Acceptability judgments and grammaticality, prospects and challenges

Jon Sprouse
Department of Linguistics
University of Connecticut

1. Acceptability judgments in Syntactic Structures

Acceptability judgments constitute a substantial portion of the empirical foundation of generative syntax. Acceptability judgments are first proposed as a proxy for grammaticalness in generative syntax in Syntactic Structures (Chomsky 1957) in the first paragraph of chapter 2:

> The grammar of L will thus be a device that generates all of the grammatical sequences of L and none of the ungrammatical ones. One way to test the adequacy of a grammar proposed for L is to determine whether or not the sequences that it generates are actually grammatical, i.e., acceptable to a native speaker, etc. We can take certain steps towards providing a behavioral criterion for grammaticalness so that this test of adequacy can be carried out. For the purposes of this discussion, however, suppose that we assume intuitive knowledge of the grammatical sentences of English and ask what sort of grammar will be able to do the job of producing these in some effective and illuminating way. (Chomsky 1957: 13)

In this quote, the precise method of collecting and interpreting acceptability judgments is left as a promissory note, so that Chomsky can get down to the core business of generative syntax, i.e., constructing a theory of grammar. My goal in this chapter is to provide a brief discussion of the prospects and challenges of using acceptability as a proxy for grammaticality, 60 years later.

This chapter is organized around two themes in current research on acceptability judgments: the methodology of acceptability judgments, and the theory of acceptability judgments. Research on judgment methodology is concerned with the act of collecting and analyzing acceptability judgments. Because methodological questions are relatively straightforward to investigate empirically (e.g., Does property X affect acceptability judgments?), it is perhaps unsurprising that generative syntacticians have made quite a bit of progress on this front over the past 60 years. Section 2 will discuss several fundamental questions about judgment methodology that have been investigated over the past 60 years. The theory of acceptability judgments is concerned with Chomsky's question of how to interpret acceptability judgments as evidence of "grammaticalness", and therefore as evidence for specific grammatical theories. Theoretical questions linking observable evidence to unobservable cognitive constructs are less straightforwardly empirical, as they involve exploring different linking hypotheses the observable and unobservable (in this case, acceptability and grammaticality). Linking hypotheses are rarely amenable to direct investigation, so progress can only be measured by the success of the theory that results from the linking hypothesis plus the empirically collected data. It goes without saying that the fundamental component of the linking hypothesis for acceptability judgments – that acceptability judgments are (relatively directly) influenced by grammaticality – has been well established by the success of the grammatical theories that have been constructed from acceptability judgments. But the answers to higher-level questions about the grammar, such as whether the grammar distinguishes two or more than

two levels of grammaticality, have remained elusive. Section 3 will discuss several higher-level questions about the theory of acceptability that are currently the focus of much research on acceptability judgments. Section 4 attempts to tie these two strands together: the past 60 years have demonstrated that acceptability judgments are a robust, replicable, and reliable data type that appears to reveal deep information about the theory of grammar; but there is still much work to be done when it comes to using acceptability judgments (and any other relevant data types from psycholinguistics) to answer higher-level questions about the theory of grammar.

2. The methodology of acceptability judgments

Methodological research on acceptability judgments only requires two assumptions to get off the ground. The first assumption is that acceptability judgments are a behavioral task just like any other behavioral task in experimental psychology. In this case, it is a task that involves explicitly asking speakers to judge whether a string of words is a possible sentence of their language (either relative to the participan't best guess at the intended meaning, or relative to an explicitly given intended meaning). The second assumption is that this particular behavioral task is influenced by the theory of grammar (though, likely not exclusively). With these two assumptions in hand, we can ask any number of questions about the impact of various methodological choices on the properties of acceptability judgment data. Here I will discuss some of the most contentious questions, and therefore some of the most interesting findings, over the past several decades of methodological work on judgments.

2.1 Differences between acceptability judgment tasks

There are any number of tasks one can use to ask participants to report acceptability judgments. To ground this first part of the discussion, I will focus on four relatively common task types in the generative syntax literature:

(i) **n-point (Likert-esque) rating scales (LS)**: participants are presented with one sentence at a time, and asked to rate the acceptability of the sentence along a scale with a finite number of ordered points (e.g., 1-7), with endpoints labeled to indicate the direction of increasing acceptability.

(ii) **two-alternative forced-choice with nominal categories (YN)**: participants are presented with one sentence at a time, and asked to categorize the sentence into one of two categories, typically labeled grammatical/ungrammatical or yes/no.

(iii) **two-alternative forced-choice comparing the sentences (FC)**: participants are presented with two sentences, and asked to indicate which of the two sentences is more (or less) acceptable.

(iv) **magnitude estimation (ME)**: participants are presented with a reference sentence (called the standard), which is assigned a numerical acceptability level (called the modulus). They are then asked to rate target sentences (one at a time) as multiples of the acceptability of a reference sentence. For example, if the reference sentence is

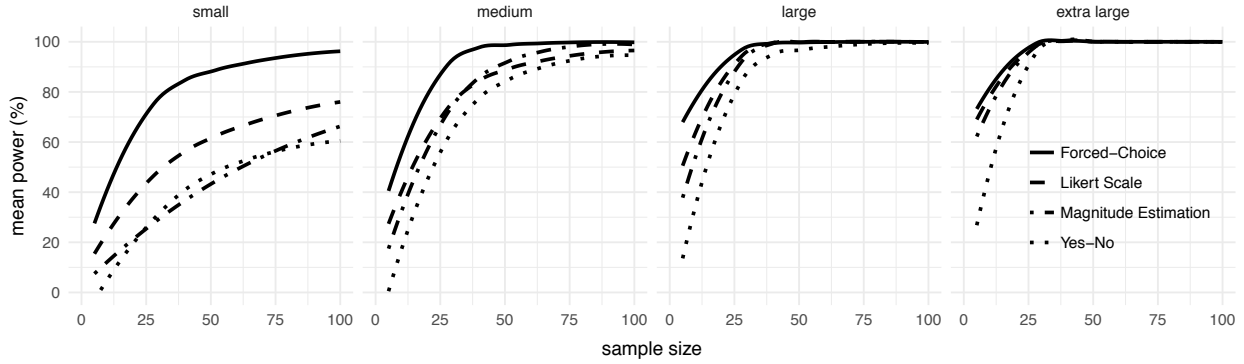
assigned an acceptability value of 100, participants might rate a target sentence that is twice as acceptable as 200.

The first question one can ask about these tasks is what kind of information each yields. The LS and ME tasks yield information about the location of a sentence along the continuum of acceptability, and therefore also yield information about the size of the difference in acceptability between two (or more) sentence types. As such, LS and ME are particularly well-suited to questions about relative acceptability. The YN task yields category information, which roughly correlates with the location of sentences along the continuum of acceptability. It is therefore particularly well-suited for questions about categorical acceptability, but ill-suited for questions about relative acceptability between sentence types that are both in the same category. It is also less well-suited for questions about effect sizes than LS and ME, because the only information it yields about the size of the difference between two sentences is the relative difference between the two category counts (number of yes's and number of no's), which is likely to be coarser-grained than scale-based information. The FC task yields information about a direct comparison between two sentences. It is therefore particularly well-suited to questions about differences between conditions. It does not yield any information about the location of the two sentences along the continuum of acceptability. And like YN, it is less well-suited to effect size questions, because the effect size information is mediated by counts. It almost goes without saying that the first criterion for choosing a task should be that it provides the type of information that will help to answer the theoretical question of interest. If the question is about relative acceptability, effect sizes, or location information, then LS and ME will likely be the best choice. If the question is about categorical information, then YN will be the best choice. If the question is about the presence of a difference between conditions, then FC will be the best choice.

The second question one can ask is how sensitive each task is to differences in acceptability between two (or more) conditions. Sprouse and Almeida 2017 investigated the sensitivity of these four tasks for 50 two-condition phenomena that span the range of effect sizes in the generative syntax literature, and for sample sizes from 5 participants up to 100 participants. They first collected ratings from 144 participants for each task, then used this real-world data to run re-sampling simulations (sampling with replacement) to estimate the proportion of (simulated) experiments that would successfully detect the difference between the two conditions in each phenomenon at each possible sample size from 5 participants to 100 participants. Figure 1 below shows the results of those re-sampling simulations. The y-axis reports the proportion of simulations that yielded a significant difference (an estimate of the statistical power of the experiment at that sample size). The x-axis reports the sample size from 5 to 100 participants. To make the plot more manageable, the phenomena were grouped into small, medium, large, and extra-large effect sizes, which are arranged by column. The colored lines report the change in the statistical power for each task as the sample size is increased. There are three very clear patterns in these results. First, the FC task is by far the most sensitive (i.e., shows the highest statistical power at smaller sample sizes and smaller effect sizes). This is not surprising given that the FC task is particularly well-suited to detecting differences between two conditions, which was exactly the definition of success in these experiments. Second, the YN task is often the least sensitive. Again, this is not surprising given how ill-suited the YN task is to detecting differences between conditions, especially in the cases where the two conditions are both in the same category (e.g., both yes or both no). Finally, the LS and ME tasks tend to track each other fairly closely for medium, large, and extra large effect sizes (but not small effect

sizes). This parallelism is potentially interesting, but in order to understand it fully, we should take a closer look at the methodological studies that led to the adoption of ME in the generative syntax literature.

Figure 1: Statistical power for four acceptability judgment tasks (adapted from Sprouse and Almeida 2017) displaying the relationship between sample size and estimated power, organized by effect size category (columns), with all four tasks plotted together. For clarity, only the (loess) fitted lines are plotted (no data points).

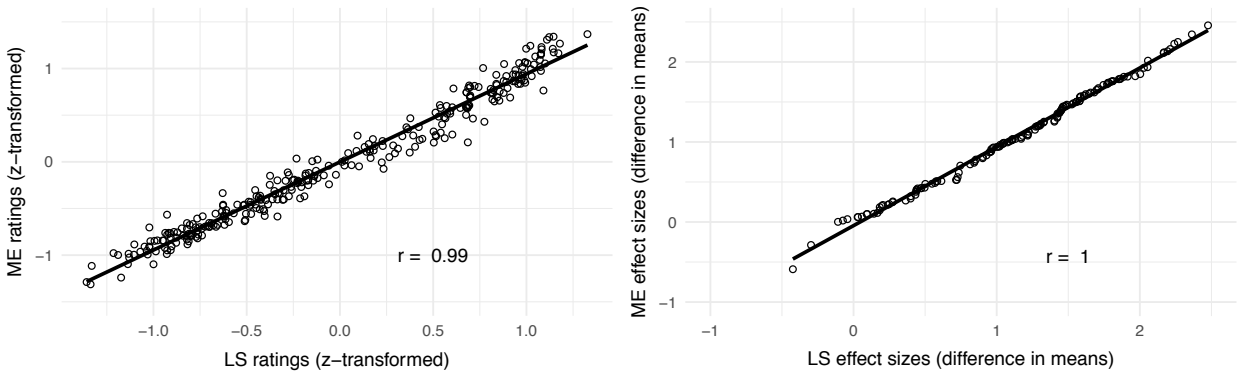


To understand the ME task it is important to first note that the LS task has two potential limitations, first discussed by Stevens (1957): (i) LS has a finite number of response options, which might lead to participants failing to report a distinction that they can nonetheless perceive; and (ii) LS assumes that participants treat the intervals between the response points as equal, but provides no mechanism to guarantee that. In an attempt to overcome these potential limitations, Stevens (1957) developed the ME task, which employs (i) a potentially infinite number of response options (any positive real number, though in practice responses are likely to be positive integers), and (ii) a reference stimulus that participants are instructed to use as a perceptual interval (thus guaranteeing that the intervals will be equal across trials within a single participant, though not necessarily between participants). Bard et al. 1996 and Cowart 1997 both observed that the LS task for acceptability judgments could potentially suffer from general LS limitations, and proposed adapting the ME task for acceptability. They both showed promising results with the ME task (as did Featherston 2005). This raised the interesting question of whether the ME task should replace the LS task as the standard rating task for acceptability judgments. Featherston 2008 was perhaps the first to question the practical reality of the theoretical superiority of the ME task, reporting that, in his experience with the task, he did not believe that participants could truly perform the ratio judgments that the ME task requires (e.g., determine that one sentence is twice as acceptable as a reference sentence). Sprouse 2011 tested this issue directly by investigating whether participants' judgments in ME were commutative (i.e., the order of two successive judgments does not matter), which is one of the fundamental mathematical assumptions of the ratio judgments in ME (Narens 1996, Luce 2002). Sprouse 2011 found that participants' judgments were not commutative, confirming Featherston's 2008 observation that participants could not make the ratio judgments that ME requires. As both Featherston 2008 and Sprouse 2011 observe, this is not unexpected given that (i) ratio judgments require a meaningful zero point, and (ii) it is not clear that acceptability has a meaningful zero point (i.e., it is not clear what it would mean to say that a sentence has zero acceptability).

Weskott and Fanselow 2011 also found that ME judgments showed higher variability than LS judgments, further undermining any claims that ME is superior to LS. It is possible that it is this increased variability that leads to ME having less statistical power than LS for small effect sizes in the Sprouse and Almeida (2017) study.

Despite the inability of participants to perform ratio judgments, it is still the case that ME judgments are robust, replicable, and reliable (as demonstrated by Bard et al. 1996, Cowart 1997, Featherston 2005, and all of the studies directly investigating the ME task). Though I know of no direct investigation of the source of ME judgments, it seems plausible that, when participants are faced with the impossible task of performing magnitude estimation of acceptability, they default to a type of LS task with an implicit scale defined by the number assigned to the reference sentence. There are at least two pieces of evidence indirectly supporting this possibility. The first is the nearly identical statistical power of the LS and ME tasks in the Sprouse and Almeida 2017 study for medium, large, and extra-large effect sizes (Figure 1 above). The second is the nearly perfect correlation between LS and ME judgments for the 300 distinct sentence types tested by Sprouse et al. 2013, both in the ratings of individual sentences, and in experimentally-defined effect sizes (the difference between a violation condition and a minimally different grammatical control condition). Though Sprouse et al. 2013 do not report that correlation in their paper, it is easily demonstrated using their data set, as in Figure 2. Under the assumption that ME becomes an LS task when used for acceptability judgments, several researchers have suggested exploring alternative tasks that preserve some of the potentially superior features of ME. Two promising tasks are Featherston's (2008) thermometer task, and rating tasks the employ a visual slider.

Figure 2: A comparison of LS and ME for individual sentence type ratings (left panel) and experimentally defined effect sizes (the difference between a violation condition and a control condition; right panel). The Pearson correlation (rounded to two decimal places) is reported in each panel. Data is from Sprouse et al. 2013.



2.2 The validity of existing acceptability judgment data

A second major topic in the methodology of judgments has focused on the potential impact of the relatively informal judgment collection methods that typify generative syntax. The fundamental concern is that informal data collection methods might lead to spurious results (e.g., false positives), either because small sample sizes and lack of statistical significance testing

might lead syntacticians to mistakenly see a signal in what is in fact noise, or because the practice of asking professional linguists for judgments might lead to effects driven by the cognitive bias of the professional linguists. This concern has arisen in one form or another since the earliest days of generative grammar (e.g., Hill 1961, Spencer 1973), it has played a central role in the two books that began the trend of employing more formal acceptability judgment experiments in the 1990s (Schütze 1996, Cowart 1997), and it has led to several high-profile discussions in the literature over the past decade and a half (see Ferreira 2005, Wasow and Arnold 2005, Featherston 2007, Gibson and Fedorenko 2013 for some criticisms of informal methods, and Marantz 2005 and Phillips 2009 for some rebuttals). To be clear, there is a straightforward method for determining whether this lack of confidence in informal methods is justified: compare the results of informal methods with the results of formal experimental methods. The results that converge between the two methods will benefit from the increase in confidence. The results that diverge can then be further investigated to determine which method is more likely giving the valid result (i.e., by manipulating the factors that give rise to concern in each method, such as the linguistic knowledge of the participants). Unfortunately, until the advent of crowdsourcing platforms like Amazon Mechanical Turk, it was nearly impossible to test the large number of phenomena that would be required to truly evaluate these concerns. The studies mentioned above do present a number of phenomena as examples of (purportedly) spurious judgments in the literature; however, in every case the phenomena were chosen with *bias* – the authors chose the phenomena because they suspected them to be spurious for one reason or another. It is impossible to estimate properties of a population from a biased sample. We simply do not know whether the dozen or so examples given in the studies mentioned above represent a small portion of the (purportedly) spurious results in the literature, or a large portion. The only way to truly address this issue is with non-biased sampling, either through the exhaustive testing of every phenomenon in a given population of phenomena, or through random sampling (which allows us to estimate a convergence rate for the population with a margin of error based on the sample and population sizes).

Two non-biased sampling studies have been conducted in English: Sprouse and Almeida 2012 tested every English data point in a recent generative syntax textbook (Adger 2003), and Sprouse et al. 2013 randomly sampled 150 two-condition phenomena from a ten-year span of a leading generative syntax journal (*Linguistic Inquiry 2001-2010*). These studies used the best practices of experimental syntax (8 items per condition, latin square designs, sample sizes over 100 participants), tested these data points using several judgment tasks (LS, ME, FC, and YN), and analyzed the results using multiple statistical methods (standard frequentist tests like t-tests and sign-tests, Bayes Factors, and mixed-effects models). Because these studies used multiple tasks and multiple types of statistical tests, the results suggest a range of convergence rates depending on the precise properties of the experiment and the precise definition of the presence of an effect: Sprouse and Almeida 2012 found that 98-100% of the data points from Adger's 2003 textbook replicated with formal experiments, and Sprouse et al. 2013 found that 86-99% of the phenomena that they randomly sampled from *Linguistic Inquiry* 2001-2010 replicated with formal experiments, suggesting an estimate of 86-99% ± 5 for the complete set of data points published in the journal during that ten-year span. To be absolutely clear, these results do not reveal which method yields the better results. These results simply quantify the difference between the two methods. We would need targeted follow-up studies that manipulate specific mechanisms that could give rise to the divergent phenomena in order to establish which method provides the more accurate results. But what we can say with these results is that the divergence

between the two methods is between 0% and 14%, depending on the population of phenomena, the judgment task (because they vary in statistical power), and the statistical test employed.

We can also use these studies to test the specific claim that the judgments of professional linguists may be impacted by theory-driven cognitive bias. An unambiguous signal of cognitive bias would be a sign reversal between the results of the formal experiments with naïve participants and the results of the informal experiments with professional linguists. Sprouse and Almeida 2012 found no sign reversals for Adger's textbook data. Sprouse et al. 2013 report a 1-3% sign-reversal rate for the *Linguistic Inquiry* data with a margin of error of ± 5 on the estimate for the population. Mahowald et al. 2016 and Häussler et al. 2016 have replicated the Sprouse et al. 2013 results without reporting an increased sign reversal rate (0-6%). Furthermore Culbertson and Gross 2009 performed direct comparisons of naïve and expert populations, and reported high inter- and intra-group correlations on 73 sentence types. Similarly, Dąbrowska 2010 found that while experts gave less variable ratings than naïve participants on several sentence types, the experts rated certain theoretically interesting syntactic violations as more *acceptable* than naïve participants, in apparent conflict with their theoretical commitments.

Taken together, these results suggest very little difference between informally collected and formally collected acceptability judgments, and very little evidence of cognitive bias influencing the judgments of (English-speaking) professional linguists. Of course, these studies all focused on one language, English, and all focused on one specific type of acceptability judgment (the type that can be given to a single written sentence, out of context, with no specific training on the intended interpretation). It is therefore logically possible that larger differences could obtain for other languages or other acceptability judgment types. But for now, the current state of evidence suggests that generative syntactic theories are built on robust, reliable, and replicable acceptability judgments, regardless of the specific method of collection.

2.3 The effect of factors other than grammaticality

A third major topic in methodology of judgments is to what extent factors other than grammaticality affect acceptability judgments. Because acceptability judgments are provided during or after the act of sentence processing, it is widely assumed that acceptability judgments will be impacted by all of the factors that influence sentence processing (complexity, ambiguity, frequency, plausibility, the disruption caused by a violation, etc.), as well as the various effects that might be caused by the judgment task itself (fatigue, repetition effects, comparison effects, etc). In other words, acceptability is assumed to be a multi-dimensional percept that is reported as a scalar value. As such, it is possible that any given acceptability effect might ultimately be due to extra-grammatical factors rather than grammaticality itself. The canonical example of this are doubly center-embedded sentences, as in (1) below. Miller and Chomsky 1963 argued that sentences like (1), which contain two relative clauses in the subject position, are unacceptable due to an inability to process the sentence, not due to a constraint in the grammar. Their primary argument was logical. They argued that grammars should not include constraints that count the number of operations that are deployed: if one relative clause can be constructed by the grammar, then two (or more) should also be able to be constructed. This analysis receives some empirical support from the fact that two (or more) relative clauses can be constructed in sequence if the relative clauses always appear in object positions (i.e., right-branching instead of center-embedded relative clauses) as in (2), and from the fact that the acceptability of doubly

center-embedded relative clauses can be increased by manipulating factors known to decrease the processing complexity of sentences as in (3).

(1) The reporter who the senator that the president insulted contacted filed the story.
(2) The president insulted the senator who contacted the reporter that filed the story.
(3) Every reporter that the senator you voted for sent a press release to managed to file a story.

Though it is logically possible that sentence processing factors could be driving the effects that generative syntacticians build into their grammars, to my knowledge, doubly center-embedded relative clauses are still the only uncontroversial example of this phenomenon. There are controversial candidates. For example, it has been proposed several times in the literature that island effects – the unacceptability that arises when the tail of a long-distance dependency is inside certain structures, such as the embedded polar question in (4) – may arise due to sentence processing complications (Deane 1991, Kluender and Kutas 1993, Hofmeister and Sag 2010).

(4) *What do you wonder whether Mary wrote __ ?

However, unlike doubly center-embedded relatives, the preponderance of evidence currently suggests that sentence processing theories cannot account for the full range of facts surrounding island effects. First, there are several properties of island effects that make any simple sentence processing based account unlikely, such as the fact that there is cross-linguistic variation in island effects (Engdahl 1982, Rizzi 1982), the fact that wh-in-situ languages like Chinese and Japanese still show a subset of island effects despite the wh-word sitting in its interpreted position (Huang 1982, Lasnik and Saito 1984), and the fact that dependencies with tails inside of island structures are grammatical (in some languages) when there is an additional gap outside of the island structure (these are called parasitic gap constructions, Endgahl 1982). Second, direct investigations of sentence processing based theories of island effects have tended to yield results that run contrary to the plausible predictions of those theories. For example, one prominent theory proposed by Kluender and Kutas 1993 is that island effects arise from limitations in working memory that prevent the parser from completing long-distance dependencies inside of island structures. One potential prediction of this theory is that the parser will not be able to complete dependencies inside of island structures. However, Phillips 2006 found that the parser could in fact complete dependencies inside of certain island structures – namely those that can host parasitic gaps – despite the fact that participants rate those sentences as unacceptable. This suggests that the unacceptability is not driven by a failure of the parser, but rather by something else, such as a constraint in the grammar. Another potential prediction is that the unacceptability of island effects will vary as a function of the working memory capacity of individual speakers. However, Sprouse et al. 2012 found that there is no correlation between working memory capacity and island effects for two types of working memory tasks and four types of island effects (a result that was replicated by Michel 2014 for additional working memory tasks). One final potential prediction is that island effects should arise for all dependencies that involve the same sentence processing mechanisms as wh-dependencies. However, Yoshida et al. 2014 demonstrated that certain (backward) binding dependencies do not respect island structures, despite the fact that those binding dependencies appear to be processed using mechanisms that are behaviorally similar to wh-dependencies (Van Gompel and Liversedge 2002, Sturt 2003,

Kazanina et al. 2007), and despite the fact that the processing of those binding dependencies involve the same cortical areas as the processing of wh-dependencies (Matchin et al. 2014). In the end, though it is logically possible that sentence processing effects could be the cause of the unacceptability for each of the sentence types that syntacticians label as ungrammatical, uncontroversial examples of sentence processing effects causing (substantial) unacceptability appear to be few and far between (with doubly center-embedded relative clauses perhaps being the only one).

In contrast to doubly center-embedded relative clauses, which suggest that sentence processing effects can substantially lower acceptability, there are also constructions that suggest that sentence processing effects can substantially increase acceptability (at least temporarily). These constructions are sometimes called *grammatical illusions*: sentences that are (widely assumed to be) ungrammatical, but are nonetheless rated as acceptable by native speakers (or at least more acceptable than one might expect). These constructions are potentially interesting as they potentially reveal the complex relationship between grammatical theories and sentence processing theories (Phillips and Lewis 2013, Lewis and Phillips 2015). However, from the point of view of detailing the effects of processing on acceptability, grammatical illusions are similar to doubly center-embedded relative clauses in that there are relatively few constructions that show this behavior. Lewis and Phillips 2015 find just three examples:

(5) More people have been to Russia than I have.
(6) The key to the cabinets are missing.
(7) The bills that no senators voted for will ever become law.

The first is the comparative illusion, first noted by Montalbetti 1984, and further explored by Townsend and Bever 2001. The comparative illusion has no meaning, and is therefore assumed to be ungrammatical (under the assumption that grammaticality entails a meaning). Yet it is nonetheless reported to be acceptable by native speakers (e.g., Wellwood et al. 2014). The second example is a phenomenon called agreement attraction. The sentence in (6) is ungrammatical: the subject of the sentence (*key*) is singular, while the verb shows plural agreement (*are*). Nonetheless, agreement attraction illusions arise in both production tasks (Bock and Miller 1991), and comprehension tasks (Wagers et al. 2009, Staub 2010). The third example is illusory negative polarity item (NPI) licensing. It is widely assumed that NPIs such as *ever* must be c-commanded by a downward entailing operator such as negation. By this assumption, the sentence in (7) should be ungrammatical: *no* is within a relative clause, and therefore does not c-command *ever*, leaving *ever* unlicensed. Nonetheless, some portion of speakers rate (7) as if it were acceptable, at least for a short while (Xiang et al. 2009, Parker and Phillips 2016). Each of these phenomena have given rise to a rich literature that goes far beyond the scope of this chapter (but see the citations above for a good starting point in each of these literatures). For our purposes, the take-home message of each of these phenomena is that sentence processing mechanisms can increase acceptability, but only in very limited cases where the implementation of grammatical constraints in an online processor creates the opportunity for errors.

The other major strand of research in the literature on extra-grammatical factors has focused on the effects of the acceptability judgment task itself. To date, there have been at least three major questions in this strand. Two of these we have covered in previous sections: the statistical power of different tasks and the consequences of naïve versus expert participant populations. The third is the effect of repetitions on acceptability judgments. Repetition effects

are a potential issue for any data type – as participants are repeatedly exposed to a stimulus, their responses to that stimulus could change (at both the behavioral and neurophysiological levels). For acceptability judgments, it has been reported that some violation types are rated higher after repeated exposures (e.g., Nagata 1988, 1989, Snyder 2000). However, instead of simply viewing repetition effects as a potential confound to be avoided in judgment experiments, repetition effects have become a special topic of interest within the acceptability judgment literature because of some initial results that suggest that repetition effects could be used as a tool to differentiate different types of phenomena, either based on categorical repetition (those that show effects and those that do not), or based on the rate of repetition effects. This is a potentially interesting tool for syntacticians to leverage; however, the empirical results of repetition studies are mixed. Whereas some studies report results that suggest potentially theoretically interesting patterns of repetition effects (e.g., Snyder 2000), attempts to replicate those results have met with mixed success (Hiramatsu 2000, Braze 2002, Sprouse 2009). The current state of evidence suggests that, to the extent that repetition effects exist, they are relatively small, and may be influenced by factors that are not relevant to grammatical theories (see, e.g., Francom 2009 for evidence that interpretability of violations may be a critical factor).

3. The theory of acceptability judgments

The primary goal of research on the theory of acceptability judgments is to determine exactly what we can (and cannot) learn about the grammar from acceptability judgments. As briefly mentioned in section 1, this requires an explicit formulation of the linking hypothesis between acceptability judgments and the theory of grammar. There is likely a lot of variation among syntacticians when it comes to beliefs about the linking hypothesis of acceptability judgments, so much so that it is impossible to do justice to all of the possible positions in a chapter of this size. Therefore my strategy in this section is to first lay out (in sections 3.1 and 3.2) a subset of the components of a linking hypothesis that I think are fairly widely assumed in the generative syntax literature (though perhaps not universally assumed), and then to use the remainder of this section to explore some of the major questions that arise in the use of acceptability judgments to make inferences about the grammar (sections 3.3 and 3.4).

3.1 Common, though perhaps not universal, assumptions about acceptability judgments

The first common assumption, and one that was implicit in the preceding sections, is that there is a percept called *acceptability* that arises for native speakers of a language during the comprehension of sentence-like word strings. Exactly how this percept arises is a matter of debate. For many, it is assumed to be an error signal of some sort. The precise mechanisms that give rise to the error signal are often left unspecified, presumably because the details of the error detection mechanisms do not (yet) impact the interpretation of acceptability judgments. Similarly, this error signal is often commonly assumed to be an automatic process (as opposed to a controlled process); though again, it is not clear how much impact the automatic/controlled distinction has on the interpretation of acceptability judgments. The second common assumption, and one that was central to the discussion in the preceding sections, is that acceptability is a scalar percept that is derived from multiple sources. This fits well with the error-signal assumption: there are multiple types of errors that can be detected, some of which can co-occur in the same sentence; these errors are then combined to form a single percept. It is also common

to assume that the combination of factors is linear, both because linear models are the simplest starting point, and because there is currently little evidence that *distinct* factors combine non-linearly. To be clear, there is some evidence that *similar* factors combine non-linearly: see Hofmeister et al. 2014 for some evidence of non-linear combination of processing factors, and Keller 2003 for evidence of non-linear combination of grammatical factors. But these could be explained as interactions of similar components, rather than a fundamental non-linearity of the system. The final assumption, which again, has figured prominently in the preceding discussions, is that the multiple sources contributing to acceptability include both grammatical factors and sentence processing factors. Though this is easy enough to state (and to encode in a general linear model), from a cognitive point of view, the fact that both grammatical and sentence processing factors influence acceptability raises difficult (but interesting) questions about what the relationship is between grammatical theories, sentence processing theories, and the cognitive systems in the human mind – a question that we turn to next in our exploration of the relationship between acceptability and grammaticality.

3.2 What is syntactic theory a theory of?

This question is obviously much larger than the question of building a theory of acceptability judgments, but it is a critical component of building such a theory because acceptability appears to be affected by both grammatical factors and sentence processing factors. Therefore the relationship that one assumes between the theory of grammar and the theory of sentence processing will directly impact the formulation of a complete linking hypothesis for acceptability judgments, and concomitantly constrain the types of inferences one can make about the grammar from acceptability judgments. There have been a number of discussions of this question in the literature recently, such as Neeleman and van de Koot 2010, Phillips and Lewis 2013, and Lewis and Phillips 2015. The discussion here will largely mirror those discussions, albeit with much less detail. The first question one can ask is whether syntactic theories are theories of a cognitive system in the human mind (mentalistic) or theories of an object outside of the human mind (non-mentalistic). Though there is a rich tradition of non-mentalistic approaches to linguistics, I am going to focus exclusively on mentalistic syntactic theories here, partly because the theory in Syntactic Structures is mentalistic, and partly because it is not clear what the linking hypothesis for acceptability judgments would be for non-mentalistic theories of syntax.

The second question we can ask is what is the cognitive system that syntactic theories describe. Is there a "grammar system" in the mind that is distinct from other cognitive systems that subserve language, such as the sentence processing system (what Lewis and Phillips 2015 call the "two-system" approach)? Or is it the case that syntactic theories and sentence processing theories are two different descriptions of a single cognitive system (what Lewis and Phillips 2015 call a "one-system" approach)? The answer to this question will determine the relationship between the grammar and sentence processing terms in the specification of a model of acceptability judgments. If there are two systems, then the terms will be separate: a complete grammatical model plus a complete sentence processing model, plus a theory of how the two interact. If there is only one cognitive system, then the theory of acceptability judgments is really a theory of error signals from sentence processing, with syntacticians attempting to partial out the component of the error signal that corresponds to the component of sentence processing that syntactic theories characterize. Though the two-system approach is reminiscent of some sentence processing models (e.g., Townsend and Bever 2001, Ferreira and Patson 2007) that posit two

stages of processing, and though the two-system view does arise from time to time in conversations among syntacticians, I know of no detailed defense of the two-system view in the generative syntax literature (except, possibly, Seely and Epstein 2006). Therefore here I will primarily entertain the one-system view, and ask the natural follow-up question: If syntactic theory is a description of the sentence processing system at some level of abstraction, which component of sentence processing is it a theory of?

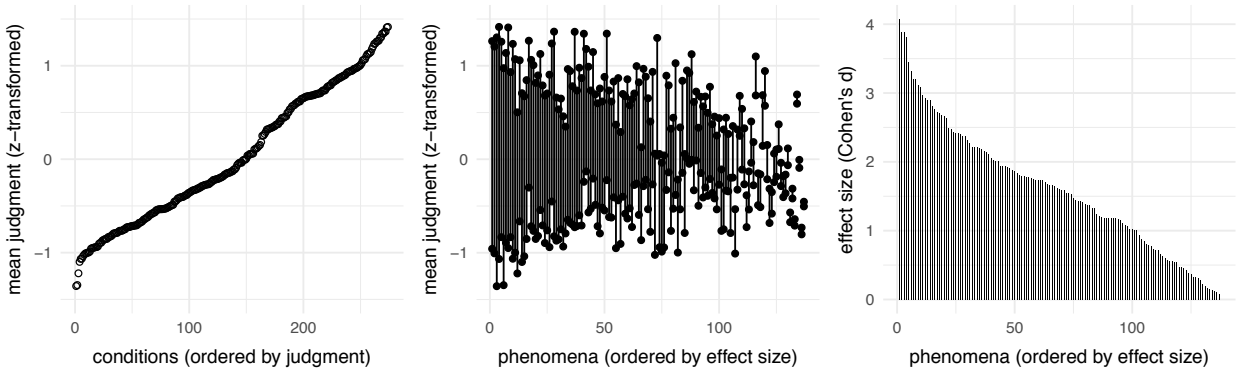
One popular approach to levels of description is Marr's (1982) famous three-level typology: the computational level describes the problem that needs to be solved, the algorithmic level describes the algorithm for solving that problem, and the implementation level describes the physical machine that instantiates the algorithm. There are a number of criticisms of Marr's levels as applied to cognition, so I will not argue that this is the correct (or only) way of thinking about theories of languages here. Instead, what I want to demonstrate is that once one uses a typology of levels (Marr's, or any other), it rapidly becomes clear that syntactic theory is not an abstraction of the complete sentence processing system; it is only an abstraction of a subcomponent of the sentence processing system. For example, it is not uncommon for syntacticians to describe syntactic theory as a computational level description (following Marr himself), i.e., a description of the problem that needs to be solved. But syntactic theory does not have components to encode many of the problems that are central to sentence processing, such as ambiguity resolution, dependency processing, memory processes, and many others. In fact, if one were to attempt to extrapolate from syntactic theories (assumed to be a computational level description) up to the algorithmic and implementation levels, the resulting machine would likely be one that could generate all and only the sentences of a given language, but with no components that could successfully parse or produce sentences incrementally. This suggests that, under the one-system approach, syntactic theories are a (computational-level) description of a subset of the sentence processing system, specifically the structure-building component, abstracting away from the precise algorithms of structure-building. This is exactly the conclusion reached by a number of syntacticians (e.g., Neeleman and van de Koot 2010, Phillips and Lewis 2013), and I suspect the most common view in generative syntax. What this means for the theory of acceptability is that the "grammar" component of the theory of acceptability will be something like an error signal from the structure-building component of the sentence processor. This in turn means that the "sentence processing" component of the theory of acceptability judgments will be everything that isn't structure-building: parsing strategies for various types of ambiguity resolution, the complexity that arises from ambiguity resolution (e.g., surprisal, Hale 2001, Levy 2008), the complexity that arises from dependency processing (Gibson 1998), the complexity that arises from working memory operations more generally (Lewis and Vasisth 2005, McElree et al. 2003), and many others components.

3.3 Gradience in acceptability judgments

Once we've settled on the relationship between grammatical theories and sentence processing theories, we can begin to investigate how it is that acceptability can be used to make inferences about the grammar. One major question in current research on acceptability judgments is what consequence, if any, does the fact that acceptability judgments are gradient have for inferences about the grammar. Acceptability judgments are gradient in at least two ways. The first is that the acceptability ratings of individual sentence types appear to form a continuum (no categorical clumping) when a large number of distinct sentence types are plotted simultaneously. The left

panel of Figure 3 demonstrates this for the 300 distinct sentence types tested by Sprouse et al. 2013 in their study of data points from *Linguistic Inquiry* by plotting the mean ratings for each sentence type (ordered by increasing acceptability). The second way that acceptability judgments are gradient is that the effect sizes of experimentally-defined phenomena also appear to form a continuum (with no categorical clumping). The middle panel of Figure 3 demonstrates this for the 150 two-condition phenomena that Sprouse et al. 2013 tested from *Linguistic Inquiry* by plotting the means of the two conditions of each phenomena (thus highlighting the difference between means as an effect size), and the right panel of Figure 3 demonstrates this by plotting a standardized effect size measure (Cohen's *d*, which is the difference between means scaled by the standard deviations of the conditions). Crucially, these two-condition phenomena were specifically designed to isolate a putative grammatical manipulation while holding other potential sources of acceptability judgment variability constant.

Figure 3: Three demonstrations of gradience. The left panel plots the mean ratings for 300 sentence types from *Linguistic Inquiry*, arranged by increasing acceptability. Although the line looks solid, it is 300 empty circles. The middle panel plots the means of the two conditions in each of 136 phenomena that replicated under a strict statistical criterion of replication in Sprouse et al. 2013. This highlights their effect sizes in natural units (z-scored ratings), arranged by decreasing effect size. The right panel plots standardized effect sizes (Cohen's *d*) for the same 136 phenomena, again arranged by decreasing effect size.



The question, then, is what is driving this gradience. Is the grammar itself gradient (i.e., is the structure-building component of sentence processing gradient)? Or is the grammar categorical, with the gradience of acceptability deriving from the other aspects of the sentence processing system?

This question is impossible to answer from acceptability judgments alone. Both categorical and gradient grammars can explain gradient acceptability judgments; they simply do so with different mechanisms. For categorical grammars, the structure-builder itself is limited to contributing a finite number of levels of acceptability (typically two, but in principle any finite number is possible). The apparent continuum that we see in the acceptability of individual sentences (left panel of Figure 3) must therefore come from some combination of the following:

(i) the effects of typical sentence processing over the portion of the sentence that can be processed typically, such as dependency complexity, ambiguity resolution complexity (e.g., surprisal), working memory, etc,

(ii) the effects of atypical sentence processing over any structure-building violations, such as processes that are designed to construct an interpretable structure out of word strings.
(iii) plausibility and real-world knowledge effects,
(iv) task effects, and
(v) any number of other components of sentence processing and acceptability judgments that we may not have explored yet.

That said, we can minimize the impact of the effects of typical processing, plausibility and real-world knowledge, task effects, and possibly even unexplored factors by using experimentally-defined phenomena (as in the right panel of Figure 3), and focusing on the effect size of the difference between them. This effect size is primarily a combination of the structure-builder error signal, the effect of atypical processing, plus whatever factors weren't perfectly controlled in the design itself. Therefore, under a binary categorical grammar the gradience we see in the right panel of Figure 3 is primarily the effect of atypical processing and any uncontrolled factors (because each phenomenon contains one grammatical and one ungrammatical sentence, so the contribution of the structure-builder error signal is the same for each phenomenon).

Gradient grammatical theories differ in their explanation of gradience in two ways. The first is obvious: instead of the structure-building component contributing only a finite number of values, truly gradient grammatical theories posit that the structure-building component can contribute a potentially infinite number of values. This means that a major component of the gradience of acceptability for individual sentences would simply be the gradient value returned by the structure-builder when it is asked to construct the target structure, and the gradience for experimentally-defined effects would be the difference in those values. In many gradient grammars the value returned by the structure-builder is grounded in some primitive of the theory, such as harmony values in harmonic grammars (with OT being a special case of harmonic grammars; Smolensky and Legendre 2006, Keller 2000) or probabilities in various probabilistic grammars (e.g. Lau et al. 2017). The second way that gradient grammatical theories could, at least in principle, differ is in the contribution of atypical processing (item (ii) above). Whereas atypical processing is logical option for categorical grammars when the structure-builder encounters a violation, it is not clear to what extent atypical processing would occur for gradient grammars; and indeed, it is not clear what would constitute "atypical" for a truly gradient structure-builder. It is therefore possible that atypical processing, whatever that might mean for a gradient grammar, would have a decreased role, if any, in determining acceptability in gradient grammars. This means that for many gradient grammars, the gradience in acceptability we see in Figure 4 is potentially a fairly direct reflection of the values of the gradient grammar (modulo any uncontrolled factors in the design).

As a quick aside, it is important to note that the components listed above are not intended to exhaust the space of possible factors influencing acceptability judgments (as indicated by item (v) above). For example, one potentially relevant idea that is sometimes discussed in the field is that minimum (linguistic) edit distances may be relevant for gradient acceptability. The idea is that, during an acceptability judgment task, participants might be implicitly comparing violation sentences to the minimally different grammatical sentences that have the same meanings. If so, it could be the case that acceptability judgments are impacted by the similarity/dissimilarity between the violation sentence and the grammatical counterpart. Similarity/dissimilarity can be quantified using a multi-dimensional distance measure, such as the number of (linguistic) edits necessary to convert from the ungrammatical sentence to the grammatical sentence. Crucially,

distance effects are very likely to correlate with atypical processing costs: as dissimilarity between an ungrammatical sentence and its grammatical counterpart increases, the distance between them increases, as does the need to do atypical processing to derive an interpretable sentence. This means that if such an implicit comparison is part of the theory of acceptability judgments, the cost that arises for minimum (linguistic) edit distance could either be an additional factor influencing acceptability, or a factor that partially (or completely) overlaps with atypical processing.

The debate between categorical and gradient grammars will only be settled by constructing large chunks of theories, spanning dozens or hundreds of phenomena. With large chunks of theories, one could probe both external predictions of the theory, such as predictions about language acquisition, and internal predictions, such as predictions about gradient acceptability. For example, for categorical grammatical theories, one way to probe internal predictions about gradient acceptability would be to evaluate how well independently-motivated aspects of sentences processing can be combined with categorical grammars to yield empirically attested patterns of gradient acceptability. Though the logic of this is straightforward enough, there are two major challenges to constructing a complete theory of acceptability judgments using categorical grammars. The first is that theories of typical sentence processing are an active area of research. It is true that there are candidate theories for calculating dependency costs (Gibson 1998), memory costs (Lewis and Vasishth 2005, McElree et al. 2003), and even complexity for the different continuations of ambiguous strings (e.g., surprisal as in Hale 2001 and Levy 2008, perhaps combined with the Hunter and Dyer 2013 method for creating probabilistic minimalist grammars); however, these theories have not yet been applied to the full range of sentence types that appear in the large acceptability judgment corpora that one would like to test (and indeed, doing so would be a monumental undertaking, as it would require creating both formal grammars with coherent coverage of the sentences types, and sentence processing models with coherent coverage of the sentence types). The second major challenge is that there is little to no research on the atypical sentence processing that arises for ungrammatical sentences. Most of the syntax and sentence processing literatures has focused on the categorical detection of errors, not the costs or processes associated with those errors. Aside from the research on doubly center-embedded sentences and grammatical illusions, it is not clear that this kind of atypical processing is a priority for either field at the moment.

Similarly, for gradient grammatical theories, one way to probe internal predictions about gradient acceptability would be to calculate the (cognitively grounded) values of the gradient grammar independently of acceptability judgments, and then ask how well those values correlate with acceptability. The major challenges with this approach will be unique to each type of gradient grammatical theory, because those challenges will be driven by the type of value that the gradient grammar is built upon. For example, for harmonic grammars, the value in question is harmony, which is a value grounded in the activation of a neural network. Since we cannot independently measure network activation in human minds, the challenge would be to empirically build harmonic grammars using the acceptability judgments of a set of phenomena that involve the same constraints as a distinct, second set of phenomena, and then see how well the grammar can predict acceptability of the second set of phenomena. As a second example, for probabilistic grammars, the value in question is a probability, likely derived from the production probabilities of sentences in natural language corpora. The challenge with this is that both classes of theories, categorical and gradient, predict a relationship between acceptability and production probabilities. For the gradient grammars, the probabilities in the grammar give rise to the

production probabilities and give rise to the acceptability judgments. For categorical grammars, the grammar plus the sentence processing system give rise to the production probabilities, and the grammar plus the sentence processing system give rise to the acceptability judgments. The difference between gradient and categorical theories would therefore be in the degree of correlation between production probabilities and acceptability judgments. One relatively simple prediction would be that the correlation would be lower for categorical grammars because the sentence processing system might cause different effects in acceptability (where the primary issue is atypical processing) and production (where the primary issue is typical processing). But without specific theories to work with, it is difficult to quantify exactly what the difference would be. In the end, this means that for many types of gradient grammatical theories, the best we can say is how well acceptability correlates with the value of interest, without being able to quantify if the correlation is exceptionally high, mediocre, or low for the space of possible grammars.

Despite these challenges, the gradience of acceptability is one of the more fruitful research questions in the acceptability judgment literature today. For one, the stakes are high: this question directly bears on the architecture of the grammar. For two, the only way to make progress on this question is to investigate corners of language that are often overlooked in generative syntax, such as atypical processing, formalizing full grammars, and directly evaluating the cognitively grounded values of gradient grammars. Finally, this question requires large-scale research projects spanning dozens, or hundreds, of sentence types. Given these properties, it is almost impossible for well-designed projects to fail to reveal something new about syntactic theory.

3.4 Absolute acceptability, experimentally-defined effects, and effect sizes

Another major question in current research on acceptability judgments is how to use the different aspects of acceptability judgments as evidence for (un)grammaticality. There are at least three aspects of acceptability that are used as evidence in syntactic theories: (i) the acceptability of individual sentence types, (ii) the presence or absence of a difference in acceptability between two (or more) minimally different sentence types (what I have called an experimentally-defined effect), and (iii) the size of the difference in acceptability between two (or more) sentence types (the size of the experimentally defined effect). The first, and most obvious, approach is to focus exclusively on the acceptability of individual sentence types. In Syntactic Structures, Chomsky assumes a transparent mapping between grammaticality and the acceptability of individual sentence types (modulo well-known counter-examples like doubly center-embedded sentences), such that sentence types at the low end of the spectrum are assumed to be ungrammatical, and sentences at the high end are assumed to be grammatical. As Chomsky anticipates, sentences near the middle of the spectrum will be problematic for this transparent mapping. In Syntactic Structures, Chomsky argues that a suitably well-defined grammar will simply predict the grammaticality of intermediate sentences, so that we do not need to rely on acceptability judgments at all: “In many intermediate cases we shall be prepared to let the grammar itself decide, when the grammar is set up in the simplest way so that it includes the clear sentences and excludes the clear non-sentences (p.14)”. In other words, a grammar that is well-specified for the clear cases should predict (some number of) the unclear cases. Based on the previous discussion, we can also add that a well-specified theory of the extra-grammatical factors that influence acceptability could also help to predict some number of the unclear cases. But in the absence of

those two well-specified theories, we must rely on syntacticians' best scientific judgment about what might be driving the acceptability of the unclear cases. Because scientific opinions differ, this can lead to syntacticians interpreting unclear cases in opposing ways (see Schütze 1996 for some examples, and see Hoji 2015 for a strict criterion for separating clear from unclear cases).

Given the potential difficulty of interpreting the acceptability of individual sentences in the middle of the spectrum, one might wonder whether focusing on experimentally-defined effects (differences between two or more conditions) might help to solve the problem. The simplest way to do this would be to look for the presence or absence of a difference between two (or more) sentence types that have been constructed to control for as many extra-grammatical properties as possible, such that any resulting difference is likely to be driven by grammatical differences. The problem with using the presence/absence of effects as a discovery procedure for grammatical effects is the phrase "likely to be drive by grammatical differences". We can never control for every possible extra-grammatical effect on acceptability in the design. This means that there is always going to be some difference between the conditions in our experiment, albeit potentially very small (meaning that to detect it with standard statistical tests, we may need relatively large sample sizes). Because there is always some difference between our conditions, this means that the interpretation of the presence/absence of an effect is in fact an interpretation of the size of the effect. Is the size of the effect the size we would expect from a grammatical effect (given the design of the experiment, and what we know about grammatical effects), or is it the size we would expect from an extra-grammatical effect (given the design of the experiment, and what we know about extra-grammatical effects). This is the same issue that we have seen throughout this section: even the experimentally-defined effects require a theory of acceptability judgments to be interpreted. What we have at the moment is a partial (and ever-growing) theory, so some amount of interpretation must be filled in by the scientific judgment of individual researchers. (To be fair, this issue is not unique to syntax. I know of no complete theory of any data type in language science, from reading times, to event-related potentials, to BOLD signals. In each case, there are partial theories that are augmented by the best scientific judgment of researchers and reviewers.)

Given the issues that arise when individual acceptability and experimentally-defined effects are used in isolation, another option is to interpret both individual acceptability and experimentally-defined effects in combination with one another. Though I know of no studies counting the different approaches to acceptability-grammaticality mappings, my impression is that this combination approach is the one most frequently adopted by syntacticians. The underlying idea is that a clear grammatical effect should yield a relatively large experimentally-defined effect, with the individual rating of the ungrammatical sentence near the low end of the spectrum of acceptability. To be clear, this approach does not eliminate the problems of intermediate individual acceptability and small effect sizes. But it does make it a bit easier to draw attention to these issues. In fact, this combination approach has uncovered some potentially interesting mismatches between individual acceptability and the presence of experimentally-defined effects, where there is a statistically significant effect, but all of the sentences in the design would still be labeled as "acceptable" in a categorical task. For example, Featherston (2005) famously observed a pattern of acceptability judgments that indicated a Superiority effect in German, despite the fact that many German native speakers label the critical sentences as "acceptable". Similarly, Almeida 2014 found a pattern of acceptability that is indicative of a wh-island effect in Brazilian Portuguese, despite the critical sentences being marginally or fully acceptable. Kush et al. *submitted* found a similar pattern for wh-islands in Norwegian. These

results raise interesting questions for syntactic theories. Are these effects driven by a grammatical violation or an extra-grammatical factor (that is specific to the critical condition)? Is it possible for the violation of a grammatical constraint to result in marginal, or even fully acceptable, sentences? Must there be a mitigating factor in these cases (e.g., a sentence processing effect that raises acceptability)? Or are these effects simply evidence that a gradient approach to syntactic theory is more likely to be correct? There are no easy answers to these questions; but the ability to quantify both aspects of acceptability has brought these questions into sharper focus, potentially yielding new evidence about the nature of syntactic theories.

Though it goes beyond the scope of this chapter, for completeness it is important to note that another method for dealing with indeterminate acceptability facts (beyond building a complete theory of acceptability, and beyond relying on syntacticians' scientific judgments) is to look for converging evidence from other sentence-processing data types, such as reading times, eye-movements, and electrophysiological measures such as event-related potentials. This literature is far too large to review in any detail here (cross-reference to other chapters?), but it is worth noting that there is an impressive literature demonstrating (i) that many grammatical constraints are respected by real-time sentence processing mechanisms, and (ii) that grammatical violations are often detected within a few hundred milliseconds of the violating word in real-time sentence processing (see Lewis and Phillips 2015 for some review; see also Sprouse and Lau 2013 for an extensive bibliography of ERP studies that detect syntactic violations relatively rapidly). Given the sheer number of grammatical effects that have been catalogued in the sentence-processing literature, it seems likely that indeterminate acceptability effects that are based on true grammatical violations would also show some sort of real-time processing consequence; however, as always, the exact nature of that consequence will depend upon developing a theory of sentence processing, and, of course, a theory of the data type in question.

4. Acceptability and grammaticality, prospects and challenges

At a methodological level, I take the current state of evidence to suggest: (i) judgment tasks are fairly sensitive, especially for the relatively large effects that characterize syntactic theory, (ii) judgments are robust, reliable, and replicable, regardless of the method used to collect them (at least for English), (iii) judgments are only affected by sentence processing effects in relatively limited circumstances (that may be revealing of the architecture of sentence processing more than the architecture of the grammar), and (iv) judgments are relatively unaffected by task effects such as repetition (at least within the scope of typical judgment experiments). With the caveat that future research could potentially overturn one or more of these trends, I find the current state of affairs incredibly encouraging for the use of acceptability judgments as a data type in generative syntax. Acceptability judgments are a well-described data type that yields surprisingly robust data. That said, the methodological landscape is as fertile as ever. There are literally dozens of topics left to explore when it comes to judgment methodology, such as the effect of the composition of the experiment, the effect of instructions (see Cowart 1997 for a first study), the effect of word and construction frequency (see Featherston 2009 for some comments on this), the effect of the size of rating scales, the comparison of informal and formal methods for other languages and data types, and many others.

At a theoretical level, many of the same challenges that the field faced in 1957 are still present today. We are still far from constructing a complete theory of typical and atypical sentence processing, or developing a complete theory of the (cognitive) values that ground

gradient grammars. Nonetheless, there are reasons to be optimistic. Though we don't have complete theories of these systems, we do have partial theories, and an ever-growing array of tools to make progress on those theories. Furthermore, the field is cognizant as ever of the challenges facing the use of acceptability judgments as evidence for grammar, and is meeting those challenges head-on by focusing on difficult topics, such as gradience in acceptability and the role of effect sizes, that will necessarily spur advances in the theory of acceptability. Some challenges still remain for the daily work of syntacticians, such as indeterminate acceptability effects and clashes between individual acceptability ratings and experimentally-defined effects, but this challenge too may spur advances, as syntacticians increasingly seek converging evidence from multiple data types. All in all, I believe it has been a productive 60 years, and can hardly wait for the next 60.

References

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

Almeida, Diogo. (2014). Subliminal wh-islands in Brazilian Portuguese and the consequences for syntactic theory. *Revista da ABRALIN, 13*(2), 55-93.

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

Bock, Kay, and Carole A. Miller. 1991. Broken agreement. Cognitive Psychology, 23, 45–93.

Braze, David. 2002. Grammaticality, Acceptability, and Sentence Processing: A psycholinguistics study. PhD dissertation. University of Connecticut.

Cowart, Wayne. 1997. Experimental Syntax: Applying objective methods to sentence judgments. Thousand Oaks, CA: Sage.

Culbertson, Jennifer, and Steven Gross. 2009. Are linguists better subjects? *British Journal for the Philosophy of Science* 60. 721-736.

Dąbrowska, Ewa. 2010. Naïve v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27.1-23.

Deane, Paul. 1991. Limits to attention: A cognitive theory of island phenomena. Cognitive Linguistics 2.1–63.

Engdahl, Elisabet. 1982. Restrictions on unbounded dependencies in Swedish. In E. Engdahl & E. Ejerhed (eds.), Readings on unbounded dependencies in Scandinavian languages, 151–174. Stockholm: Almqvist & Wiksell.

Featherston, Sam. 2005. Magnitude estimation and what it can do for your syntax: some wh-constraints in German. Lingua 115.1525-1550.

Featherston, Sam. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33.269-318.

Featherston, Sam. 2008. Thermometer judgments as linguistic evidence. Was ist linguistische evidenz?, ed. by C. M. Riehl and A. Rothe. Aachen: Shaker Verlag.

Featherston, Sam. 2009. Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenshaft* 28.127-132.

Ferreira, Fernanda. 2005. Psycholinguistics, formal grammars, and cognitive science. *The linguistic Review* 22.365-380.

Ferreira, Fernanda, and Nikole Patson. 2007. The 'good enough' approach to language comprehension. Language and Linguistics Compass, 1, 71–83.

Francom, Jerid. 2009. Experimental syntax: Exploring the effect of repeated exposure to

anomalous syntactic structure – evidence from rating and reading tasks. PhD. dissertation, University of Arizona.
Gibson, Edward. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1-76.
Gibson, Edward & Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. Language and Cognitive Processes 28(1-2). 88-124.
Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics, pages 159–166.
Häussler, Jana, Tom Juzek, and Thomas Wasow. 2016. To be grammatical or not to be grammatical – Is that the question? Poster presented at the Annual Meeting of the Linguistic Society of America.
Hill, Archibald A. 1961. Grammaticality. *Word* 17.1-10.
Hiramatsu, Kazuko. 2000. Accessing linguistic competence: Evidence from children's and adults' acceptability judgments. PhD. dissertation, University of Connecticut.
Hofmeister, Philip, and Ivan A. Sag. 2010. Cognitive constraints and island effects. *Language*, *86*(2): 366.
Hofmeister, Philip, Laura Staum Casasanto, and Ivan A. Sag. 2014. "Processing effects in linguistic judgment data: (Super-)additivity and reading span scores." *Language and Cognition*, vol. 6(1), 111--145.
Hoji, Hajime. 2015. *Language Faculty Science*. Cambridge University Press.
Huang, C.-T. James. 1982. Move WH in a language without WH-movement. The Linguistic Review 1:369–416.
Hunter, Tim, and Chris Dyer. 2013. Distributions on minimalist grammar derivations. In *Proceedings of 13th Meeting on the Mathematics of Language (MoL 2013)* (pp. 1-11).
Kazanina, Nina, Ellen F. Lau, Moti Lieberman, Masaya Yoshida, and Colin Phillips. 2007. The effect of syntactic constraints on the processing of backward anaphora. Journal of Memory and Language, 56, 384–409.
Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD. dissertation, University of Edinburgh.
Keller, Frank. (2003). A psychophysical law for linguistic judgments. In *Proceedings of the 25th annual conference of the Cognitive Science Society* (pp. 652-657)
Kluender, Robert and Marta Kutas. 1993. Subjacency as a processing phenomenon. *Language and cognitive processes*, *8*(4), 573-633.
Kush, Dave, Terje Lohndal, and Jon Sprouse. (*submitted*). Investigating Variation in Island
Effects: A Case Study of Norwegian. *Natural Language and Linguistic Theory*.
Lasnik, Howard and Mamoru Saito. 1984. On the nature of proper government. Linguistic Inquiry 15:235–289.
Lau, Jey Han, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science*.
Levy, Roger. 2008. Expectation-based syntactic comprehension. Cognition, 106:1126–1177.
Lewis, Shevaun and Colin Phillips. 2015. Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, *44*(1), 27-46.
Lewis, Richard L., and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. Cognitive Science, 29, 1–45.
Luce, R. Duncan. 2002. A psychophysical theory of intensity proportions, joint presentations,

and matches. Psychological Review 109.520–32.
Mahowald, Kyle, Peter Graff, Jeremy Hartman, and Edward Gibson. 2016. SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language*, *92*(3), 619-635.
Marantz, Alec. 2005. Generative linguistics within the cognitive neuroscience of language. *The linguistic Review* 22.429-445.
Marr, David. 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. The MIT Press.
Matchin, William, Jon Sprouse, & Greg Hickok. 2014. A structural distance effect for backward anaphora in Broca's area: an fMRI study. *Brain and Language* 138: 1-11.
McElree, Brian, Stephani Foraker, and Lisbeth Dyer. (2003). Memory structures that subserve sentence comprehension. Journal of Memory and Language, 48, 67–91
Michel, Daniel. 2014. Individual cognitive measures and working memory accounts of syntactic island phenomena. PhD dissertation, University of California San Diego.
Miller, George and Noam Chomsky. 1963. Finitary models of language users. In Luce, R.; Bush, R. and Galanter, E. (eds.) *Handbook of Mathematical Psychology, Vol 2*. New York: Wiley. 419-93.
Montalbetti, Mario, M. 1984. After binding. On the interpretation of pronouns. Ph.D. dissertation, MIT, Cambridge, Mass.
Nagata, Hiroshi. 1988. The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. *Journal of Psycholinguistic Research* 17.1-17.
Nagata, Hiroshi. 1989. Effect of repetition on grammaticality judgments under objective and subjective self-awareness conditions. *Journal of Psycholinguistic Research* 3.255-269.
Narens, Louis. 1996. A theory of ratio magnitude estimation. Journal of Mathematical Psychology 40.109–29.
Neeleman, Ad and Hans van de Koot. 2010. Theoretical validity and psychological reality of the grammatical code. In H. De Mulder, M. Everaert, Ø. Nilsen, T. Lentz, & A. Zondervan (eds.), Theoretical validity and psychological reality, pp. 183-212. Amsterdam: John Benjamins.
Parker, Dan and Colin Phillips. 2016. Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, *157*, 321-339
Phillips, Colin. (2006). The real-time status of island phenomena. *Language*, *82*(4), 795-823.
Phillips, Colin. 2009. Should we impeach armchair linguists? In *Proceedings from Japanese/Korean Linguistics* 17. S. Iwasaki, H Hoji, P. Clancy, and S.-O. Sohn, eds. Stanford, CA: CSLI Publications.
Phillips, Colin and Shevaun Lewis. 2013. Derivational order in syntax: Evidence and architectural consequences. *Studies in Linguistics*, *6*, 11-47.
Rizzi, Luigi (1982). Violations of the wh-island constraint and the subjacency condition. In Luigi Rizzi (ed.), Issues in Italian syntax, 49–76. Dordrecht: Foris
Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
Seely, T. Daniel and Samuel D. Epstein, (2006) *Derivations in Minimalism*, Cambridge University Press: Cambridge.
Smolensky, P., & Legendre, G. 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1*. MIT Press.
Snyder, William. 2000. An experimental investigation of syntactic satiation effects. *Linguistic*

Inquiry 31.575-582.
Spencer, Nancy J. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2.83-98.
Sprouse, Jon. 2009. Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry* 40.329-341.
Sprouse, Jon. 2011. A test of the cognitive assumptions of Magnitude Estimation: Commutativity does not hold for acceptability judgments. *Language* 87.274–288.
Sprouse, Jon and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, *2*(1).
Sprouse, Jon and Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, *48*(03), 609-652.
Sprouse, Jon, Matthew Wagers, and Colin Phillips. 2012. A test of the relation between working-memory capacity and syntactic island effects. *Language*, *88*(1), 82-123.
Sprouse, Jon, Carson T. Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, *134*, 219-248.
Staub, Adrian. 2010. Reponse time distributional evidence for distinct varieties of number attraction. Cognition, 114, 447–454.
Stevens, Stanley S. 1957. On the psychophysical law. Psychological Review 64.153-181.
Sturt, Patrick. 2003. The time-course of the application of binding constraints in reference resolution. Journal of Memory and Language, 48, 542–562.
Townsend, David J. and Thomas G. Bever. 2001. Sentence comprehension: The integration of habits and rules. Cambridge, MA: MIT Press.
Van Gompel, Roger P. G., and Simon P. Liversedge. 2003. 'The influence of morphological information on cataphoric pronoun assignment', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 128–139.
Wagers, Matthew, Ellen Lau, and Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. Journal of Memory and Language, 61, 206–237.
Wasow, Thomas and Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115.1481-1496.
Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2014). Deconstructing acomparative illusion. Ms. Northwestern University, University of Southern California, and University of Maryland.
Weskott, Thomas and Gisbert Fanselow. 2011. On the Informativity of Different Measures of Linguistic Acceptability. *Language* 87, 249–273.
Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. Brain and Language, 108, 40–55.
Yoshida, M., Kazanina, N., Pablos, L., & Sturt, P. (2014). On the origin of islands. *Language, Cognition and Neuroscience*, *29*(7), 761-770.