A user's view of the validity of acceptability judgments as evidence for syntactic theories
Jon Sprouse

Abstract
My primary goal in this chapter is to explicitly discuss the validity of acceptability judgments as a data type from the perspective of a user of acceptability judgments. My hope is that such a discussion might help both syntacticians and other language researchers chart a path forward for investigating the validity of acceptability judgments. My view is that acceptability judgments have most, if not all, of the hallmarks of a valid data type. Syntacticians have a plausible theory of the source of acceptability judgments, a theory of how to leverage judgments for the construction of syntactic theories using experimental logic, and a set of evaluation criteria that are similar to those used for other data types in the broader field of psychology. At an empirical level, acceptability judgments have been shown to be relatively reliable across tasks and participants, to be relatively sensitive (at least to syntactic phenomena), and to be relatively free from theoretical bias. Therefore, I would argue that acceptability judgments are at least as valid as other data types that are used in the broader field of language science. That said, I also note in these discussions that most of the current evidence either comes from subjective evaluations of syntactic theories, or experimental studies that have focused primarily on English. Therefore it is possible that future studies could challenge these conclusions.

1. Introduction

Acceptability judgments are ubiquitous in syntactic research, but their use is not without controversy. For some researchers (typically generative syntacticians), acceptability judgments often serve as the primary evidence for the construction of theories. But for other researchers (often language researchers outside of the generative syntax tradition), acceptability judgments are used as pilot data at best, because acceptability judgments are not viewed as valid data for the construction of syntactic theories. At an abstract level, empirical variety is a virtue – researchers should be free to choose the data type that most appropriately addresses their question of interest. But at a practical level, the fact that there is an empirical split between two otherwise closely related groups of researchers (often working on identical or substantially related questions) is stifling for the field. It is clear that it will take work along multiple parallel lines to resolve this division. This volume represents an important step in that direction. In this chapter, I would like to contribute to this endeavor by attempting to lay my view – as a user of acceptability judgments – of the validity of acceptability judgments as explicitly as I can. My hope is that by making this view explicit, and by discussing the areas where we do and do not have empirical results to support this view, this chapter, like the others in this volume, can help facilitate future discussions about, and perhaps even future studies of, the validity of acceptability judgments.

For organizational purposes, I will divide the primary discussion in this chapter into three components. The first is a discussion of the theory underlying acceptability judgments: what their theoretical purpose is, what their (cognitive) source is, how they are used (logically) in theory construction, and how their success (or failure) is evaluated within generative syntax. My hope is to show that users of judgments often have a theory of judgments that would plausibly yield data that is as valid as other data types in language research, and to provide a foil for discussions of alternative theories of judgments. The second component is a discussion of our current state of knowledge about the empirical properties of acceptability judgments: their

reliability, their sensitivity, and their susceptibility to theoretical bias. My hope is to show why it is that users of judgments often believe that they show the empirical properties that one would expect of a useful data type –reliability and sensitivity that rivals (or surpasses) that of other data types in language research. The third component is a discussion of the practical question facing generative syntacticians: should we continue to use acceptability judgments knowing that some language researchers might not accept them as valid evidence (and therefore not accept the resulting theories), or should we adopt other methods that other researchers appear to already accept as valid. My take on this is that, we do not yet have the systematic evidence that we need to make that assessment definitively, as neither group of researchers has done the work that it would require. With the little evidence that we do have, my impression is that there is no scientific reason to prefer acceptability judgments over other sentence processing methods like reading times, eye-movements, or scalp voltages, but that there is some evidence that judgments might be preferred for practical reasons like yielding less variability than other measures for the phenomena of interest to syntacticians.

Before beginning the discussion, two quick disclaimers may be appropriate. The first is there are at least two questions in the literature about the validity of acceptability judgments: (i) the general question of whether judgments are a valid data type, and (ii) the more specific question of how best to collect acceptability judgments. I will focus on the general question of the validity of acceptability judgments in this chapter because I believe it is the more pressing challenge. There is no chance for collaboration between linguists and other cognitive scientists if those cognitive scientists do not believe the data underlying the theory is valid. I will not have much to say about the nitty-gritty details of judgment collection (even though much of the empirical evidence that I will discuss is also relevant for questions about the validity of traditional judgment methods, and has appeared in journal articles focused on that question). My thoughts on this topic are relatively prosaic – I think that researchers should be free to use the method that is most appropriate for their specific research question, which entails making scientific judgments about the various factors that influence data quality. The second disclaimer is that, as the title suggests, this chapter skews more toward an opinion piece than a typical research article. I do not attempt to accurately represent anyone else's position, or anyone else's interpretation of the empirical results, only my own. My hope is that this will help to facilitate future discussions of the validity of acceptability judgments, perhaps by spurring others to make their opinions equally explicit, or by identifying areas where additional empirical studies might be valuable.

2. A theory of acceptability judgments

In this section I will attempt to lay out an explicit theory of acceptability judgments from the perspective of a judgment user, drawing heavily on previous work (e.g., Schütze 1996, Cowart 1997), and my own impressions from the field. I will divide the theory into four components (that are by no means exhaustive): (i) a statement of the goal of syntactic theory, (ii) a proposal for the source of acceptability judgments, (ii) a discussion of the logic that is used to convert judgments into evidence for syntactic theories, and (iv) a discussion of the criteria that are used to evaluate the success or failure of judgments. To my mind, laying this out makes it clear that syntacticians have a theory of acceptability judgments that is at least as well-worked out, and plausibly valid, as the theories underlying other data types in language research.

2.1 What is the goal of syntactic theory?

Generative syntacticians use acceptability judgments to build syntactic theories. The theory of judgments should connect to this goal in directly. To my mind, there are two fundamental assumptions driving generative syntactic theory: (i) that there is an underlying combinatorial math to human syntax, (ii) that this math is (at some level) a description of a cognitive ability. I take the goal of syntactic theory to be the specification of that math in a way that can (one day) be integrated into broader theories of language as a cognitive ability (including a theory of language acquisition, a theory of language processing, a theory of language use, etc). To study this math, Chomsky (1957) argued that syntacticians must first divide word strings into those that are possible in the language, and those that are impossible. Chomsky assumed that the underlying math yielded a binary classification, or two discrete sets of word strings (grammatical and ungrammatical). We now know this is an open empirical question – it is possible that the underlying math could yield more than two sets, or fuzzy membership in two or more sets, or even a truly continuous spectrum of word strings. But the fundamental goal remains the same – to classify word strings (in some way) so that syntacticians can investigate the properties of these word strings that are relevant to the underlying combinatorial math. Chomsky (1957) suggested that acceptability judgments might be a good method (potentially among many) for making this classification.

2.2 What is the cognitive source of acceptability judgments?

If pushed to give a one sentence definition of acceptability judgments, I would probably say something like this: acceptability judgments are the conscious report of the perception of an error signal that arises automatically during the processing of a sentence. There are a number of important claims in that definition. The first is that there is an error signal that arises during sentence processing. I think all syntacticians assume that there are multiple factors that impact that error signal – grammar (phonology, morphology, syntax, semantics, pragmatics, etc), language processing (parsing strategies, working memory, predictive processes), real world knowledge (e.g., plausibility), task effects, etc. I think syntacticians sometimes assume that there is a single unitary error signal that is itself a composite of the multiple factors that influence it, but as Colin Phillips once pointed out (p.c.), that claim has never been tested empirically. It is possible that speakers can distinguish different sources of errors systematically. Nothing in the way that judgments are currently used hinges on this assumption as far as I can tell.

The second claim is that the error signal is generated automatically – it cannot be consciously disengaged. I believe this is a critical assumption for most syntacticians. If the error signal were consciously driven, similar to a learned skill, I believe that syntacticians would be less inclined to use judgments as the foundation of syntactic theory. I know of no explicit research into the automaticity of the error signal underlying judgments. One paradigm that has been used in the ERP literature to investigate automaticity involves repeating a condition over and over to see if the response is suppressed (suggesting it is under some amount of control, though it is not clear if this is conscious control or not) or if it persists (suggesting it is automatic). Hahne and Friederici (1999) showed that repeated exposure to one type of ungrammatical sentence in German leads to the suppression of the P600 response, suggesting it is controlled, but no suppression of the ELAN response, suggesting it is automatic. The judgment satiation literature could potentially be viewed as analogous, but I do not think that the analogy

goes through. In judgment satiation the judgments change after repeated exposure (typically increasing in acceptability); but they are not suppressed. Judgment satiation seems more consistent with the idea that one of the components contributing to the error signal has changed (perhaps even the syntactic component), rather than the idea that the error signal itself has been suppressed.

The third claim is that judgments are a conscious report of a perception (of the error signal). I believe that this is also a critical assumption for syntacticians. Syntacticians, like most cognitive scientists, reject introspection in the Wundtian sense – we do not believe that humans have conscious access to cognitive mechanisms, therefore the claim is not that judgments are a direct report of the syntactic representations or syntactic mechanisms underlying language. Instead, syntacticians, like most other cognitive scientists, believe that humans have conscious access to percepts (like the brightness of light), making it valid to ask participants to report those perceptions, as long as appropriate methodological controls are in place (section 3 below). Schütze 1996 has a terrific discussion of this issue in his seminal book on the topic of acceptability judgments. In his discussion, he distinguishes an introspection-based definition of judgments from a perception-based definition by formulating the research questions that each would answer:

Introspection:	What must be in the minds of participants for the sentence to have the [syntactic] status that they claim it has?
Perception:	What must be in the minds of participants in order for them to react this way to the sentence?

In my experience, the second formulation (perception) more accurately reflects the research questions that syntacticians attempt to answer. However, I do see why some language researchers get the impression that syntacticians are attempting to use the first formulation (introspection). I believe this is an illusion that arises because syntacticians tend to assume a relatively direct mapping between syntactic well-formedness and acceptability judgments. I believe that this is partly due to the assumption that syntactic well-formedness has relatively large effects on acceptability (while other factors have smaller effects; see sections 3 and 4), and partly due to the fact that syntacticians use experimental logic to control for effects from other factors and thereby isolate the effect of syntactic well-formedness – a topic I turn to presently.

2.3 What is the logic that is used to convert acceptability judgments into evidence?

Syntacticians use the same logic that all cognitive scientists use – experimental logic. This is because syntacticians are interested in establishing a causal relationship between one or more syntactic factors and the resulting acceptability judgments. The simplest case of experimental logic is the minimal pair – for syntax, that would be a pair of sentences that share all possible (judgment-affecting) properties except one, the syntactic property of interest. More complex cases, like multi-factorial designs, can be used when it is impossible to hold all of the factors that might influence judgments constant, or when the syntactician is interested in quantifying the effects of multiple properties and their interactions simultaneously. In this way, the only difference between syntax and other domains of cognitive science is in the mechanisms of interest and the data types used to investigate them.

I am aware that the presentation of acceptability judgments in the syntax literature is not always transparent about the use of experimental logic. At times there are sentences that appear in isolation, as if they are not part of any experimental logic. I think there are (at least) two ways that this occurs. The first is through an implicit use of experimental logic. Though the target sentence appears in isolation, there are typically one or more other conditions implicated in the claim being made. These other sentences may occur explicitly in earlier passages of the text (due to the flow of the scientific narrative), or may be implicit, with the author assuming that other syntacticians can generate the relevant conditions from the theory itself. I admit that the style of presentation in syntactic articles can be opaque in this way. It may be a consequence of the sheer scope of syntax articles – each article typically contains a large number of data points, and explores a relatively large number of hypotheses. This in turn may be a consequence of the ease with which judgments can be collected, making it feasible to test a large number of hypotheses in one project. The second way that sentences can appear in isolation is if they are not acting as evidence in support of a syntactic theory. For example, judgments can be used as a criterion to identify constructions that might warrant additional study, perhaps because they are below a certain threshold in acceptability. Though identifying potential phenomena is undoubtedly a critical part of the scientific enterprise, it is logically distinct from gathering evidence in support of a theory. Once a sentence is identified as warranting additional study, syntacticians will invariably use experimental logic to identify the cause of the low acceptability, either explicitly or implicitly.

I have, in conversation but not print, encountered syntacticians who wonder to what extent there is a logical connection between the type of syntactic theory that one assumes and the type of evidence that one is able to use. The specific question seems to be whether binary syntactic theories, which divide strings into two types (grammatical and ungrammatical), might be able to use judgments of single sentences, whereas gradient theories of syntax, which divide strings into more than two types (either some finite number, or an infinite number), might somehow be more amenable to the comparison of multiple sentences. I do not see how binary syntactic theories can make use of the acceptability of standalone sentences (as discussed above); nor do I see how gradient syntactic theories could make better use of experimental logic than binary syntactic theories. The question of whether syntactic theory distinguishes two or more types of strings appears to me to be orthogonal to the question of how to build causal theories from acceptability judgments. It is, of course, true that the two types of theories have different mechanisms available to them to explain the fact that acceptability itself is a continuous measure: binary theories must rely on extra-syntactic factors to explain the gradience of acceptability, whereas gradient theories can explain the gradience directly with syntactic mechanism (in addition to extra-syntactic factors). But again, this empirical issue appears orthogonal to the question of the logic that syntacticians can use to identify those mechanisms. To my mind, that is always experimental logic.

2.4 What are the criteria for evaluating the success (or failure) of acceptability judgments?

The question of how to evaluate the success (or failure) of acceptability judgments is intimately tied (if not identical to) the question of validity from psychometrics. A test is valid if it measures what it is intended to measure. Though validity is rarely discussed explicitly in the syntax literature, it is my impression that syntacticians do critically evaluate the success of acceptability judgments as a data type. It is also my impression that syntacticians evaluate acceptability

judgments using the same criteria that other language researchers use for other data types. These criteria are indirect. It is not currently possible to directly measure the error signal that gives rise to acceptability judgments, the same way that it is not possible to measure the processing mechanisms that underlie reading times, or the neural computations that underlie scalp potentials. No data type in language science is held to that kind of direct validity requirement (and, in fact, it would defeat the purpose of having these data types, since at that point, we could just measure the underlying cognitive mechanism directly). Instead, the criteria that syntacticians and other language researchers use build on the logic that a valid data type will have certain properties, and invalid data types will not. In this section I will mention three criteria that syntacticians appear to use to evaluate the validity of acceptability judgments.

The first, and perhaps most important, criterion is that acceptability judgments can be used to create internally consistent syntactic theories. Syntactic theories make predictions about one phenomenon based on the mechanisms proposed for another phenomenon. They succeed in explaining multiple phenomena with a relatively small number of theoretical constructs. They make predictions about the space of cross-linguistic variation. They interact with other domains of language like phonology, morphology, semantics, acquisition, and sentence processing in exactly the way that one would expect of a syntactic theory. We would not expect this kind of internal consistency if syntactic theories were built on random, or unrelated, data. This is not to say that there are not debates about the details of syntactic theories. And this is not to say that the theories constructed from acceptability judgments are necessarily theories of syntax; it is logically possible that all of these properties could hold in a world where acceptability judgments do not provide evidence about syntax at all, but rather provide evidence about some other cognitive system. Nonetheless these properties increase the probability that acceptability judgments are providing meaningful information about syntax. It is my impression that this is precisely the same criterion used for other measures in language science. Reading times, eye-movements, and scalp potentials are considered valid measures of comprehension processes because the resulting theory has the properties we would expect of a theory of comprehension processes.

A second criterion is the fact that acceptability judgments generally correlate with other measures, like sentence processing measures, corpus/production measures, and language acquisition measures. To be clear, there is a nuance to this correlation. It is not the case that syntacticians believe that acceptability judgments perfectly correlate with these other measures (otherwise, syntacticians could simply use these other measures directly; or psycholinguists and acquisitionists could use acceptability judgments directly). Nor do syntacticians claim that the relationship between acceptability judgments and these measures will be simple. Nonetheless, there is quite a bit of evidence spanning these literatures that there is a general correlation. As Phillips 2009 points out, we can even see it in the types of research that the field publishes. It is not a publishable research result to find that a sentence with low acceptability also yields a reading time effect, or an ERP effect, or has low frequency in a corpus. The publishable result is when we find the opposite – a sentence has low acceptability but no sentence processing effects or high frequency. This suggests that the field has accepted the general correlation between acceptability judgments and other measures as the most probable configuration of the language faculty. Of course, given this general correlation, one could ask whether there would be value in syntacticians adopting other data types in addition to, or perhaps instead of, acceptability judgments. That is a question that I will turn to in section 4.

There is a third criterion that is worth mentioning, if for no other reason than it receives relatively little discussion in the language research literature, and that is face validity. Face validity is when a measure appears, on its face, to measure the property of interest. Face validity is probably one of the weaker criteria for validity, as one could imagine measures with high face validity that are not ultimately valid, and measures with low face validity that are ultimately valid. But face validity is also a core component of most measures in language research, because researchers typically invent tasks that seem as though they will give the information that the researcher wants. Sometimes the task works, and sometimes it doesn't. The evaluation of the task usually involves other criteria, but the first criterion is almost always face validity. As such, it is perhaps unsurprising that most measures in language research, including acceptability judgments, have high face validity.

3. The empirical properties of acceptability judgments

Whereas the previous section focused on the fundamental assumptions that form the theory of acceptability judgments, which are typically only testable indirectly, this section focuses on three empirical properties that we would expect from a valid data type that are directly testable: reliability, independence of bias, and sensitivity.

3.1 The reliability of acceptability judgments

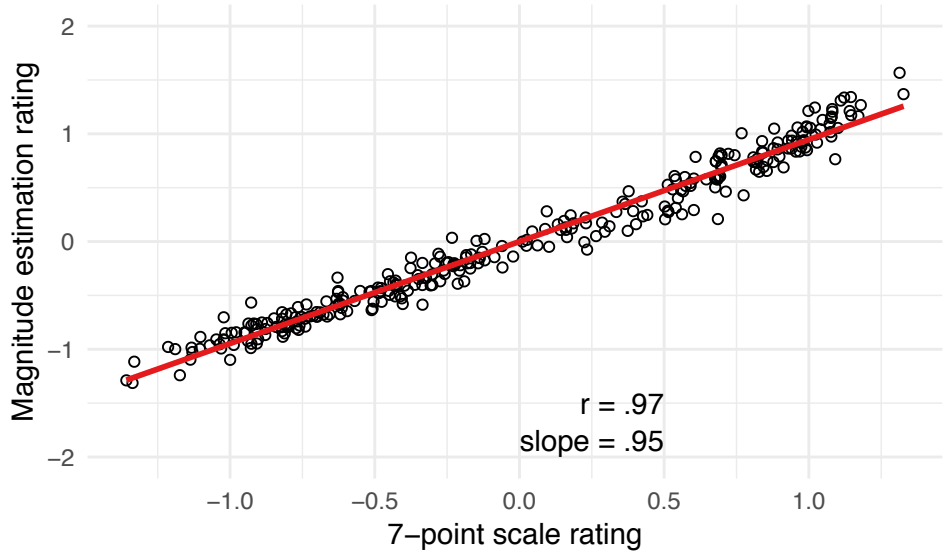
Reliability is the propensity to yield the same results when repeated under the same conditions. One of the most frequent empirical questions about acceptability judgments raised in the literature is to what extent they are reliable. I take this to indicate that most language researchers expect relatively high reliability from measures. The specific concerns about the reliability of acceptability judgments that the relatively informal collection methods that are typical in syntax might lead to unreliability, because the informal methods may be contaminated by confounds of various sorts, such as theoretical bias if professional linguists are used as participants, the outsized contribution of specific lexical items if relatively few tokens of each condition are tested, the distortion of judgments if there are no fillers to mask the theoretical goal of the experiment, or the misinterpretation of results if inferential statistics are not used as part of the data analysis (e.g., Edelman and Christiansen 2003, Ferreira 2005, Wasow and Arnold 2005, Featherston 2007, Gibson and Fedorenko 2013, and Brannigan and Pickering 2017). These concerns are also often accompanied by the claim that the more formal collection methods that are typically used in psycholinguistics would not suffer from these confounds, and would therefore lead to higher reliability. As such, this concern is not about the reliability of judgments in general, but rather a concern about the reliability of the set of informally collected judgments that have been published in the literature (and, by extension, a concern about the theories that have been constructed from those judgments).

There has been quite a bit of recent work exploring the reliability of acceptability judgments, at least in English. Two of my own studies, Sprouse and Almeida 2012 and Sprouse et al. 2013, investigate the impact of informal collection methods by re-testing two large sets of informally collected judgments using formal methods. Sprouse and Almeida 2012 re-tested all of the data points from Adger's 2003 syntax textbook *Core Syntax*, and found a replication rate of 98%-100%, depending on the definition of replication. Sprouse et al. 2013 re-tested a random sample of 300 data points forming 150 two-condition phenomena taken from the journal

Linguistic Inquiry (2001-2010), and found a replication rate of 88%-99%, depending on the judgment task and the definition of replication, with a margin of error of ±5 for the full population of data points in the journal over that time period (because the sample was random). The *Linguistic Inquiry* results were replicated directly in Sprouse et al. 2013, and then replicated again by Mahowald et al. 2016 using a different sample of data points from the same time period of the journal. The interpretation of these replication rates is subjective. Speaking for myself, I find these replication rates to be exceedingly high (compare Open Science Collaboration 2015 for estimated replication rates in other areas of experimental psychology in the range of 36%-53% using similar definitions of replication). Therefore, to my mind, these results suggest that the differences between informal and formal judgment collection methods have relatively little impact on the resulting acceptability judgments. This suggests that acceptability judgments are remarkably reliable, at least for English.

Sprouse et al. 2013 also yields information about between-task reliability, as we tested three distinct tasks: a 7-point scale task, the magnitude estimation task, and a two-alternative forced-choice task where participants selected the more acceptable of a pair of conditions. Figure 1 below shows the (between-subjects) correlation between the judgments of the 300 individual conditions using the 7-point scale and magnitude estimation tasks. The correlation is nearly perfect.

Figure 1: Correlation between acceptability judgments using the 7-point scale and magnitude estimation tasks for the 300 sentence types randomly sampled from *Linguistic Inquiry* (2001-2010) by Sprouse et al. 2013. The ratings are z-score transformed.



Langsford, Perfors, Hendrickson, Kennedy, and Navarro 2018 tested both between-task reliability and within-participant (test-retest) reliability using a subset of the materials from Sprouse et al. 2013, adding in a yes-no task, and a forced-choice task based on the Thurstone method, which tests random pairs of sentences rather than the theoretically-constrained pairs in the Sprouse et al. 2013 test. Langsford et al. report impressively high rates of reliability for both between-task and within-participant reliability.

Taken together, all of these studies suggest that judgments are likely fundamentally reliable – they appear to be reliable across both informal and formal collection methods, across

samples of participants and items, across judgment tasks, and across time within the same participants. The primary limitation of these findings is that they have focused almost exclusively on English, leaving open the possibility that reliability may vary by language. There are a number of large-scale studies in progress by a number of research teams in other languages that may address this question in the near future.

3.2 Theoretical bias in acceptability judgments

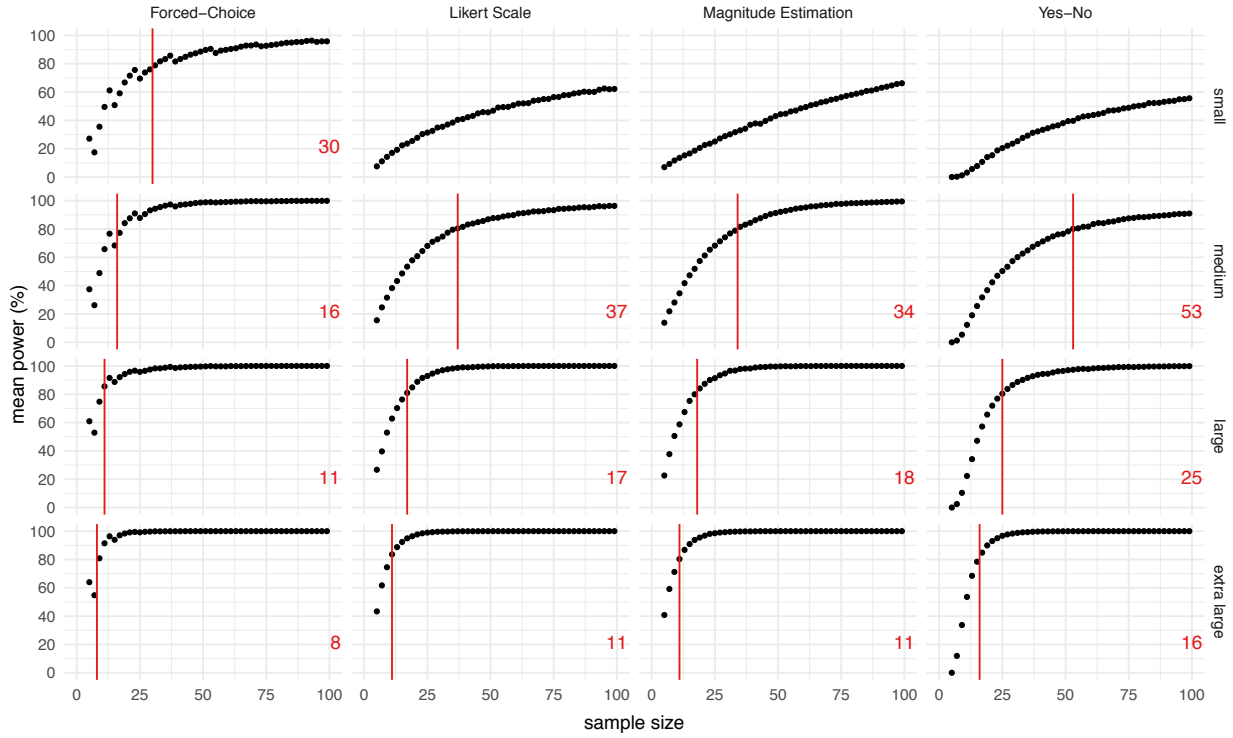
One of the potential confounds frequently mentioned in discussions of informal collection methods is the fact that syntacticians often report their own judgments, and the judgments of their students and colleagues. Given that linguists have the potential to recognize the experimental manipulation, and perhaps even the hypotheses under consideration, this raises the possibility of theoretical bias contaminating the reported judgments. To be clear, the concern is typically not that linguists will purposefully report judgments that confirm (or disconfirm) their theoretical beliefs, but rather that linguists' theoretical knowledge will subconsciously influence their judgments. The reliability results reported in the previous subsection put a potential upper-bound on the effect of this bias: 0-2% for the Adger textbook data set, and 1-12% for the *Linguistic Inquiry* data set. We can also look more closely at those results. We can ask what sorts of replication failures we would expect to find if theoretical bias were present in the judgments in the syntax literature. One possible prediction is that we would expect to find sign reversals: a change in direction of the effect between the informally collected judgments and the formally collected judgments. To be clear, sign reversals can arise for reasons other than theoretical bias (e.g., low statistical power in either the informal or formal experiments; see Sprouse and Almeida 2017 for a mathematical discussion of that). But the prediction here is that theoretical bias could be one generator of sign reversals between experiments involving professional linguists and experiments involving naïve participants, therefore the presence of sign reversals would be a potential indicator of theoretical bias. In Sprouse et al.'s 7-point scale results, there were 2 statistically significant sign reversals, 9 null results, and 137 statistically significant replications (2 of the 150 phenomena were not analyzed because of errors in the experimental materials). In the forced-choice results, there were 3 statistically significant sign reversals, 6 null results, and 139 statistically significant replications. Thus it seems that not only are there relatively few replication failures in the *Linguistic Inquiry* data set (6-7% of the sample), but within those replication failures there are very few sign reversals (1-2% of the sample). In short, there is very little evidence for theoretical bias in this data set.

3.3 The sensitivity of acceptability judgments

Another expectation for valid data types is that they will be sensitive to the phenomena that they are intended to measure. Sprouse and Almeida 2017 provide some information about the sensitivity of acceptability judgments to the syntactic phenomena: they estimated the statistical power (the ability to detect an effect when an effect is truly present) of four judgment tasks (7-point scale, magnitude estimation, forced-choice, and yes-no; the rows of Figure 2), for 50 of the phenomena from the *Linguistic Inquiry* data set, by running 1000 re-sampling simulations for each sample size from 5 to 100 participants (the x-axis of Figure 2), and calculating the percentage of results that reached statistical significance in those 1000 simulations (the y-axis of Figure 2). The results are summarized in Figure 2 below, organized by effect size as measured

using Cohen's *d*, a standardized effect size measure, and classified by Cohen's (1988) criteria for small (*d*=.2), medium (*d*=.5), and large (*d*=.8) effect sizes (the rows in Figure 2).

Figure 2: Empirically estimated power relationships from Sprouse and Almeida 2017, arranged by task (columns) and effect size (rows), and using null hypothesis tests. The x-axis is sample size, and the y-axis is estimated power (based on 1000 re-sampling simulations for each sample size). The vertical line and number indicates the sample size at which 80% power is first reached.

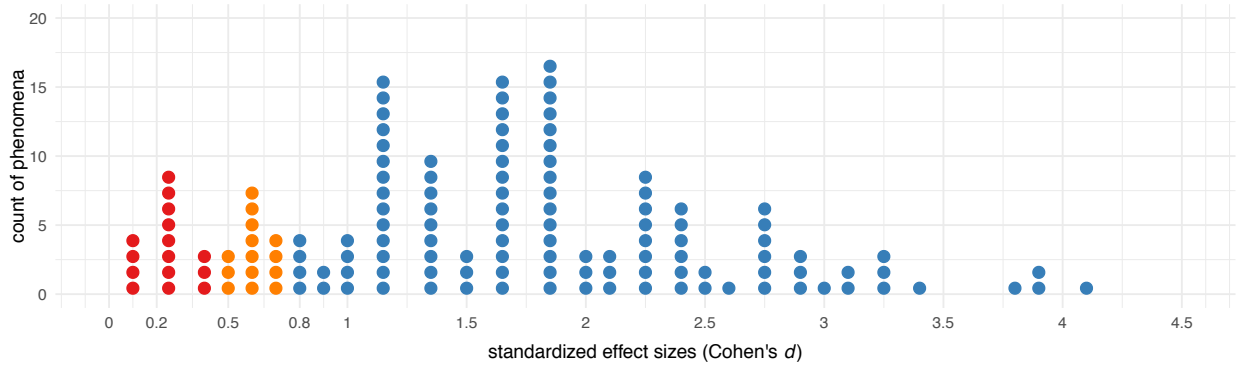


The general trend that emerges in Figure 2 is that judgment tasks are remarkably sensitive when it comes to syntactic phenomena. The forced-choice task is the most sensitive, reaching 80% power for small effect sizes (using null hypothesis tests) at 30 participants, medium effect sizes at 16 participants, and large effect sizes at 11 participants.

As with the results in the previous two sections, the interpretation of these results is subjective. To put these results into context, one possibility would be to compare the sensitivity of acceptability judgments to other data types in the broader field of experimental psychology. Unfortunately, the two fields have thus far focused on different measures of the statistical power in their respective fields. For acceptability judgments, we have the power curves for specific tasks across a range of effect sizes and sample sizes, but no measure of the power of the published studies in the literature (because sample sizes and tasks are rarely reported for informally collected judgments, making such calculations impossible). In the broader field of psychology, we have measures of the power of the published studies in the literature (see Ioannidis and Szucs 2017 for a recent study, and a review of previous studies), but we do not have power curves for specific tasks (because there are so many distinct tasks in the field, it would likely be impractical to measure a substantial number of them). For now, we can say two things. The first is that acceptability judgment tasks reach Cohen's (1988) suggested target level of power of 80% for syntactic phenomena with relatively reasonable sample sizes, particularly

for effect sizes that are medium or larger. The second is that the syntactic phenomena that have been published in *Linguistic Inquiry* tend to be medium or larger. As Figure 3 shows, 87% of the phenomena randomly sampled from *Linguistic Inquiry* have a Cohen's *d* greater than .5. This suggests that for the vast majority of syntactic phenomena in the current literature, acceptability judgment tasks can reach good power with a reasonable sample size.

Figure 3: The count of phenomena based on standardized effect size (Cohen's *d*) for the statistically significant replications from the 7-point scale task from Sprouse et al. 2013. The dots are colored based on the thresholds in Cohen's (1988) suggestions for the interpretation of effect sizes: red is smaller than the medium threshold (<.5), orange is smaller than the large threshold (<.8), and blue is equal to or larger than the large threshold (≥.8).



We do not currently have any systematic information about the sensitivity of judgment tasks to non-syntactic phenomena (processing, frequency, plausibility, task effects, etc.). The critical issue is that we do not know either the size of these effects, or the amount of variability in judgments to them. One might wonder why syntacticians, who presumably are not interested in studying non-syntactic phenomena, should care about the sensitivity of judgments to non-syntactic phenomena. My impression is that syntacticians often construct acceptability judgment experiments as if they believe that syntactic properties have a larger effect on judgments, while non-syntactic properties (processing complexity, frequency, etc) have a smaller effect on judgments. One way this arises is that syntacticians typically explicitly control for known syntactic factors, and also factors from other areas of grammatical theory (phonology, morphology, semantics, pragmatics) when designing their informal judgment studies, but do not always control for factors that are more traditionally part of psycholinguistics, like processing complexity, frequency of words or constructions, and task effects. I don't want to give the impression that syntacticians completely ignore these issues, just that there is a general trend for the two literatures, syntax and psycholinguistics, to focus on potential confounding factors that are more central to their respective theories. My impression is that debates in the literature about how well-controlled acceptability judgment experiments are typically hinge on the factors that are being controlled, not whether control in general is being applied. A systematic study of the size of non-syntactic effects would help to resolve this issue. A second way that this belief appears to arise in the field is that syntacticians sometimes mention the possibility of using effect size as a heuristic for identifying potential syntactic effects, with larger effect sizes indicating a potential syntactic effect (or perhaps a grammatical effect), and smaller effect sizes indicating a potential language processing effect. I do not believe syntacticians would use this as evidence

toward a theory, but, as a heuristic for identifying phenomena to study in more detail, it is appealing. We are not in a position to evaluate the viability of this heuristic – we do not have systematic information on the effect sizes of non-syntactic effects. (But we do have evidence that some of the effects that syntacticians appear to care about are small, so the heuristic cannot be used as an absolute criterion).

4. The choice to continue to use acceptability judgments

For this final section, I would like to ask a difficult question – At what point should syntacticians decide to abandon acceptability judgments in favor of other data types? As the previous sections have made clear, I do not personally believe that this is necessary. But I can imagine that some language researchers may remain unconvinced, perhaps for empirical reasons, or perhaps for reasons relating to their own assumptions about the source of acceptability judgments. Even if syntacticians believe that acceptability judgments are valid, if this disagreement prevents the dissemination of results, or the collaboration among researchers from otherwise allied fields, one might ask whether there could be practical value in adopting other methods, either substantially or completely. In this section I would like to take this question seriously. There are two ways in which it would make sense to switch methods – if there were a scientific reason, or if there were a practical reason.

4.1 The scientific question

Can other data types provide the type of information that syntacticians need to construct and evaluate syntactic theories? I think the answer is unequivocally yes. Syntacticians already appear to believe that the automatic error signal that gives rise to acceptability judgments also impacts other comprehension measures such as reading times, eye-movements, scalp potentials, and hemodynamic responses. This is part and parcel of the argument for predictive validity – judgments tend to correlate with effects in these other measures. Therefore, in principle, syntacticians could simply use these other measures instead of acceptability judgments.

Given that syntacticians likely believe that other measures could provide evidence for syntactic theories, one might wonder if there are scientific reasons why they haven't adopted the other measures to a greater degree. The issue, in my opinion, is that these methods will require specifying a relatively detailed processing theory that can be combined with a syntactic theory to make predictions about these other measures. Since syntacticians are not primarily interested in sentence processing theories, it would be more desirable to have a measure that can provide information about syntactic theories without requiring the specification of a detailed processing theory. While I am sympathetic to this issue (see especially Stabler 1991 for a discussion of this, and some caveats about how difficult it may be to derive syntactic predictions from processing theories), I do worry that this concern may be overstating the difference between judgments and other data types in this regard. Judgments do require a processing theory, because judgments are a type of processing data. It only appears that judgments are different because syntacticians do not typically discuss the processing theory explicitly in work using acceptability judgments. This is possible because judgments only provide one measure, at one time point (typically at the end of the presentation of the sentence), rather than word-by-word (or millisecond-by-millisecond) measures. But full sentence processing must still be accounted for in the experimental logic. Syntacticians can infer that the end-of-sentence judgment reflects a syntactic effect and not some

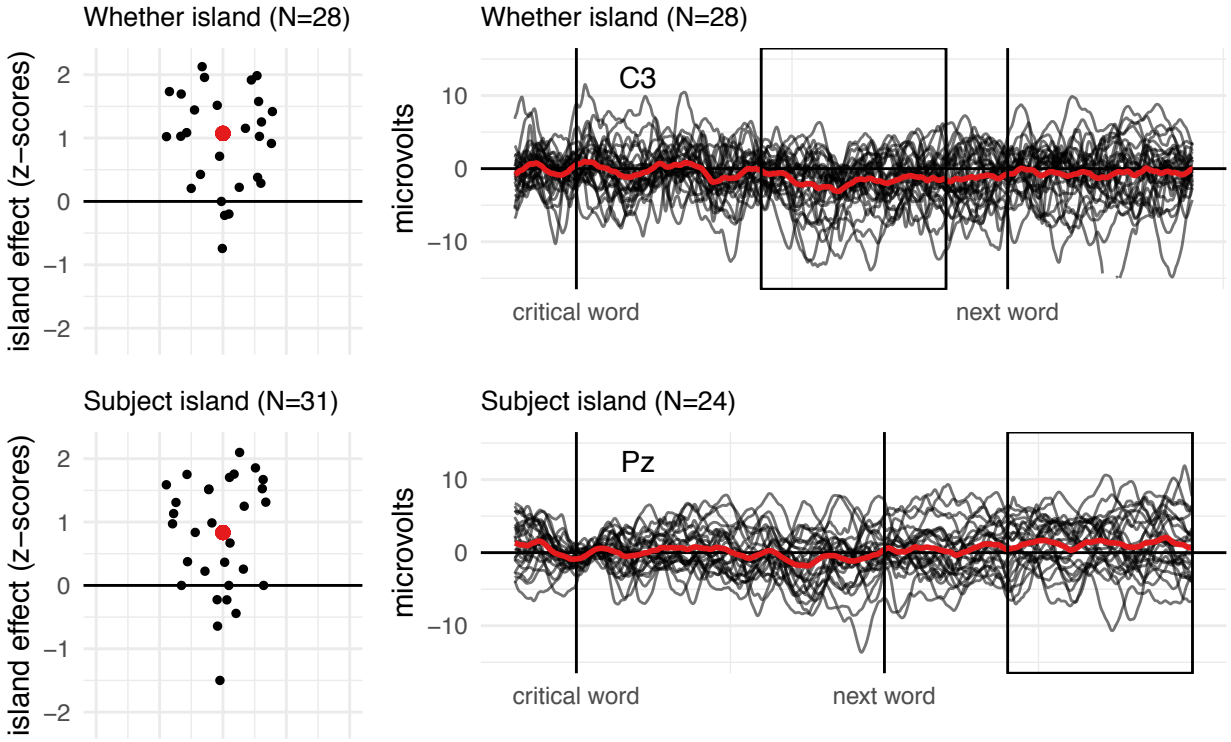
other sentence processing effect because they designed the experiment to manipulate syntax and hold other properties constant. Similarly, syntacticians avoid having to specify a precise temporal prediction for the error signal by assuming that error signals that arise at points during sentence processing will impact judgments made at the end of the sentence. The fact that syntactic theories are as successful as they are (section 2), and that judgments are as reliable and sensitive as they are (section 3), suggests that that these assumptions are substantially correct. There appears to be no scientific reason for syntacticians to abandon acceptability judgments completely.

Though my personal belief is that there no scientific reason for syntacticians to abandon judgments, there are certainly scientific reasons for some subset of syntacticians to explore other measures in service of syntactic theory. To my mind this has never been in doubt – psycholinguists and experimental linguists have been exploring the link between syntax and sentence processing since the earliest days of the field. To be clear, it is not the case that all syntacticians must do this – it seems healthy for the field for different researchers to specialize in different aspects of the syntactic enterprise based on their own personal interests.

4.2 The practical question

Do the potential benefits of switching to other data types (in terms of encouraging collaboration across the broader field of language science) outweigh the potential practical costs? I think the answer is that we do not currently have enough information to make a definitive assessment. The little bit of information that we do have suggests that there would be quite a number of practical costs involved in such a switch. I think this helps to explain why syntacticians have generally not switched away from acceptability judgments as a response to concerns about the dissemination of results and possibility of collaboration. On the one side of the equation, we know that judgments are much cheaper, and much easier to collect, than most (if not all) of the other data types in sentence processing. There is typically no need for special equipment or consumables, and it is typically possible to investigate a relatively large number of conditions at once. On the other side of the equation, we have some hints that sentence processing measures like EEG may ultimately be noisier than judgments when it comes to detecting the effects that syntacticians are interested in. For example, Figure 4 compares *whether*-island effects (*What do you wonder whether Mary read__?) and subject island effects (*What did the advertisement for __ interrupt the game?) using both acceptability judgments and ERPs. For each method, I have plotted the effect for each participant: for judgments, it is the difference in acceptability between the island effect sentence and a control sentence; and for ERPs, it is the difference in scalp voltage at the critical word (for *whether*-islands it is the embedded subject, e.g., *Mary*; and for subject islands it is the verb, e.g., *interrupt*). The black points and lines each represent a participant; the red points and lines represent the mean of the participants.

Figure 4: A comparison of by-participant effects for two island effects (*whether* and subject islands), for both acceptability judgments (left panels) and event-related potentials (right panels). By-participant effects are in black; grand means are in red. The acceptability judgment effects are jittered roughly according to the probability density of the effects (a sina plot). The boxes in the ERP plots indicate the time-window of the significant effect: a negativity for *whether*-islands (Kluender and Kutas 1993), and a positivity for subject islands (Neville et al. 1991).



For acceptability judgments, all but a few participants show a clear positive effect. It is quite easy to see the effect, even without calculating a grand mean, and without using statistical tests. For ERPs, the situation is quite different. It is not as easy to see a clear effect without calculating a grand mean, or without using statistical tests. Crucially, all four samples have roughly the same number of participants. The issue here is an inherent property of these measures: acceptability judgments have relatively less variability, whereas ERPs have relatively more variability (at both the trial and participant level; see Luck 2014 for an introduction to ERPs that describes some of the physiological reasons for this).

It may be possible to comb through the sentence processing literature and extract information about effect sizes and variability for a number of syntactic effects (agreement violations, case violations, phrase structure violations, etc). However, combing the existing literature will only get us so far, because only a subset of effects of interest to syntacticians have been tested using sentence processing methods to date. The obvious next step would be to test a larger number of syntactic effects using sentence processing measures. Combined with the projects mentioned in previous sections, the end result would be a systematic 2x2 investigation of effects and methods: a test of both syntactic and non-syntactic effects using both acceptability judgments and other sentence processing methods. With that information we may be in a

position to definitively assess the cost/benefit ratio of each of the methods for both of the sets of phenomena. Without that information, we must either rely on the impressions of individual researchers based on the small amount of data that we do have, or rely on the broader evaluation metrics that were discussed in section 2.

5. Conclusion

My primary goal in this chapter was to explicitly discuss the validity of acceptability judgments from the perspective of a user of acceptability judgments, in the hope that such a discussion might help both syntacticians and other language researchers chart a path forward for investigating the validity of acceptability judgments. Along the way I have also tried to make my current personal opinion explicit as well – I believe that acceptability judgments have most, if not all, of the hallmarks of a valid data type. Syntacticians have a plausible theory of the source of acceptability judgments, a theory of how to leverage judgments for the construction of syntactic theories using experimental logic, and a set of evaluation criteria that are similar to those used for other data types in the broader field of psychology. At an empirical level, acceptability judgments have been shown to be relatively reliable across both tasks and participants, to be relatively sensitive (at least to syntactic phenomena), and to be relatively free from theoretical bias. As the facts currently stand, I would argue that acceptability judgments are at least as valid as other data types that are used in the broader field of language science. That said, I have also noted that most of our evidence either comes from subjective evaluations of syntactic theories (section 2), or experimental studies that have focused primarily on English (section 3). Therefore it is possible that future evaluations or future experimental studies could challenge these conclusions. I have also argued that there is no scientific reason to prefer acceptability judgments over other data types, therefore the general choice in the field to use judgments over other sentence processing measures appears to be a purely practical one. Judgments are unquestionably cheaper and easier to deploy, and there is some (admittedly limited) evidence that acceptability judgments involve less noise, and therefore yield larger effect sizes, for syntactic phenomena than other sentence processing measures do. But we do not yet have the full set of data that we would need to determine the optimal practical choice(s) – a systematic (2x2) study of both syntactic and non-syntactic phenomena using both acceptability judgments and other sentence processing methods.

References

Adger, David. 2003. Core syntax: A minimalist approach. Oxford University Press.

Branigan, Holly P., and Martin J. Pickering. 2017. An experimental approach to linguistic representation. Behavioral and Brain Sciences 40: E282

Chomsky, Noam. 1957. Syntactic structures. The Hague: Mouton.

Cohen, Jacob. 1988. Statistical power analysis for the behavioral sciences, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Cowart, Wayne. 1997. Experimental syntax. Sage.

Edelman, Shimon, and Christiansen, Morten H. 2003. How seriously should we take Minimalist syntax? Trends in Cognitive Science 7: 60–61.

Featherston, Sam. 2007. Data in generative grammar: The stick and the carrot. Theoretical linguistics 33: 269-318.

Ferreira, Fernanda. 2005. Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review* 22: 365-380.

Gibson, Edward, and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. Language and Cognitive Processes 28: 88-124.

Hahne, Anja, & Friederici, Angela D. 1999. Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. Journal of Cognitive Neuroscience 11: 194–205.

Szucs Denes, Ioannidis John P. A. 2017. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biology 15(3): e2000797. https://doi.org/10.1371/journal.pbio.2000797.

Kluender, Robert, and Marta Kutas. 1993. Subjacency as a processing phenomenon. Language and cognitive processes 8: 573-633.

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.

Luck, Steven J. 2014. An introduction to the event-related potential technique. Cambridge, MA: MIT Press.

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.

Neville, Helen, Janet L. Nicol, Andrew Barss, Kenneth I. Forster, and Merrill F. Garrett. 1991. Syntactically based sentence processing classes: Evidence from event-related brain potentials. Journal of Cognitive Neuroscience 3: 151-165.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science 349: 943.

Phillips, Colin. 2009. Should we impeach armchair linguists. Japanese/Korean Linguistics 17: 49-64.

Schütze, Carson T. 1996. The empirical base of linguistics: Grammaticality Judgments and Linguistic Methodology. Chicago: The University of Chicago Press

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.

Sprouse, Jon, and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. Glossa 2: 1.

Stabler Edward P. 1991. Avoid the Pedestrian's Paradox. In: Robert C. Berwick, Steven P. Abney, Carol Tenny (eds). Principle-Based Parsing. Studies in Linguistics and Philosophy, vol 44: 199-237. Springer, Dordrecht.

Wasow, Thomas, and Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115: 1481-1496.