiority Condition violation*
The manager tried to figure out if the waiter wondered what who had ordered.

c. *Acceptable control*
The manager tried to figure out if the waiter wondered who had ordered what.

Example (12a) is a contextually supported version of (10). If Chomsky's judgements are correct, then (12a) should be rated as more acceptable than (12b)—the unacceptable version (object preceding subject) of the classic Superiority comparison—and it should be roughly as acceptable as (12c)—the acceptable version (subject preceding object) of the classic Superiority comparison (see Clifton et al., 2006; Fedorenko & Gibson, in press, 2010a, 2010b, for relevant quantitative data).

6.2.1. Experiment

6.2.1.1. Participants. We posted surveys for 60 workers on Amazon.com's Mechanical Turk. The workers were paid for their participation. Participants were asked to indicate their native language at the beginning of the survey, but payment was not contingent on the participants' responses to this question.

6.2.1.2. Design and materials.[10] The materials for the critical experiment consisted of 15 sets of sentences appearing in supportive contexts. The materials were constructed based on the materials used in Experiment 1 of Fedorenko and Gibson (2010a). The contexts introduced two sets of entities (*restaurant patrons* and *their food orders* in (12) above), agents and patients in relation to some verb (*ordering* in (12)). The context also introduced another agent (the second-verb agent—*a waiter* in (12)) who was uncertain about the pairings between the embedded-verb agents and the embedded-verb patients (e.g., the waiter was uncertain about who ordered what in (12)). Finally, there was a top-level verb agent (*the manager*) who was interested in finding out what the second-verb agent (*the waiter*) was uncertain about. This context licenses all three of the target sentences.

The items were constructed such that the nouns and verbs were as variable as possible across the items, up to the constraints imposed by the experimental design. The verbs were most constrained across the items. Each target sentence contained three verbs: two verbs that take embedded questions as complements (e.g., *figure out* and *wondered* in (12)) and a transitive verb in the most embedded clause. The most embedded verb was varied across items: 14 verbs were used across the 15 items, with the verb *presented* being the only repeated verb. Because there is a limited set of verbs that take embedded questions as complements, we repeated these verbs across items, but never within an item (that is, the top-level verb and the first embedded verb were always distinct within an item). The distribution of these verbs across items was as follows: find out (5 uses); know (4); wonder (3); remember (3); ask (3); check (3); figure out (3); learn (2); recall (2); decide (1); be uncertain (1).

The motivation for presenting these materials in supportive contexts was two-fold. First, multiple-wh-questions have several potential interpretations, including at least (a) the pair-list/n-tuple-list interpretation, (b) the echo-reprise interpretation, (c) the reference-reprise interpretation (e.g., Bolinger, 1978; see e.g., Fedorenko & Gibson, 2010a, for a recent summary; Pesetsky, 2000). Only the n-tuple-list interpretation has been argued to elicit the

[10] For summaries of issues relevant to basic experimental design for language research, see e.g., Ferreira (2005) and Myers (2009).

subject/object asymmetry in examples like (5)–(6). Fedorenko and Gibson (2010b) have provided quantitative evidence for this claim. To ensure that the participants retrieve the intended n-tuple-list interpretation in reading the critical materials, we used contexts which force this interpretation. Second, complex (multiple-clause) sentences may be difficult to interpret when presented in isolation, especially when they require constructing a complex discourse (as in the current set of sentences). We therefore wanted to minimise the cognitive load associated with constructing a discourse by providing appropriate contexts, thereby increasing our chances of detecting between-condition differences due to the critical syntactic structure manipulations.

In addition to the 15 critical items, the experiment included 12 items from the control experiment, and 10 filler items. The items for the control experiment were taken directly from Fedorenko and Gibson (2010a). They manipulated the degree of centre-embedding of relative clauses in a target sentence, as in (13).

(13) Example of an item from the control (centre-embedding) experiment.

Context: The lawyer took the case of a client who started having some complications after a recent surgery. He questioned all the doctors in the clinic who had anything to do with the client. He already knew that the neuropsychologist consulted the cardiologist sometime before the surgery and that after that, the neuropsychologist warned the surgeon. However, a few dates were missing from the medical history of the client.

a. *No embedding:*
The lawyer tried to clarify when the cardiologist consulted the neuropsychologist who warned the surgeon who performed the surgery.

b. *One level:*
The lawyer tried to clarify when the neuropsychologist who the cardiologist consulted warned the surgeon who performed the surgery.

c. *Two levels:*
The lawyer tried to clarify when the surgeon who the neuropsychologist who the cardiologist consulted warned performed the surgery.

As discussed above, we included these materials because they had been evaluated previously by the local participant pool from our lab, as discussed in Fedorenko and Gibson (2010a). This allowed us to compare the results

from the current paradigm, using Amazon.com's Mechanical Turk, to the more traditional paradigm in which participants are recruited to a lab where they fill out paper surveys (see also Frank & Gibson, in press; Frank, Tily, Arnon, & Goldwater, 2010; Munro et al., 2010). If we replicate the results of Fedorenko and Gibson (2010a), then this would help to validate the new paradigm for obtaining experimental linguistic data (which is more time- and resource-efficient than the traditional paradigm).

As discussed in Fedorenko and Gibson (2010a), Chomsky and Miller (1963) observed that the acceptability of a structure decreases with the level of embedding. A plausible reason for the difference among the structures with more/fewer levels of embedding has to do with working memory limitations (e.g., Gibson, 1998; Lewis & Vasishth, 2005). Fedorenko and Gibson (2010a) observed that the structures with two levels of centre-embedding were rated as significantly less acceptable than the other two conditions, with no differences between the less complex conditions.

The filler items were also similar to the critical experimental materials in terms of their general structure, such that they consisted of a several sentences long context and a critical sentence following it. The critical sentence contained an embedded question in all of the items. An example of a filler item is shown in (14).

(14) Example of a filler item:

> Context: The chemist was preparing to conduct a new experiment. He prepared all the necessary chemicals and beakers. His lab assistant was supposed to come soon and help him in carrying out the experiment. The chemist could not find the lab notebook with the notes on the experiment, which was conducted last week. The chemist was trying to remember where the lab assistant kept the notebook.

Three lists were created following a Latin-Square design, such that each participant saw only one version of each item. Two different random orders of each of these lists were generated, for a total of six lists. Each list was presented to 10 participants.

6.2.1.3. Procedure. The participants were given the following instructions:
Please read each short paragraph, then answer the questions immediately following, and provide the requested rating.
Please read the paragraph context before answering the questions and giving the rating!

The context was preceded by the word "CONTEXT:", and the target sentence was preceded by the words "TARGET SENTENCE:".

In order to ensure that the participants read the contexts and the critical sentences carefully and understood what the sentences were intended to convey, we included two questions for each trial: the first about the content of the context, and the second about the content of the target sentence. For example, the context question following example (12) was as in (15), and the question about the target sentence was as in (16):

(15) Did the waiter get tired and confused?
(16) Was the manager trying to figure out something about what the waiter was wondering?

The answers to questions (15) and (16) were both "yes". "Yes" and "no" responses were balanced across items such that each list had equal numbers of yes and no answers. Participants entered their answer to the question by clicking the appropriate ("Yes" or "No") radio button on the Mechanical Turk web page.

Following the two questions, the participant was asked to provide a rating for the sentence in the context (preceded by the heading "Rating of TARGET SENTENCE in CONTEXT") by clicking a radio button beside the appropriate rating. There were five choices for each sentence: "Extremely unnatural", "Somewhat unnatural", "Possible", "Somewhat natural", and "Extremely natural". These responses were converted to numerical scores from 1 (Extremely unnatural) to 5 (Extremely natural) for the analyses.

The experiment took approximately 20 minutes to complete.

6.2.2. Results

The data from non-native English speakers and from people outside the USA were not analysed, leaving the survey data from 56 participants. In addition, the data from participants who either left more than 10% of the survey unanswered or answered fewer than 75% of the comprehension questions correctly were not analysed, thus omitting the data from five additional participants. The data from 51 participants remained.

6.2.3. Analyses

All analyses reported here were conducted with the lme4 package (Bates, Maechler, & Dai, 2008) for the statistical language R (R Core Development Team, 2008). Significance (p) values were obtained using the function pvals.fnc in the R package languageR (Baayen, 2008). This function uses Markov Chain Monte Carlo simulation to approximate the posterior distribution over possible values of all of the parameters in the model given the data, and reports p values showing for each parameter the proportion of that distribution which crosses zero.

6.2.4. Comprehension question accuracy

The dependent measure of interest in both the critical and the control experiments was sentence rating. Comprehension questions were included in order to ensure that the participants understood what the sentences were intended to convey. This goal was achieved. Whereas three of the initial 60 participants answered the comprehension questions at a rate of below 75% across all the items (and were therefore not included in the analyses), the remaining participants answered the comprehension questions at a rate of 95.1% across conditions. The rates for each condition for each experiment are presented in Table 1. There were no reliable between-condition differences in accuracies in either of the two experiments.

6.2.5. Ratings

The means for the three conditions in the multiple-wh extraction experiment are presented in Table 2. We performed a linear mixed-effects regression including condition as a dummy-coded factor and participants and items as random intercepts. This analysis demonstrated that the vacuous movement condition was rated as significantly worse than either the Superiority violation condition ($\beta = .486$ points on a 5-point scale; $p < .0001$) or the acceptable control condition ($\beta = 1.859$ points on a 5-point scale; $p < .0001$).

The means for the three conditions for the control centre-embedding experiment are presented in Table 3. We performed a linear mixed-effects regression including condition as a dummy-coded factor and participants and items as random intercepts. This analysis demonstrated that the two levels of centre-embedding condition was rated as significantly worse than

TABLE 1
Comprehension question accuracy across conditions for the two experiments and the filler items

Multiple-wh extraction experiment	*Accuracy in percentage*
(11a) Vacuous movement	97.8
(11b) Superiority Condition violation	96.7
(11c) Acceptable control	97.4
Centre-embedding experiment	
(12a) Zero	93.1
(12b) One	92.1
(12c) Two	90.4
Filler	97.9

TABLE 2
Mean sentence ratings in the multiple-wh extraction experiment: ''1'' = ''Extremely unnatural'' and ''5'' = ''Extremely natural'' in the given context

Multiple-wh extraction experiment	*Average rating (Standard error)*	
(11a) Vacuous movement	2.17	(0.14)
(11b) Superiority Condition violation	2.65	(0.14)
(11c) Acceptable control	4.03	(0.14)

either the zero ($\beta = 1.18$ points on a 5-point scale; $p < .0001$) or one level of centre-embedding conditions ($\beta = 1.08$ points on a 5-point scale; $p < .0001$), but that there was no significant difference between the zero-level and one-level of centre-embedding conditions ($\beta = .10$ points on a 5-point scale; $p = .283$).

6.2.6. Discussion

The control experiment closely replicated the results from Fedorenko and Gibson (2010a). In particular, the means across the three condition for Fedorenko and Gibson's experiment were 4.1, 4.0, and 2.9 for the zero, one and two levels of centre-embedding conditions, on a 7-point scale of acceptability (a slight difference from the current experiment, which used a 5-point scale). Like the current experiment, the condition with two levels of centre-embedding was rated as significantly less acceptable than the other two conditions, with no difference between zero and one level of embedding. This replication, together with similar replications on Mechanical Turk of experiments in the lab from Frank et al. (2010), Munro et al. (2010), and Frank and Gibson (in press), establishes that the method of obtaining experimental linguistic data through Amazon.com's Mechanical Turk is viable.

The results of the experiment that evaluated Chomsky's (1986) empirical claim about examples like (10)/(12a) demonstrate that, contrary to Chomsky's claim that such sentences are acceptable, these examples are much *less*

TABLE 3
Mean sentence ratings in the centre-embedding experiment. ''1'' = ''Extremely unnatural'' and ''5'' = ''Extremely natural'' in the given context

Level of centre-embedding	*Rating (Standard error)*	
(13a) Zero	2.77	(0.12)
(13b) One	2.67	(0.12)
(13c) Two	1.59	(0.12)

acceptable than the unacceptable versions of the Superiority Condition (11)/(12b). Furthermore, these examples are not nearly as acceptable as the acceptable version of the traditional Superiority comparison (12c). This pattern of data is not predicted by Chomsky's theorising in this domain and therefore puts into question all the theoretical work that has built on these contrasts. An experiment would have clarified these issues years ago and would have prevented theorising aimed at accounting for nonexistent effects.

A question remains as to why the vacuous movement sentences (12a) are less acceptable than the unacceptable Superiority Condition sentences (12b). One possibility is that in (12a) the wh-pronoun *what* is separated from its role-assigning verb *ordered* by an intervening clause, whereas no such additional clause separates the wh-pronoun *what* from its role-assigning verb *ordered* in (12b). Connecting wh-fillers with embedded positions has been shown to be sensitive to the linear distance of the dependency (Gibson, 1998, 2000; Gordon, Hendrick, & Johnson, 2001; Grodner & Gibson, 2005; Lewis & Vasishth, 2005; Lewis, Vasishth, & Van Dyke, 2006; Warren & Gibson, 2002). The additional long-distance dependency in (12a) may therefore make it more complex than (12b), leading to greater unacceptability.

A second possibility, raised by an anonymous reviewer, is that the contexts that we constructed here (including the questions that were asked of the comprehender) may favour an interpretation in line with (12b) or (12c), but less so for the vacuous movement conditions (12a). However, we think that this is an unlikely possibility because the contexts were specifically created in order to license the long-distance vacuous movement extractions. In particular, the contexts had to contain two people, each of whom was uncertain about what happened in a set of events, one embedded within another. This kind of context is exactly what is needed to license the vacuous movement conditions (and also happens to license the double wh-extraction conditions as a by-product). We therefore think that it is unlikely that the contexts are somehow biased against the vacuous movement interpretation.

7. CONCLUDING REMARKS

In this paper we have discussed several instances from the syntax literature in which a questionable intuitive judgement has led to an incorrect generalisation, which has then led to unsupported theorising that many researchers have followed. As we have discussed in Section 3.5, we believe that these errors are due to cognitive biases on the part of the experimenter and the people who they consult: (a) the experimenter will often have a confirmation bias favouring the success of the predicted result, with the consequence that he/she will tend to justify ignoring intuitions from speakers that do not fit the

predicted pattern; (b) the experimenter, and in some cases the people providing the judgements, will be unconsciously affected by the hypotheses in making their judgements about the relevant experimental materials. Furthermore, as discussed in Section 3.6, social factors may affect behaviour in nonquantitative evaluations of hypotheses, so that, for example, people providing the judgements may be biased because they subconsciously want to please the experimenter. These are all unconscious biases that affect all people (including an author of this article, as we have demonstrated in Case study 1).

Although Phillips (2009) claimed that examples like the ones discussed in this and other papers do not exist, we hope that it is now apparent that such examples do exist. Indeed, most experiments that we have run which attempt to test a hypothesis in the syntax or semantics/pragmatics literature have resulted in a pattern of data that had not been hypothesised before the experiment (e.g., Breen, Fedorenko, Wagner, & Gibson, 2010; Fedorenko & Gibson, in press; Patel, Grosz, Fedorenko, & Gibson, 2009; Scontras & Gibson, 2010). There are probably many more such instances in the literature, especially for materials with multiple clauses and/or temporary or global ambiguity. However, as discussed above, even if there were no examples which led to incorrect generalisations and unsound theorising, it would be healthy for the fields of syntax/semantics to have more rigorous data collection and reporting standards, for several reasons:

(1) The fact that the current methodology is invalid has the unwelcome consequence that researchers with stronger methodological standards (including those in related fields) will often ignore the current theories and empirical claims from the field of linguistics. These researchers would be more likely to pay attention to the developments in the field if the methodology were valid, leading to faster progress.
(2) Because quantitative investigations are less subject to cognitive biases, gathering quantitative data is likely to frequently falsify hypotheses. This in turn will lead researchers to entertain alternative hypotheses, which is critical in making progress in any field of science.
(3) Whereas intuitions are useful in getting started in cognitive psychology, many reliable effects aren't accessible via intuitions and may require more sensitive measures. The use of rigorous data collection standards across multiple methods is likely to increase the chances of detecting potentially subtle distinctions which nevertheless may be critical to distinguish among different theoretical positions.

But as we have discussed above, gathering quantitative data does not guarantee easy solutions to open questions about the nature of language. Even when quantitative data are gathered, there can be confounding

influences and incorrect interpretations of data. However, following existing standards from cognitive science for presenting quantitative analyses will minimise the presentation of data that are spurious, driven by cognitive biases on the part of the researchers.

Finally, a question that is often put to us is whether it is necessary to evaluate *every* empirical claim quantitatively. A major problem with the fields of syntax and semantics is that many papers include *no* quantitative evidence in support of their research hypotheses. Because conducting experiments is now so easy to do with the advent of Amazon.com's Mechanical Turk, we recommend gathering quantitative evidence for all empirical claims. However, it would clearly be a vast improvement to the field for all research papers to include at least *some* quantitative evidence evaluating their research hypotheses.

REFERENCES

Allen, V. L. (1965). Situational factors in conformity. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 133–175). New York: Academic Press.

Aoun, J., Hornstein, N., Lightfoot, D., & Weinberg, A. (1987). Two types of locality. *Linguistic Inquiry*, *18*, 537–577.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership and men* (pp. 177–190). Pittsburgh, PA: Carnegie Press.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

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

Bates, D., Maechler, M., & Dai, B. (2008). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-27. Retrieved from http://www.lme4.r-forge.r-project.org/

Bolinger, D. (1978). Asking more than one thing at a time. In H. Hiz (Ed.), *Questions* (pp. 107–150). Dordrecht: D. Reidel.

Bond, R., & Smith, P. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin*, *119*, 111–137.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7/8/9), 1044–1098.

Bresnan, J. (2007). A few lessons from typology. *Linguistic Typology*, *11*, 297–306.

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.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain & Cognition*, *5*, 41–66.

Chomsky, N. (1955/1975). *The logical structure of linguistic theory*. Chicago, IL: University of Chicago Press.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A festschrift for Morris Halle* (pp. 232–286). New York: Holt, Rinehart & Winston.

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

Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. D Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, (pp. 269–321). New York: Wiley.

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

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

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

Dabrowska, E. (2010). Naïve v. expert intuitions: An empirical study of acceptability judgments. *The Linguisitic Review*, *27*, 1–23.

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

Evans, J., St., B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*, 295–306.

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

Fanselow, G., Schlesewksy, M., Vogel, R., & Weskott, T. (2010). *Animacy effects on crossing wh-movement in German*. Manuscript submitted for publication.

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

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

Fedorenko, E., & Gibson, E. (in press). Adding a third wh-phrase does not increase the acceptability of object-initial multiple-wh-questions. *Syntax*.

Fedorenko, E., & Gibson, E. (2010a). *Syntactic parallelism as an account of superiority effects: Empirical investigations in English and Russian*. Manuscript submitted for publication.

Fedorenko, E., & Gibson, E. (2010b). *Wh-phrase order effects in multiple-wh-questions are contingent on the interpretation*. Manuscript submitted for publication.

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

Frank, M. C., & Gibson, E. (in press). Resource dependence of artificial rule learning. *Language Learning and Development*.

Frank, M. C., Tily, H., Arnon, I., & Goldwater, S. (2010). Beyond transitional probabilities: Human learners impose a parsimony bias in statistical word segmentation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 760–765), Portland, OR, August 11–14, 2010.

Fukuda, S., Michel, D., Beecher, H., & Goodall, G. (2010). *Comparing three methods for sentence judgment experiments*. LSA Annual Meeting, Baltimore, MD, January 8, 2010.

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

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, & W. O'Neil (Eds.), *Image, language, brain*. Cambridge, MA: MIT Press.

Gibson, E., Desmet, T., Grodner, D., Watson, D., & Ko, K. (2005). Reading relative clauses in English. *Cognitive Linguistics*, *16*, 313–354.

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

Gibson, E., & Fedorenko, E. (2010b). *Control structures in syntax research: A response to Culicover and Jackendoff*. Manuscript submitted for publication.

Gibson, E., Fedorenko, E., Ichinco, D., Piantadosi, S., Twarog, N., & Troyer, M. (2009). *Quantitative investigations of syntactic representations and processes*. Talk presented at the workshop on formal vs. processing explanations of syntactic phenomena, University of York, UK, April 28, 2009.

Gibson, E., & Pearlmutter, N. (1998). Constraints on sentence comprehension. *Trends in Cognitive Society*, *2*, 262–268.

Gibson, E., & Thomas, J. (1996). The processing complexity of English center-embedded and self-embedded structures. In C. Schütze (Ed.), *Proceedings of the NELS 26 sentence processing workshop, MIT occasional papers in linguistics 9* (pp. 45–71). Cambridge, MA: MIT Press.

Gibson, E., & Thomas, J. (1997). *Processing load judgments in English: Evidence for the Syntactic Prediction Locality Theory of syntactic complexity*. Unpublished manuscript, MIT, Cambridge, MA.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1411–1423.

Grewendorf, G. (2007). Empirical evidence and theoretical reasoning in generative grammar. *Theoretical Linguistics*, *33*, 369–380.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, *29*(2), 261–290.

Hakes, B., Evans, J., & Brannon, L. (1976). Understanding sentences with relative clauses. *Memory and Cognition*, *4*, 283–296.

Holmes, V. (1973). Order of main and subordinate clauses in sentence perception. *Journal of Verbal Learning and Verbal Behavior*, *12*, 285–293.

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

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

Krifka, M. (2001). *Lexical representations and the nature of the dative alternation*. Talk presented at the University of Amsterdam, November 11, 2001.

Kuno, S., & Robinson, J. (1972). Multuple wh-questions. *Linguistic Inquiry*, *3*, 463–487.

Labov, W. (1975). Empirical foundations of linguistic theory. In R. Austerlitz (Ed.), *The scope of American linguistics* (pp. 77–134). Lisse: Peter de Ridder Press.

Labov, W. (1996). When intuitions fail. *Chicago Linguistic Society*, *32*(2), 77–106.

Lasnik, H., & Saito, M. (1984). On the nature of proper government. *Linguistic Inquiry*, *15*, 235–289.

Levin, B. (1993). *English verb classes and alternations. A preliminary investigation*. Cambridge, MA: MIT.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 375–419.

Lewis, R., Vasishth, S., & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454.

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

Milgram, S. (1963). Behavioral study of obedience''. *Journal of Abnormal and Social Psychology*, *67*, 371–378.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., … Tily, H. (2010). Crowdsourcing and language studies: The new generation of linguistic data. *Proceedings of Workshop at NAACL 2010 'Creating Speech and Text Language Data with Amazon's Mechanical Turk*, June 6, 2010.

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

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

Patel, P., Grosz, P., Fedorenko, E., & Gibson, E. (2009). *Experimental evidence against a strict version of the formal link condition on E-Type pronouns*. Poster presented at the 22nd CUNY conference on sentence processing, University of California, Davis, March 26, 2009.

Pesetsky, D. (1987). Wh-in-situ: Movement and unselective binding. In E. Reuland & A. ter Meulen (Eds.), *The representation of (In)definiteness* (pp. 98–129). Cambridge, MA: MIT Press.

Pesetsky, D. (2000). *Phrasal movement and its kin*. Cambridge, MA: MIT Press.

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

Pinker, S. (1989). *Learnability and cognition. The acquisition of argument structure*. Cambridge, MA: MIT.

R Core Development Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: *R Foundation for Statistical Computing*. Available from http://www.R-project.org

Richards, N. (2001). *Movement in language*. Oxford: Oxford University Press.

Roberts, I. G. (2008). *The mystery of the overlooked discipline: Modern syntactic theory and cognitive science*. Unpublished manuscript, University of Cambridge.

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

Scontras, G., & Gibson, E. (2010). *A quantitative investigation of the imperative-and-declarative construction in English*. Manuscript submitted for publication.

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

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, *85*, 79–112.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140.

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

Weskott, T., & Fanselow, G. (2008). Variance and informativity in different measures of linguistic acceptability. In N. Abner & J. Bishop (Eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics (WCCFL)* (pp. 431–439). Somerville, MA: Cascadilla Press.

Weskott, T., & Fanselow, G. (2009). Scaling issues in the measurement of linguistic acceptability. In S. Featherston & S. Winkler (Eds.), *The fruits of empirical linguistics. Vol. 1: Process* (pp. 229–245). Berlin, New York: Mouton de Gruyter.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, *104*, 444–466.

APPENDIX: MATERIALS FOR THE EXPERIMENT IN CASE STUDY 3

Each condition was presented in the supportive context at the beginning of each item. Condition (a) is the vacuous movement condition, condition (b) is the Superiority violation condition, and condition (c) is the acceptable control condition.

1. The waiter was trying to keep track of the orders for all of the tables that he was serving. He suddenly became very tired and confused. He then collapsed on the floor. The manager of the restaurant came quickly to try to understand what the waiter had been confused about.
   (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.
   (c) The manager tried to figure out if the waiter wondered who had ordered what.
2. At Christmas the employees of a company played a game of Secret Santa. Each employee brought a wrapped gift addressed to another employee and secretly placed it under the

Christmas tree. At the party, everybody unwrapped their gifts. The manager looked interested and watched everybody attentively.

(a) One of the employees asked what the manager tried to find out who brought.
(b) One of the employees asked if the manager tried to find out what who brought.
(c) One of the employees asked if the manager tried to find out who brought what.

3. A big conference for the American Cancer Society was held at the Marriott Hotel. As it often happens, many of the attendees lost various items by leaving them in the conference hall or in the corridors during the breaks. Luckily, the hotel had a lost-and-found service, which kept all the items. During the last day of the conference, several people were standing in line at the lost-and-found office trying to locate their misplaced belongings.
   (a) The hotel manager came to check what the maid was trying to figure out who was missing.
   (b) The hotel manager came to check if the maid was trying to figure out what who was missing.
   (c) The hotel manager came to check if the maid was trying to figure out who was missing what.
4. A marketing research company was conducting an analysis of how effective event sponsorship is. Specifically, they were interested in how much the sponsoring of sports events boosts the companies' profits. The intern who was hired to work on this project was told to compile a list of important sports events in the last 5 years and the names of the sponsors.
   (a) The supervisor wanted to know what the intern found out who sponsored.
   (b) The supervisor wanted to know if the intern found out what who sponsored.
   (c) The supervisor wanted to know if the intern found out who sponsored what.
5. At the university, several professors and a group of graduate students were organising a workshop. A professor from the physics department was in charge of the schedule. He had to invite some speakers from various universities, and he also had to assign some papers to be presented by graduate students.
   (a) One of the students asked what the professor had decided who would present.
   (b) One of the students asked if the professor had decided what who would present.
   (c) One of the students asked if the professor had decided who would present what.
6. For Christmas, the parents of children attending kindergarten were planning a party where one of the fathers would dress as Santa Claus and would give out presents to the kids. All the parents told him about the kinds of presents their children wanted and about the kinds of presents they promised the children Santa Claus would bring.
   (a) Before the party, one of the parents wanted to know what the designated Santa Claus was likely to recall who wanted.
   (b) Before the party, one of the parents wanted to know if the designated Santa Claus was likely to recall what who wanted.
   (c) Before the party, one of the parents wanted to know if the designated Santa Claus was likely to recall who wanted what.
7. It was the end of the semester at the culinary school and the final grades were due soon. In the French cuisine class, the instructor was evaluating students' work by having each student bring a French dish they prepared, one student for each classroom session. This way, at the beginning of each class there was a tasting session where the instructor and the students in the class evaluated the dish on several dimensions. The instructor didn't take careful notes throughout the semester, hoping that his memory would serve him well.
   (a) The students wondered what the instructor would remember who cooked.
   (b) The students wondered if the instructor would remember what who cooked.
   (c) The students wondered if the instructor would remember who cooked what.

8. The owner of a bicycle rental shop hired three assistants for the summer, because he was having an unusually high number of customers. On the first day of work, the assistants rented out many bikes, but they didn't keep careful records of the transactions.
   (a) The owner asked what the assistants remembered who rented.
   (b) The owner asked if the assistants remembered what who rented.
   (c) The owner asked if the assistants remembered who rented what.
9. Some architecture students at the local school of design were taking a tour of the city. They visited many different areas and saw many interesting buildings designed by famous architects.
   (a) Their professor wanted to check what the students had learned who designed.
   (b) Their professor wanted to check if the students had learned what who designed.
   (c) Their professor wanted to check if the students had learned who designed what.
10. Each student in a psychology class had to make a short presentation at some point in the semester. The professor was sick for a few weeks, and another professor was substituting for him. However, the substitute professor forgot to write down the presentations that took place in each class.
   (a) The professor was uncertain what the substitute professor could recall who presented.
   (b) The professor was uncertain if the substitute professor could recall what who presented.
   (c) The professor was uncertain if the substitute professor could recall who presented what.
11. The students in the 6th grade went on a day trip to the New England Aquarium. At the end of the tour everybody rushed to the gift shop to buy some souvenirs. Some of the students didn't bring any money and had to borrow from their friends. However, later, some of the kids were confused about who borrowed money from them or from whom they borrowed money.
   (a) The supervisor tried to find out what the kids had figured out who borrowed.
   (b) The supervisor tried to find out if the kids had figured out what who borrowed.
   (c) The supervisor tried to find out if the kids had figured out who borrowed what.
12. Several doctors were working in the cardiology department. It was a busy day and each doctor saw many patients. The doctors prescribed different medications to different patients. However, the secretary was sloppy and got confused about all the different prescriptions.
   (a) One of the doctors wanted to know what the secretary was trying to find out who prescribed.
   (b) One of the doctors wanted to know if the secretary was trying to find out what who prescribed.
   (c) One of the doctors wanted to know if the secretary was trying to find out who prescribed what.
13. The undergraduate student advisor in the biology department was keeping track of which classes the biology majors took each semester. The department recently relocated to a new building and some of the files got lost.
   (a) The advisor wondered what the head tutor recalled who took.
   (b) The advisor wondered if the head tutor recalled what who took.
   (c) The advisor wondered if the head tutor recalled who took what.
14. Peter was moving to a new bigger apartment in the same building and he asked some of his friends to help him carry furniture and boxes on Sunday. He had five people helping him and it took them about 5 hours to move everything. When everything was moved, Peter noticed that one chair and a small bookshelf were nowhere to be found.

(a) Peter tried to find out what his friends remembered who moved.
(b) Peter tried to find out if his friends remembered what who moved.
(c) Peter tried to find out if his friends remembered who moved what.

15. Andrew had a literature test tomorrow at school. He had to know about the life of famous writers and about their literary works.
(a) His father checked what Andrew had learned who wrote.
(b) His father checked if Andrew had learned what who wrote.
(c) His father checked if Andrew had learned who wrote what.