Acceptability ratings cannot be taken at face value

Carson T. Schütze
UCLA

3rd Draft 2019.07.11

Abstract

This chapter addresses how linguists' empirical (syntax) claims should be tested with non-linguists. Recent experimental work attempts to measure rates of convergence between data presented in journal articles and the results of large surveys. The chapter presents three follow-up experiments to one such study (Sprouse, Schütze, and Almeida 2013), arguing that this method may underestimate the true rate of convergence because it leaves considerable room for naïve subjects to give ratings that do not reflect their true acceptability judgments of the relevant structures. To understand what can go wrong, the experiments were conducted in two parts: the first part had visually presented sentences rated on a computer, replicating previous work. The second part was an interview where the experimenter asked each subject about the ratings they gave to particular items, in order to determine what interpretation/parse they had assigned, whether they had missed any critical words, etc.

Keywords

acceptability judgments, syntax, linguists vs. non-linguists, rate of convergence, interview

1. Introduction

1.1 Motivation

In the on-going debates over the empirical base of linguistic theory**,** there have been increasing attempts in recent years to conduct experimental tests of acceptability on naïve speakers. One way this has often been done is to take sentences directly from the linguistics literature and present these to subjects via computer to rate, e.g. on a 1–7 Likert scale or using magnitude estimation, without the researcher further engaging with the subjects (Sprouse and Almeida 2012; Sprouse, Schütze, and Almeida 2013; Munro et al. 2010; Song, Choe, and Oh 2014; Mahowald et al. 2016; Häussler and Juzek 2017; Langsford et al. 2018; Linzen and Oseki 2018). Amazon Mechanical Turk (AMT) has been a frequent tool in conducting such studies. While the

overall result is that linguists' judgments are mostly replicated,[1] there is debate over the importance of the number and nature of cases when they are not. For purposes of this chapter it does not matter what conclusions one wishes to draw, if any, from those previous studies: all that is relevant is that one believes that gathering judgment data from naïve speakers is sometimes useful. Given that, my goal is to make the point in the title of the chapter: finding that subjects' ratings on a set of experimental stimuli do not align with the published judgments of linguists does not necessarily represent a genuine data discrepancy. I provide empirical evidence that, at least in many instances, subjects' responses have resulted from factors other than (un)acceptability of the structure that the linguists were actually interested in. How and why this occurs will be expounded in detail. The lesson is that, for the field to make progress, we need to go beyond observing and counting mismatches in the naïve way we have been doing, and strive to understand their causes. Furthermore, such understanding is quite unlikely to be achieved simply by conducting more large-scale crowdsourced acceptability experiments.

What I advocate and demonstrate in this chapter is that as the field continues to collect large amounts of quantitative data in this manner, as it surely will, it needs in parallel to also be collecting data in a very different way for a very different purpose: to answer the "Why?" question. Why are naïve speakers rejecting sentences that linguists claim are acceptable or accepting ones that linguists claim are ill-formed? If it turns out, for example, that the reason is that subjects' ratings are based on some irrelevant alternative parse of a sentence, this could lead to construction of less ambiguous materials for a subsequent rating experiment. If it turns out that subjects are not interpreting a critical word in the way intended, this could lead to the use of a

[1] Linzen and Oseki (2018), one of the two non-English studies cited, replicated 11 out of 18 (61%) Hebrew contrasts and 14 out of 18 (77%) Japanese contrasts, but these relatively low numbers reflect biased samples that were deliberately chosen to focus on judgments that the authors believed were incorrect, in contrast to, e.g., Sprouse et al. or Mahowald et al., who tested random samples. Lizen and Oseki's conclusion is worth quoting: "We stress that our results do not suggest that there is a 'replicability crisis' in Hebrew or Japanese linguistics." Song et al. (2014)'s Korean study replicated 102 out of 118 (86%) contrasts exhaustively sampled from two volumes of a journal. It seems premature to draw any conclusions about published judgments in English versus other languages.

preceding sentence to provide context in a subsequent rating experiment. Of course, the first attempted "fix" might not be perfectly successful. If the results do not change, or change for only a subset of respondents, we need to keep asking why subjects are giving the ratings they are. What I describe in this chapter is a kind of experiment that seeks to answer that question, and thus complements rating questionnaire experiments by helping us to interpret their results and to successively refine them to the point where they (ideally) tell us only about subjects' grammars. At a gross level, this kind of experiment is very familiar: psychologists know it as an interview, linguists know it as elicitation. What is (perhaps) novel is how I propose that the field deploy this tool alongside its other (relatively) new tool, the large-scale judgment survey. I suggest that they be used in tandem: a survey experiment yields an unexpected result, you explore it with an interview experiment, develop hypotheses, test these in another survey experiment, if the results do not resolve the issue, you repeat the process.

Thus, the chapter seeks to make a general point using a specific example. The general point is that collecting lots of numerical ratings from subjects (of anything, not just sentences) is useful only to the extent that you are confident you know what they are basing those ratings on. If you are not very confident, you should find out, and often the best way is to ask them (generally via a separate experiment). Now to the specific case at hand. In syntactic argumentation, we are trying to pinpoint one very specific thing that is making a type of sentence go bad (or at least, go worse than a very similar type of sentence), but any particular example sentence will have oodles of other properties that naïve subjects could react to, so we generally should have low confidence that we know the reasons behind their ratings. If those ratings do not match linguists' claims, is the most likely explanation that the intuitions of linguists were wrong? That is certainly possible, but I will demonstrate that alternatives are abundant, and my hunch is that as a whole they are more likely. The reader may disagree, but the take-home message is that this is an answerable empirical question, and therefore one we should strive to answer each time it comes up, rather than assuming the worst-case answer (i.e., that linguists got it wrong).

1.2 The approach

To better understand how subjects are actually responding in crowdsourced experiments of the sort cited above, subjects in the lab first underwent an abbreviated version of one of those experiments, interacting only with a computer. The abbreviated experiments were partial

replications of Sprouse, Schütze, and Almeida (2013), details of which appear below[2]; due to my involvement in that study, I already had ideas about what aspects of the stimuli might have been problematic. Then each subject was interviewed about their responses, to (hopefully) reveal why they reacted to the sentences in the way that they did. This is not unlike how linguists have traditionally gathered data from native speaker consultants, or indeed, from each other. When linguists elicit judgments, they may start by asking "Can you say X?", but a Yes/No answer is rarely the end of the matter. An interactive discourse typically ensues in which the "subject" can ask a range of relevant questions. If the subject happens to be another linguist, these questions can be formulated in technical terms, e.g., "Do you want *she* to refer to Mary? Do you want the modal to scope over negation? Do you care if the elided pronoun gets a strict or sloppy interpretation? Can I put focus on word X? Are you asking the reason for the telling or the buying? What's a scenario where you would want the sentence to be true? What's a discourse in which you would want the sentence to be felicitous?" Naïve speakers may seek such information too, or the linguist may offer it up front. The point is that it is common for a sentence being judged not to be fully "self-contained," in the sense that it lacks information relevant to rendering a judgment *that bears on the issue the linguist is interested in*. What is going wrong in crowdsourced judgment experiments, I contend, is that some sentences are being tested for which such extra information is relevant, but obviously there is no way for subjects to ask for or receive it.

Having subjects come to the lab was crucial: most often, linguistic consultations take place in person (or via an audio-visual computer link such as Skype), where the use of spoken language provides a much richer signal to work with than purely written materials. By contrast, to my knowledge no one has attempted systematic large-scale acceptability studies on naïve subjects using auditory presentation, presumably because they would immediately face the conundrum of showing that the prosody used for (potentially) unacceptable sentences is appropriate, i.e., it makes them sound neither better nor worse than they "deserve to sound." This is a challenging methodological problem that I hope the field will begin to tackle. In the

[2] Choosing that study as a starting point was a matter of convenience, and should certainly not be read as a claim of superiority: indeed, it should become clear that I am particularly well positioned to identify shortcomings in the materials.

meantime, my results will reinforce the value of prosodic information in linguist–subject interactions; this could not have been demonstrated with written follow-up questionnaires.

1.3 Roadmap

Section 2 recaps the critical details of the study that I am following up on. Section 3 provides the methodological details of the new experiments. Section 4 reports major qualitative findings. Section 5 considers consequences and future directions.

2. Background

2.1 What we did

Sprouse, Schütze, and Almeida (2013, henceforth SSA) tested English syntactic examples sampled randomly from 10 years' worth of *Linguistic Inquiry* articles, restricted to sentences whose acceptability (we hoped) could be assessed without having to present any supplementary information concerning interpretation, such as that conveyed by referential indices, the (typically struck-through) interpretation of elided material, etc. The stimuli consisted of 148 "pairwise phenomena," i.e., purported contrasts in acceptability between two sentences that were (ideally) identical in all respects other than the syntactic issue at hand, e.g.,

(1) a. *Who do you wonder what bought?
b. What do you wonder who bought?

In the original articles, one member of the sentence pair had an annotation indicating some degree of degradation ("?", "*", etc.) while the other was unannotated. For ease of reference I call these the "bad" and "OK" members of an item pair.[3] In a few instances where the OK

[3] When linguists present pairs of this kind in articles, they are not always explicit as to whether they are claiming only that the OK member of the pair is substantially better than the bad one, or also claiming that there is "nothing wrong" with the OK member. SSA essentially ignored the latter potential claim, i.e., the linguists' judgments were considered confirmed if subjects' judgments reliably differed between the bad and OK tokens (in the right direction), no matter the absolute rating of the OK tokens. It is not obvious to me how to operationalize claims of "nothing wrong" with a sentence, actually, nor is it clear how important such claims might be for linguistic theory.

member of the pair was not explicitly provided but was clear from context, SSA supplied it themselves.

SSA tested eight tokens of each pairwise phenomenon: generally the one(s) that appeared in the article (the "original token(s)") verbatim, plus ones we made up that were intended to have the same structural properties but different open-class content (the "new tokens"). (As we shall see in §4, the challenges in inventing new tokens that have just the same properties as the original turn out to be myriad.) One goal in creating these new tokens was to seek empirical evidence that the original linguists' assertions were empirically true for the full range of structures that the associated theoretical claims would apply to, rather than being an accident of idiosyncratic properties of the particular example(s) presented in the article. (In retrospect, SSA might have been overly ambitious in this regard—see again §4.) Marantz (2005) notes that linguists are seldom explicit about the fact that examples in articles typically are stand-ins for large (often infinite) sets of sentences that are claimed to behave identically, and that the crucial empirical claims are about such sets as a whole. Examples in articles are invitations for the reader to verify the claims they are meant to illustrate, often by pondering relevantly related examples beyond—or in place of—those actually provided. It would rarely be of any interest for linguistics to know that one particular sentence is (un)acceptable if such (un)acceptability were not reflective of a large class of relevantly similar sentences.[4] Consequently, it is actually not relevant for syntactic theory whether native speakers disagree with the explicitly-reported token judgments in an article (this could be due, i.a., to quirks of particular words whose lexical entries might vary across speakers), so long as they agree that the relevant *class* of sentences generally patterns as claimed. We should guard against a potential sort of naïve falsificationism in this regard.

SSA conducted their experiments via AMT, testing the full stimulus set in three different tasks (with 312 subjects each): single sentence presentation with Magnitude Estimation (ME)

[4] Interestingly, the same is not true in the sentence processing literature, where it is of great interest that structurally identical sentences may be treated very differently by the parser as a function of lexical idiosyncrasies, e.g., the contrast between the trivial (ia) and the intractable (ib).

(i) a. The mine buried in the sand exploded.
b. The horse raced past the barn fell.

responses; single sentence presentation with 1–7 Likert Scale (LS) responses; and paired sentence presentation with Forced Choice (FC) responses ("Which is better, a or b?"). In the first two tasks, a given subject saw only one member of each token pair, while in the third a subject saw both members together.

2.2 What we found and what we can(not) conclude

Results were assessed mainly in terms of whether each pairwise phenomenon showed a significant contrast in the direction claimed by the linguists or not; we ignored issues such as whether "?" sentences were better than "*" sentences. For the vast majority of the 148 pairwise phenomena, our naïve subjects did replicate the published judgments—specifically, this was true for 93% of the phenomena in the ME and LS tasks and 95% in the FC task. We referred to these as rates of "convergence" between the judgments of linguists and naïve speakers. (The non-convergent phenomena showed a variety of patterns detailed below.) The question that then occupied us, and has been taken up in subsequent literature, is whether these convergence rates are cause for celebration or consternation—that is, do they indicate that the informal way in which linguists have traditionally gathered and presented data is reasonable or do they cry out for the field to adopt "more rigorous" standards? Opinions have differed on this point,[5] but most responses to these and similar results have made an assumption that the present chapter is designed to question: people have generally taken for granted that the nonconvergent phenomena

[5] In reviewing an earlier draft of this chapter, Ted Gibson emphasized two points that he states he and his colleagues have made in their responses to SSA (Gibson, Piantadosi, and Fedorenko 2013; Mahowald et al. 2016). In fact, the former paper was responding in part to an early draft of SSA written before the published experiment was even run and mostly to Sprouse and Almeida (2013), and both points are arguments against the suggestion that syntacticians do not need to run formal experiments, a suggestion that appears nowhere in SSA. Furthermore, since this chapter is about interpreting the results of such experiments, I obviously am not making that suggestion.

constitute bad data, i.e., that linguists' claims about them must have been wrong.[6] For the sake of discussion, let us suppose that if naïve speakers' judgments of these phenomena "genuinely" differed (in a sense I make clear) from those reported by linguists, we indeed ought to reject the linguists' judgments and adjust our theories to account for the naïve judgments instead.[7] I want to be emphatic that we would not be justified in taking that step yet, because the premise is most likely false: the nonconvergent data probably do not represent genuine judgment disagreements.

The crucial point is that there is a logical gap between the experimental observations and the conclusion: knowing that subjects gave nonconvergent responses to a set of strings does not entail that their grammars do not conform to the pattern claimed in the source articles—it could be that subjects' responses reflected factors other than the acceptability of those sentences on the intended interpretations. This would obviously be true if, for example, subjects systematically misread a set of stimuli—and I show that this can happen. More generally, we cannot establish genuine judgment disagreements until we are certain that subjects read and understood the stimuli in the manner relevant to the linguists, and that the low ratings (individually) or dispreferences (in pairs) for bad items were due to the property that the linguists identified as their flaw, rather than some completely orthogonal issue. The experimental results to follow suggest that it is common (though they cannot indicate how common) for at least one of these preconditions not to be met: subjects' responses will be demonstrated in many cases to be based on a structural parse different from the one of interest to the linguist, and/or on disliking some property of the sentence that has nothing to do with its purported grammatical violation.

---

[6] Thus, Gibson et al. (2013) repeatedly use the phrase "error rate" to refer to the nonconvergence rate. By contrast, Mahowald et al. (2016: 626) are more cautious: "It should not automatically be concluded that…these sentences [represent] failures on the part of the researchers."

[7] As Jon Sprouse and I have discussed (separately and jointly) in previous work, it is not obvious that this move would be appropriate. Among other things, we would want to know whether linguists agree among themselves about the judgments in question or whether the original authors were outliers, whether we might be dealing with genuine dialectal differences, etc.

Consequently, the results of SSA and similar studies may at best[8] indicate a lower bound on what the true rate of convergence is (for whatever range of phenomena they examined): there could well be some genuine judgment disagreements among the nonconvergent items, but I suspect the majority are spurious: only further experimentation designed to avoid the confounds I exemplify below can answer this question. To be frank, the reason that I can illustrate many ways in which judgment tasks can go wrong is partly because there were abundant flaws in the materials. But those flaws, I would argue, are less the result of carelessness and more the result of ambition—the desire to test the claims actually being made in the source articles, rather than the more straightforward but less informative alternative of testing the properties of the particular sentence that happened to be chosen to illustrate a given claim, or truly trivial variants thereon.[9]

2.3 Why we shouldn't be surprised

Upon reflection, it is not surprising that the results of crowdsourcing experiments like those of SSA are challenging to interpret. It is naïve to think that AMT could be the panacea for linguistic data gathering that some seem to have hoped: in addition to the domain-independent problems of not knowing much about who the subjects are or where, when, under what conditions and with what intentions they may be engaging in our tasks, the nature of language in general and grammaticality judgments in particular induces domain-specific problems. Subjects are reading isolated sentences. There is no context to help them zero in on the intended meaning, only scant hints as to how the sentence should sound, and they have no opportunity to seek clarification or even indicate confusion: all they can do is pick a number between 1 and 7, or pick one of the

[8] I say "at best" because I cannot exclude the possibility that some *convergent* results were not genuine judgment *agreements*, i.e., subjects might have given the "right" answers for the wrong reasons. Indeed, we shall see some potential cases of this type.

[9] SSA could have simply replaced *John* with *Bill, Bob, Fred…* and *cat* with *dog, rabbit, hamster*, and so on, but generally we aimed higher. We might well have achieved slightly greater convergence rates if we had taken the easier road, but we will learn more in the long run because we did not. Indeed, to the extent that the materials in the present experiments diverge from those of SSA, the intent was to be still more ambitious, creating more opportunities to (fail and hence) learn.

sentences. The communication bandwidth between experimenter and subject is incredibly narrow compared to how linguists traditionally gather data, be it from a colleague down the hall, a class of undergrads, or a fieldwork consultant. While the general problems with AMT may simply contribute noise, the domain-specific ones can lead to systematic confounds.

Furthermore, examples that appear in linguistics articles are generally not designed for purposes of presentation to naïve subjects, but rather for the primary readership of such articles: other linguists who understand what issue(s) an example is intended to bear on and who will therefore ignore aspects of the example irrelevant to those issues. (Indeed, examples are sometimes further shaped by a desire to make them entertaining to fellow linguists.) I suspect it is common for linguists reading articles to consciously observe that applying their "raw judgments" to such examples may yield results that do not conform to the author's claims, but that they can construct related examples that factor out orthogonal problems in order to verify the crucial claims. (By contrast, when psycholinguists design materials for experiments, great care is generally taken to avoid ambiguity, unnecessary processing complexity, pragmatic implausibility, garden pathing, uncommon words, etc.) Trying to test examples verbatim from articles was, in retrospect, asking for trouble.

From this perspective, convergence rates of 93–95% across a broad sample of linguistic phenomena are almost miraculous, in my view—testimony to the robustness of the underlying cognitive structures. Still, we clearly must take the nonconvergences seriously, and I do not advocate excluding platforms like AMT from the linguistic enterprise, though they surely can never replace the need for face-to-face encounters.

3. Experimental methods

3.1 Properties common to all three experiments

3.1.1 Participants

All subjects were UCLA undergraduates from the psychology department's subject pool who had not taken a linguistics course; they received course credit for participation. They self-identified as native speakers of North American English (not necessarily monolingual, because that restriction would have excluded too large a proportion of the subject pool).

3.1.2 Procedure

Subjects came to our lab, where the experiment was conducted in a quiet room. As dictated by subject pool constraints, it lasted at most one hour. After informed consent was granted, the first phase of the experiment used a computer to collect judgments on sentences presented visually on the screen without interaction with the experimenter. The second phase consisted of an interview in which the experimenter presented some of the sentences seen in the first phase on paper and asked subjects about their reactions to them. This phase was audio-recorded and later transcribed by the same experimenter, who also took contemporaneous notes that could be used to add clarification to the transcripts. Between the two phases a computer program was run to generate the list of sentences for subsequent discussion; two copies of this list were printed, so that the experimenter and subject could each read it easily. Items were numbered for ease of reference, and the subject's response from the first phase was indicated. The program used heuristics to identify sentences that were likely to be informative to discuss, based on a given subject's responses, and it ordered them with those likely to be of greatest interest first. (This was done because of the time constraint: interviews often ended before all sentences on the printout could be discussed. To further maximize the breadth of information gathered, the experimenter had the discretion to skip items exemplifying a phenomenon already discussed.)

The details of the heuristics were complex and are not claimed to have any scientific validity, but they did embody some bias that should be kept in mind when considering the results that follow. For items whose mean outcome from SSA was convergent, responses that were probably replicating that result[10] were usually considered of little interest, while opposite responses were likely to be flagged for further investigation. The reverse was true for items that were divergent in SSA. Consequently, if there are items that naïve subjects generally judge in

[10] The caveat "probably" is necessary because of the logic of the design of the LS experiments. In such experiments a given subject sees only one member of each minimal pair token (e.g., sentence (1a) or (1b) above), but the aggregate results consider the direction of difference in mean ratings between members of minimal pairs. Thus, the heuristics can only look at the absolute (raw or normalized) ratings of individual sentences (comparing them, e.g., to those of SSA) or compare condition means of tokens drawn from different minimal pairs of the same pairwise phenomenon.

agreement with linguists but for the wrong reasons (which we might call "false convergence"), the current study is less likely to uncover them (but we will see a couple of potential cases).

In the interview phase, discussion of each item began with the experimenter asking the subject to read the sentence(s) out loud. This was intended both to reveal any potential misreadings (e.g., skipped words) and to elicit prosody that could indicate parsing choices the subject might have made. Nontrivial misreadings were corrected; trivial misreadings, such as *definitely* for *definitively* or *subject* for *suspect*, were not. How the discussion proceeded thereafter was left to the experimenter's discretion, but usually included a request to paraphrase the sentence, and if the subject had given it a low rating (or it was the dispreferred member of a pair in the FC task), a request to identify which part or what aspect of the sentence sounded bad to the subject. There could then be follow-up questions that would further elucidate the subject's responses.

3.1.3 Materials

The stimuli were created from the items used by SSA, in particular, from 34 of the 148 unique pairwise phenomena. Many items were retained verbatim, while some underwent changes of various extents intended to address concerns that became apparent subsequent to running that study.[11] (Unfortunately, those changes themselves sometimes introduced new problems, as I confess in §4.) The 34 phenomena break down as follows:

- 4 pairwise phenomena that came out significantly in the unexpected direction in at least one of the three tasks in SSA;
- 2 pairwise phenomena that came out numerically in the unexpected direction in at least one of the three tasks;
- 6 pairwise phenomena that came out numerically but nonsignificantly in the expected direction;
- 10 pairwise phenomena that came out significantly in the expected direction in some but not all tests;

[11] Many of the flaws are discussed in the online supplement to SSA at http://sprouse.uconn.edu/papers/SSA.Materials.xlsx. In our defense I can only point out the enormity of our task: we had to create 148×7×2 = 2072 items for those experiments.

12 pairwise phenomena that consistently came out significantly in the expected direction.[12]

The number of token pairs per phenomenon ranged from one to six in Experiments 1 and 3 (based on suspicions that differences among tokens might have been important),[13] but was fixed at four in Experiment 2. The two systematic changes made vis-à-vis SSA were that items in the present study could have one or two words printed in all capitals to indicate intended prosodic emphasis and could have commas added, in cases where this could be done identically in both members of a minimal pair and was thought to be helpful in bringing out the intended reading. In addition to the target items just described, each experiment included catch items that were believed to be uncontroversially acceptable or uncontroversially unacceptable (also labeled "OK" and "bad" below), intended to assess how closely subjects were paying attention. Since the targets represent a wide variety of sentence types, no additional fillers were included.

[12] All of these descriptions suppress some detail: SSA ran three different kinds of significance tests on each task's results, which did not always agree, plus for some phenomena the numerical results fell in opposite directions across tasks, particularly when comparing FC to the other two, so the outcomes resist brief summary. Suffice it to say that I included all phenomena that raised doubts about the original linguists' claims, plus phenomena where particular tokens seemed to be behaving suspiciously. Note that, of the subsets listed above, both the 4+2 and the 6 could be seen as failures to replicate linguists' claims (nonconvergence), though the 6 could also be consistent with convergence without sufficient statistical power. Viewing all of them as replication failures would add up to 12/148 = 8.1% nonconvergence; SSA's nonconvergence percentages (their Table 6) range from 1% to 14% because they were broken down by task and statistical measures.

[13] In a few cases, I made refinements to particular tokens from one experiment to the next. Since results below are pooled across experiments, some lists of items contain tokens that did not appear in the same experiment.

3.1.4 Instructions

In addition to describing how the experiment would be carried out, the instructions included information on how to interpret "grammaticality", which read as follows, with slight variations for the FC task (this was different from the wording used by SSA):

> After you read each sentence, you will rate it based on how "grammatical" it sounds to you. For the purpose of this experiment, a *grammatical* sentence is one that would seem natural for a native speaker of English to say. In contrast, an *ungrammatical* sentence is one that would seem unnatural for a native speaker of English to say. The response scale goes from 1 (definitely **un**grammatical) to 7 (definitely grammatical).
>
> To make it clear what we would like you to base your responses on, here are some <u>things that we are NOT asking you to rate</u>:
>
> 1) How understandable the sentence is: A sentence may be perfectly easy to understand even though native speakers of English would agree that it is ungrammatical. For example, if you heard someone say "What did you ate yesterday?", you would have no difficulty answering the question, but you would still give the sentence a low rating.
>
> 2) How the sentence would be graded by an English teacher or writing tutor.

3.2 Properties unique to individual experiments

3.2.1 Participants

Experiment 1 involved 23 participants, 3 of whom had their responses excluded based on their ratings of the catch items.[14] Experiment 2 involved 16 participants, with none excluded on the same basis. Experiment 3 involved 20 participants, also with none excluded.[15]

3.2.2 Procedure

Experiments 1 and 2 elicited ratings of individual sentences on a 1–7 Likert scale (LS task).[16] Subjects pressed a number key on the computer keyboard corresponding to one of the seven

[14] Exclusion criteria were a mean rating for the two OK items lower than the mean rating for the two bad items, or neither of the bad items being rated lower than 3.

[15] The exclusion criterion was making the expected choice on fewer than 6 of the 8 catch item pairs.

possible ratings. On the screen, potential ratings were labeled only by number, but to the left of "1" was the string "definitely ungrammatical" and to the right of "7" was the string "definitely grammatical" to remind subjects of the direction of the scale. To familiarize subjects with the task and attempt to anchor the response scale, four practice trials were presented prior to the experiment proper.

Experiment 3 elicited forced-choice responses to minimal pairs of sentences (FC task). Subjects clicked a radio button next to the sentence they considered more grammatical. Sentences were displayed one above the other, counterbalancing which position was occupied by the OK member of the pair. I was concerned that this presentation mode might induce a strategy whereby subjects could identify where the two sentences diverged by visual matching without actual reading, and then base a response upon only the mismatching substrings. To deter this, one member of each sentence pair was indented (the choice counterbalanced against both vertical position and bad/OK status), e.g., (2) or (3).

(2) Who do you wonder whether saw John? ❍
Who do you wonder whether John saw? ❍

(3) Who do you wonder whether saw John? ❍
Who do you wonder whether John saw? ❍

To familiarize subjects with the task, four practice trials were presented prior to the experiment proper.

3.2.3 Materials

Experiment 1 contained 63 target sentences, representing 33 of the 34 pairwise phenomena, plus 4 check sentences. Experiment 2 contained 88 target sentences, representing 22 of the 34 pairwise phenomena, plus 4 check sentences. In experiments 1 and 2, each subject saw only one member of each token pair (half OK, half bad); item order was randomized for each subject. Experiment 3 contained 63 target sentence pairs, representing 33 of the 34 pairwise phenomena, plus 8 check pairs. In this experiment, four lists were generated to accommodate the four

[16] LS was chosen over the ME task for two reasons: it is easier for subjects to understand, and the results are less noisy statistically (SSA; Weskott and Fanselow 2011).

presentation variants of each sentence pair (OK vs. bad on top, crossed with indented vs. unindented on top); item order was randomized for each list.

4. Results

The purpose of the discussions that follow is not to try to establish what the "true" acceptability of the target sentences is, but rather to illustrate the ways in which numeric ratings by themselves can be misleading if we know nothing about the basis on which the ratings were offered. The forthcoming presentation of results is characterized by an absence of statistics, both descriptive and inferential. This is deliberate and necessary. As can be gleaned from the preceding discussion, neither the full set of items in the first phase of the experiment nor the subset discussed during the second interview phase in any sense constituted random or balanced samples relative to the original full SSA random sample. Thus, quantitative summary statistics would not be meaningful. Results have been subjectively selected for reporting based on their potential to inform our understanding of how subjects approached the tasks in general and the assessment of specific sentence structures in particular, with the aim of representing the wide variety of challenges involved. I attempt to give some sense of how common a given type of reaction to an item was, keeping in mind that there was not much control over how likely a given subject was to be interviewed about a given item.

4.1 General observations

Before delving into details, a few general observations are in order, all of which will be exemplified among the case studies to follow in section 4.2. One concerns misreadings: the most common error in out-loud reading was omission of short words, and when this was pointed out to subjects they typically remarked that they had made the same error during the first phase of the experiment, meaning the computer-elicited responses were not judgments of the intended string. The potentially dire consequences of such errors were brought home by the fact that several subjects gave high ratings (e.g., "7") to the following original catch item:

(4) Who do you suspect I would be capable of committing such a crime?

presumably because they missed the word *I*. The lesson is that, when possible, crucial contrasts should be expressed using more and/or longer words rather than hinging on one short one,[17] but as discussed below this was sometimes not possible.

A second observation concerns the attempt to manipulate prosody by capitalizing words intended to receive emphasis. Subjects' out-loud readings sometimes did not follow these "hints," either because they forgot the relevant instruction or because their attempt to implement it yielded an unexpected result. In such cases the experimenter normally would eventually say the target string with the intended prosody (after trying to determine how the subject originally understood the sentence), often yielding a reaction indicating that this had revealed an interpretation the subject had not previously been considering. Again, this means the computer-elicited ratings were not judgments of the intended structure. This highlights the importance of implicit prosody in silent reading tasks, and the need to find (more reliable) methods for conveying intended prosody when this is important.

A third observation concerns word choice. In the LS experiments subjects sometimes identified the flaw that triggered a low rating as a particular word that they found unfamiliar or so uncommon that "No one would say that." Given the close matching within minimal pairs of items, these words virtually always were identical in the bad and OK versions of a sentence, making these responses orthogonal to the point that the original linguists were illustrating. While subjects occasionally remarked on these words in the FC task too, they obviously could not use them as a basis for preferring one member of a pair over the other. Methodologically, it might thus seem advantageous to present sentences as minimal pairs rather than singletons to avoid such irrelevant responses. (Of course, such presentation need not be combined with an FC response task—one could still elicit numeric ratings of each sentence individually.) However, pairwise presentation also draws conscious attention to the manipulation of interest, which might have unintended consequences, as we shall see.

A fourth observation is that, as is evident below, some minimal pairs express identical (intended) meanings while others, by virtue of the syntactic contrast they address, cannot possibly do this. In the latter situation, responses in both tasks could be influenced by differences in the plausibility of the scenarios described, and we shall see evidence that this happened. Such

[17] And indeed *I* was replaced with *she* once this pattern was noticed.

responses could in principle lead to false convergence or false divergence. The way to avoid this would be to pretest a large number of candidate sentence pairs to identify a subset for which plausibility ratings are matched.

A fifth and final observation is that, contra the impression one might get from just the first case study below, instances when subjects felt they did not understand a sentence they read were rare; much more commonly, they arrived at an understanding that was different from the one the linguists had intended.

4.2 Case studies

In this section, pairwise phenomena are always presented with the bad member labeled (a) and the OK member labeled (b), but no annotations of ill-formedness are included. Examples marked with "†" are original tokens (modulo CAPS, which were never in the original). The citation refers to the article whence the phenomenon was drawn by SSA's random sampling procedure (where it may have represented an original claim by the authors or reported a previous or "standard" judgment). Items have been grouped into subsections by the properties of the sentences. For each pairwise phenomenon, the relevant set of tokens is listed, followed by discussion of subjects' responses.

4.2.1 Antecedent-Contained Deletion

(5) a. John wants for everyone YOU do to have fun.† (Fox 2002)
b. John wants for everyone to have fun that YOU do.[18]
(6) a. Sophia is anxious for everyone YOU are to arrive.
b. Sophia is anxious for everyone to arrive that YOU are.
(7) a. Ben is hopeful for everyone YOU are to attend.
b. Ben is hopeful for everyone to attend that YOU are.
(8) a. We want for everyone that you do to have fun.
b. We want for everyone to have fun that you do.
(9) a. Sophia is excited for everyone that you are to arrive.
b. Sophia is excited for everyone to arrive that you are.

[18] This is a non-minimal pair syntactically in that the complementizer in the relative clause is overt in (b) but silent in (a). Fox makes no comment on this, and presumably he could just as easily have had overt *that* in (a), so I included it in some tokens. The same is true for the subsequent data set based on Bhatt and Pancheva (2004).

(10) a. The coach is thrilled for every player that you are to play.
b. The coach is thrilled for every player to play that you are.
(11) a. Valerie is excited for everyone YOU are to graduate.
b. Valerie is excited for everyone to graduate that YOU are.

These are instances of Antecedent-Contained Deletion (ACD) modifying the subject of an embedded infinitival clause. Both items in a pair have the same intended meaning, e.g. for (5), *John wants (for) [$_{DP}$ everyone OP$_i$ that you want [t_i to have fun]] to have fun*; thus, *do* is intended to be the dummy auxiliary introducing an elided VP, and the emphasis on *you* is meant to contrast it with *John*[19] and encourage deaccenting of *do*. These were undoubtedly the most difficult structures in the experiments for subjects to figure out. To the extent that they were able to paraphrase the sentences (often they said they were unable to), their paraphrases were almost never the intended ones. For example, for (5a) subjects offered "John wants everyone that you do something for, or do something with, to have fun" and "John wants everyone to have the same amount of fun, so, all the fun that you had, John wants that for everybody". These paraphrases make clear that the intended antecedent for VP Ellipsis was not identified: in the first, *do* was apparently interpreted as a main verb and there was no ellipsis at all, while in the second, the elided material seems to be *(have) fun*, i.e., *John wants everyone to have the same amount of fun that you have*. Likewise for (6a), a subject offered a no-ellipsis interpretation where *everyone you are* contains a copular relative clause and means "every aspect of you." (7a) got paraphrased as "Ben is hopeful on behalf of everyone that you will attend".

But the (b) versions fared little better: one subject paraphrased (8b) as "[We] want everyone to have the same amount of fun as you have". (9b) was paraphrased by two subjects as "Sophia is as excited as you are for everybody to arrive"; a third rendered it as "Sophia is excited…uh…something about that you're arriving." Even a subject who paraphrased (5b) accurately gave it a "2" rating because "it just starts off funny…I wouldn't say a sentence like that."

(9a) grammatically allows an unintended parse that involves no ellipsis, on which *that you are to arrive* is a CP meaning "that you are going/expected to arrive", such that the whole sentence means roughly *Sophia feels excitement on behalf of everyone about the fact that you*

[19] It was felt that putting the matrix subject in all caps in such examples was not critical: out of the blue, a nonpronoun in this position seems likely to get a pitch accent anyway.

will be arriving. Several subjects reported this interpretation (which often got a high rating).[20] (10a) similarly was wrongly interpreted by multiple subjects as *The coach is thrilled for [every player [who$_i$ you are going/expected to play (with/against)* t$_i$*]]*, which was then sometimes rated low because it seemed odd that a coach would be thrilled for opposing players rather than their own players. In the FC task, a subject who read both members of (11) (and supplied emphasis where intended) gave them different paraphrases, taking (11a) to mean "Out of all graduating people, Valerie is glad that you're graduating too" while taking (11b) to mean "Valerie is more excited for you to graduate than other people."

(12) a. I expect that everyone YOU do will visit Mary.[†] (Bhatt and Pancheva 2004)
b. I expect that everyone will visit Mary that YOU do.
(13) a. I suspect that all the couples that HE does will kiss.
b. I suspect that all the couples will kiss that HE does.

This pairwise phenomenon differed from the previous ACD structure by virtue of the relativized DP being the subject of a finite clause. Several subjects reported that they could not discern the intended meaning of the sentences. Some reported that it became clear once the experimenter read the example with the intended prosody. One subject took (12a) to mean "I expect that everyone you visit will visit Mary," which seems to involve the antecedent for VPE coming after the ellipsis site, and requires ignoring *Mary* as the object in that antecedent. One thought (13b) might mean "Everyone he kisses, those couples will kiss," which seems to involve taking the elided VP to be headed by *kiss* (and transitive, despite the fact that the antecedent would be intransitive). Another paraphrased (13b) as "I suspect that all the couples will kiss the way that he does."

4.2.2 Extraction from PP within DP

(14) a. Who did they find a parent of guilty? (Bruening 2010)
b. Who did they send a parent of to an unpleasant meeting?

[20] I suspect the overtness of *that* in (9a) and the absence of contrastive capitalization made this parse more readily available. The fact that not all tokens in this paradigm were given the CAPS treatment was intended to probe for such effects.

(15) a. Who did the citizens elect an enemy of mayor?
b. Who did the coach trade an enemy of to another team?[21]
(16) a. Who did officials proclaim an associate of the winner?
b. Who did officials dispatch an associate of to the embassy?

Both members of each pair are intended to include a DP containing a PP with a stranded preposition, e.g., *[a parent of t_{who}]*; at issue is whether extraction from the subject of a small clause (e.g., complement of *find*) is harder than from the (direct) object of *send*. Unfortunately, (14a) is conducive to a garden path effect in a way that (14b) is not: the word following *of* could be the beginning of an overt complement to *of*, as in *[a parent of [guilty children]]*. The correct parse would typically differ from the incorrect one by containing a prosodic break after *of*, but it was not obvious how to convey that prosody in visual presentation. Tellingly, when some subjects read the sentences out loud the first time there was sometimes no such break, as if *a parent of guilty* were a constituent. It is apparently on that basis that some of them rated the sentence "1", perhaps because that would leave no obvious position for the trace of *who*. These were therefore cases of getting the right result for the wrong reason (false convergence). (Some subjects were subsequently able to identify the correct interpretation after the experimenter read the sentence with the intended prosody.) Another sort of misparse seems to be reflected in a subject's paraphrase of (15a) as "Who did the citizens elect that was an enemy of the mayor?". When SSA created (16), we apparently failed to notice that it, unlike the other tokens, is not just temporarily but in fact globally ambiguous: instead of the trace of *who* being the complement of *of*, it could be the subject of the small clause: *Who*$_i$ *did officials proclaim* [$_{SC}$ t_i *[an associate of the winner]]*. At least two subjects seem to have gone for this parse, on the basis that their prosody grouped the last five words as a constituent.

4.2.3 Control

(17) a. I told Mr. Smith that I am able to paint the fence together.† (Landau 2010)
b. I told Mr. Smith that I wonder when to paint the fence together.
(18) a. Jennie discovered that her boyfriend dared to kiss in front of his parents.
b. Jennie discovered that her boyfriend hoped to kiss in front of his parents.

[21] The pair in (15) differ on the noun in the matrix subject because SSA felt this maximized the plausibility of each sentence (at the cost of a lexical mismatch).

The intended meaning for both sentences in (17) was that *together* should refer to Mr. Smith and I, that is, two people together would do the painting. The hypothesis was that the PRO subject under *wonder when* could take those two DPs as a joint antecedent because as an interrogative predicate it allows Partial Control, whereas *able,* a modal predicate, requires Exhaustive Control and would only allow its complement subject PRO to be interpreted as *I*, providing no grammatical antecedent for *together*. The same contrast was predicted for (18) on the basis that *dare* is implicative (Exhaustive Control) while *hope* is desiderative (Partial Control). Subjects who gave (17a) a high rating did so because of a different interpretation on which *paint the fence together* means "paint the fence such that it would be/stick together", where *together* is a resultative predicated of the object. Subjects described this meaning as involving, e.g., painting on glue to fix the fence. Other subjects apparently treated *together* as a object depictive, meaning that the fence is "all in one piece" or "already assembled." For (18), several subjects in the FC task reported a preference for (18a) over (18b) on the basis that the latter describes a much less plausible scenario: given that the boyfriend wants to kiss in front of his parents and assuming that the parents will be displeased by this, *dare* acknowledges the danger and might suggest a rationale (demonstrating his love is more important?), while *hope* does neither.

(19) a. The bed was slept in naked. (Landau 2010)
b. The bed was slept in wearing no clothes.
(20) a. The club was entered shirtless.
b. The club was entered wearing no shirt.

The intended meaning of (19) was that the implicit agent, the one who did the sleeping, was naked/wearing no clothes. Landau's claim was that the participial in (19b) should be able to be interpreted with the "weak" implicit agent as its subject because it heads a Control structure *[PRO wearing no clothes]*, whereas the bare adjective in (19a) should not because it combines with the main clause by predication. However, several subjects gave the (b) sentences very low ratings because they took them to mean that the bed was wearing no clothes, the club was wearing no shirt, etc., a meaning they (reasonably) found anomalous. Other reasons for low ratings of (19b) included "You never say 'the bed was slept in', you just say 'I go to bed'."

Examples like (20), where the same noun (*shirt*) appears suffixed by *-less* and following *no*, seem closely synonymous; this triggered responses in FC preferring (a) because it is less

wordy/more concise.[22] This is a case where drawing attention to the difference between the members of a minimal pair might have had an undesirable consequence: it is possible that if subjects find such pairs equally well-formed they fall back on concision as a basis for establishing a preference. (Of course it is possible that subjects in the LS experiments were also responding on the basis of concision, despite only seeing one member of such pairs, by thinking up (a) as a synonym for (b). Indeed, thinking of other strings that would express the meaning of a target string appears to be a common strategy—see §4.2.9.)

4.2.4 Floating quantifiers

(21) a. All the winners are unlikely to have all been notified already. (Costantini 2010)
b. The winners are unlikely to have all been notified already.

The alleged problem with the (a) example is two instances of *all* "floated" off the same DP chain, but numerous subjects rated such sentences very high ("6" or "7") in the first phase; when asked to re-read them in the second phase, many recalled not having noticed this redundancy, i.e., their initial high rating was based on a misreading of the string. (This is an example of the "one short word" problem for which there is no obvious work-around.) However, other subjects seemed perfectly happy with the two *all*s, reading the (a) sentence correctly and affirming their initial high ratings (e.g. "7"), warranting exploration of a potentially genuine data disagreement.

4.2.5 Embedded tense and aspect

(22) a. George remembered that he would have made the phone call by the time he leaves work. (Martin 2001)
b. George remembered that he will have made the phone call by the time he leaves work.

[22] The present design does not allow us to distinguish whether this was actually the basis for subject responses in the first phase of the experiment or whether they came up with this as a way to rationalize their preference after the fact during the second phase. Either way, Landau's claim is called into question: one would probably not expect such responses if the (a) version violated the grammar in a way that the (b) version did not.

The adjunct *by the time he leaves work* was intended to be interpreted as part of the embedded clause in both examples; this should lead to a tense clash within that clause in (22a) but not in (22b). However, if the adjunct is instead construed as modifying the matrix clause, it causes a tense clash (with *remembered*) in both examples. Some subjects reported this interpretation and consequently rated (22b) low.

4.2.6 Complementizer omission

(23) a. It seems as of NOW David had left by December. (Bošković and Lasnik 2003)
b. It seems as of NOW that David had left by December.

In (23) the original claim was that omitting the complementizer *that* when a temporal PP like *as of now* intervenes between *seem* and its complement is degraded. But many subjects gave (23b) a very low rating ("2" or "3"), apparently for independent reasons: they tended to comment on the seeming temporal inconsistency between *seems* and *now* versus *had left* and *by December*. (This problem was introduced by SSA: the article's original example used simple past throughout. However, it contained another confound that SSA were attempting to avoid: *It seemed at that time David had left*† should be fine if the PP is parsed as part of the embedded clause; Bošković and Lasnik indicated with bracketing that this string was purportedly bad only if the PP is parsed as part of the matrix.)

(24) a. Brittany knew that Morty had lost his wallet, but Morty had lost his KEYS, Brittany DIDN'T believe.[23] (Bošković and Lasnik 2003)
b. Brittany knew that Morty had lost his wallet, but that Morty had lost his KEYS, Brittany DIDN'T believe.

The problem in (24) was omission of the complementizer in the second conjunct of (a): a bare TP, unlike the CP in (b), allegedly cannot be fronted. However, the omission in (24a) allows

[23] In SSA we followed the example from the article, which contained only the structure following *but*, e.g. *(That) John likes Mary, Jane didn't believe*†. But I was very dubious that subjects would parse such strings with the first clause as the fronted complement of the second, so I "modeled" the underlying structure in the new first half of the sentence, using contrast to attempt to motivate the fronting in the second half.

a garden path whereby *Morty had lost his keys* is directly conjoined with the material preceding the first comma; if one adopts that parse, *Brittany didn't believe* is missing a (not quite obligatory) complement for *believe,* and its relationship to the rest of the sentence is obscure. It was this unintended parse that was rated fairly low by some subjects—another example of getting the right result for the wrong reason.

(25) a. What the conductor believes is the train will crash. (Bošković and Lasnik 2003)
b. What the conductor believes is that the train will crash.
(26) a. What the bidders hope is they will get the house.
b. What the bidders hope is that they will get the house.

The alleged badness of complementizer omission in these examples is due to the embedded clause being pseudoclefted. One subject misread (25a), supplying the missing *that*, and rated it "7" on this basis. Another rated (26b) "2" because they did not like the use of the pseudocleft, which seemed unnecessarily wordy out of context.

(27) a. They AFFIRMED, but we DENIED, Mary would tell the truth.
(Bošković and Lasnik 2003)
b. They AFFIRMED, but we DENIED, that Mary would tell the truth.

Here complementizer omission is alleged to be blocked by Right Node Raising. But (27b) was sometimes rated low because of an apparent tense mismatch: subjects were expecting *Mary told the truth*; other subjects rated it low because they thought *affirmed* was an odd verb to use in this context. (In SSA, following the original authors, clauses were conjoined with *and* rather than *but* and there were no commas, e.g., *They suspected and we believed Peter would visit the hospital*[†]. But such strings do not force a Right Node Raising parse, given that *suspect* is compatible with null complement anaphora: subjects could have parsed them as *[They suspected Ø] and [we believed (that) Peter would visit the hospital]*.)

(28) a. MARY believed Tommy drank his milk, and JANE he ate his vegetables.
(Bošković and Lasnik 2003)
b. MARY believed that Tommy drank his milk, and JANE that he ate his vegetables.

In this instance, gapping of *believed* is claimed to preclude omission of the following complementizer.[24] One subject initially skipped *he* while reading (28a), which created a perfectly grammatical string yielding a high rating. (In SSA, as in the original article, there was no comma and there was a proper name in place of *he*, e.g., *Mary believed Peter finished school and Bill Peter got a job*†. Given that many first names in English can also be last names, such strings could be parsed without gapping (which requires highly marked prosody): *Mary believed [[Peter finished school] and [[$_{DP}$ Bill Peter] got a job]*, which is why I made the changes.)

4.2.7 Head–complement intervention

(29) a. What did the children say at that time that the band played?
(Bošković and Lasnik 2003)
b. At that time, what did the children say that the band played?

(29a) was meant to be parsed as *What$_i$ did the children say [$_{PP}$ at that time][$_{\text{CP}}$ that the band played* t$_i$*]?*, with an adjunct intervening between the verb and its complement. This might be bad if that adjunct forces the embedded CP to be extraposed, since subsequent extraction from it might be blocked (e.g. by Freezing). However, subjects found a different parse: *What$_i$ did the children say* t$_i$ *[$_{PP}$ at that time [$_{\text{CP}}$ OP$_j$ that the band played* t$_j$*]]*, taking the CP as a relative clause, which unsurprisingly got high ratings. (This problem was introduced by SSA: Bošković and Lasnik's original example had an obligatorily transitive verb in place of *played*, but only half of SSA's tokens preserved that property.)

(30) a. The video showed definitively the suspect to be in the kitchen. (López 2001)
b. The video showed the suspect definitively to be in the kitchen.
(31) a. The lawyers proved decisively the defendant to be innocent.
b. The lawyers proved the defendant decisively to be innocent.
(32) a. We proclaimed in the newspaper Ralph to be generous.
b. We proclaimed Ralph in the newspaper to be generous.

[24] The absence of the complementizer from the first conjunct in (28a) introduces a potentially confounding difference from (b), but I suspect that Bošković and Lasnik felt (28a) would degrade from an asymmetry between its first and second conjunct if the first complementizer had been overt.

These are ECM structures containing an expression that is meant to modify the matrix verb, which in the (a) versions intervenes between that verb and the embedded subject, violating the adjacency requirement on Case assignment. The (b) versions are supposed to be acceptable because the embedded subject has raised to a matrix object Case position. Two subjects gave (30b) very low ratings because they were unfamiliar with the word *definitively*. One subject rated (31b) "1" because *decisively* "wasn't too relevant, …it adds so little, and it's interrupting so much, that it doesn't have to be there." A FC subject pronounced (32b) as if *Ralph in the newspaper* were a DP; since no such parse is available for (32a), this subject was comparing apples and oranges. (Unfortunately, all of these problems arose from attempts to generalize beyond López's original example, where the intervenor was a PP argument.)

4.2.8 Superiority

(33) a. I know that the teacher bought some gifts and I know that she plans to give them to her favorite students, but I don't know to whom she will give what. (Richards 2004)
b. I know that the teacher bought some gifts and I know that she plans to give them to her favorite students, but I don't know what she will give to whom.

Such pairs were meant to show the effect of Superiority: *Wh*-moving the theme while the goal-PP stays in situ should be better than *Wh*-moving the goal while the theme stays in situ, if the theme is higher than the goal. But many subjects rated both sentence types low simply because they did not like the word *whom*, as opposed to *who*. (However, other subjects might well have disliked the choice of *who*; future studies need to include items with both variants.) On the other hand, several other subjects rated the (a) versions "7", potentially raising a challenge to the empirical generalization. However, SSA's and Richards's original examples were matrix questions (*To whom did you give what?*[†]); I worried that out of the blue such multiple *wh*-questions would be so odd for naïve subjects that the results would be meaningless, so I decided to add contextualizing preambles. Unfortunately, in so doing it is possible that I allowed for *whom* and *what* to receive (quasi-)D-linked interpretations (*which students/gifts*), which are known to evade Superiority.

4.2.9 A-movement

(34) a. There is unlikely a fleet of enemy ships to appear. (Hazout 2004)
b. There is unlikely to appear a fleet of enemy ships.
(35) a. There is likely a bookshelf to stand against the wall.
b. There is likely to stand a bookshelf against the wall.

These examples challenge the suggestion that the infinitival complement to a raising predicate has an EPP feature that could drive overt movement to its Spec-TP, deriving the (a) order from the (b) order.[25] But it should be noted that the base (b) orders are themselves rather marked out of context (Hazout's original even more so: *There is likely a man to appear*†). Sentences like (34b) were rated low by many subjects who expressly would have preferred they be worded differently, e.g., "A fleet of enemy ships is unlikely to appear", or "There is a fleet of enemy ships that is unlikely to appear", or "It is unlikely for a fleet of enemy ships to appear". In the FC task, items like (35) yielded responses that highlighted the possibility of an unintended parse for the (a) version, involving a purposive infinitival adjunct and a non-raising use of *likely*, paraphraseable as *There is probably a bookshelf that is (meant) to stand against the wall.* No such parse is available for the (b) variant, so such responses were apples-to-oranges comparisons. (Use of *unlikely* for *likely* could have avoided this issue.)

4.2.10 Pro-forms

(36) a. The politicians said that we should use less gas, but the actual doing of so has proved very challenging. (Haddican 2007)
b. The politicians said that we should use less gas, but the actual doing of it has proved very challenging.
(37) a. Alex said we should take Sunset Blvd, but the actual doing of so was slowed down by heavy traffic.
b. Alex said we should take Sunset Blvd, but the actual doing of it was slowed down by heavy traffic.

The claim here was that *the doing of it* is fine while *the doing of so* is degraded. However, one subject rated (36b) rather low because *proved* should have been *proven*. Another subject rated

[25] Even if it does, examples like (34a) would be expected to be well-formed only if the DP surfacing there can get Case via its relationship with the upstairs expletive *there*.

(37b) low because they wanted to replace *doing of it* with *driving*. (Both of these "irrelevant" responses might have been avoided in the FC task, where *proved* and *doing of* would have appeared in both alternatives.)

5. Conclusions and future directions

The general conclusion from the results presented above should be obvious: judgments collected from naïve subjects in computer-based acceptability experiments cannot be taken at face value, given the way virtually all such experiments are currently conducted. We need to know the reasons behind subjects' responses in order to assess whether they are germane to the linguistic question that the judgments are meant to address, and most extant studies make no attempt to ascertain those reasons. Although this research stemmed from attempts to confirm linguists' published judgments, I believe the conclusion is not restricted to such studies: we should never underestimate subjects' creativity in finding ways of looking at sentences that would not have occurred to us, or in being bothered by aspects of sentences that we find mundane. This is not to deny that there have been judgment experiments focused on narrow theoretical questions where the construction of stimuli was more constrained and careful than in the experiments presented here, allowing concomitantly greater confidence in the results. But the only way to fully dispel concerns of the sort I have raised is to run a version of the experiment in which subjects' reactions can be probed in the ways I have suggested (and perhaps additional ways I have not thought of), seeking to ascertain what structure + interpretation they were judging and, unless the sentence was found unobjectionable, what they disliked about it. We may think of this as "pilot" work, but I would encourage reporting the results of such work, at least for the final version of stimuli. (In the ideal case, this could be done in a single sentence, e.g., "In open-ended discussions, the pilot subjects never reported irrelevant parses/interpretations and never offered

irrelevant reasons for judging the sentences unacceptable."[26]) Then we can be (more) confident that the results of a large-scale acceptability questionnaire administered by computer bear on the linguistic issues we intend them to.

A couple of more specific methodological conclusions suggest themselves. I was pleasantly surprised at how much information could be gleaned simply from having a subject read a stimulus sentence out loud: I highly recommend not skipping this seemingly trivial step. Also, I believe I have demonstrated the value in presenting stimuli as minimal pairs rather than singletons (which, as noted, does not preclude eliciting LS or ME ratings of each sentence): the problems that this solves seem to greatly outnumber the few it may occasionally create. (Of course, all results should eventually be replicated using multiple methods.) As for the use of crowdsourcing platforms such as AMT, the results I have presented should invite the field to think about how we can enrich our interactions with subjects via these platforms. If we are willing to allow more time (and money) to be spent gathering each data point, we can be more expansive both in terms of the information we send (e.g., providing greater guidance on how to

[26] I do not mean to suggest that we generally expect subjects to be able to articulate what is wrong with unacceptable sentences. But from conducting the interviews described above, I believe we can interpret their judgments as being relevant ("for the right reasons") based, e.g., on asking which portion of the sentence they think sounds wrong/strange/etc., how they would change the sentence to make it sound better, and in the case where just a single (purportedly) bad sentence has been presented, subsequently asking for a judgment on the OK counterpart. Of course, the experimenter conducting such an interview must have extensive linguistic background: this is not a task for a first-year research assistant.

read the sentence, for instance by embedding it in a context[27]) and the information we receive (e.g., eliciting not just numerical ratings/rankings but qualitative feedback). Finally, it should be clear that cutting and pasting examples from linguistics articles directly into experimental stimuli will generally yield uninterpretable results. Naïve subjects need stimuli that are exquisitely crafted, controlled, normed, and piloted: this is no less true for testing the empirical claims of linguistic theory than for testing hypotheses about language processing.

Acknowledgements

These experiments would not have been possible without the help of undergraduate research assistant Ethan Chavez, who was heavily involved in all aspects of the design, in addition to running most of the subjects. Thanks to Jon Sprouse, Jesse Harris, and Colin Wilson for discussions about these experiments, an audience at UConn for feedback, and Henry Tehrani for the implementation of Experiment 3. Thanks to Sam Schindler for his patience and encouragement, and to non-anonymous reviewer Ted Gibson for feedback on an earlier draft. This research was supported by a UCLA Academic Senate COR grant.

[27] Ted Gibson suggests that "all of the problems" that I observe in this paper "could be solved by providing an adequate context." I am not so optimistic: I find it unlikely that contexts will reduce the propensity to skip over short words, to give low ratings based on unfamiliar/rare words, etc. More importantly, how could we determine which contexts are "adequate" for removing confounds except by something like the interview method I have described? (If I re-ran the SSA study with contexts and found 100% convergence, which reaction is more likely from skeptics who share the concerns of Gibson and colleagues: a) Great news, linguists have perfectly reliable judgments after all!; or b) I want to understand how those contexts are affecting the way subjects interpret the target sentences?)

References

Bhatt, Rajesh and Roumyana Pancheva (2004). 'Late merger of degree clauses', *Linguistic Inquiry* 35, 1–45.

Bošković, Željko and Howard Lasnik (2003). 'On the distribution of null complementizers', *Linguistic Inquiry* 34, 527–546.

Bruening, Benjamin (2010). 'Ditransitive asymmetries and a theory of idiom formation', *Linguistic Inquiry* 41, 519–562.

Costantini, Francesco (2010). 'On infinitives and floating quantification', *Linguistic Inquiry* 41, 487–496.

Fox, Danny (2002). 'Antecedent-contained deletion and the copy theory of movement', *Linguistic Inquiry* 33, 63–96.

Gibson, Edward and Evelina Fedorenko (2010). 'Weak quantitative standards in linguistics research', *Trends in Cognitive Sciences* 14, 233–234.

Gibson, Edward and Evelina Fedorenko (2013). 'The need for quantitative methods in syntax and semantics research', *Language and Cognitive Processes* 28, 88–124.

Gibson, Edward, Steve Piantadosi, and Kristina Fedorenko (2011). 'Using Mechanical Turk to obtain and analyze English acceptability judgments', *Language and Linguistics Compass* 5, 509–524.

Gibson, Edward, Steven T. Piantadosi, and Evelina Fedorenko (2013). 'Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013)', *Language and Cognitive Processes* 28, 229–240.

Haddican, Bill (2007). 'The structural deficiency of verbal pro-forms', *Linguistic Inquiry* 38, 539–547.

Häussler, Jana and Tom Juzek (2017). 'Hot topics surrounding acceptability judgement tasks', in Sam Featherston, Robin Hörnig, Reinhild Steinberg, Birgit Umbreit, and Jennifer Wallis (eds.), *Proceedings of Linguistic Evidence 2016: Empirical, Theoretical, and Computational Perspectives*. University of Tübingen, https://publikationen.uni-tuebingen.de/xmlui/handle/10900/77066.

Hazout, Ilan (2004). 'The syntax of existential constructions', *Linguistic Inquiry* 35, 393–430.

Landau, Idan (2010). 'The explicit syntax of implicit arguments', *Linguistic Inquiry* 41, 357–388.

Langsford, Steven, Amy Perfors, Andrew T. Hendrickson, Lauren A. Kennedy, and Danielle J. Navarro (2018). 'Quantifying sentence acceptability measures: Reliability, bias, and variability', *Glossa: A journal of general linguistics* 3: 37, 1–34.

Linzen, Tal and Yohei Oseki (2018). 'The reliability of acceptability judgments across languages', *Glossa: A journal of general linguistics* 3: 100, 1–25.

López, Luis (2001). 'On the (non)complementarity of θ-theory and checking theory', *Linguistic Inquiry* 32, 694–716.

Mahowald, Kyle, Peter Graff, Jeremy Hartman, and Edward Gibson (2016). 'SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments', *Language* 92, 619–635.

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

Martin, Roger (2001). 'Null Case and the distribution of PRO', *Linguistic Inquiry* 32, 141–166.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily (2010). 'Crowdsourcing and language studies: The new generation of linguistic data', in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 122–130. Stroudsburg, PA: Association for Computational Linguistics.

Richards, Norvin (2004). 'Against bans on lowering', *Linguistic Inquiry* 35, 453–463.

Schütze, Carson T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press. Reprinted 2016 by Language Science Press, Berlin (Classics in Linguistics #2).

Schütze, Carson T. (2009). 'Web searches should supplement judgements, not supplant them', *Zeitschrift für Sprachwissenschaft* 28, 151–156.

Schütze, Carson T. (2011). 'Linguistic evidence and grammatical theory,' *Wiley Interdisciplinary Reviews: Cognitive Science* 2, 206–221.

Schütze, Carson T. and Jon Sprouse (2013). 'Judgment data', in Robert J. Podesva and Devyani Sharma (eds.), *Research methods in linguistics*, 27–50. New York: Cambridge University Press.

Song, Sanghoun, Jae-Woong Choe, and Eunjeong Oh. (2014). 'FAQ: Do non-linguists share the same intuition as linguists?' *Language Research* 50, 357–386.

Sprouse, Jon and Diogo Almeida (2012). 'Assessing the reliability of textbook data in syntax: Adger's *Core Syntax*', *Journal of Linguistics* 48, 609–652.

Sprouse, Jon, Carson T. Schütze, and Diogo Almeida (2013). 'A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010', *Lingua* 134, 219–248.

Weskott, Thomas and Gisbert Fanselow (2011). 'On the informativity of different measures of linguistic acceptability', *Language* 87, 249–273.