Research Methods in Linguistics

Chapter 3: Judgment Data

Carson T. Schütze
University of California, Los Angeles

Jon Sprouse
University of California, Irvine

REVISION, May 2012

1. Preliminaries

This chapter covers what have been traditionally called grammaticality judgments in linguistics (which are more aptly referred to as acceptability judgments—see below). We examine such judgments from several angles, with the goal of assisting researchers in deciding whether and how to use this kind of data. Our goal in this chapter is to provide an introduction to the major themes that arise when using acceptability judgments as a data source for the construction of linguistic theories. Importantly, this chapter will not be a step-by-step guide for constructing a particular experiment, as the curious reader can find several fine introductions to the mechanics of experiment construction and analysis elsewhere (e.g., Chapters 7 and 8, and Cowart 1997). Instead, we intend this chapter to be an introduction to the theory underlying the methodology of acceptability judgment collection. Most of what follows will involve discussion of syntactic well-formedness judgments, because that is where the greatest amount of research about judgment data has been focused, but we believe that many of our remarks are also relevant for judgments at other levels of linguistic representation. Specific considerations regarding other sorts of judgments can be found elsewhere in this volume. For example, judgments about the lexicon and phonotactic well-formedness are generally gathered in the language documentation process (see Chapter 4); judgments about morphological processes might be gathered using the experimental methods that predominate in psycholinguistics (see Chapter 8); judgments about sociolinguistic variables might be gathered via a survey (see Chapter 6). For considerations specific to semantic judgments, see Matthewson (2004) and Chemla and Spector (2011).

This first section is comprised of issues that researchers should consider in deciding whether to use judgment data and how to collect it in general. The subsequent three sections look in more detail at issues of choice of task (section 2), experimental design (section 3), and data interpretation (section 4). A brief conclusion completes the chapter.

1.1 The nature of judgment data

Speakers' reactions to sentences have traditionally been referred to as grammaticality judgments, but this term is misleading. Since a grammar is a mental construct not accessible to conscious awareness, speakers cannot have any impressions about the status of a sentence with respect to that grammar; rather, in Chomsky's (1965) terms one should say their reactions concern *acceptability*, that is, the extent to which the sentence sounds "good" or "bad" to them. Acceptability judgments (as we refer to them henceforth) involve explicitly asking speakers to "judge" (i.e., report their spontaneous reaction concerning) whether a particular string of words

is a possible utterance of their language, with an intended interpretation either implied or explicitly stated. The primary assumption underlying acceptability judgment experiments is that *acceptability* is a *percept* that arises (spontaneously) in response to linguistic stimuli that closely resemble sentences (i.e., strings of words). Acceptability is just like other percepts (e.g., brightness, loudness, temperature, pain) in that there are no methods for directly measuring the percept as it exists within a participant's mind. Instead, experimenters must rely on indirect measurement methods. One common method in the study of perception is to ask participants to *report* their perceptions along some sort of scale (e.g., Stevens 1956, 1957). In this way, an acceptability judgment is in fact a *reported perception of acceptability* (Chomsky 1965; Schütze 1996; Sprouse and Almeida 2012). As with all reported perceptions, acceptability judgments are a type of behavioral response that requires a (likely cognitive) explanation. Similarly, acceptability judgments can be used as evidence for making inferences about the cognitive systems that give rise to them, which syntacticians assume includes the grammatical system of the human language faculty (among other cognitive systems).

It has sometimes been suggested that claims made on the basis of acceptability judgment data do not necessarily bear on how the human language faculty is actually constructed unless their "psychological reality" has been tested via some experimental procedure using another dependent measure such as time, error rate, electrophysiological response, etc. (Edelman and Christiansen 2003). This view belies a misunderstanding (Dresher 1995): acceptability judgments are themselves data about human behavior and cognition that need to be accounted for; they are not intrinsically less informative than, say, reaction time measures — in fact, many linguists would argue that they are more informative for the purposes of investigating the grammatical system. The use of the term "psychological reality" in this sense seems to be vacuous, as both acceptability judgments and other behavioral and electrophysiological responses are behaviors that can bear on the cognitive systems that subserve language.

Another objection to judgment data is that they demand awareness of language as an object of attention and evaluation, i.e., metalinguistic awareness. This is claimed to make them artificial and undermine their external validity (e.g., Bresnan 2007). At one level, this is certainly true: reported perceptions require the participant to be aware of their perception and consciously report it using the responses made available to them. However, reported perceptions have long been considered a valid data type for the construction of cognitive theories because reported perceptions tend to be systematic in ways that can lead to the construction of falsifiable theories (e.g., Stevens 1957). This is no less true of acceptability judgments (reported perceptions of

acceptability), which have led to the construction of grammatical theories that make falsifiable predictions about cross-linguistic variation, language acquisition, and even language processing.

Relatedly, though acceptability judgments are sometimes described as "introspections" or "intuitions," it should be clear that a reported perception is entirely distinct from both of these notions (Carr 1990; Schütze 1996). The terms intuition and introspection come from an early tradition of experimental psychological research pioneered by Wilhelm Wundt that assumed that individuals have (or can learn to have) direct access to cognitive systems. However, by the time of the cognitive revolution, few (if any) psychologists still believed that direct access to cognitive systems is possible. Modern linguistic theory, as a direct product of the cognitive revolution, has never assumed that speakers have direct access to the grammatical system, just the behavioral outputs of that system (see also Chomsky 1965; Schütze 1996).

1.2 The need for judgment data

Judgment data play a crucial role in linguistic investigation because they provide information not readily available from other kinds of data. Most importantly, they provide evidence (under certain assumptions) about the grammaticality of utterances that have never been naturally produced. (There are no known brain measures that are sensitive to all and only the ungrammatical sentences, and failure to appear in even a very large corpus (such as the Web) is not evidence for ungrammaticality, nor is appearance evidence for grammaticality—see Schütze 2009.) Acceptability judgments provide evidence about the status of phenomena that occur so rarely in spontaneous language use that we could not otherwise learn about them. And acceptability judgments sometimes demonstrate knowledge of language in speakers whose behavior on other tasks does not evince the same degree of knowledge: Linebarger, Schwartz and Saffran (1983) showed this with respect to syntax for people with agrammatic aphasia, and Toribio (2001) showed that balanced bilinguals who (for ideological reasons) do not exhibit code-switching behavior nevertheless can provide judgments of the well-formedness of code-switched sentences. A further advantage of judgment data over spontaneous usage data is that the latter will include some proportion of production errors (slips of the tongue/pen/keyboard, etc.), the vast majority of which will be judged as ill-formed by the very speakers who produced them, and which therefore should not be generated by the grammar. Unlike analyzing corpora, collecting judgments allows the researcher to question speakers about what they have said. (See also the discussion of stimulated recall in Chapter 7.) And judgments can be collected in language communities where the use of expensive laboratory equipment is infeasible, and for which there are no corpora available. In light of all of these considerations, the increasingly common suggestion that acceptability judgments should be eliminated as a source of evidence

for linguistics (e.g., Sampson 2007) would be not only counter-productive, but in fact lethal to the field's progress.

1.3 Formal and informal judgment collection

While the elicitation of acceptability judgments is itself a behavioral experiment—the speaker is asked for a voluntary response to a stimulus—the majority of judgment collection that has been carried out by linguists over the past 50 years has been quite informal by the standards of experimental cognitive science. Some have defended this practice on the grounds that it has worked sufficiently well in the past and has led to rapid development of the field (Phillips and Lasnik 2003; Phillips and Wagers 2007; Phillips 2009), while others have criticized linguistics for its informal approach (Ferreira 2005; Wasow and Arnold 2005; Gibson and Fedorenko 2010a, 2010b; Keller 2000; Featherston 2007), suggesting the field may be on shaky empirically ground as a result. The former group have sometimes suggested that following the recommendations of the latter group would entail wasted time and effort that would be better devoted to theoretical matters. We consider it an empirical question whether linguistics would arrive at different conclusions if it followed the more formal (and more time-consuming) experimental structure of nearby fields. We will therefore review recent experimental work that has sought to address the question directly, in the hopes of providing researchers with the information to decide for themselves how to go about collecting their data.

There are five major respects in which typical informal linguistic judgment gathering tends to differ from standard practice in psychology. It typically involves (i) relatively few speakers (fewer than ten), (ii) linguists themselves as the participants, (iii) relatively impoverished response options (such as just "acceptable," "unacceptable," and perhaps "marginal"), (iv) relatively few tokens of the structures of interest, and (v) relatively unsystematic data analysis. The first three issues—sample size, the naïveté of the participants, and response scales—have been explicitly studied; we discuss them in sections 3.3, 3.4, and 2, respectively. (See also the discussion of sample size in Chapter 5.) As we shall see, it is not obvious what the "best" choice is in each case, because all methods appear to provide relatively reliable results. The latter two issues—number of tokens and statistical analysis—we take to be fairly uncontroversial; they are discussed in sections 3.2.3 and 4.1, respectively. (See also Chapters 14–16 for more discussion of statistics.) For now, we look at some case studies that compare formally and informally collected judgment data.

Gibson and Fedorenko (2010b) report such comparisons for seven sentence types taken from previous literature. The informally reported judgments for the relevant comparisons suggest that there are differences among the sentence types, but in their formal experiments Gibson and Fedorenko find no significant differences. (However, see section 3.3 for more on two of the contrasts they tested.) This, they argue, proves that it is possible that the informal methods that have characterized data collection in syntactic theory have led to unsound theorizing. In contrast, Sprouse and Almeida (in press) adopted the following approach in an effort to determine how different the data underlying syntactic theory would be if formal experiments were used to establish a representative set of data points that form part of the foundation of generative syntactic theory. They tested 469 data points from an introductory syntax textbook (Adger 2003) in formal experiments using 440 naïve participants, the magnitude estimation and yes-no tasks (see section 2), and three different types of statistical analyses (traditional null hypothesis significance tests, linear mixed effects models (Baayen et al. 2008), and Bayes factor analyses (Rouder et al. 2009)). The results of that study suggest that the maximum replication failure rate between the informal and formal judgments for those 469 data points is 2%. When it comes to the data being used as the basis for ongoing research, i.e. examples in journal articles, Sprouse, Schütze, and Almeida (submitted) randomly sampled 292 sentence types forming 146 two-sentence phenomena from *Linguistic Inquiry* published between 2001and 2010. By re-testing this random sample in formal experiments, they were able to estimate a minimum replication rate for data points published in *Linguistic Inquiry* (2001-2010) with a margin of error of ±5%. They found that 95% of the phenomena replicated using formal experiments, suggesting a minimum replication rate for journal data of 95% ±5. Taken together, these studies suggest that replacing informal with formal judgment data would have very little impact on the shape or empirical coverage of syntactic theory (see also Featherston 2009 and Phillips 2009 for similar conclusions).

2. Judgment tasks

Judgment tasks can be divided into two categories: non-numerical (or qualitative) tasks and numerical (or quantitative) tasks. This distinction has direct implications for the types of research questions that they can be used to answer. As we will see, non-numerical tasks such as the forced-choice (section 2.1) and yes-no task (section 2.2) are designed to detect *qualitative* differences between conditions, but in the process they sacrifice some of the information about the *size of the difference*. In contrast, the numerical tasks such as Likert scaling (section 2.3) and magnitude estimation (section 2.4) are designed to provide information about the *size of the*

difference, but in the process they may lose power to detect small differences between conditions.

2.1 Forced choice task

In a forced-choice (FC) task, participants are presented with two (or more) sentences, and instructed to choose the sentence that is most (or least) acceptable (perhaps by filling in a corresponding circle or radio button). In this way, FC is explicitly designed to qualitatively compare two (or more) conditions, and directly answer the qualitative question *Is there a difference between these conditions?* (The assumption is that if there is

Figure 1: An example of the two-alternative forced-choice task

What do you think that John bought?	O
What do you wonder whether John bought?	O

There are two major benefits to FC tasks. First, FC tasks are relatively easy to deploy, since each trial in an FC task is an isolated experiment unto itself. In other words, participants do not need to see any sentences other than the two (or more) being directly compared in order to complete the trial accurately. (See section 3.2.4 for the need to use fillers in quantitative tasks.) The second benefit of FC tasks is increased statistical power to detect differences between conditions (see section 3.3). FC tasks are the only task explicitly designed for the comparison of two (or more) conditions; the other tasks compare conditions indirectly through a response scale (either yes-no, or a numerical scale).

There are two primary limitations of FC tasks. First, they can only indirectly provide information about the *size of the difference* between conditions, in the form of the proportion of responses (e.g., 80% choose condition 1 over condition 2, versus 65% choose condition 3 over condition 4—see Myers 2009b). Therefore, if the nature of the research question is simply to ascertain the existence of a predicted acceptability contrast, the FC task seems to be the optimal choice, but if the research question is quantitative in nature, it may be better to use one of the numerical tasks. Second, the task provides no information about where a given sentence stands on the overall scale of acceptability. For linguistic purposes, this is often important: a difference between two sentences both of which are at the high or low end of the acceptability spectrum

may call for a different kind of explanation than a difference between two sentences in the middle of the spectrum.

2.2 Yes/No Task

In the Yes-No (YN) task, participants are presented with one sentence at a time and instructed to judge the sentence as a member of one of two categories: acceptable/yes or unacceptable/no. The YN task is similar to the FC task in that it is primarily a qualitative task; however, there are also substantial differences. The YN task is designed to answer the question *Does this sentence belong to the yes-category or the no-category?* In this way the YN task probes the relationship between a single sentence and the two categories presented to the participant (rather than the relationship between two sentences as in the FC task). However, it is not clear whether all speakers use the same category boundary between yes-no, nor whether the yes-no boundary in any given speaker maps to the theoretically relevant grammatical/ ungrammatical boundary, assuming there is such a boundary.

Figure 2: An example of the yes-no task

What do you wonder whether John bought?	O Yes O No

The primary advantage of the YN task is that that it is quick to deploy. Moreover, as with the FC task, several researchers have demonstrated that the YN task can be used to compare the relative difference between conditions, by computing the proportion of yes-responses for each condition (Myers 2009b, Bader and Häussler 2010).

The primary disadvantage of the YN task is that it is likely less sensitive than the FC task at detecting qualitative differences between two conditions (because the difference is always relative to the category boundary) and likely less sensitive than the quantitative tasks at establishing numerical estimates of the difference between conditions (because the difference is indirectly computed through proportions).

2.3 Likert scale task

In a Likert scale (LS) task, participants are given a numerical scale, with the endpoints defined as acceptable or unacceptable, and asked to rate each sentence along the scale. The most commonly used scales usually consist of an odd number of points (such as 1–5 or 1–7) because odd numbers contain a precise middle point; however, if the research goals require it, a

preference can be forced by choosing an even number of points. One of the primary benefits of LS is that it is both numerical and intuitive. The former means that LS can be used to answer questions about the *size of a difference* between conditions by leveraging inferential statistical tests such as ANOVA and linear mixed-effects modeling.

Figure 3: An example of a Likert Scale task

What do you wonder whether John bought?	O 1	O 2	O 3	O 4	O 5	O 6	O 7

The primary limitations of LS are all related to the use of the numerical scale. For example, the scale itself suggests that the intervals between points are uniform: the interval between 1 and 2 is one unit, the interval between 2 and 3 is one unit, etc. However, because participants can only use the limited number of response points (i.e., there is no 3.5 on the scale), it is impossible to ensure that the intervals are truly uniform, i.e., that subjects treat the difference between 1 and 2 the same as the difference between 4 and 5. This problem is compounded when aggregating across participants in a sample. In practice, this risk can be minimized by including anchoring examples at the beginning of the experiment to establish some of the points along the scale (see section 3.2.1). Furthermore, participants' responses can be z-score transformed (see section 4.1.1) prior to analysis to eliminate some additional forms of bias such as scale compression (e.g., using only points 3–5 on a 1–7 scale) or scale skew (e.g., using only the high end of the scale).

2.4 Magnitude estimation task

In the magnitude estimation (ME) task, participants are given a reference sentence and told that the acceptability of the reference sentence is a specific numerical value (e.g., 100). The reference sentence is called the *standard* and the value it is assigned is called the *modulus*. Participants are then asked to rate additional sentences as a proportion of the value of the standard. For example, a sentence that is twice as acceptable as the standard would be rated 200.

Figure 4: An example of the magnitude estimation task

Standard: Who thinks that my brother was kept tabs on by the FBI?
Acceptability: 100

Item: What do you wonder whether John bought?
Acceptability: ____

ME was developed by Stevens (1957) explicitly to overcome the problem of potentially non-uniform, and therefore non-meaningful, intervals in the LS task (in the domain of psychophysics). In the ME task, the standard is meant to act as a unit of measure for all of the other sentences in the experiment. In this way, the intervals between sentences can be expressed as proportions of the standard (the unit of measure). This offers the theoretical possibility of substantially more accurate ratings (Bard et al. 1996; Cowart 1997; Keller 2000; Featherston 2005a, 2005b) than the LS task. In addition, the response scale in ME is the entire positive number line, which means that participants can in principle report a potentially infinite number of levels of acceptability (Bard et al. 1996; Keller 2000), as opposed to the (typically small) finite number in the LS. As a numerical task, an ME experiment requires the same design properties as an LS task (see section 3). The choice of the standard can affect the amount of the number line that is available for ratings: a highly acceptable standard set at a modulus of 100 means that nearly all ratings will be between 0 and 100, whereas a relatively unacceptable standard means that nearly all ratings will be above 100. For this reason, and in order to prevent certain types of response strategies, it is normal practice to employ a standard that it is in the middle range of acceptability.

Unfortunately, a series of recent studies of the ME task have called into question many of its purported benefits. First, although the availability of any positive real number as a response would in theory allow participants to rate every stimulus differently, in practice this is not at all what they do. Rather, they use a small set of (typically whole) numbers repeatedly, and (many or all of) the members of that set often stand in a salient relationship to one another that does not seem to depend on the stimuli (e.g., multiples of five or ten). Second, one of the primary assumptions of the ME task is that participants truly use the reference sentence as a unit of measurement. In order for this to be true, participants must be able to make a ratio comparison of two sentences (e.g., the acceptability of sentence B is 1.5 times the acceptability of sentence A). Adapting a series of techniques developed in the psychophysics literature (Narens 1996; Luce 2002), Sprouse (2011b) tested this assumption directly, and found that participants could not make ratio comparisons of the acceptability of two sentences. This failure of the primary assumption of the ME task suggests that participants may be treating the ME task as a type of LS task, only with an open and infinite response scale. Why this is true is still an open question,

although one possibility is that the lack of a meaningful zero point for acceptability (i.e., the concept of absolutely no acceptability) prevents participants from making ration judgments. This finding accords well with the results of a direct comparison between ME and LS tasks for several sentence types in German that was conducted by Weskott and Fanselow (2011): they found that there is no evidence of increased sensitivity of ME over LS, though there is increased variance, which is likely due to the increased number of response options in ME.

The burgeoning consensus among researchers is that the true value of ME lies in the increased number of levels of acceptability that participants can report—though this might come at the cost of higher variance and is not unique to ME (cf. section 2.5)—and the sociological impact on the field of using a task that is perceived as more sophisticated than LS. Countervailing drawbacks include the fact that magnitude estimation is less intuitive for many participants than traditional scales (and hence more time consuming and labor intensive for experimenters), and some participants do not apply the task to sentence judgments in the intended way and their data must be discarded.

2.5 The Thermometer task

Some researchers have proposed new tasks that are intended to combine the intuitive nature of point scales with the sensitivity of ME. For example, Featherston (2009) has proffered a "thermometer task" in which participants are given two reference sentences with associated acceptability values, such as 20 and 40 (analogous to freezing and boiling points). They can then choose values for target sentences along the real number line relative to those two points by treating it as a linear scale: for example, a target whose acceptability is halfway between the acceptability of the two reference sentences would be rated 30.

2.6 The fundamental similarity of acceptability judgment tasks

Before concluding this section, it is important to note that at a fundamental level, all of the acceptability judgment tasks are the same: the participants are asked to perform the same *cognitive task*, that is to report their perceptions of acceptability. Because the cognitive task is the same, the data yielded by each task is likely to be very similar (modulo small differences in the response scale discussed above), especially when the criterion for comparison is the detection of differences between conditions. Indeed, this is exactly what has been found by several recent studies that have directly compared the various judgment tasks. For example, Bader and Haüssler (2010) compared ME and YN tasks for several sentence types in German, and found that both tasks detected differences between the conditions (at the chosen sample sizes). Similarly,

Weskott and Fanselow (2011) compared the ME, LS, and YN tasks for several other sentence types in German, and found that all three tasks detected differences between the conditions (at the chosen sample sizes). Though there are likely to be differences between tasks with respect to statistical power (e.g., Sprouse and Almeida submitted), when it comes to simply *detecting a difference* between conditions at relatively large sample sizes (e.g., 25 participants), the fact that the cognitive task is identical across these measures strongly suggests that choice of task is relatively inconsequential.

3. Designing judgment experiments

Chapters 7 and 8 of this volume provide general discussion of many issues in experimental design. There are also several excellent resources for interested readers to learn the mechanics of creating multiple lexicalizations, distributing items according to a Latin Square, pseudorandomizing items, etc. (for example, see Cowart 1997, Kann and Stowe 2001). In this chapter we focus on methodological issues that are particularly germane to the design of judgment experiments.

3.1 Instructions

While there is no standard way of wording the instructions for a judgment experiment, there is general agreement that we want to convey to speakers that certain aspects of sentences are **not** of interest to us and should not factor into their responses. These include violations of prescriptive grammar rules, the likelihood that the sentence would actually be uttered in real life, and the truth or plausibility of its content. See Chapter 6 for more on these effects. We also want to avoid the question of the sentence being understandable, since uncontroversially ungrammatical sentences are often perfectly comprehensible, e.g. *What did he wanted?* It is common to instruct participants to imagine that the sentences were being *spoken* by a friend, and ask whether the sentences would make them *sound* like a native speaker of their language. Crucially this formulation invokes the spoken modality even with written surveys, and attempts to guide the participant toward a notion of acceptability that is tied to native-speaker ability rather than frequency or plausibility.

One question that is often asked by researchers who are new to acceptability judgments is to what extent the instructions of the experiment can influence the results. The consensus among experienced acceptability judgment experimentalists is that the exact nature of the instructions (modulo the issues discussed in the previous experiment) matters relatively little. To put this another way, the experimenter has relatively little control over how participants choose to

respond to the sentences presented to them. Cowart (1997) suggests that this means that experimenters should focus on controlling the experiment (materials, fillers, etc.) rather than controlling the behavior of the participant. Unfortunately, because most experienced experimenters do not believe that there is much effect of instructions on acceptability judgments, the formal data on this subject is relatively limited. Cowart (1997) compared what he calls "intuitive" instructions like those described in the previous experiment with "prescriptive" instructions that explicitly asked participants to evaluate the well-formedness of sentences in the context of an undergraduate term paper, and found no substantive difference in the pattern of acceptability for several sentence types (though there was one significant absolute difference in the ratings of one of the sentence types).

3.2 Materials

3.2.1 Practice items

Acceptability judgment tasks are generally considered intuitively natural for participants. As such, explicit practice sessions are generally unnecessary to familiarize participants with the task. However, there are a few specific instances where certain types of practice items may be helpful.

In the LS task, it is common to provide *anchor* items for certain points on the scale, to help ensure that every participant uses the scale the same way (thus minimizing scale bias, see section 4.1.1). An anchor item is a single sentence token that the researcher assigns to a single point on the rating scale. It is not necessary to provide an anchor for every point on the scale. Instead, it is common to provide an anchor for the lowest point (to establish a floor) and for the highest point (to establish a ceiling). Some researchers also provide an anchor for the midpoint of the scale. It is also common to include five to ten items at the very beginning of the survey whose sole purpose is to help the participants become familiar with using the scale. These items are not marked in any way, so the participant is unaware that they are distinct from the rest of the experiment. These items generally cover the full range of acceptability, so that by the end of the sequence the participant will have used every point along the scale at least once. These items are technically fillers in that they will not be analyzed in the service of an experimental hypothesis, but they may be more profitably thought of as *unannounced* practice items.

In the ME task, it is common to include an initial (announced) practice phase in which participants conduct a simple ME task with line lengths, to ensure that participants understand the basic premise of the ME task. This practice phase is usually short, perhaps five to ten items.

After the practice phase is concluded, participants are introduced to the idea of using ME to rate the acceptability of sentences. Given recent evidence that participants may not be making ratio judgments and instead may be treating ME tasks as a type of rating task similar to LS tasks (Sprouse 2011b), it is probably also a good idea to include unannounced practice items with ME tasks as well.

3.2.2 Factorial designs

If you have chosen to conduct a formal experiment, it is likely that your hypothesis requires quantifying relative differences in acceptability, above and beyond simply establishing that two sentences are different (see section 2 for more about the relationship between tasks and the types of information that they provide). In such cases, it is generally useful to consider using fully-crossed factorial designs (see also Myers 2009b and Chapter 7). For example, imagine that you are interested in testing the effect of D-linking on Complex Noun Phrase Constraint (CNPC) violations. You would start by comparing the acceptability of a CNPC violation with non-D-linked wh-words (*what*) to the same configuration with D-linked wh-phrases (*which book*) as in (1):

(1) a. What did you make the claim that John bought?

b. Which book did you make the claim that John bought?

Imagine that you find that (1b) is more acceptable than (1a). Can you claim that D-linking improves the acceptability of CNPC violations? Not really. It may be that D-linking improves the acceptability of *all* sentences, even those that do not contain a CNPC violation. To test this, you need to compare two additional sentences:

(2) a. What did you claim that John bought?

b. Which book did you claim that John bought?

Now the question is whether the difference between (1a) and (1b) is smaller than, equal to, or larger than the difference between (2a) and (2b). This will tell us whether D-linking has a specific effect on CNPC violation, or whether it has the same effect on all extractions from embedded clauses. The four sentences in (1) and (2) form a *factorial design*, as there are two factors (embedded clause type and wh-phrase type), each with two levels (± island, ± D-linking), that give rise to the four conditions. Factorial designs are the best tool an experimenter has for isolating the factors that could give rise to relative differences in acceptability.

3.2.3 Multiple lexicalizations

Most hypotheses in linguistics are not about individual sentences but about types of sentences, i.e. all sentences that have a particular structural property. This fact is sometimes obscured when reading linguistics articles, where often just one or two examples are presented. However, these are almost always intended to be representative exemplars. The assumption is that the author has considered a range of possible lexicalizations to verify the generality of their claim, and is simply saving space by not reporting all of them. The same procedure should apply in conducting formal experiments. Whenever possible, it is desirable to create multiple lexicalizations of each condition (ideally eight or more) and distribute them evenly among the participants, in an effort to minimize the contribution of particular lexical items, facts about real-world plausibility, etc. to the results. In experiments with one sentence per trial rather than a pair of sentences to compare, we use a distribution procedure to ensure that no one participant sees the same lexicalization of related conditions. The most common distribution procedure is called a Latin Square (for details of the mechanics, see Kaan and Stowe 2001 and Chapter 7).

3.2.4 Fillers

In most experiments it is beneficial to include filler items (i.e., sentences that are not related to the research question). These can serve at least three purposes. First, they can reduce the density of the critical comparisons across the whole experiment, reducing the chances that participants will become aware that a particular sentence type is being tested, which could trigger conscious response strategies. Second, they can be used to try to ensure that all the possible responses (Yes and No, or points along a scale) are used about equally often. This helps to protect against scale bias, which occurs when one participant decides to use the response scale differently from other participants, such as only using one end of the scale (skew), or only using a limited range of responses (compression). (See also section 4.1.1 for statistical approaches to mitigating the effect of scale bias.) Third, they can be used to investigate a separate research question.

3.3 Sample size and statistical power

Informal judgment experiments of the sort that linguists carry out every day tend to be conducted on relatively few participants (almost always fewer than 10),[1] whereas formal

[1] Sometimes this is by necessity. In the case of languages spoken in remote locations and languages with few remaining speakers, collecting data from just one or two speakers may be all that a linguist can practically do (see Chapter 4). Nothing in what follows is meant to lessen the value of such linguistic fieldwork.

judgment experiments tend to use samples of 20 or more. Whether differences in sample size are relevant for the reliability of the results is an empirical question that can only be answered relative to the sentence types under investigation. Sprouse and Almeida (submitted) analyzed the relationship between sample size and the probability of detecting a significant difference (also known as *statistical power*) for 47 two-sentence phenomena from *Linguistic Inquiry* 2001-2010 (Sprouse, Schütze, and Almeida submitted) for all four judgment tasks: ME, LS, YN, and FC.

Sprouse and Almeida (submitted) found that (i) the FC task is substantially more powerful than the other three tasks at detecting differences between conditions, especially for small and medium-sized effects, (ii) the ME and LS tasks are approximately equally powered, albeit less powerful than the FC task, and (iii) the YN task is the least powerful of the four. Sprouse and Almeida provide several types of comparisons to illustrate these power differences, but perhaps the most striking is in terms of empirical coverage. Following the conventions of experimental psychology, Sprouse and Almeida assume that experimenters should strive for at least 80% power (i.e., an 80% chance of detecting a true difference when one exists) in their experiments. They then ran re-sampling simulations on their results to empirically estimate the number of phenomena in *Linguistic Inquiry* (2001-2010) that would be detected with 80% power for every possible sample size between 5 and 100 participants. The results suggest that the FC task would be well-powered (i.e., reach 80% power) for the detection of 70% of the phenomena published in *Linguistic Inquiry* (2001-2010) with only 10 participants each providing only one judgment per phenomena (i.e., 10 observations total). With only 15 participants (each providing one judgment per phenomenon), the empirical coverage of the FC task rises to 80% of the phenomena in *Linguistic Inquiry*. In contrast, 10 participants in the ME and LS tasks lead to less than 60% coverage of the phenomena in LI. The ME and LS tasks require 30-35 participants to reach the 80% coverage that the FC task achieves with only 15 participants. Finally, the YN task only achieves 40% coverage with 10 participants, and requires 40 participants to reach 80% coverage. Of course, these power estimates are lower bounds, inasmuch as they assume that each participant provides only one judgment per condition. Increasing the number of judgments per condition will also increase statistical power, thereby decreasing the required sample sizes.

As a concrete example of the importance of understanding the relationship between sample size, task, and statistical power, let's take a closer look at two effects that have been reported in the linguistics literature using linguists' judgments, but have failed to replicate with larger, formal experiments. The first is the center embedding effect from Frazier (1985), attributed to Janet Fodor, where linguists' judgments suggested that doubly center-embedded sentences can be made more acceptable by deleting the second VP, as in (3b).

(3) a. *The ancient manuscript that the graduate student who the new card catalog had confused a great deal was studying in the library was missing a page.

b. ?The ancient manuscript that the graduate student who the new card catalog had confused a great deal was missing a page.

Formal experiments reported by Gibson and Thomas (1999) using a LS task failed to corroborate this difference. However, Sprouse and Almeida (2012) found that this is likely due to the relatively large sample sizes that are required to detect this difference in numerical rating tasks: they report that at least 78 participants (giving one judgment each) are required to detect this difference with 80% power with the ME task. The fact that the FC task, which is likely the task used by Fodor and Frazier (1985) to detect the center embedding effect, tends to be more powerful than numerical rating tasks at detecting differences (Sprouse and Almeida submitted) is one possible explanation for the failure to replicate in Gibson and Thomas (1999).

A similar situation is reported by Gibson and Fedorenko (2010b). They note that Gibson (1991) reported a contrast between doubly embedded object relative clauses in subject versus object position, as in (4), using informal judgments provided by himself and other linguists:

(4) a. *The man that the woman that the dog bit likes eats fish.

b. ?I saw the man that the woman that the dog bit likes.

However, Gibson and Fedorenko report that subsequent experiments using LS tasks have failed to replicate this result (unfortunately, they do not report the details of these experiments). Sprouse and Almeida (2012) tested this contrast in a FC task with 99 naïve participants, and then ran power analyses like those in Sprouse and Almeida (submitted) to determine a target sample size. They found that a sample size of 11 is required to detect the difference in (4) with 80% power using the FC task. Although they do not have data for numerical tasks, based on the power analyses in Sprouse and Almeida (submitted), phenomena that require 11 participants in the FC task tend to require 30-35 participants in the LS task. If the experiments reported by Gibson and Fedorenko (2010b) used fewer than 30-35 participants, then the lack of replication of the Gibson (1991) informal results could simply be due to relative power differences between the FC and LS tasks.

There are two important lessons in these case studies. First, it is critical to understand the relationship between sample size, task, and statistical power when designing an experiment. Although it may seem impossible to estimate a required sample size *before* collecting the data, it

is possible to use existing power studies such as Sprouse and Almeida (submitted) to estimate the sample size required for a given phenomenon by comparing your judgments of the size of the difference in your conditions to the phenomena that they tested. Second, it is important to realize that the failure to find an effect in a formal experiment does not mean that there is no effect to be found: the experiment may simply have been underpowered.

3.4 Naïve versus expert participants

One of the most contentious aspects of judgment data is whether they should be collected from trained linguists versus naïve speakers. It would not be especially surprising if it turned out that linguists do not have the same judgments as non-linguists—see below for empirical evidence on this point. Even if that is true, however, it does not follow that using linguists' judgments is bad for the field—that would depend on how and why linguists behave differently. This is a harder question to answer empirically, and in our opinion it remains an open one. A priori, one can image at least two ways in which judgments from the two populations might diverge. One is that linguists as participants will likely be aware of the theoretical consequences of their judgments, and may be subconsciously biased to report judgments consonant with their theoretical viewpoints (Edelman and Christiansen 2003; Ferreira 2005; Wasow and Arnold 2005; Gibson and Fedorenko 2010a, 2010b). On the other hand, professional linguists may provide a sort of expert knowledge that increases the reliability, and possibly the sensitivity, of their judgments over non-linguists' judgments (see Newmeyer 1983, 2007, as well as Fanselow 2007, Grewendorf 2007, and Haider 2007 for possible examples in German, and Devitt 2006, 2010, Culbertson and Gross 2009, Gross and Culbertson 2011 for a discussion of what could be meant by 'expert knowledge'). Valian (1982) makes a case in favor of using such expert linguistic judgments, based on an analogy to wine tasting, which relies on the acquired ability to detect subtle distinctions that inexperienced wine drinkers simply cannot make. Linguists may have similarly heightened sensitivity, or they may be more practiced at factoring out aspects of sentences that irrelevant to their grammatical status.

There are several examples of demonstrated differences between populations in the literature. For example, Spencer (1973), Gordon and Hendrick (1997), and Dąbrowska (2010) all report differences in ratings between linguists and non-linguists, Culbertson and Gross (2009) report differences between participants who have completed a formal experiment previously and participants who have not, and Dąbrowska (2010) reports differences between generative linguists and functional linguists in the ratings of CNPC violations. However, we know of no studies that have conclusively established the cause of the differences (which would require

careful parametric manipulations of the relevant grouping factors over a series of experiments), and no studies that have demonstrated that these differences would lead to major differences in theoretical conclusions (indeed, many of the differences appear to be in absolute ratings but not in the relative pattern of acceptability – the latter generally being the data upon which theories are built).

4. Interpreting judgment data

4.1 Statistical analysis

As in most of experimental psychology, the analysis of judgment data involves two steps: pre-processing, which covers operations performed prior to statistical tests, and the statistical tests themselves.

4.1.1 Data pre-processing

The pre-processing of numerical judgment data generally involves two steps. The first is common to all data in experimental psychology: the identification of participants who did not perform the task correctly, and the identification of extreme outliers in the responses. We will not discuss this basic step further as we assume that readers can consult general experimental textbooks for the logic and mechanics of participant and outlier removal (e.g., Stowe and Kaan 2001), though it should be noted that there are as yet no generally agreed upon procedures for participant and outlier removal for acceptability judgments. The second step is common to many scale-based data types: each participant's responses are transformed using the *z-score transformation* to eliminate some of the potential scale bias that was mentioned above. The z-score transformation allows us to express each participant's responses on a *standardized* scale. It is calculated as follows: For a given participant P, calculate the mean and standard deviation of all of P's judgments. Next, subtract each of P's judgments from the mean. Finally, divide each of these differences by P's standard deviation. The resulting set of responses (z-scores) represent a *standardized* form of P's responses, as each response is expressed in standard deviation units from P's mean. The process is repeated for each participant so that every participant's responses are reported on a scale based on standard deviation units. The z-score transformation is a linear transformation, which means that it maintains all of the relationships that exist within the data (i.e., it adds no distortion).

Many researchers, including us, believe that the z-score transformation should be used routinely for both LS and ME judgment data. However, from time to time, some researchers disagree. The most common criticism of the z-score transformation for LS data is that LS data is

not continuous, whereas the z-score transformation transforms these bounded responses into a continuous scale for each participant. However, if you plan to run parametric statistical tests on LS data (e.g., *t*-tests, ANOVAs, linear mixed effects models), then you are already assuming that you can treat LS data as continuous for practical purposes. So there is no harm to applying the z-score transformation first, and there are many benefits. If you do not wish to treat LS data as continuous, then you should run non-parametric statistical tests. These tests convert each participant's data into ranks before analysis, which actually eliminates scale bias in the process, so there is no reason to run a z-score transformation prior to non-parametric tests. However, non-parametric tests are generally less sensitive than parametric tests, so this is less ideal than the use of z-score transformations and parametric tests.

The most common criticism of the use of z-score transformations for ME data is that ME data should be *log-transformed* instead. The purported rationale behind the log-transformation with ME data is that it will eliminate right-tail outliers that arise because the scale in ME tasks is open ended to the right and bounded to the left. However, the log-transformation is a powerful transformation that is normally not recommended for simple outlier removal. It is a non-linear transformation, which means it distorts the relationships within the data, therefore it should only be used when absolutely necessary. The log-transformation is intended to be used when the distribution of the data is log-normal, which is a type of logarithmic distribution, as the log transformation (by definition) transforms a log-normal distribution into a normal distribution. Unfortunately, this means that if the log-transformation is applied to non-log-normal distributions, then it will transform them into non-normal distributions. In our experience, judgments are never distributed log-normally (and are very often distributed normally), so the log-transformation is inappropriate.[2]

4.1.2 Statistical tests

The current best practice in the experimental syntax literature is to use *linear mixed effects models* for the analysis of numerical judgment data (LS and ME), and to use *logistic mixed effects models* for the analysis of non-numerical judgment data (FC and YN) (see Baayen 2007, Baayen et al. 2008, and Chapter 16). However, as mentioned above, from time to time some researchers worry that parametric statistical tests should not be used to analyze judgment data, particularly LS data. The concern usually revolves around the response scale: many believe

[2] We are not sure why many researchers assume that the log-transformation should be standard practice for ME experiments, but one possibility is that it has arisen due to the presence of log-transformations in early psychophysical studies, which were used for reasons not relevant to current judgment experiments.

that LS tasks fail to meet the assumption of parametric tests that the responses are on an interval or ratio scale. While it is important to take the assumptions of statistical tests seriously, the actual situation is more complex. Parametric tests involve several assumptions (including random sampling from the parent population, normality of the parent populations of each condition, and homogeneity of the variances of the conditions) that are rarely met in psychological research. The question then is when it is tolerable to violate the assumptions and when it is not. A full discussion of this question is beyond the scope of this chapter (for interesting reviews of the use of null hypothesis significance testing in psychology see Hunter and May 1993, Nickerson 2000, Gigerenzer et al. 2004, and references therein). At a practical level, the nearly universal use of parametric tests in psychology suggests that the field has decided (consciously or not) that it is willing to tolerate the potential consequences of the violations of parametric tests. Hunter and May (1993) evaluate this decision in relation to the alternative—the adoption of non-parametric tests, which do not carry the same assumptions as parametric tests. They argue that the application of many standard parametric tests (e.g., *t*-tests and *F*-tests) in scenarios where the assumptions are not met (e.g., lack of random sampling) actually approximates the application of non-parametric tests (e.g., randomization tests).[3]

4.2 Interpreting variation across participants

Finding a statistically significant effect for some set of participants does not mean that every participant demonstrated the effect. In practice, given sufficient statistical power, very few participants need to show the effect in order for the sample as a whole to show a significant effect. What should one make of such variability? What if 75% show the effect and 25% do not? What if only 25% show the effect, and 75% do not? (As Raaijmakers (2003) points out, statistical significance can still be achieved in such circumstances.) What if some of those who do not show the expected effect actually show the opposite effect? There seem to be three different approaches to this problem:

1. Variation as noise: On this view, since all measurement involves noise, only the central tendency of the sample matters, and it is expected that not every participant or every item

[3] There are differences between the inferences licensed by parametric and non-parametric tests. For example, when all of the assumptions are met, parametric tests can be used to make inferences about population parameters from the samples in the experiment. Non-parametric tests, which do not assume random sampling, can only be used to make inferences about the sample(s) in the experiment itself. As Hunter and May point out (see also Nickerson 2000), it is relatively rare for experimental psychologists to be interested in population parameters; instead, they tend to be concerned with establishing a significant difference between two samples within a well-controlled experiment. So even this consequence of the parametric/non-parametric distinction may be relatively benign within experimental psychology.

in the sample will show the difference. This interpretation is the default assumption in experimental psychology and much of the experimental syntax literature.

2. Variation as dialect/idiolect: On this view, if a large enough proportion of participants do not show the predicted effect, this might be evidence for a different grammar for that subset of participants. In psychology this is usually not a possible interpretation, because the population of interest is all humans; in linguistics, the population of interest is all speakers of a given language, so it is always a logical possibility that the participants who do not show an effect have a different grammar (or perhaps control additional lexical variants in the sense of Adger's 2006, 2007 combinatorial variability approach) from the speakers who do show the effect (den Dikken et al. 2007). Unfortunately, it is nearly impossible to establish the existence of a dialectal/idiolectal difference in a single experiment; conclusive evidence generally requires systematic parametric manipulations of potential dialectal/idiolectal grouping factors across several experiments. (See Chapter 5 for considerations in sampling participants, and Gervain 2003 for the potential use of cluster analysis for the detection of dialects/idiolects.)

3. Variation as disconfirmation: On this view, given a strong hypothesis that ungrammatical sentences should be overwhelmingly judged to be unacceptable, a large enough proportion of participants that fail to show the predicted effect will be taken as evidence that the theoretical prediction is disconfirmed. If so, the difference (among those who do show it) is not due to the grammar. The assumption here is that a truly grammatical effect should not show a high degree of variability, whereas extra-grammatical effects may (Hoji 2010). Some criticisms of informal experiments rest upon this assumption (Wasow and Arnold 2005; Gibson and Fedorenko 2010a, 2010b).

In the literature one can find instances of all three approaches—the field has evidently not reached a consensus on which one is appropriate, or indeed if the answer ought to vary as a function of the question being asked. One way to address the problem is to seek converging evidence from a wide array of types of data whenever possible. The assumption behind this is that random noise will not be consistent across tasks, while grammar-based variation should. Less obvious is the question of whether extra-grammatical sources of variation are expected to be consistent across tasks.

4.3 Interpreting gradience

The freedom provided by magnitude estimation and related tasks to distinguish a theoretically infinite number of levels of acceptability and to quantify the distances between those levels has been a catalyst for some researchers to replace a categorical model of grammar in which there are two distinct categories, grammatical and ungrammatical (possibly with distinctions among the latter), with a gradient model of grammar in which grammaticality is a continuous property. This possibility has recently been explored in several different ways, such as the Optimality Theory approach of Keller (2000), the Generative Grammar approach of Featherston (2005c), and the probabilistic approach of Bresnan (2007). While it is not surprising that judgment tasks yield continuous acceptability values, what is nontrivial is that respondents are consistent in their use of the intermediate levels of acceptability, suggesting that they are indeed tapping into a robust cognitive system that yields gradient results. The key question is whether those gradient results are a reflection of grammatical knowledge on its own, or grammatical knowledge in combination with factors that affect language processing, decision making, etc. and are already known to display gradient behavior (working memory load, semantic plausibility, lexical and syntactic frequency, prototypicality, etc.).

It is not uncommon to encounter those who believe continuous acceptability necessitates a continuous (or gradient) syntactic system. However, there is no necessary link between the nature of acceptability and the nature of the syntactic system. For example, Armstrong et al. 1983 and Barsalou 1987 demonstrate that participants can give systematic gradient judgments about concepts that we know to be categorical, such as the concept of *even number*. This observation does not entail that our knowledge of mathematics fails to make a perfectly sharp distinction between even and odd numbers. Rather, our judgments can evidently be sensitive to factors other than our underlying competence. One possibility is that instead of rating the extent to which some number is even, participants may (not necessarily consciously) reinterpret the task as seeking a rating of how representative or typical the properties of a particular number are as compared to the set of even numbers as a whole. Putting it another way, when asked for gradient responses, participants will find some way to oblige the experimenter; if doing so is incompatible with the experimenter's actual question, they apparently infer that the experimenter must have intended to ask something slightly different. By the same logic, gradient acceptability judgments are perfectly compatible with a categorical model of competence. The (admittedly difficult) question facing the experimenter is whether gradient acceptability judgments are the result of the nature of the grammar, the result of gradient processing factors, or simply an artifact of asking participants to provide gradient responses.

5. Conclusion

In closing, we wish to emphasize two points. First, the correct interpretation of acceptability judgment data will ultimately require a theory of the judgment task itself (cf. Schütze 1996, p. 175). This will minimally include a theory of grammar, a theory of parsing, a theory of partial parsing in the case of ungrammatical sentences, a theory of rating tasks, and possibly other components. A priori we cannot know which of these components is the source of any given property of judgment data (e.g. gradience)—this is a classic "black-box" problem in cognitive science: several different unobservable systems contribute to the observable behavior. Second, the experimental and analytical techniques discussed in this chapter are no substitute for human thought. In particular, the fact that a carefully conducted experiment yields a significant result is not ipso facto important for any particular theories of grammar, processing, or what have you—it is up to the researcher to interpret it. Likewise, the fact that a carefully conducted experiment fails to yield a significant result does not mean that an effect does not exist—it could simply indicate a flaw in the design, including a lack of sufficient power. Determining what results mean is part of the art of doing science, not a task that the machinery of experimentation can do on its own.

References

Adger, D. 2003. *Core Syntax: A Minimalist Approach*. Oxford University Press.

Adger, D. 2006. Combinatorial variability. *Journal of Linguistics* 42: 503–530.

Adger, D. 2007. Variability and modularity: A response to Hudson. *Journal of Linguistics* 43: 695–700.

Armstrong, S.L., Gleitman, L.R. and Gleitman, H. 1983. What some concepts might not be. *Cognition* 13: 263–308.

Baayen, R. H. 2007. *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H., Davidson, D. J. and Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59: 390–412.

Bader, M. and Häussler, J. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46: 273–330.

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

Barsalou, L.W. 1987. The instability of graded structure: Implications for the nature of concepts. In U. Neisser (ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*, pp. 101–140. Cambridge: Cambridge University Press.

Bresnan. Joan 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In *Roots: Linguistics in Search of Its Evidential Base*, ed. Sam Featherston and Wolfgang Sternefeld. Berlin: Mouton de Gruyter, 77–96.

Carr, P. 1990. *Linguistic realities: An autonomist metatheory for the generative enterprise*. Cambridge: Cambridge University Press.

Chemla, E. and Spector, B. 2011. Experimental evidence for embedded scalar implicatures. *Journal of Semantics* 28: 359–400.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Cohen, J. 1965. Some statistical issues in psychological research. In B. B. Wolman (ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

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

Culbertson, J., and Gross, J. 2009. Are linguists better subjects? *British Journal of the Philosophy of Science* 60: 721–736.

Dąbrowska, E. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27: 1–23.

den Dikken, M., Bernstein, J., Tortora, C. and Zanuttini, R. 2007. Data and grammar: means and individuals. *Theoretical Linguistics* 33: 335–352.

Devitt, Michael 2006. Intuitions in linguistics. *British Journal for the Philosophy of Science* 57: 481–513.

Devitt, Michael 2010. Linguistic intuitions revisited. *British Journal for the Philosophy of Science* 61: 833–865.

Dresher, E. 1995. There's no reality like psychological reality. *Glot International* 1(1): 7.

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

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

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

Featherston, S. 2005b. Universals and grammaticality: wh-constraints in German and English. *Linguistics* 43: 667–711.

Featherston, S. 2005c. The Decathlon Model of empirical syntax. In *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*. M. Reis & S. Kepser (eds). Berlin: Mouton de Gruyter. 187-208.

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

Featherston, S. 2008. Thermometer judgments as linguistic evidence. In C. M. Riehl and A. Rothe (eds.), *Was ist linguistische Evidenz?* Aachen: Shaker Verlag.

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

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

Gervain, Judit. 2003. Syntactic microvariation and methodology: problems and perspectives. *Acta Linguistica Hungarica* 50: 405–434.

Gibson, E. and Fedorenko, E. 2010a. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14: 233–234.

Gibson, E. and Fedorenko, E. 2010b. The need for quantitative methods in syntax. *Language and Cognitive Processes.* In press.

Gigerenzer, G., Krauss, S. and Vitouch, O. 2004. The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gordon, P.C. and Hendrick, R. 1997. Intuitive knowledge of linguistic co-reference. *Cognition* 62: 325–370.

Grewendorf, G. 2007. Empirical evidence and theoretical reasoning in generative grammar. *Theoretical Linguistics* 33: 369–381.

Gross, Steven and Culbertson, Jennifer 2011. Revisited linguistic intuitions. *British Journal of the Philosophy of Science* 62: 639–656.

Haider, H. 2007. As a matter of facts – comments on Featherston's sticks and carrots. *Theoretical Linguistics* 33: 381–395.

Hoji, Hajime 2010. Hypothesis testing in generative grammar: Evaluation of predicted schematic asymmetries. *Journal of Japanese Linguistics* 26: 25–52.

Hunter, M. A. and May, R. B. 1993. Some myths concerning parametric and nonparametric tests. *Canadian Psychology/Psychologie canadienne* 34: 384–389.

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

Kruschke, J. K. 2010. *Doing Bayesian Data Analysis*. Academic Press.

Linebarger, M.C., Schwartz, M.F. and Saffran, E.M. 1983. Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition* 13: 361–392.

Luce, R. D. 2002. A psychophysical theory of intensity proportions, joint presentations, and matches. *Psychological Review* 109: 520–532.

Matthewson, Lisa 2004. On the methodology of semantic fieldwork. *International Journal of American Linguistics* 70: 369–415.

Myers, James 2009a. Syntactic judgment experiments. *Language & Linguistics Compass* 3: 406–423.

Myers, James 2009b. The design and analysis of small-scale syntactic judgment experiments. *Lingua* 119: 425–444.

Narens, L. 1996. A theory of ratio magnitude estimation. *Journal of Mathematical Psychology* 40: 109–129.

Newmeyer, F. J. 1983. *Grammatical theory: Its limits and its possibilities*. Chicago: University of Chicago Press.

Newmeyer, F. J. 2007. Commentary on Sam Featherston, 'Data in generative grammar: The stick and the carrot.' *Theoretical Linguistics* 33: 395–399.

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

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

Phillips, C., and Lasnik, H. 2003. Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Sciences* 7: 61–62.

Phillips, C. and Wagers, M. 2007. Relating Structure and Time in Linguistics and Psycholinguistics. In G. Gaskell (ed.), *Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press.

Raaijmakers, J. G. W. 2003. A further look at the 'language-as-fixed-effect fallacy.' *Canadian Journal of Experimental Psychology* 57: 141–151.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. and Iverson, G. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* 16: 225–237.

Sampson, G.R. 2007. Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory* 3: 1–32.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology.* Chicago: University of Chicago Press.

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

Spencer, N. J. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2: 83–98.

Sprouse, J. 2011a. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43: 155–167.

Sprouse, J. 2011b. A test of the cognitive assumptions of Magnitude Estimation: Commutativity does not hold for acceptability judgments. *Language* 87: 274–288.

Sprouse, J. and Almeida, D. 2012. The role of experimental syntax in an integrated cognitive science of language. *The Cambridge Handbook of Biolinguistics*, Kleanthes Grohmann and Cedric Boeckx (eds.).

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

Sprouse, J. and Almeida, D. (submitted). Power in acceptability judgment experiments.

Sprouse, J., Schütze, C.T. and Almeida, D. (submitted). Assessing the reliability of journal data in syntax: Linguistic Inquiry 2001–2010.

Stevens, S. S. 1956. The direct estimation of sensory magnitudes: Loudness. *American Journal of Psychology* 69: 1–25.

Stevens, S. S. 1957. On the psychophysical law. *Psychological Review* 64: 153–181.

Stowe, L. and Kaan, E. 2001. Developing an experiment. Ms., Rijksuniversiteit Groningen and the University of Florida.

Toribio, Almeida Jacqueline 2001. Accessing Spanish-English code-switching competence. *International Journal of Bilingualism* 5: 403–436.

Valian, V. 1982. Psycholinguistic experiment and linguistic intuition. In Simon, T. W. and Scholes, R. J. (eds.), *Language, mind, and brain*, pp. 179–188. Hillsdale, NJ: Lawrence Erlbaum.

Wasow, T. and Arnold, J. 2005. Intuitions in linguistic argumentation. *Lingua* 115: 1481–1496.

Weskott, T. and Fanselow, G. 2011. On the Informativity of Different Measures of Linguistic Acceptability. *Language* 87: 249–273.