Syntactic judgment experiments

James Myers
Graduate Institute of Linguistics
National Chung Cheng University

Lngmyers *at* ccu.edu.tw
September 19, 2008

Later published as:
Myers, J. (2009). Syntactic judgment experiments. *Language & Linguistics Compass, 3* (1), 406-423. doi:10.1111/j.1749-818X.2008.00113.x

Abstract

Informal judgments of sentence acceptability have long been the primary source of evidence about grammaticality in syntax, and have been controversial just as long. In the past decade there has been growing interest in collecting and analyzing acceptability judgments according to the formal protocols of experimental psycholinguistics, an approach sometimes called experimental syntax. This article reviews the major issues relevant to this approach, namely the relative reliability, validity, sensitivity, and convenience of formal vs. informal methods.

1. From informal judgments to formal judgment experimentation

Theoretical syntacticians typically test empirical claims with what may be called informal judgments: they create one or two example sentences and decide (perhaps with the help of a colleague or two) whether they seem acceptable in their native language. Informal judgments inspire horror in many scholars, since they so clearly violate the methodological protocols standard in the rest of the empirical cognitive sciences (cognitive psychology and neuroscience). The procedure seems to be a kind of psycholinguistic experiment, with the sentences as stimuli and the judgments as responses (as noted even by critics like Labov 1975), but too few sentences and speakers are tested to assess generality, the experimental participant is usually the linguist him- or herself (risking unrepresentativeness and experimenter bias), and judgments are usually made on a binary good/bad scale (despite the general recognition that sentence acceptability varies gradiently; Chomsky 1965: 10-1). Moreover, it's not clear how acceptability judgments relate to other data sources like corpora or processing experiments, let alone grammaticality itself (i.e. the status of a sentence accorded by the internal mental grammar).

Linguists have been debating the role of acceptability judgments in syntax ever since Chomsky (1957) put judgments at the center of his methodology. In a highly influential review of these debates, Schütze (1996) concludes that syntacticians must learn to collect and analyze judgments in proper psycholinguistic fashion. Today, more attention is being paid to linguistic judgments than ever before. A crude measure of this growing interest is shown in Figure 1, which plots the per-year percentages (1973-2007) of abstracts in the CSA Linguistics and Language Behavior Abstracts database mentioning "acceptability/grammaticality judg(e)ment(s)" or "judg(e)ment(s) of acceptability/grammaticality".

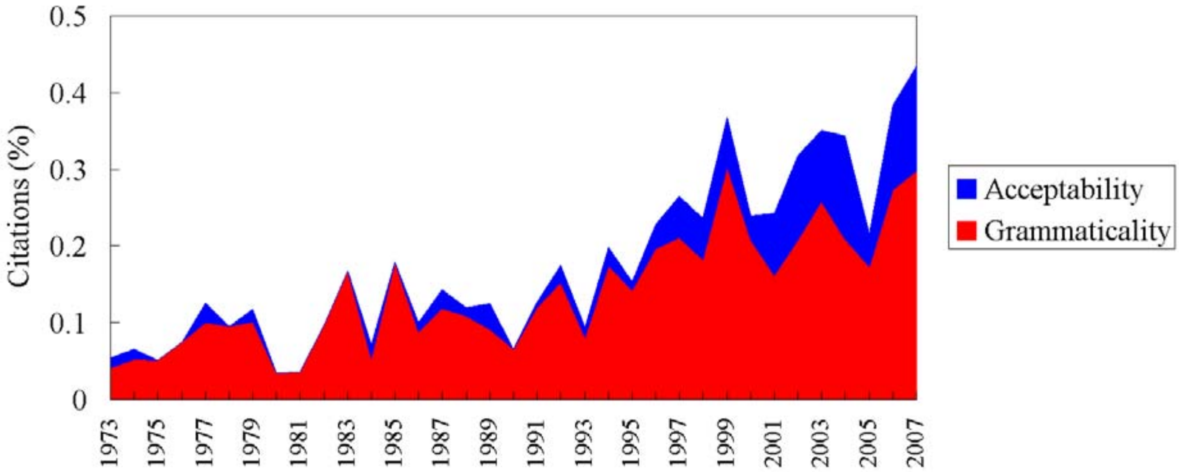

Figure 1. Percentage of abstracts in LLBA database referring to acceptability or grammaticality judgments (see text for details).

There has recently also been a shift in focus from mere criticism of informal judgments (and ad hoc experimental debunking) to positive action towards methodological reform. The first prominent example of this approach was Cowart (1997), still the best hands-on guide to formal judgment experimentation (often called, after Cowart's book, experimental syntax). In the past decade numerous studies grounding theoretical syntax on formal judgment experiments have appeared in journals (e.g. McDaniel and Cowart 1999; Bernstein, Cowart, and McDaniel 1999; Keller and Alexopoulou 2001; Sorace and Keller 2005; Featherston 2005a,b, 2007; Francis and Matthews 2006; Alexopoulou and Keller 2007; Sprouse 2007b), dissertations (e.g. Keller 2000; Hiramatsu 2000; Braze 2002; Whitman 2002; Kovalik 2004; Sprouse 2007a), and anthologies (e.g. Kepser and Reis 2005; Featherston and Sternefeld 2007).

All sciences trend towards greater empirical rigor. Before formal judgment experimentation becomes widely adopted by theoretical syntacticians, however, two major challenges must be faced. The first is the view among many linguists and cognitive scientists that judgments, whether informal or formal, are inherently inferior to data from corpora (e.g. Labov 1975, 1996; Sampson 2007) or processing experiments (e.g. Edelman and Christiansen 2003; Bornkessel-Schlesewsky and Schlesewsky 2007). The second challenge is the contrary view among other linguists that informal judgments generally already provide a sufficiently rich source of data for grammatical theory, reducing the urgency of methodological reform (see e.g. Chomsky 1965: 20-1; Phillips forthcoming).

Both views make sense; judgments should be supplemented by alternative data sources, and informal judgments have earned a legitimate place in the syntactician's toolbox, as will be shown below. Nevertheless, the growing enthusiasm for formal judgment experimentation makes sense too. The remainder of this review explains why, organizing the discussion around four key issues: reliability (and its assessment), validity (whether judgments actually reflect grammar rather than something else), sensitivity (whether informal judgments miss important information about grammar), and convenience (of formal vs. informal methods).

2. Reliability

Contrary to what many critics have assumed, informal judgments do, in fact, tend to be reliable (replicable). The real problem is that unlike formally collected judgments, their reliability cannot be assessed.

2.1 Can informal judgments be trusted?

Note first that informal judgments are not a Chomskyan invention; for centuries linguists have taken their reliability (and validity) for granted. Householder (1973) traces the citation of unacceptable sentences back almost two thousand years to Greek grammarian Apollonius Dyscolus, and informal judgments are also a standard tool in fieldwork (Henry 2005) and in the preparation of psycholinguistic stimuli (sometimes formal judgments are also used, as in Kazanina et al. 2007). As Newmeyer (1983, 2007) observes, even advocates of methodological reform may use informal judgments in their own research; Schütze (2001) is an example. The general reliability of judgments is also admitted by Labov (1975, 1996) in otherwise highly critical reviews.

One positive argument for the reliability of informal judgments is that they can be immediately checked by any native speaker (Phillips and Wagers 2007: 740-1); by contrast, it is rare for psycholinguistic experiments to be replicated exactly. Unsurprisingly, then, when informal judgments are tested in formal experiments they are often reconfirmed. Examples include the *that*-trace effect in English (Chomsky and Lasnik 1977) tested by Cowart (1997), the Superiority Condition in English (Chomsky 1973) tested by Clifton, Fanselow, and Frazier (2006), and the various Chinese judgment claims made in Li (1998), tested by Myers (forthcoming).

However, formal experiments do not always confirm informal judgments. Examples include the putative "amnestying" of Superiority violations claimed by Kayne (1983), which was not replicated by Clifton et al. (2006), and the claimed lack of *that*-trace effects in German (e.g. Haider 1983), which were nevertheless observed experimentally by Featherston (2005a).

The general reliability of informal judgments is at best a contingent empirical fact. Perhaps English judgment claims do tend to be trustworthy, given the large number of English-speaking linguists critically examining them. But this doesn't mean that a Thai-speaking linguist could readily convince an Urdu-speaking rival that her personal judgments undermine his theory. A formal judgment experiment, with an explicit, replicable methodology, has a better chance of resolving such data disputes. This promise of objectivity (cf. the subtitle of Cowart 1997) underlies the choice of McDaniel and Cowart (1999) and Bernstein et al. (1999) to base their theoretical claims on formal judgment experiments, even though their own informal judgments point in the same direction.

2.2 Reliability and statistics

Unlike informal judgments, formal experimentation is designed with reliability in mind. This is partly because every quantitative result comes with an evaluation of its trustworthiness, usually expressed in terms of statistical significance, or the chance probability of getting a result at least as large as observed. This probability, in turn, is related to the probability of future replications (see e.g. Killeen 2005).

More fundamentally, the quantitative analysis of many data points necessarily provides a more reliable picture of latent patterns than the qualitative analysis of a few isolated data points. Featherston (2007) and Myers (2007) independently observe that the "ideal speaker-listener" of Chomsky (1965: 3), long a target of scorn from critics of traditional syntactic methodology, actually exists, but only after averaging across multiple speakers and sorting out random variation from the systematic.

The point that true reliability only emerges in the aggregate is made particularly dramatically by Cowart (1997: 26-7, 33-5) (Featherston 2007: 286-9 gives a similar argument). Cowart's *that*-trace experiments involved sentences like those in (1). From the theoretical literature we expect sentence (1c), which contains a sequence of *that* and the gap (trace) left by the fronted *wh*-element, to be judged as less acceptable than the other three (diacritics like * and ? are avoided throughout this review, due to ambiguity over what they are intended to express; see section 3).

(1) a. Who was the nurse imagining _ would find her?
b. Who was the nurse imagining she would find _ ?
c. Who was the nurse imagining that _ would find her?
d. Who was the nurse imagining that she would find _?

Twenty sets of sentences showing the same contrasts were judged by three large groups of American college students on a numerical scale, where higher values indicate greater subjective acceptability. The results are shown in Figure 2 (results for (1) are in cell 1).

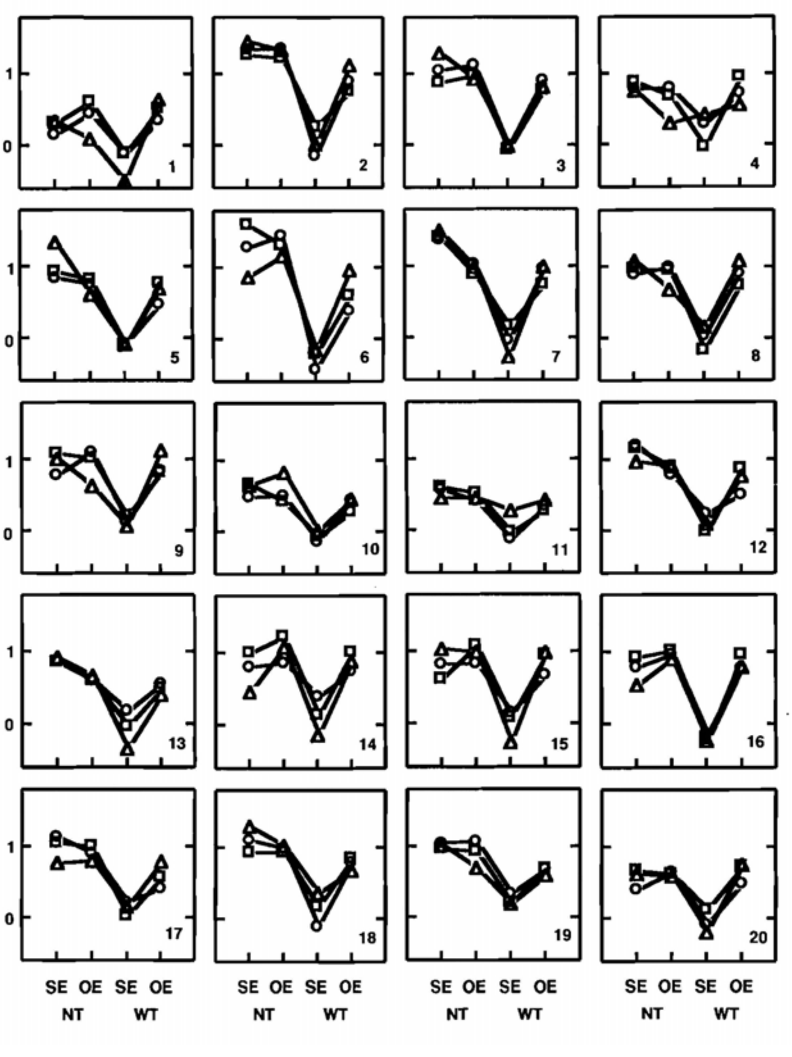

Figure 2. Judgments in a *that*-trace experiment. Each numbered cell represents a different set of four sentences: NT = no *that*, WT = with *that*, SE = subject extraction, OE = object extraction. Judgments have been standardized to have a mean of zero and standard deviation of one. The three lines in each cell represent different American universities. (Adapted from Cowart 1997: 164, figure 36; copyright 1997 by Sage Publications, Inc., reprinted by permission of Sage Publications, Inc., via Copyright Clearance Center.)
[NOTE: They didn't know I was going to post this preprint online, so don't tell them, OK?]

What is immediately striking about this figure is that it reflects a great deal of apparently random variability, across both sentence sets and participant groups, and yet simultaneously also shows clear systematicity, in that the lines virtually always dip in the same place. These dips correspond to *that*-trace violations like (1c), correctly predicted to receive the lowest acceptability scores.

Since theoretical linguists tend to be unfamiliar with the notion that systematicity can emerge out of apparent randomness, they often take cross-speaker variability to reflect systematic idiolects. The problem is that there is virtually no evidence that judgment idiolects actually exist (Labov 1975, 1996; Featherston 2007; though cf. Han, Lidz, and Musolino 2007). Instead, within a speech community, variability across speakers in a judgment experiment can be considered as essentially random (at least until proven otherwise), due not to grammar itself but to noise in the judgment-making process.

Curiously, theoretical linguists often deal with variability across sentence sets in a very different way, by attempting to select just the clearest cases. Thus if skeptics note that the sentences in set 11 in Figure 2 (with *hear* as the main verb) show a rather weak judgment contrast, a typical response would be to dismiss such examples and point instead to clearer cases like set 2 (*suppose*). Yet this search for clear cases is a textbook example of confirmation bias (Nickerson 1998). After all, skeptics could just as easily point to the contradictory judgments in set 4 (*announce*). A more appropriate strategy is to create as heterogeneous a sample of sentences (fitting the experiment's design) as is feasible. Such a sample won't be truly random, but heterogeneity should reduce the risk of selection bias. Moreover, publishing the selection procedure and the entire sentence sample will make it possible for readers to evaluate representativeness for themselves.

In short, while informal judgments often happen to be reliable, only sampling and statistical analysis allow reliability to be assessed.

3. Validity

Informal judgments also have a legitimate claim to validity. While judgments reflect many factors in addition to grammar, theoretical linguists generally understand the importance of proper experimental design in dealing with them. Moreover, judgments provide an important complement to data sources like corpora and processing experiments. The main threats to validity come from the very informality of informal judgments; designs are applied inconsistently, and handling certain variables requires sampling and other more sophisticated procedures.

3.1 A methodological truism

Phillips and Lasnik (2003: 61) correctly note that "it is a truism in linguistics, widely acknowledged and taken into account, that acceptability ratings can vary for many reasons independent of grammaticality." However, the devil is in the details. First there is the key distinction between acceptability, which is a (report of a) consciously accessible "sensation" (Schütze 1996: 52), and grammaticality, which is not directly accessible to consciousness (see Nisbett and Wilson 1977 for reasons to mistrust introspection into internal mental processes). The banality of the acceptability/grammaticality distinction, which merely exemplifies the distinction between the behavioral and the mental, has been repeatedly emphasized (e.g. Chomsky 1965: 11-2; Newmeyer 1983: 50-3; Schütze 1996: 20-34).

Unfortunately, many linguists, even those who should know better, persist in using the misleading phrase "grammaticality judgment" (e.g. Schütze 1996; Featherston 2007; Penke and Rosenbach 2004). As Newmeyer (2007) points out, a grammaticality judgment, strictly speaking, is a theoretical claim, not evidence at all; for this reason, Wasow and Arnold (2005) call such judgments mere secondary intuitions. A similar conceptual blurring happens with judgment diacritics (e.g. * vs. ?), which sometimes represent acceptability and sometimes grammaticality (diacritics can also influence how readers judge the acceptability of example sentences, casting doubt on replicability; Luka 1998). Fortunately, it seems that "acceptability judgment" is slowly becoming more common in the literature, as reflected in Figure 1.

Now consider another part of the truism, that acceptability is affected by many factors other than grammaticality. Potential extra-grammatical influences include sentence parsing (Chomsky 1965: 10-5), superficial analogy (Chomsky 1970: 27-9), discourse context (Newmeyer 1983: 55-7), and lexical semantics (Newmeyer 1998: 189-90). Often the influence of non-grammar on acceptability is negative, as in the presumably grammatical but difficult to parse center-embedded structure in (2) (Chomsky 1965: 11).

(2) The man who the boy who the students recognized pointed out is a friend of mine.

However, non-grammar can also improve acceptability, at least temporarily. A famous example is given in (3) (from Montalbetti 1984: 6), which sounds fine until one realizes that it is meaningless (for other examples see Bever, Carroll, and Hurtig 1976; Gibson and Thomas 1999; Cann, Kaplan, and Kempson 2005).

(3) More people have been to Berlin than I have.

Since the effects of non-grammar on acceptability may potentially go in either direction, distinguishing non-grammar from grammar is partly a matter of definition, and therefore not resolvable through empirical evidence alone (as Pinker 1994: 347 notes in another context, "This is a debate for dictionary-writers, not scientists"). This problem is not unique to linguistics; all concepts shift as we learn more about the world (Churchland 2002: 129-34). The truly scientific goal is to figure out how the world naturally carves up, whatever we end up calling the pieces.

The fact that acceptability is not identical with grammaticality has sometimes been held to be a serious weakness of judgments (e.g. Edelman and Christiansen 2003: 60 cite their "meta-cognitive overtones"). However, there is no behavioral task that can't be challenged in a similar way. The lexical decision task, the psycholinguistic workhorse in which participants decide if a stimulus is a real word, is just as artificial (see critiques by Henderson 1989 and Hung, Tzeng, and Ho 1999), yet it has yielded decades of reliable results in support of coherent and apparently insightful models (the closest we can get to validity in science).

Nevertheless, it is true that psycholinguists rarely rely so heavily on a single method; true validity cannot be ensured unless grammar is disentangled from the judgment task itself (as Bever 1970 and many others have noted). Thus it is necessary to consider briefly the major alternatives to acceptability judgments: corpora and processing experiments.

Corpora record grammatical knowledge as realized in language production; thus they provide no information about the parsing and comprehension system (which judgments tap into). It is true that it is possible for speakers to judge a construction as unacceptable, yet go on to produce it themselves (Labov 1975, 1996). Yet if non-grammar can push judgments in both directions, it's not obvious why speakers cannot also systematically produce utterances that are, technically, ungrammatical (whatever "grammar" turns out to be). Thus judgments and corpora complement each other, rather than one source being intrinsically superior to the other (see e.g. Hoffman 2006).

Processing experiments can play a similarly complementary role. Since grammar is presumably not a fleeting effect best caught on the fly, non-speeded tasks seem to be a reasonable first choice (see Derwing and de Almeida forthcoming for non-speeded alternatives to the standard judgment task). Nevertheless, grammar is revealed in speeded (reaction time) tasks as well, helping to shed additional light on its nature. For example, Phillips and Wagers (2007) review data suggesting that the sentence processor ignores logically possible, but ungrammatical, parses. Formal experimentation also makes it possible to study the judgment-making process itself, using non-speeded judgments (e.g. Nagata 2003; Luka and Barsalou 2005; Sprouse forthcoming a, b), speeded judgments (e.g. Garnham, Oakhill, and Cain 1998), and brain imaging (e.g. Bornkessel-Schlesewsky and Schlesewsky 2007).

3.2 Taking non-grammar into account

Returning now to the truism of Phillips and Lasnik (2003), consider finally how theoretical linguists "take into account" extra-grammatical influences on acceptability judgments. Psychologists and other scientists deal with such "nuisance" variables through the use of carefully constructed experimental designs, and, surprisingly perhaps, so do theoretical linguists. In particular, linguists are familiar with the key notion that experiments are inherently comparative (Phillips and Lasnik 2003; Myers forthcoming): given that gradient acceptability provides no absolute benchmarks (Cowart 1997; Featherston 2007), judgments are only interpretable in terms of contrasting sets like (1).

The logic of experimental design offers three general techniques for dealing with nuisance variables: matching, factoring, and sampling. Informal judgment methodology uses the first two, though not always consistently, but is inherently incapable of using the third.

Linguists use the matching technique when choosing stimulus sets like (1), where sentences are essentially identical except for the factor(s) of theoretical interest. In particular, they are matched on lexical content (unless itself theoretically relevant) and discourse context (usually, none). While this logic is sound, it is not always applied consistently. Linguists do sometimes cite isolated judgments, relying on an absolute acceptability-grammaticality yardstick that doesn't exist (Myers forthcoming cites examples of this sort in Li 1998). Even when making contrasts, linguists also tend to overuse minimal pairs, probably because the notion is so familiar from phonology, even though crossed factors are often more appropriate. This is the case in (1), where (1a,b) and (1c,d) contrast in *that* while (1a,c) and (1b,d) contrast in extraction site. The *that*-trace effect predicts an interaction between these two factors. A direct contrast between (1a) and (1c), for example, would not be able to disentangle a *that*-trace effect from a mere *that* effect.

Like other scientists, linguists also sometimes include nuisance variables as factors in the experimental design itself. Thus it is common, even for those who assume a sharp distinction between syntax and pragmatics, to test sentence acceptability in different discourse contexts, as a way of highlighting alternative structure-dependent readings (an example can be seen in Li 1998: 694, fn 3).

In another notable application of this technique, Chomsky (1970: 27-9) claims that (4a) is grammatical while (4b) is not. Conceding that some speakers find them equally acceptable, Chomsky then suggests that such speakers are actually forming (4b) "by analogy" (p. 27) with (4a).

(4) a. his criticizing the book before he read it
b. his criticism of the book before he read it

In order to disentangle grammar from the claimed effects of analogy, Chomsky adds a new experimental factor represented by the predicate context in (5a,b), where gerundive nominals like *his criticizing the book* are forbidden (presumably by the grammar). The (putatively correct) prediction is that even speakers who accept (4b) will still reject (5b), presumably because the ungrammatical (5a) does not exist to be analogized to. Some bumps in the argument are being smoothed over here (see Harris 1993: 141-2 for a sharp critique), but Chomsky's factorial strategy is essentially sound.

(5) a. His criticizing the book before he read it is to be found on page 15.
b. His criticism of the book before he read it is to be found on page 15.

While matching and factoring can be appropriate tools for dealing with nuisance variables, linguists sometimes confuse them with what might be called prefiltering. This is where linguists attempt to filter out what their expert knowledge tells them is *not* grammar, before judging the acceptability of what remains. For example, linguists who assume that discourse context and lexical semantics are external to syntax proper may try to ignore them when evaluating sentence acceptability.

Prefiltering is highly problematic. Not only does it rely on the doubtful notion of "grammaticality judgments," but the requirement for expert knowledge also implies, incorrectly, that non-linguist judgments cannot be reliable. Even more seriously, prefiltering removes the only common ground one may share with an opponent, namely the acceptability judgment itself (e.g. not all syntacticians would agree that discourse and semantics should be set aside). Finally, it is logically impossible to prefilter one's own unconscious biases.

The third technique commonly used for dealing with nuisance variables is sampling, unavailable with informal methodology. Sampling improves not only reliability, as we have seen, but also validity, by reducing the confounding of experimental factors with nuisance variables, even if unknown. For example, sampling across enough speakers will tend to cause unconscious idiosyncratic biases to cancel each other out. Similarly, testing multiple sentence sets increases the probability that any lexical, semantic, or pragmatic influences imperfectly matched within each set may be washed out across them.

Even aside from their use of sampling, formal judgment experiments can deal with a greater variety of nuisance variables than can informal judgments. For example, acceptability can shift through repeated judgments of similar sentences (Luka and Barsalou 2005, Snyder 2000). The best way to disentangle this effect from truly syntactic factors is to randomize the presentation order of sentences (and perhaps also factor out order statistically, as advocated in Myers 2007, forthcoming). There is also a concern that judging sentences in matched sets may elicit task-specific strategies (e.g. imagining a judgment contrast that doesn't actually exist). The standard way to handle such problems in psycholinguistics is to break up the matched sets across different groups of participants (Cowart 1997 gives simple instructions for doing this semi-automatically).

Thus even though the informal judgment method deserves credit for taking the validity problem seriously, formal judgment experiments do a better job at separating the wheat from the chaff. Formal experimentation also helps elucidate the judgment-making process itself, which may ultimately be as important to syntax as optics is to astronomy.

4. Sensitivity

The greater sensitivity of formal judgment experiments has the potential to reveal hitherto unknown properties of grammar, though it is not yet clear which of the properties uncovered so far are best considered part of grammar per se.

4.1 Sources of sensitivity

Formal experimentation improves sensitivity not only through improved reliability and consistently applied experimental designs, but in two other ways as well. First, larger sample sizes (more speakers and sentence sets) make it easier to detect weak but consistent effects through the noise. Second, the measurement scale itself can be chosen to increase the amount of encodable information. Acceptability judgments are traditionally measured on a binary good/bad scale, but they can also be measured on ordinal (Likert) scales (e.g. 1 = impossible and 5 = perfect) or on open-ended continuous-valued scales.

Continuous-valued scales are typically considered the gold standard in the formal judgment literature (Bard, Robertson, and Sorace 1996; Cowart 1997; Sorace and Keller 2005; Featherston 2005a,b, 2007). The oldest technique for eliciting continuous-valued judgments is magnitude estimation, originally proposed by Stevens (1956) for quantifying sensations in psychophysics. In the syntactic version (introduced in Bard et al. 1996), the participant is first given a reference sentence and asked to assign to it a freely chosen number. All subsequent sentences in the experiment are given numbers representing acceptability proportional to the reference.

However, as Featherston (2008) points out, replicability is threatened in magnitude estimation experiments by the dependence on a single reference sentence. Participants also often ignore the instructions and end up using something more like an ordinal or linear scale (as already noted by Stevens 1956). For reasons like these, Featherston (2007, 2008) advocates the use of an open-ended scale anchored linearly by two reference sentences with pre-assigned scores, thus combining (it is hoped) the sensitivity of a continuous-valued scale with the simplicity of an ordinal scale.

Though continuous-valued scales are mathematically the most sensitive, it is possible to detect relatively subtle judgment contrasts via ordinal and binary scales, by aggregating across speakers (Weskott and Fanselow 2008; Myers forthcoming). In any case, not all of the patterns revealed by improved sensitivity need be theoretically important. For example, note the small (but statistically significant) difference in acceptability for no-*that* vs. *that* sentences in Figure 2 (the first vs. second pairs of data points in each cell; Cowart 1997: 19). There is also a difference in acceptability between subject and object extraction in no-*that* sentences (first vs. second data points), though it is so weak that it only becomes statistically significant in a sample of over 1,100 speakers (Cowart 1997: 123). While these facts were hitherto unknown, it is not immediately obvious what they teach us about grammar per se (though see below).

4.2 Uses of sensitivity

Other subtle judgment patterns have a stronger claim to theoretical relevance, however. Take German *that*-trace effects, which, as noted earlier, had been thought not to exist prior to experimental work by Featherston (2005a). Informal judgments were apparently unable to detect an effect of extraction site in *that* sentences; that is, German sentences parallel to the English (1c) and (1d) seemed equally bad. Nevertheless, when tested in a magnitude estimation experiment, German *that*-trace violations like (1c) consistently received the lowest acceptability scores among sentences parallel to (1). This was so despite sentences like (1d) also receiving below-average scores.

Subtle properties of judgments can also be exploited as diagnostic tools. Featherston (2007) illustrates this in a study in which two German constructions were tested in a variety of predicate contexts. Gradient variation in acceptability for the two constructions tracked each other quite closely (when one was slightly higher or lower, so was the other), suggesting that the constructions are related.

Another subtle property of judgments that has been explored in formal experimentation is the shifting in judgments over the course of an experiment, alluded to earlier. Such shifting was first demonstrated experimentally by Snyder (2000), who suggested that the increased probability of judging ungrammatical sentences as acceptable, a phenomenon he called syntactic satiation, can be used as a diagnostic tool. On the one hand, since grammar itself doesn't shift so quickly, satiation may help distinguish grammar from non-grammar (see Goodall 2005). On the other hand, though not itself a product of grammar, satiation may interact differently with different grammatical components (see Hiramatsu 2000). Recently, however, Sprouse (forthcoming b) has argued that satiation is merely an artifact of the binary judgment scale used in earlier studies. Judgment shifts have been observed with ordinal (Luka and Barsalou 2005) and continuous-valued scales (Ko 2007), but so far only in grammatical sentences.

Based on sensitive judgment experiments, Sorace and Keller (2005) distinguish between hard constraints, which elicit very low acceptability scores when violated, and soft constraints, violations of which elicit only moderately low acceptability. Soft constraints, they claim, are also more sensitive to context (and thus perhaps also more prone to satiation, though this link isn't made in their work).

4.3 Gradient acceptability vs. gradient grammar

Several advocates of formal judgment experimentation have argued that gradient acceptability is best modeled via an inherently gradient grammar. This is the position taken in Bard et al. (1996), Keller (2000), Sorace and Keller (2005), and Featherston (2005a,b, 2007); a formal model of gradient grammar is illustrated in Keller (2000) and Sorace and Keller (2005).

However, at least some of the time, gradient acceptability effects must result from processing, not grammar, as in judgment shifting. Gradient effects may potentially also arise through the interaction between processing and a categorical grammar; this is the logic underlying the experiments in Hiramatsu (2000), and Alexopoulou and Keller (2007) seem to hold a similar position. While syntactic constraints may have different strengths in different languages (as noted above, *that*-trace effects are stronger than *that* effects in English, while the reverse holds in German), Alexopoulou and Keller (2007) argue that such differences can be analyzed as side effects of structural differences between grammars. Sprouse (2007b) even argues that acceptability itself may be less gradient than often assumed.

Nevertheless, as noted earlier, debates over the nature of grammar are partially definitional (e.g. the distinction between hard and soft constraints may prove useful even if not all such constraints are truly "in grammar"). Regardless of one's views on the nature of grammar, the improved sensitivity of formal judgment experiments reveals hitherto unknown generalizations and raises hitherto unaskable questions.

5. Convenience, and the future

Linguists have never denied that their methods could be improved (e.g. Chomsky 1965: 18-21; Newmeyer 1983: 66-7). Rather, the main argument for the methodological status quo has always been that the benefits of formal experimentation are not (yet) offset by the decrease in convenience. Hence another way to advance methodological reform is to simplify it.

Convenience is a powerful force. Phillips and Lasnik (2003: 61) credit the informality of informal judgments with "mak[ing] it possible to obtain large numbers of highly robust empirical results in a short period of time, from a vast array of languages"; Labov (1975: 81) and Cowart (1997: 1-2) agree. Yet the inconvenience of formal judgment experimentation has been exaggerated, ironically by the reformers themselves, who focus too much on the differences from informal judgments and too little on the similarities.

As highlighted above, informal judgment methodology already tends to follow key principles of experimental psycholinguistics, including the distinction between behavior and mind, the use of contrastive conditions, the matching of items on nuisance variables, and factorial logic. Aside from a certain lack of consistency, the only "essential" element (Featherston 2007: 282) still missing is sampling across speakers and sentence sets, necessary to provide reliability assessments and to control unknown threats to validity. Fortunately, simplifying sampling and statistics is possible through education, automation, and small-sample techniques.

Syntacticians can educate themselves by examining the experimental studies cited above, and especially by perusing Cowart (1997); judgment experiments are also discussed in the introductory statistics text by Johnson (2008). Software automating the more tedious aspects of experimental design and analysis is readily available; Cowart's book shows how to exploit Microsoft Excel for this purpose, and Johnson's book teaches R, the free statistics package fast becoming the de facto standard in quantitative linguistics (R Development Core Team 2008; Baayen 2008). Cowart (1997: 50-1) also recommends creating unbiased sentence samples by selecting verbs from a thesaurus (electronic or otherwise) and building sentences around them. Free software for running experiments includes WebExp, widely used for collecting judgments over the Web (Mayo, Corley and Keller 2005), and MiniJudge, which guides the novice through experimental design and analysis (Myers 2007, forthcoming).

Syntacticians are already accustomed to eliciting binary good/bad judgments from a smattering of native speakers. Although formal experiments traditionally involve large sample sizes and numerical judgment scales, neither are necessary to achieve statistical significance; as Phillips and Wagers (2007: 741) point out, "unbiased agreement among half a dozen friends will do." Myers (forthcoming) gives explicit instructions for running small-sample tests on binary judgments collected in a variety of simple experimental designs.

Naturally, even with automation, formal judgment experimentation slows down syntactic research compared with its traditional breakneck pace. Yet as software improves and a new generation of syntacticians start their careers by learning not just theory but basic empirical methods as well, it seems inevitable that formal experimentation will become a familiar tool in the theoretician's toolbox, with the extra effort more than fully repaid by the increase in reliability, validity, and sensitivity of judgment data.

Works Cited

Alexopoulou, Theodora, and Frank Keller. 2007. Locality, cyclicity and resumption: at the interface between the grammar and the human sentence processor. Language 83.110-60.

Baayen, R. H. 2008. Analyzing linguistic data: A practical introduction to statistics. Cambridge, UK: Cambridge University Press.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. Language 72.32-68.

Bernstein, Judy B., Wayne Cowart, and Dana McDaniel. 1999. Bare singular effects in genitive constructions. Linguistic Inquiry 30.493-502.

Bever, Thomas G. 1970. The cognitive basis for linguistic structures. Cognition and the development of language, ed. by John R. Hayes, 279-362. New York: John Wiley and Sons.

---., John M. Carroll, and Richard Hurtig. 1976. Analogy; or, Ungrammatical sequences that are utterable and comprehensible are the origins of new grammars in language acquisition and linguistic evolution. An integrated theory of linguistic ability, ed. by Thomas G. Bever, Jerrold Katz, and D. Terence Langendoen, 149-82. New York: Crowell.

Bornkessel-Schlesewsky, Ina, and Matthias Schlesewsky. 2007. The wolf in sheep's clothing: against a new judgement-driven imperialism. Theoretical Linguistics 33.319-33.

Braze, Forrest David. 2002. Grammaticality, acceptability and sentence processing: a psycholinguistic study. Storrs: University of Connecticut dissertation.

Cann, Ronnie, Tami Kaplan, and Ruth Kempson. 2005. Data at the grammar-pragmatics interface: the case of resumptive pronouns in English. Lingua 115.1551-77.

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

---. 1965. Aspects of the theory of syntax. Cambridge, MA: MIT Press.

---. 1970. Remarks on nominalization. Reprinted 1972 in Studies on semantics in Generative Grammar, 11-61. The Hague: Mouton. [Pagination in the text refers to the 1972 reprint.]

---. 1973. Conditions on transformations. A Festschrift for Morris Halle, ed. by Stephen Anderson and Paul Kiparsky, 232-86. New York: Holt.

---., and Howard Lasnik. 1977. Filters and control. Linguistic Inquiry 8.425-504.

Churchland, Patricia S. 2002. Brain-wise: studies in neurophilosophy. Cambridge, MA: MIT Press.

Clifton, Charles, Jr., Gisbert Fanselow, and Lyn Frazier. 2006. Amnestying superiority violations: processing multiple questions. Linguistic Inquiry 37.51-68.

Cowart, Wayne. 1997. Experimental syntax: applying objective methods to sentence judgments. London: Sage Publications.

Derwing, Bruce L, and Roberto G. de Almeida. Forthcoming. Non-chronometric experiments in linguistics. Experimental and quantitative linguistics, ed. by David Eddington. Munich: Lincom.

Edelman, Shimon, and Morten H. Christiansen. 2003. How seriously should we take Minimalist syntax? Trends in Cognitive Science 7.60-61.

Featherston, Sam, 2005a. That-trace in German. Lingua 115.1277-302.

---. 2005b. Magnitude estimation and what it can do for your syntax: some wh-constraints in German. Lingua 115.1525-50.

---. 2007. Data in generative grammar: the stick and the carrot. Theoretical Linguistics 33.269-318.

---. 2008. A standard scale of well-formedness: why syntax needs boiling and freezing points. Paper presented at the International Conference on Linguistic Evidence 2008, Tübingen, Germany.

---., and Wolfgang Sternefeld (eds.) 2007. Roots: linguistics in search of its evidential base. Berlin: Mouton de Gruyter.

Francis, Elaine J., and Stephen Matthews. 2006. Categoriality and object extraction in Cantonese serial verb constructions. Natural Language and Linguistic Theory 24.751-801.

Garnham, Alan, Jane Oakhill, and Kate Cain. 1998. Selective retention of information about the superficial form of text: ellipses with antecedents in main and subordinate clauses. The Quarterly Journal of Experimental Psychology 51A.19-39.

Gibson, Edward, and James Thomas. 1999. Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. Language and Cognitive Processes 14.225-48.

Goodall, Grant. 2005. The limits of syntax in inversion. Chicago Linguistics Society 41.

Haider, Hubert. 1983. Connectedness effects in German. Groninger Arbeiten zur Germanischen Linguistik 23.82-119.

Han, Chung-hye, Jeffrey Lidz, and Julien Musolino. 2007. V-raising and grammar competition in Korean: evidence from negation and quantifier scope. Linguistic Inquiry 38.1-47.

Harris, Randy Allen. 1993. The linguistics wars. Oxford: Oxford University Press.

Henderson, Leslie. 1989. On mental representation of morphology and its diagnosis by measures of visual access speed. Lexical representation and process, ed. by William D. Marslen-Wilson, 357-91. Cambridge, MA: MIT Press.

Henry, Alison. 2005. Non-standard dialects and linguistic data. Lingua 115.1599-617.

Hiramatsu, Kazuko. 2000. Accessing linguistic competence: evidence from children's and adults' acceptability judgments. Storrs: University of Connecticut dissertation.

Hoffmann, Thomas. 2006. Corpora and introspection as corroborating evidence: the case of preposition placement in English relative clauses. Corpus Linguistics and Linguistic

Theory 2.165-95.

Householder, Fred W. Jr. 1973. On arguments from asterisks. Foundations of Language l0.365-76.

Hung, Daisy L., Ovid J. L. Tzeng, and Ho, Chia-yun. 1999. Word superiority effect in the visual processing of Chinese. Journal of Chinese Linguistics Monograph Series 13.61-95.

Johnson, Keith. 2008. Quantitative methods in linguistics. Oxford, UK: Blackwell Publishers.

Kazanina, Nina, Ellen F. Lau, Moti Lieberman, Masaya Yoshida, and Colin Phillips. 2007. The effect of syntactic constraints on the processing of backwards anaphora. Journal of Memory and Language 56.384-409.

Kayne, Richard S., 1983. Connectedness. Linguistic Inquiry 14.223-49.

Keller, Frank. 2000. Gradience in grammar: experimental and computational aspects of degrees of grammaticality. Edinburgh: University of Edinburgh dissertation.

---., and Theodora Alexopoulou. 2001. Phonology competes with syntax: experimental evidence for the interaction of word order and accent placement in the realization of information structure. Cognition 79.301-72.

Kepser, Stephan, and Marga Reis. 2005. Linguistic evidence: empirical, theoretical and computational perspectives. Berlin: Mouton de Gruyter.

Killeen, Peter R. 2005. An alternative to null-hypothesis significance tests. Psychological Science 16.345-353.

Ko, Yuguang. 2007. Grammaticality and parsibility in Mandarin syntactic judgment experiments. Unpublished National Chung Cheng University master's thesis.

Kovalik, Ludovic Mihai. 2004. Acceptability assessments of elliptical sentences in context. Stillwater: Oklahoma State University dissertation.

Labov, William. 1975. Empirical foundations of linguistic theory. The scope of American linguistics, ed. by Robert Austerlitz, 77-133. Lisse: Peter de Ridder.

---. 1996. When intuitions fail. Chicago Linguistics Society 32.77-105.

Li, Yen-Hui Audrey. 1998. Argument determiner phrases and number phrases. Linguistic Inquiry 29.693-702.

Luka, Barbara J. 1998. Example sentences in syntax articles: the influence of asterisks and affinity towards author. Chicago Linguistics Society 34.269-280.

---., and Lawrence W. Barsalou. 2005. Structural facilitation: mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. Journal of Memory and Language 52.436-59.

Mayo, Neil, Martin Corley, and Frank Keller. 2005. WebExp2 experimenter's manual. University of Edinburgh ms. <www.webexp.info/>

McDaniel, Dana, and Wayne Cowart. 1999. Experimental evidence for a minimalist account of English resumptive pronouns. Cognition 70.B15-B24.

Montalbetti, Mario M. 1984. After binding: on the interpretation of pronouns. Cambridge, MA: MIT dissertation.

Myers, James. 2007. Generative morphology as psycholinguistics. The mental lexicon: core perspectives, ed. by Gonia Jarema and Gary Libben, 105-28. Amsterdam: Elsevier.

---. Forthcoming. The design and analysis of small-scale syntactic judgment experiments. Lingua.

Nagata, Hiroshi. 2003. Judgments of grammaticality of Japanese sentences violating the principle of Full Interpretation. Journal of Psycholinguistic Research 32.693-709.

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

---. 1998. Language form and language function. Cambridge, MA: MIT Press.

---. 2007. Commentary on Sam Featherston, 'Data in generative grammar: The stick and the carrot'. Theoretical Linguistics 33.395-99.

Nickerson, Raymond S. 1998. Confirmation bias: a ubiquitous phenomenon in many guises. Review of General Psychology 2.175-220.

Nisbett, Richard. E., and Timothy Decamp Wilson. 1977. Telling more than we can know: verbal reports on mental processes. Psychological Review 84.231-59.

Penke, Martina, and Anette Rosenbach. 2004. What counts as evidence in linguistics? An introduction. Studies in Language 28.480-526.

Phillips, Colin. Forthcoming. Should we impeach armchair linguists? Japanese/Korean Linguistics 17, ed. by Shoichi Iwasaki. Palo Alto, CA: Center for the Study of Language and Information.

Phillips, Colin, and Howard Lasnik. 2003. Linguistics and empirical evidence: reply to Edelman and Christiansen. Trends in Cognitive Science 7.61-2.

---., and Matt Wagers. 2007. Relating structure and time in linguistics and psycholinguistics. Handbook of psycholinguistics, ed. by Gareth Gaskell, 739-56. Oxford: Oxford University Press.

Pinker, Steven. 1994. The language instinct. New York: William Morrow.

R Development Core Team. 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <www.R-project.org>.

Sampson, Geoffrey 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.

---. 2001. On the nature of default case. Syntax 4.205-38.

Snyder, William. 2000. An experimental investigation of syntactic satiation effects. Linguistic Inquiry 31.575-82.

Sorace, Antonella, and Frank Keller. 2005. Gradience in linguistic data. Lingua

115.1497-524.

Sprouse, Jon. 2007a. A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge. College Park: University of Maryland dissertation.

---. 2007b. Continuous acceptability, categorical grammaticality, and experimental syntax. Biolinguistics 1.117-28.

---. Forthcoming a. The differential sensitivity of acceptability to processing effects. Linguistic Inquiry.

---. Forthcoming b. Revising satiation: Evidence for an equalization response strategy. Linguistic Inquiry.

Stevens, S. Smith. 1956. The direct estimation of sensory magnitudes - loudness. American Journal of Psychology 69.1-25.

Wasow, Thomas, and Jennifer Arnold. 2005. Intuitions in linguistic argumentation. Lingua 115.1481-96.

Weskott, Thomas, and Gisbert Fanselow. 2008. Scaling acceptability -- different measures, same results. Paper presented at the 27th West Coast Conference on Formal Linguistics, UCLA.

Whitman, Neal P. 2002. Category neutrality: a type-logical investigation. Ohio State University PhD thesis.