Acceptability judgments

1. Introduction

The goal of experimental syntax, at least to my mind, is straightforward: to use experimental methods to collect data that is relevant for the construction and evaluation of syntactic theories. For data types that can only be collected using a formal experiment, such as reaction times or EEG, the work of experimental syntax is simply the work of leveraging these methods for questions in theoretical syntax. However, things appear to be a bit more complicated when the data type in question is acceptability judgments, as acceptability judgments can be collected both relatively informally, as is typical in much of the syntax literature, or relatively formally, as is typical in the experimental syntax literature. I take the coexistence of these two methods of judgment collection to imply that the goal of experimental syntax with respect to acceptability judgments is not simply to collect acceptability judgments, because that is what is done in all syntactic work, but rather to explore ways in which the formal collection of judgments can add new insights over and above those that derive from informal methods. Therefore my goal in this chapter is to identify four areas in which formal judgment experiments have made substantial contributions – two that lean toward methodological issues, and two that lean toward theoretical issues, and to review the current state of the evidence that we have for each of those areas. To be clear, this chapter is not intended as an exhaustive review of all possible areas in which formal judgment experiments could potentially make a contribution, but is rather intended as a starting point for thinking about the kinds of questions in theoretical syntax that might benefit from formal acceptability judgment experiments. My hope is that these questions will help to inspire new questions, and new work, in the growing field of experimental syntax.

Before delving into the primary content of this chapter, I would like to briefly mention a few assumptions (and/or decisions) that I am making to get the chapter off of the ground. The first is that I assume, following many working syntacticians, that acceptability judgments are in principle valuable for the construction and evaluation of syntactic theories. I will, therefore, not attempt to motivate the use of acceptability judgments in general (see Schütze 1996 for a comprehensive discussion of this). The second is that I will assume a relatively minimal linking hypothesis between acceptability judgments and the cognitive properties of sentence processing, specifically that acceptability judgments are a behavioral response that arise when speakers comprehend sentences, and that these behavioral responses are impacted by a large number of cognitive factors, such as the grammaticality of the sentence, the processing dynamics of the sentence, the sentence processing resources required by the sentence, the meaning of the sentence, the plausibility of the sentence relative to the real world, and even the properties of the specific task that is given to the participants. I believe wholeheartedly that a more precise linking hypothesis would be helpful for using judgments as evidence in syntax; however, I also believe that the minimal linking hypothesis above is more than sufficient to begin to explore the value of formal acceptability judgment experiments in syntax. The third is that I will assume that there is no substantive qualitative difference between "informal" and "formal" judgment experiments. Both are experiments in the sense that they involve the manipulation of one variable (syntactic structure) to reveal a causal relationship with another variable (acceptability), therefore both involve all of the components that typify psychology experiments: a set of conditions, a set of items in each condition, a set of participants, a task for the participants to complete using the items, and a process for analyzing the results of the task. The difference appears to me to be primarily quantitative, in that "formal" experiments tend to involve more conditions, more items per condition, more participants, and more complex analysis processes. To my mind, the labels

"informal" and "formal" simply point toward different ends of this quantitative spectrum. What this means in practice is that when I say that formal experiments are valuable in some way, what I intend is that increasing the number of conditions, items, participants, and/or increasing the complexity of the analysis can yield insights that fewer conditions, items, participants, and/or less complex analyses cannot. The labels "informal" and "formal" are a more concise way to express this idea. My fourth assumption is that Schütze 1996 is the definitive review of experimental syntax work that was published before 1996. Therefore, in order to provide something new for the field, I will focus here on work published after 1996. Finally, this chapter is not a how-to for constructing formal judgment experiments. The goal is for this to be the chapter one reads, either before or after reading a how-to, for inspiration about the types of questions one can ask with the method. I will provide some references for learning acceptability judgment methods in the annotated bibliography for this section of the handbook.

2. The validity and reliability of acceptability judgments

Perhaps the most frequently asked question in the experimental syntax literature is to what extent the informally collected judgments that have been published in the literature can be trusted to form the empirical basis of syntactic theory. This question has arisen since the earliest days of generative grammar (Hill 1961, Spencer 1973); it played a central role in the two books that ushered in the most recent wave of interest in experimental syntax (Schütze 1996, Cowart 1997); and it has given rise to a number of high-level debates in the experimental syntax literature over the past 15 years or so (see Edelman and Christiansen 2003; Ferreira 2005; Wasow and Arnold 2005; Featherston 2007; Gibson and Fedorenko 2013 for some concerns about informal methods, see Marantz 2005 and Phillips 2009 for some rebuttals, and Myers 2009 for a proposal that attempts to split the difference between informally collected judgments and full-scale formal experiments). The existence of this question is understandable. First, informally collected judgments form the vast majority of the data points published in the (generative) syntax literature. Second, the properties of informal collection methods are not identical to the properties of the formal experimental methods that are often used in other domains of cognitive science: informal methods often involve a smaller number of participants, those participants are often professional linguists instead of naïve participants, the participants are often presented a smaller number of items, and the results are often only analyzed descriptively (without inferential statistics). If one believes that the properties of formal experiments are what they are for a reason, then it is logically possible that the differences between informal methods and formal experiments could matter. The consequences of this cannot be understated. If there are systemic problems with informally collected judgments, then there is likely to be systemic problems with (generative) syntactic theories.

This question touches upon a number of issues in psychometrics and the broader philosophy of measurement. The first question is – What do we mean when we say that data can be "trusted" to form the basis for a theory? Psychometric theories have identified a number of properties that good measurement methods should have. Here I will mention two (and only in a coarse way, setting aside subtypes of these properties): validity and reliability. A measurement method is *valid* if it measures the property it is purported to measure. A measurement method is *reliable* if it yields consistent results under repeated measurements (with unchanged conditions). The concerns about informal methods that have figured most prominently in the literature appear to be a combination of concerns about validity and reliability, such as the concern that small sample sizes will lead to an undue influence of random variation, the concern that a small number of experimental items will lead to an undue influence of lexical properties, and the

concern that the participation of professional linguists will lead to theoretical bias. In each case, the concern seems to be that informally collected judgments will not reflect the true acceptability of the sentence (validity), and furthermore that the judgments themselves will be inconsistent over repeated measurements (reliability).

This leads to a second question – How does one establish validity for the measurement of a cognitive property like acceptability? The direct method for establishing validity is to compare the results of the measurement method with a second, previously validated, measurement method. This is obviously unavailable for most cognitive properties – if cognitive scientists had a method to directly measure the cognitive property of interest, we would not bother with the unvalidated measurement method. That leaves only indirect methods of validation. One indirect method is to ask whether the theory that results from the data has the properties of a good scientific theory. This, of course, interacts with broader issues in the philosophy of science about what properties a good theory would have, so I will not attempt to provide an exhaustive list. But two possible criteria are: (i) making potentially falsifiable predictions, and (ii) explaining multiple phenomena with a relatively small number of theoretical constructs. In the case of acceptability judgments, I would argue that the resulting theory of syntax does have these properties. Another indirect method is to ask whether other data types provide corroborating evidence, modulo the linking theories between the data types and the underlying theory. In the case of acceptability judgments, we can ask whether the resulting syntactic theory can be linked to a sentence processing theory in a way that makes potentially falsifiable predictions about other psycholinguistic measures, such as reading times, eye-movements, or EEG, and ultimately whether these measures corroborate the syntactic theory. I would argue that the current results in the literature connecting syntactic theories and sentence processing theories are promising. That said, indirect methods cannot guarantee validity. It is logically possible that acceptability judgments could give rise to a theory that has all of the hallmarks of a good theory, but that does not ultimately explain human syntax (perhaps it is about probability, or plausibility, or even prescriptive grammatical rules).

The final question is – How does one establish reliability? In principle, establishing reliability is relatively straightforward, as it simply entails replicating the measurement. The exact replication can vary based on the type of reliability one is interested in: between-participant (or inter-rater) reliability asks whether the same judgments are obtained with different sets of participants; within-participant (or test-retest) reliability asks whether one set of participants will give the same judgments at two different times; between-task reliability asks whether different judgment tasks will yield the same judgments (either between-participant or within-participant). In practice, establishing the reliability of informal methods is complicated by their informality. By definition, informal methods control the various properties of the judgment collection process less strictly than formal methods, making a strict replication difficult if not impossible. One way to circumvent this problem is to compare the results of informal methods, perhaps as reported in the syntactic literature, with the results of formal experiments. This would be a type of between-task reliability for informal and formal methods, and to the extent that the two sets of results converge, it would establish a kind of reliability for informal methods. Many of the results reported below test precisely this kind of reliability. But it is important to note that while convergence between the two methods can be interpreted as establishing a type of reliability for both, divergence between the two methods can be interpreted in three ways: it could be the case that informal methods are unreliable, or it could be the case that formal methods are unreliable, or both. It is tempting to assume that formal methods enjoy some sort of priority in such a conflict (i.e., that they reveal the ground truth), but as many linguists have pointed out, it is easy to imagine experimental materials that lead to unreliable judgments from non-linguist

participants, but not from linguist participants (such as garden path sentences like *The horse raced past the barn fell*). Resolving the source of the divergence between two methods requires follow-up experiments that manipulate specific hypotheses for the divergence. To my knowledge, though there have been many studies of the convergence/divergence between informal and formal methods, there have been no systematic studies of the source of the divergences that do arise (presumably because, as we will see presently, there are relatively few divergences between the methods).

In reviewing the evidence collected so far on the convergence between informal and formal methods for judgment collection, it is useful to make a distinction between studies that sampled the data points to re-test with bias, and studies that sampled the data points to re-test randomly. Biased sampling means that the data points were chosen because of some property that they have; in these studies, this is typically the belief that the specific data points are invalid or unreliable. Typically this belief comes from debates in the literature about the status of the data point, or the researchers' own (informally collected) judgments. Biased sampling studies can be used to establish that the convergence between informal methods and formal methods is not perfect by showing that the data points in question do not replicate using formal methods. But biased sampling cannot be used to estimate a specific convergence rate. A biased sample could either overestimate or underestimate the actual convergence rate by virtue of the biased selection procedure: a procedure that focuses on selecting known invalid or unreliable data points will almost certainly underestimate the true convergence rate; similarly, a procedure that focuses on selecting likely uncontroversial data points (e.g., a judgment for *This is a pen.*) is likely to overestimate the convergence rate. There are only two options for determining the true convergence rate: an exhaustive comparison of all data points, which would establish the convergence rate with certainty, or a random sampling procedure, which would estimate the convergence rate within a margin of error determined by the size of the random sample relative to the size of the population in question.

Biased sampling studies dominated much of the debate about the validity and reliability of informally collected judgments until relatively recently, presumably because of the time and financial cost associated with testing large numbers of data points prior to the creation of crowdsourcing platforms like Amazon Mechanical Turk. Here I will briefly review some of the more prominent biased sampling studies. Wasow and Arnold 2005 tested a claim from Chomsky 1957 that the ordering preference between NPs and particles in verb-particle constructions is based on the complexity of the NP, not the length. They found that the judgments follow Chomsky's reported judgments when averaged over the entire sample of participants, but that there are some individual participants who do not report Chomsky's judgment pattern. A similar result was obtained by Langendoen, Kalish-London, and Dore 1973 when they tested the claim by Fillmore 1965 that wh-movement is impossible out of the first object position of a ditransitive verb (**Who did you show __ the woman?*). 87 out of 99 responses in their experiment indicated a second object interpretation, in line with Fillmore's claim, but 22 indicated a first object interpretation contrary to Fillmore's claim. Though the results corroborate Fillmore's claim in the aggregate (the proportion is highly significant by sign test), it is possible to interpret the participants who show the opposite pattern as presenting a potential problem for Fillmore's claim (e.g., Gibson and Fedorenko 2013). These two studies demonstrate the difficulty of defining convergence between informal and formal methods. If one assumes that judgments are variable in the way that other behavioral methods are variable, such that the correct way to analyze the results is to look at the sample means, then these two examples are convergences. If, instead, one assumes that judgments will show less variability, perhaps because of a belief in a grammar that creates a stark, binary contrast between grammatical and ungrammatical sentence types then the

participants who fail to show the predicted pattern constitute a divergence. To my knowledge, these two interpretations have not been investigated in detail for these data points, or for any data points that have been claimed to be divergences between informal and formal methods in the judgment literature.

Another prominent biased sampling study is Clifton, Fanselow, and Frazier's 2006 test of Kayne's (1983) claim that a third wh-word can rectify what would otherwise be a superiority violation (i.e., *What can who do about it when?* is better than *What can who do about it?*). In a formal experiment, they found that the two sentences had identical ratings. Fedorenko and Gibson 2010 replicate this finding. However, as Clifton and Frazier (2006) note, this example illustrates the complexity inherent in defining a theoretically relevant data point. If one assumes that Kayne's claim is that the three-wh condition should be more acceptable than a two-wh condition, then this is a divergence. But, if one assumes that Kayne's claim is that the three-wh condition is more acceptable than it would be predicted to be given that it is a superiority violation, then this is in fact a convergence. Clifton and Frazier demonstrate elsewhere in their study that the number of wh-words in a sentence leads to a linear decrease in acceptability: one wh-word is more acceptable than two, and two wh-words is more acceptable than three. It is thus surprising that the three-wh superiority condition is equal in acceptability to the two-wh superiority condition, suggesting that there is a relative increase in acceptability in this configuration over what would otherwise be expected. To my knowledge, this difference in interpretation of the Kayne effect has not been investigated further.

The most recent biased sampling study is Linzen and Oseki's 2018 study of Hebrew and Japanese. They searched several issues of top theoretical journals for "subtle contrasts" that they found "potentially questionable" based on their own native speaker judgments. They identified 14 questionable judgments for each language, and retested them in formal experiments. For Hebrew they found that 7 replicated and 7 failed to replicate. For Japanese they found that 10 replicated and 4 failed to replicate. Like previous biased sampling studies, this study demonstrates that there are some number of divergences between informal and formal methods. It also demonstrates that judgments by professional linguists can be used to identify questionable judgments (though it is not possible to calculate the accuracy of this without information about how many data points were considered during the sampling stage). As is the case with all biased sampling studies, it is impossible to use these numbers to estimate the divergence rate for the two methods. These 11 divergences could represent a small sample of a much larger number of divergences, or they could represent a substantial proportion of the divergences in the literature.

Exhaustive sampling and random sampling studies have recently supplanted biased sampling studies in the literature of reliability. This makes sense, given that (i) exhaustive and random sampling studies can provide the overall convergence rate between methods, and (ii) online crowdsourcing platforms have made exhaustive and random sampling studies much more practical. Sprouse and Almeida 2012 exhaustively tested all of the English acceptability judgment data points in Adger's 2003 textbook *Core Syntax*. They defined convergence relatively conservatively – statistical significance (using both null hypothesis testing and Bayes factor analysis) in the direction reported by Adger – but crucially assumed that there would be variability in judgments. They found that 98% of the data points replicated in their formal experiments. Sprouse, Schütze, and Almeida 2013 randomly sampled 300 data points (forming 150 two-condition phenomena) from articles published in the journal Linguistic Inquiry between 2001 and 2010, and tested these 150 two-condition phenomena using three different judgment tasks. Using the same relatively conservative criteria as Sprouse and Almeida 2012, they found that 95% of the sampled data points replicated in their formal experiments. Given the size of their sample relative to the number of data points published between 2001 and 2010, the 95%

result can be used as an estimate for the overall convergence rate for judgments in LI 2001-2010 with a margin of error of ±5. Sprouse et al. internally replicated their own results with a different sample of participants (for between-participants reliability), and found the same 95% convergence rate. Mahowald, Graff, Hartman, and Gibson 2016 closely replicated this finding (observing a 92% convergence rate), crucially with a novel random sample of phenomena, and with a crowdsourced approach to materials construction (students from a large psychology class at MIT created the materials) rather than relying on professional linguists to create the materials (Mahowald et al. use these results to provide guidelines for the sample size of experiments, a topic that we discuss in more detail in the next section). I have heard that there are similar random sampling studies either completed or underway in Korean, Japanese, and Spanish; but at the time of publication, I have been unable to find published versions of these projects.

Exhaustive and random sampling studies provide us with several pieces of information that may be relevant for assessing the validity and reliability of informally collected judgments. First, they provide convergence rates between the two methods that we can attempt to interpret. The interpretation of these rates is ultimately subjective, so it is up to individual researchers to determine what level of convergence is required to increase confidence in the validity or reliability of informal methods (see, for example, the discussion formed by Gibson and Fedorenko 2013, Sprouse and Almeida 2013, and Gibson, Piantadosi, and Fedorenko 2013). My subjective opinion is that the observed convergence rates of 92%-98% are impressive. I know of no other area of cognitive science that has replication rates this high (see Open Science Collaboration 2015 for estimated replication rates in other areas of psychology, all in the range of 36%-53% using similar definitions of replication as the one used in the judgment studies). Second, these results begin to establish a set of known divergences: Linzen and Oseki find 7 Hebrew and 4 Japanese divergences; under one count, Sprouse and Almeida 2012 find 6 (English) divergences in Adger 2003; under one count, Sprouse et al. 2013 find 9 (English) divergences in Linguistic Inquiry. Under one count, Mahowald et al. find 8 English divergences. Though these are relatively few compared to the hundreds of data points that have been investigated, they still deserve follow-up work, as the divergence itself does not tell us which result better reflects reality. Relatedly, these results provide some information about the types of divergences that we find. The vast majority of the divergences involve a null result in the formal experiments. These null results are ambiguous between a true divergence and low statistical power for the size of the effect to be detected. Very few of the divergences involve a sign reversal – an effect in the formal experiment that is opposite in direction to the effect reported using informal methods. This suggests that the most egregious sources of differences between the two methods, such as a contamination of theoretical bias from professional linguists, are relatively rare (see also Dabrowska 2010 for evidence that the differences in judgments between professional linguists and non-linguists on a rating scale at most differ quantitatively, not qualitatively). Finally, it should be noted that the vast majority of this work has been done in English, and has focused on standard acceptability judgments that do not involve co-reference, prosody, or multiple interpretations. Though the results of these studies have been encouraging, there is very clearly a need for studies in other languages, and a need for studies on different judgment types.

3. The differences among acceptability judgment tasks

A second set of questions that formal judgment experiments are particularly well-suited to explore concerns the differences among judgment tasks. The results of the exhaustive and random sampling convergence studies discussed in section 2 suggest that, at least when it comes

to detecting differences between sentence types, informal methods provide valid and reliability results. Furthermore, the relatively narrow range of results in those studies suggests that all of the formal experimental tasks that those studies employed provide roughly similar information. This is reassuring, as it suggests that all of the major methods for collecting acceptability judgments tap into the same underlying cognitive states and/or cognitive processes. But there are still a number finer-grained questions that one could ask about the differences between tasks in order to optimize the use of formal experimental methods to address specific hypotheses in the syntactic literature. In this section, I will review a number of the most prominent questions that have been asked in the literature, as well as point out some of the questions that have not, to my knowledge, been systematically investigated yet.

The first question one could ask is whether different tasks provide different pieces of information about acceptability. The answer is almost certainly yes given that different tasks ask participants to provide acceptability judgments in different ways. Here I will mention three classes of tasks to illustrate these differences. *Rating tasks* ask participants to rate individual sentences along a scale with some number of points (such as 1-7). Rating tasks can provide information about the rating of individual sentences along what is assumed to be a linear rating scale with regular intervals between the numbers on the scale. This information can be used to describe the absolute acceptability of the sentence, or the ratings of two sentences can be used to calculate the size of the difference between them. *Categorization tasks* ask participants to assign sentences to some number of nominal categories (without the assumption of regular linear distances between the categories). The most common example is the two-alternative forced-choice task that asks participants to rate sentences as acceptable or unacceptable. Categorization tasks can be used to determine the category membership of sentences in cases where theories make predictions about some number of acceptability categories. But they provide only coarse information about the absolute acceptability of the sentences. They also provide only coarse information about the size of the difference between two sentence types (in the form of the difference in the proportion of responses to each sentence). *Selection tasks* ask participants to choose a sentence out of a set of sentences. The most common example is a two-alternative forced-choice task that asks participants to select the more acceptable sentence in a pair. Selection tasks are extremely sensitive to the presence or absence of a difference between sentence types (because the explicit task is to detect a difference), but provide no information about individual acceptability, provide no information about categories, and only provide very coarse information about the size of differences between sentence types (through the proportion of selections). Given these differences, at the most general level, the choice of task should reflect the type of information that is necessary to test the hypothesis under consideration.

The second question one could ask is whether there are optimizations to be made within each class of tasks, or in other words, if there is an optimal instantiation of a specific task. Bard, Robertson, and Sorace 1996 and Cowart 1997 initiated a new line of research in judgment methodology by asking just this question about rating tasks. As both note, the psychophysicist Stanley Smith Stevens observed that typical rating tasks potentially suffer from (at least) two drawbacks. The first is that the number of response options is typically finite. If the number of options is too small, participants may be able to perceive differences among the stimuli of the experiment that they cannot report using the response scale. The second is that the rating scale assumes that the intervals between the response points are uniform, that is, the interval between 1 and 2 is the same size as the interval between 4 and 5. Though this may turn out to be true, it cannot be guaranteed for every participant or every stimulus. Stevens (1956) proposed a new type of rating task called *magnitude estimation* to eliminate these two potential drawbacks. Bard et al. 1996 and Cowart 1997 adapted Stevens' magnitude estimation task to acceptability

judgments. In the most typical version of magnitude estimation of acceptability, one sentence, called the standard, is presented to the participants along with a number, called the modulus, which represents the standard's acceptability. The standard is typically chosen such that it is somewhere in the middle range of acceptability, and the modulus is typically set to be a number that is easy to divide and multiply, such as 100. Participants are then shown other sentences without any acceptability numbers associated with them. They are told to rate each sentence as a multiple or fraction of the standard. If the sentence is twice as acceptable, the rating should be 200; if the sentence is half as acceptable, the rating should be 50. The fundamental idea is that the standard becomes a type of perceptual measurement unit that participants use to measure the acceptability of the sentences in the experiment. Because the response scale is the positive real number line, participants can report as many distinctions among stimuli as they want. And because there is just one interval (the acceptability of the standard), there is no risk of unequal interval distances. Both Bard et al. 1996 and Cowart 1997 demonstrated the potential utility of magnitude estimation (and also some of the potential drawbacks of finite response scales). Their results inspired a number of syntacticians to explore the utility of magnitude estimation across a number of phenomena (e.g., Keller 2000, Sorace 2000, Featherston 2005a).

Ultimately, the rising popularity of magnitude estimation led a number of researchers to directly test its purported advantages over rating tasks. Weskott and Fanselow 2011 compared the results of magnitude estimation and a standard rating task for three phenomena in German and found that magnitude estimation leads to higher variability, and consequently smaller standardized effect sizes and less statistical power. They conclude that this may be a consequence of the larger response scale in magnitude estimation. Sprouse 2011 adapted methods from the psychophysics literature to test one of the fundamental assumptions of magnitude estimation – that participants can make ratio judgments of the acceptability of a sentence (i.e., that it is a multiple of the acceptability of the standard). The results suggest that participants cannot make ratio judgments of acceptability. This runs contrary to the results for other types of psychophysical judgments in the literature, which has historically demonstrated that participants can make ratio judgments of physical stimuli like brightness and loudness. The impossibility of ratio judgments of acceptability is likely because acceptability has no true zero point that represents the absence of all acceptability. Without a true zero, ratio judgments are impossible. The Sprouse 2011 results suggest that participants cannot do true magnitude estimation with acceptability; therefore Stevens' arguments that magnitude estimation is superior to other rating tasks simply do not apply to acceptability judgments. Participants in magnitude estimation experiments must be covertly converting the magnitude estimation task into some other type of rating task that they can actually complete, and this covert rating task likely suffers from the same potential drawbacks as other rating tasks. This in turn leads to a further interpretation of the Weskott and Fanselow 2011 results: in both cases participant are completing a rating task; but in the case of magnitude estimation, the rating task that they are adopting is leading to more variability and lower statistical power (perhaps due the unbounded response scale, or perhaps due to other issues that arise when participants are asked to perform a task that they cannot cognitively complete). As such, there is no argument, either logical or empirical, to support the use of magnitude estimation over other rating tasks for acceptability judgments.

The third question that one could ask is whether different tasks yield different levels of sensitivity to acceptability judgment differences. Bard et al. 1996 began this line of questioning by comparing several conditions in both a rating task and magnitude estimation. They report that magnitude estimation allows participants to report more levels of acceptability than a rating task with 5 points. Bader and Haüssler 2010 compared magnitude estimation to a two-alternative categorization task for 16 sentence types (forming one 2x2 design and two 2x3 designs) in order

to create a signal detection model for acceptability judgments. In the process, they report that the two tasks yield the same pattern of statistically significant results at two sample sizes: 24 and 36 participants. As previously mentioned, Weskott and Fanselow 2011 compared magnitude estimation and rating tasks for three phenomena, and found more variability for magnitude estimation, which in turn decreases statistical power. Sprouse and Almeida 2017 tested 47 two-condition phenomena taken from Linguistic Inquiry (from Sprouse et al. 2013) using all four types of tasks: a 7-point rating task, and two-alternative categorization task, a two-alternative selection task, and magnitude estimation. They used resampling simulations to estimate statistical power for each phenomenon and each task at a range of sample sizes from 5 to 100 participants. Their results suggest that the selection task has the most statistical power for detecting differences between two conditions (as expected given the logic of the task), that the rating task and magnitude estimation tasks have similar statistical power (both less than the selection task), and that the categorization task has the least statistical power (again, as expected given that it will, by definition, have trouble detecting differences between two conditions that fall into the same category). Their results can also be used to estimate the number of participants necessary for a given level of statistical power for a wide range of effect sizes. Langsford, Perfors, Hendrickson, Kennedy, and Navarro 2018 extend these findings (using the Sprouse et al 2013 materials) in two ways: (i) they focus on test-retest reliability (instead of statistical power), and (ii) they include an additional task from psychophysics, the Thurstone method (Thurstone 1927). The Thurstone method is a combination of a two-alternative selection task in which the pairs of sentences are random, and a modeling procedure which converts judgments of the random pairings into an ordering among all of the test items along an inferred acceptability scale. Their results suggest that the Thurstone method is not superior to the best performing traditional methods, as traditional rating tasks and selection tasks demonstrate the best combination of within-participant and between-participant test-retest reliability. That said, the high degree of correlation between the Thurstone method, which makes very few assumptions about the nature of acceptability, and the traditional rating task, which makes many assumptions by virtue of the structure of the rating scale, adds an extra dimension of validation to the traditional rating task. In fact, spread throughout all of the papers discussed in this paragraph is quite a bit of information about the correlation among the various methods. In all cases, the methods appear to be yielding highly correlated results, modulo the differences in the types of information the tasks yield, and minor differences in reliability.

Though there has obviously been much work on the effects of different tasks, there are still a number of properties of judgment tasks that have not been systematically investigated. What is the contribution of task instructions? Cowart 1997 has some initial results on this, suggesting that there is relatively little impact of the instructions, but I know of no other systematic studies. What is the contribution of individual versus paired presentation of sentences? This is typically confounded in the difference between rating tasks and selection tasks, but could in principle be separated. What is the effect of the number of judgments per condition per participant on statistical power? The existing studies all used one judgment per condition, so they offer only a minimum estimate. What is the effect of the number of fillers (unrelated, and typically unanalyzed items) in the experiment? It is often assumed that the judgments of individual sentences can be pushed higher or lower by including them with different types of fillers (e.g., extremely unacceptable fillers could lead to higher ratings for otherwise unacceptable sentences). What is the effect of the number of distinct items created for each condition? This interacts with the debate in the statistical literature about random effects and item generalizability (e.g., Clark 1973, Wike and Church 1976, Raaijmakers, Schrijnemakers, and Gremmen 1999, Barr, Levy, Scheepers, and Tily 2013). These questions

also interact with the proposal by Myers 2009 that there may be a middle ground between typical informal experimental methods and fully formal experimental methods that he calls *small scale* experiments. The answers to these questions could both inform the construction of full-fledged formal experiments, and also help determine exactly how small scale experiments can be for different syntactic questions.

4. The source of acceptability judgment effects

As briefly discussed in section 1, acceptability judgments are typically assumed to be impacted by a number of factors. This means that the source of any given acceptability judgment effect is ambiguous: the effect could be due to a syntactic constraint violation, a violation in a different part of the grammar, some component of sentence processing (beyond the recognition of a syntactic violation), word or construction frequency, sentence plausibility, or any number of other factors that impact sentence comprehension. The primary tool for dealing with this ambiguity is experimental design – constructing conditions to isolate a potential syntactic effect to the exclusion of all other possible types of judgment effects. Experimental design can be leveraged this way with either informal or formal judgment methods; and, in fact, many syntax articles that use informally collected judgments include explicit manipulations designed to exclude extra-syntactic explanations for acceptability effects. That said, formal judgment experiments can contribute to the investigation of the source of judgment effects in two ways: by formalizing the process of designing an experiment to isolate the factors that contribute to acceptability (through factorial logic), and, in the instances where it is impossible to logically separate syntactic effects from other possible effects, by quantifying judgments in a way that allows us to test predictions that involve data types beyond offline acceptability judgments. In this section, I will use island effects as an example phenomenon to illustrate these two properties of formal judgment experiments.

Factorial logic is a formalization of the process that all experimentalists use to test for the presence of an effect (including syntacticians who use informal experiments to collect acceptability judgments). The term *factor* means a property that can be manipulated, such as some dimension of the structure of a sentence; the term *level* is used to refer to the specific values that a factor can take. Factors can be continuous (an infinite number of levels) or categorical (a finite number of levels). The goal of factorial logic is to isolate effects using subtraction logic. As a concrete example, we can look at a factorial design for island effects. As a first definition, we can define island effects as the low acceptability that arises when the tail of a long-distance dependency is contained within a specific structure, called an island structure. The *whether*-island sentence in (1d) below is a classic example: the tail of the wh-dependency is contained within the embedded *whether* clause. The space of possible sources for this effect is large; it is bounded by the list of factors that we believe contribute to acceptability judgments (i.e., the linking hypothesis that was briefly discussed in section 1 and in the previous paragraph). For this example we will consider two possible sources. The first possibility is that there is a constraint in the grammar that specifically targets the syntactic structure of (1d) and rules it ungrammatical (which then leads to unacceptability when coupled with an appropriate linking hypothesis between ungrammaticality and acceptability). A second possibility proposed by Kluender and Kutas 1993 is that the sentence in (1d) is grammatical, but that the low acceptability is the result of the combination of two types of processing complexity: the complexity associated with processing a long-distance dependency, and the complexity associated with processing the island structure itself (in this case, the embedded *whether* clause). The Kluender and Kutas (1993) theory suggests that there are (at least) two acceptability

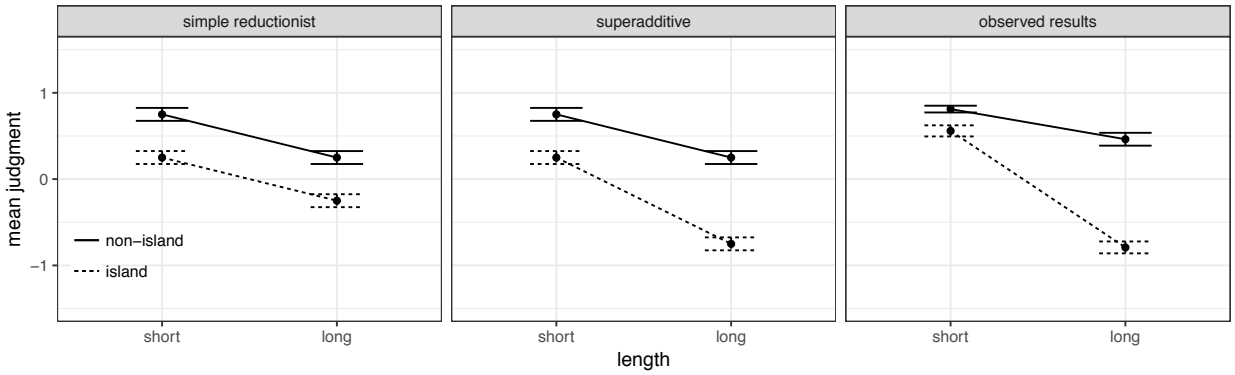
judgment effects that any investigation of island effects will want to quantify: the effect of a long-distance dependency, and the effect of parsing the island structure. We can use factorial logic to isolate these two effects with two factors. The first we can call DEPENDENCY LENGTH. At its simplest, we can define two levels for DEPENDENCY LENGTH: short and long. The subtraction between these two will isolate the effect of a long-distance dependency. The second we can call STRUCTURE. Again, at its simplest, we can define two levels for STRUCTURE: non-island and island. The subtraction between the two will isolate the effect of parsing an island structure. With two factors, each with two levels, we have a 2x2 design (each digit in this label represents a factor, and each value of the digit represents the number of levels), which yields four conditions, as in (1):

(1)	a.	Who __ thinks that Mary wrote a book?	short | non-island
	b.	What do you think that Mary wrote __?	long | non-island
	c.	Who __ thinks that Mary wrote a book?	short | island
	d.	What do you wonder whether Mary wrote __?	long | island

The factorial design in (1) lends itself to subtraction logic in the following way. The subtraction of (1a-1b) isolates the effect of processing a long-distance dependency by subtracting the short and long levels of DEPENDENCY LENGTH while holding STRUCTURE constant (and not involving the potentially ungrammatical sentence). The subtraction of (1a-1c) isolates the effect of processing the island structure by subtracting the non-island and island levels of STRUCTURE while holding DEPENDENCY LENGTH constant. Armed with these quantities, it becomes possible to test two competing classes of theories of what it means to be an island effect. If an island effect is completely reducible to the combination of the dependency length and structure processing effects, then we would predict that the difference (1a-1d) should be a linear sum of the differences (1a-1b) and (1a-1c); in other words: (1a-1d) = (1a-1b) + (1a-1c). In contrast, if island effects are more than just the linear combination of these two processing effects, as predicted by grammatical theories, as well as more complex sentence-processing based theories, we would expect that the difference (1a-1d) will be larger than the linear sum of these differences; in other words: (1a-1d) = (1a-1b) + (1a-1c) + X, where X is some additional effect that is not isolated by any of the factors. In statistical terms, we would say that the simple reductionist theory in which island effects are the result of linearly combining the two processing effects predicts two main effects, one for dependency length and one for structure, but no interaction of the two factors. The class of more complex theories predicts that there will be a superadditive interaction between the two factors, such that combining the long and island levels leads to a larger effect than the linear sum of the two factors alone.

The left panel of Figure 1 illustrates the prediction of the simple reductionist theory. The parallel lines indicate that the acceptability of the long|island condition is the linear sum of the two factors. The center panel of Figure 1 illustrates a superadditive interaction in which the acceptability of the long|island condition is lower than one would expect based on the two factors alone. Finally, the right panel of Figure 1 shows the observed results of a real acceptability judgment experiment using this design (with 32 participants, participants rated two token of each condition, 9 practice items, and 14 fillers using a 7-point rating task) .

Figure 1: The two predictions of the 2x2 design for *whether*-islands (left panel and center panel), and the observed results of an actual experiment (right panel).



The fact that we observe a superadditive interaction in the real experiment suggests that the simple reductionist theory is incorrect. *Whether*-island effects are not the simple linear sum of the effects of long-distance dependencies and island structures. This result suggests that we must explore the class of more complex theories to explain *whether*-island effects. A number of studies have used this factorial logic to explore the variation in the presence and absence of different types of island effects across the world's languages. Here is a non-exhuastive list covering an interesting subset of languages: Arabic: Tucker, Idrissi, Sprouse, and Almeida 2019; Danish: Christensen, Kizach, and Nyvad 2013; English: Sprouse 2007, Sprouse, Wagers, and Phillips 2012; Goodall 2015; Hofmeister, Culicover, and Winkler 2015, Atkinson, Apple, Rawlins, and Omaki 2016; Italian: Sprouse, Caponigro, Greco, and Cecchetto 2016; Japanese: Sprouse, Fukuda, Ono, and Kluender 2011; Korean: Kim and Goodall 2016; Norwegian: Kush, Lohndal, and Sprouse 2018; Slovenian: Stepanov, Mušič, and Stateva 2018.

Before we explore the class of more complex theories, it is worth noting that factorial designs like the one discussed here can be used to define a necessary condition for the existence of a syntactic explanation for an acceptability judgment effect. A syntactic explanation entails that the acceptability effect cannot be fully explained by the other factors that impact acceptability (e.g., processing effects, semantic effects, task effects). In statistical terms, a syntactic explanation entails that there is no factorial design consisting of solely of non-syntactic factors that leads to linear additivity; all such designs will yield superadditivity (because, definition, they do not contain a factor to capture the syntactic constraint). Another way to view this is that, given a factorial design consisting exclusively of factors that lie outside of the theory of syntax, linear additivity as illustrated in the left panel of Figure 1 is deductive evidence for a simple reductionist approach to acceptability judgment effects, while superadditivity as illustrated in the other two panels of Figure 1 is ambiguous between a syntactic explanation or a complex interaction of extra-syntactic factors. In short, supreadditivity in these designs is a necessary, but not sufficient, condition for a syntactic explanation. This is not a new observation. Careful work in theoretical syntax has always incorporated both factorial logic and this necessary condition for the existence of syntactic explanations. This is easily seen in the syntax literature where factorial designs are common albeit rarely described using factorial terminology (see also Myers 2009 for a discussion), and where researchers often demonstrate that non-syntactic factors cannot completely explain the judgment effect. The point here is simply that formal experiments allow us to make this logic explicit, and potentially allow us to isolate a larger number of factors simultaneously, insofar as they make quantifying judgments, and keeping track of multiple conditions, a bit easier.

At this point it is clear that the superadditive pattern for *whether*-islands in Figure 1 could either be due to a syntactic constraint targeting condition (1d), or it could be due to an interaction between the processing of long-distance dependencies and the processing of island structures

(which only arises in condition 1d). Teasing these two explanations apart is not trivial. This brings us to the second benefit of formal judgment experiments when it comes to identifying the source of acceptability judgment effects – the ability to test hypotheses that go beyond offline acceptability judgments. One example of this is looking for relationships between acceptability judgments and other data types that might be indicative of a causal relationship between extra-syntactic factors (in the case of islands, sentence processing factors) and the acceptability effect. A number of studies have explored this approach. Stowe 1986 was perhaps the first, using self-paced reading to demonstrate (i) that the parser attempts to complete long-distance dependencies at the first available gap location (a strategy that Frazier and Flores d'Arcais 1989 later named the active filling strategy), and (ii) that the parser does not attempt to complete dependencies inside of finite subject islands. Phillips 2006 reviews a number of studies that extend Stowe's finding using different islands and different data types (eye-tracking, ERPs). Phillips 2006 also demonstrates that the behavior of the parser is even more sophisticated than previously thought, as it not only suppresses the active filling strategy inside of islands, but also allows active filling inside of islands that can participate in parasitic gap constructions (Engdahl 1983; see also Culicover 2001, for a review of parasitic gaps). The crucial fact here is that these islands do give rise to unacceptability when there is only one gap in the construction (inside of the island). Phillips' result suggests that this unacceptability cannot be due to the inability of the parser to complete a dependency inside of these islands, because there is reading time evidence that the parser does complete the dependency. This dissociation between acceptability and first-pass sentence processing means that the resulting unacceptability must be due to some later process, such as a check against the grammatical requirements of a parasitic gap configuration (i.e., a second gap outside of the island).

Sprouse, Wagers, and Phillips 2012 take a slightly different approach. Instead of investigating real time sentence processing effects, they investigate the relationship between acceptability judgments and individual variation in working memory capacity. The rationale behind this is a specific proposal by Kluender and Kutas 1993 that island effects are the result of limited working memory resources, such that the simultaneous processing of long-distance dependencies and island structures taxes working memory resources beyond their capacity, leading to the perception of unacceptability. Sprouse et al. interpret this proposal to predict that there should be a detectable inverse relationship between working memory capacity and the size of island effects as quantified by the superadditive interaction in acceptability judgments – as working memory increases, the size of the island effect should decrease. They tested a large number of participants on two working memory tasks (serial recall and *n*-back) and four island types, yet found no evidence of a relationship (see also Michel 2014 for a replication using a third working memory task, serial recall).

Yoshida et al. 2014 test a second prediction of the Kluender and Kutas working memory theory, namely that, if island effects are due to working memory capacity, island effects should arise for any dependency that triggers the same working memory requirements as wh-dependencies. It has long been established in the syntactic literature that island effects, as defined as acceptability judgment effects, do not arise for binding. But Yoshida et al. demonstrate that island effects also have no impact on the real-time processing of binding dependencies. Their study builds upon previous work showing that the parser attempts to resolve binding dependencies in which an anaphor appears before its antecedent (called cataphora, or backwards anaphora) at the first possible opportunity during real-time processing (van Gompel and Liversedge 2001, Sturt 2003, Kazanina, Lau, Lieberman, Yoshida, and Phillips 2007). This process shares a number of similarities with active gap filling, including recruiting the same areas of the left inferior frontal gyrus in the brain (Matchin, Sprouse, and Hickok 2014). Despite

these similarities, Yoshida et al. demonstrate that island structures do not suppress the search for an antecedent for binding during real-time processing, contrary to the plausible prediction of the working memory approach to island effects.

Another classic example of testing predictions beyond offline acceptability judgments can be found in the syntactic satiation literature. Syntactic satiation is the phenomenon by which acceptability judgments for a specific sentence type appear to increase with repeated exposures to that sentence type. Syntactic satiation has long been informally reported by professional linguists, particularly over the course of weeks or months working on a specific phenomenon. The satiation studied in the judgment literature is different than this, as the goal has been to induce satiation in non-linguist participants over short (single experiment) timescales, thus potentially linking the effect to syntactic priming, as recently discussed in Do and Kaiser 2017, or implicit learning, as in Luka and Barsalou 2005. The first systematic study that I am aware of is Snyder 2000. He tested seven different sentence types instantiating distinct syntactic violations, and found that three violation types showed evidence of satiation over the course of five exposures, but that four other violation types did not. Though these results were not designed to probe the source of the satiation effects (as it was just a first study), Snyder suggested that one possibility might be that satiation differs based on the source of the acceptability judgment effect. His specific claim was that judgment effects that satiate may be due to sentence processing sources, while judgment effects that do not satiate may be due to grammatical sources. The underlying idea appears to be either that the difficult parsing processes themselves might get easier with time, or that licit grammatical representations might somehow become easier to construct with time. In either case, the prediction would be that acceptability judgment effects due to sentence processing should satiate, whereas effects due to grammatical violations should not. The complication with the satiation literature is that satiation effects have proved particularly variable across experiments. Several authors have attempted to replicate and extend Snyder's results, with mixed results (see Hiramatsu 2000, Sprouse 2009, Francom 2009, Goodall 2011, and Do and Kaiser 2017 for English islands among other violation types; see Christensen et al. 2013 for Danish islands). This has led some authors to raise the possibility that source of satiation effects may not be as theoretically deep as Snyder suggested, perhaps instead reflecting: (i) task effects such as identifiability (and therefore correctability) of the violations (Francom 2009), (ii) response strategies intended to equalize the number of responses along the provided response scales (Sprouse 2009), (iii) perhaps mere exposure effects (Luka and Barsalou 2005), (iv) or syntactic priming (Do and Kaiser 2017).

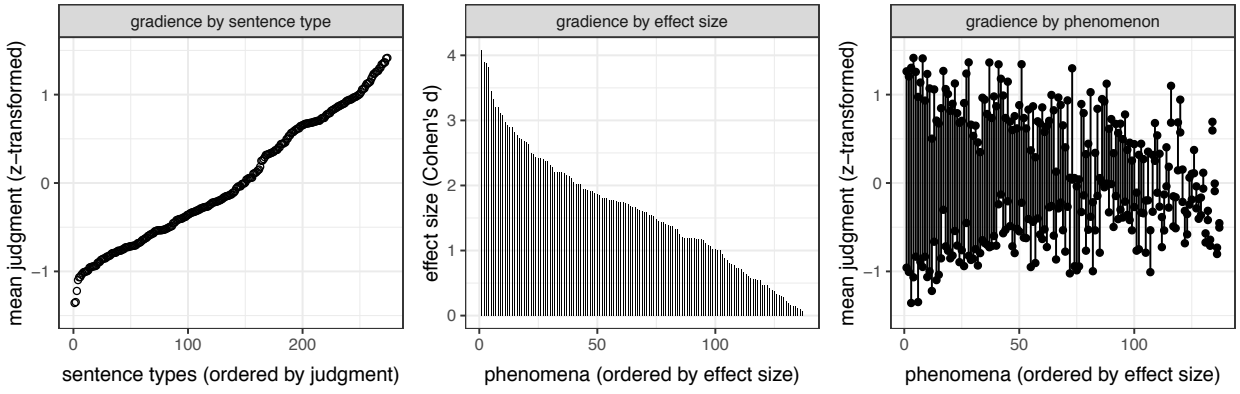
Though this section has reviewed a number of studies that have attempted to identify the source of judgment effects, there are still a number of directions that need to be explored. For syntactic theory as a whole, there is relatively little formal experimental research into the source of acceptability judgments for the vast majority of phenomena. This could be because there are no other potential candidate theories to explain the judgment effects; that is, it could be the case that island effects are a special case in that there are potential extra-syntactic explanations, like the Kluender and Kutas 1993 working memory theory. Or this could be because experimental methods have only recently been adopted on the scale necessary to do this kind of research. For island effects, there is still relatively little formal experimental research into grammatical explanations outside of syntax, like the semantic approaches of Szabolcsci and Zwarts 1993 or Abrusán 2011, or the pragmatic approaches of Erteschik-Shir 1973 or Goldberg 2007. At the level of sentence processing, there are still a number of open questions about the processing of long-distance dependencies, including to what extent active gap filling is suppressed in islands other than subject islands, and to what extent there are second-pass gap-filling mechanisms in constructions like parasitic gaps. And when it comes to satiation, there is still much work that

needs to be done to both solidify the empirical results and develop a causal theory of satiation (at both short, single experiment timescales, and at the longer timescales experienced by professional linguists).

5. The architecture of the grammar

Any acceptability judgment collection method, be it relatively informal or relatively formal, can contribute to investigations of the architecture of the grammar. Acceptability judgments are, after all, the primary data type of syntactic theory; and the architecture of the grammar is, after all, the primary object of study of syntactic theory. The special relevance of formal judgment methods for questions about the architecture of the grammar is that formal methods allow us to quantify acceptability judgments with a level of precision that makes it readily apparent that acceptability is a continuous measure. Figure 2 shows this in three ways. The left panel shows the mean acceptability for the 300 individual sentence types from Sprouse et al. 2013, arranged in order of increasing acceptability. It is clear that there are no substantial step-like breaks in these means to indicate quantized acceptability. The center panel shows the mean effect size for the 150 pairwise phenomena from Sprouse et al. 2013, arranged in order of increasing size. Again, there are no substantial step-like breaks to indicate quantized effect sizes. The right panel shows the same pairwise phenomena, ordered by effect size, but with the ratings of each sentence type represented by the end points of the lines. There is nothing particularly novel in these results. The continuous nature of acceptability has been acknowledged since the earliest days of generative syntax (Chomsky 1957). However, the rise of formal experimental methods for judgment collection has made it easier than ever to demonstrate the continuous nature of acceptability, which in turn has led to a renewed interest in determining its source. This is a complex question – so much so that there is an entire chapter dedicated to it in this volume (chapter X). I will leave a detailed review of this topic for that chapter. Here I simply want to touch upon three of the questions that drive current research into continuous acceptability and what it may, or may not, reveal about the architecture of the grammar.

Figure 2: Three demonstrations of the continuous nature of acceptability judgments.



The first question we can ask is whether the grammar itself is categorical, yielding some discrete number of levels of grammaticality, or continuous, yielding an infinite number of levels of grammaticality. Both types of grammars can explain continuous acceptability, albeit in different ways. For categorical grammars, continuous acceptability must be entirely the result of continuous extra-grammatical factors, such sentence processing mechanisms, plausibility, or

even task effects (as Armstrong, Gleitman, and Gleitman 1983 showed, even concepts that are categorical by definition, such as *even number*, will receive continuous judgments under certain rating tasks). For continuous grammars, continuous acceptability is the result of a combination of continuous extra-grammatical factors (sentence processing mechanisms, plausibility, task effects) and the infinite levels of grammaticality made available by the grammar. What this means in practice is that continuous acceptability itself is not dispositive of the two architectures. Instead, we must build relatively complete theories of both types, to see which better explains the phenomena we wish to explain (continuous acceptability judgments, sentence processing facts, language acquisition facts, etc.).

Assuming that we want to pursue the strategy of constructing both grammar types to evaluate their ability to predict continuous acceptability, the second question we can ask is how to make grammatical theories continuous. One option is to directly add abstract weights to the constraints that already exist in familiar grammatical theories. Two famous examples of this direct approach are Keller's (2000) Linear Optimality Theory (see also Keller 2006, which includes a nice comparison of Linear Optimality Theory to other grammar types, such as Harmonic Grammar and Stochastic Optimality Theory), and Featherston's (2005b) Decathlon model. Though these direct approaches to gradience are relatively successful at explaining continuous acceptability judgments, the direct approach leads to two complications. The first is that, even for continuous grammars, continuous acceptability is not solely driven by the grammar; it is also driven by continuous extra-grammatical factors. This means that we cannot determine constraint weights directly from acceptability judgments. We still need a relatively complete theory of extra-grammatical factors to help us determine how much of the continuous acceptability is due to the grammar, and how much is due to the grammar. The second complication is that, if abstract constraint weights vary cross-linguistically, they must be learned by children acquiring the language. We already know from the acquisition literature that learning syntactic constraints is a complex problem in its own right. Adding weights to these constraints compounds the acquisition problem. Not only does this dramatically increase the space of possible grammars that children must explore (the space of all possible combinations of constraints and all possible combinations of weights), it also raises difficult questions about what would count as evidence for the acquisition of constraint weights. As researchers, we can measure acceptability from native speakers in an experiment, and then use that information to develop the weights for a continuous grammar; but children likely do not have access to the (presumably internal) acceptability judgments of the speakers around them. This suggests that purely abstract weights are likely only viable under a strongly nativist theory in which either the constraints are innately specified, the weights are innately specified, or both.

A second option is to avoid purely abstract constraint weights, and instead link the constraint weights to continuous quantities that exist in the language system for independent reasons. The space of possible quantities that could be used to ground constraint weights is relatively large, and therefore beyond the scope of this chapter. However, I would like to mention that one robust area of research involves using frequencies of occurrence (as estimated through natural language corpora) to derive probabilities that can be added in various ways to different grammar types to yield gradient outputs. Classic examples of this can be found in the phonology literature, where Stochastic Optimality Theory (e,.g., Boersma and Hayes 2001) and Maximum Entropy grammars (e.g, Jäger 2007) provide two ways to convert frequencies into grammar internal quantities that can yield gradience like we see with continuous acceptability. Classic examples of this can also be found in the computational linguistics literature, where various grammar types, such as context free grammars, have long been augmented with probabilities to capture the gradience that we see in production frequencies. Hunter and Dyer

2013 have recently extended this work to develop probability distributions for minimalist grammars (in the sense of Stabler 1997). And Bresnan 2007 has demonstrated the utility of systematically investigating the psychological factors that can influence gradience in production through a case study of the dative alternation in English. To my knowledge, none of these frameworks have been implemented as comprehensively in service of explaining continuous acceptability as the direct approach to abstract constraint weights, but these previous studies do show that it is possible in principle to ground continuous acceptability in an independently motivated quantity.

Assuming that we wish to pursue the strategy of explaining continuous acceptability through something like a continuous grammar, the final question we can ask is whether the best strategy is to look for ways to add continuous properties to existing grammatical architectures (e.g., Optimality Theory, Minimalism), or whether we should instead begin to consider new grammatical architectures altogether. The integrated connectionist/symbolic (ICS) cognitive architecture developed by Paul Smolensky and colleagues (most recently in the Harmonic Grammar framework; Smolensky and Legendre 2006) has long been at the forefront of this line of research. The ICS architecture provides a linking theory between continuously valued neural networks and categorical symbolic grammatical theories like Optimality Theory and Harmonic Grammar that are more familiar within linguistics. This raises the possibility of grounding quantities like continuous acceptability in low-level quantities of the cognitive architecture itself, such as connection weights or Harmony. Moving in a slightly different direction, Lau, Clark, and Lappin (2017) suggest that adding probabilities directly in the grammar can not only help to capture the continuous nature of acceptability, but can also make grammatical architectures that linguists have typically dismissed as inadequate to capture human grammars more viable, such as n-gram models (see Chomsky 1956) and simple recurrent neural networks (see Marcus 2001) that are trained solely on surface word strings from a naturally occurring corpus. There is some debate about how successful these models are at explaining syntactic phenomena (e.g., Sprouse, Yankama, Indurkhya, Fong, and Berwick 2018), but the broader point still holds – the addition of continuous quantities to grammatical theories could motivate a re-evaluation of the adequacy of different grammatical architectures. The recent explosion of work in machine learning using neural networks (that the Lau et al. paper is part of) is poised to dramatically expand the range of possible theories that linguists might consider for explaining acceptability judgments (see also Warstadt, Singh, and Bowman 2018 for a neural net that was trained to perform categorical judgments).

6. Conclusion

Acceptability judgments have been the primary data type in (generative) syntactic theory for over 60 years, and will likely continue to be, at the very least, a substantial component of the empirical base of syntactic theory for many years to come. This is because they provide the kind of information that syntacticians need to construct syntactic theories – information about which sentence types are licensed by the grammar and which are not (that is, assuming a linking hypothesis in which grammar is one of the factors influencing acceptability judgments). Formal experimental methods for judgment collection provide another useful tool for syntacticians to probe the nature of the grammar. Formal methods can be used to resolve both methodological questions, such as questions about validity, reliability, and sensitivity, and theoretical questions, such as questions about the source of acceptability effects and the architecture of the grammar. This chapter has reviewed some of the work that has been done to date on these questions. But to my mind, the work of using formal experimental methods has really just begun. It will be up to

the next generations of syntacticians to figure out how to leverage the power of formal experimental methods, across new empirical domains and new theoretical questions, in order to push the boundaries of syntactic theory.

References

Abrusán, Márta. 2011. Wh-islands in degree questions: A semantic approach. Semantics & Pragmatics 4: 1-44.

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

Armstrong, Sharon Lee, Gleitman, Lila R., Gleitman, Henry. 1983. What some concepts might not be. Cognition 13: 263–308

Atkinson, Emily, Aaron Apple, Kyle Rawlins, and Akira Omaki. 2016. Similarity of *wh*-phrases and acceptability variation in wh-islands. Frontiers in Psychology 6: 2048

Bader, Markus, and Jana Häussler. 2010. Toward a model of grammaticality judgments." Journal of Linguistics 46: 273-330.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. Language: 32-68.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language 68: 255-278.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. Linguistic Inquiry 32(1): 45–86.

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

Chomsky, Noam. 1956. Three models for the description of language. IRE Transactions on Information Theory 2: 113-124.

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

Christensen, Ken Ramshøj, Johannes Kizach, and Anne Mette Nyvad. 2013. Escape from the island: grammaticality and (reduced) acceptability of wh-island violations in Danish. Journal of Psychonlinguistic Research 42: 51-70.

Clark, Herbert H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of verbal learning and verbal behavior 12: 335-359.

Clifton Jr, Charles, Gisbert Fanselow, and Lyn Frazier. 2006. Amnestying superiority violations: Processing multiple questions. Linguistic Inquiry 37: 51-68.

Cowart, Wayne. 1997. Experimental syntax. Sage.

Culicover , Peter. 2001. Parasitic gaps: a history. In Peter Culicover and Paul Postal, editors, Parasitic Gaps. 3-68. Cambridge, MA; MIT Press.

Dąbrowska, Ewa. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. The linguistic review 27: 1-23.

Do, Monica L., and Elsi Kaiser 2017. The Relationship between Syntactic Satiation and Syntactic Priming: A First Look. Frontiers in Psychology 8: 1851.

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

Engdahl, Elisabet. 1983. Parasitic gaps. Linguistics and Philosophy 6: 5-34.

Erteschik-Shir, Nomi. 1973. On the nature of island constraints. PhD dissertation, MIT.

Featherston, Sam. 2005a. Magnitude estimation and what it can do for your syntax: some whconstraints in German. Lingua 115: 1525–1550.

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

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

Fedorenko, Evelina, and Gibson, Edward. 2010. Adding a third wh-element does not increase the acceptability of object-initial multiple-wh-questions. Syntax 13: 183-95.

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

Fillmore, Charles J. 1965. *Indirect Object Constructions in English and Ordering of Transfrmations*. The Hague: Mouton.

Francom, Jerid Cole. 2009. Experimental Syntax: exploring the effect of repeated exposure to anomalous syntactic structure--evidence from rating and reading tasks. PhD dissertation. The University of Arizona.

Frazier, Lyn, and Giovanni B. Flores d'Arcais. 1989. Filler driven parsing: A study of gap filling in Dutch. *Journal of memory and language* 28: 331-344.

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

Gibson, Edward, Steven T. Piantadosi, and Evalina Fedorenko. 2013. Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). Language and Cognitive Processes 28: 229-240

Goldberg, Adele. 2007. Constructions at work: The nature of generalization in language. Oxford University Press.

Goodall, Grant. 2011. Syntactic Satiation and the Inversion Effect in English and Spanish Wh-Questions. *Syntax* 14: 29-47.

Goodall Grant. 2015. The D-linking effect on extraction from islands and non-islands. *Frontiers in psychology*, *5*, 1493. doi:10.3389/fpsyg.2014.01493

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. The University of Connecticut.

Hofmeister, Philip, Peter Culicover, and Susanne Winkler.2015. Effects of processing on the acceptability of "frozen" extraposed constituents. Syntax 18: 464-483.

Hunter, Tim, and Chris Dyer. 2013. Distributions on Minimalist grammar derivations. Proceedings of the 13th Meeting on the Mathematics of Language.

Jäger, Gerhard. 2007. Maximum entropy models and stochastic Optimality Theory. *Architectures, rules, and preferences: variations on themes by Joan W. Bresnan. Stanford: CSLI*: 467-479.

Kayne, Richard S. 1983. Connectedness. Linguistic inquiry 14: 223-249.

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.

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

Keller, Frank. 2006. Linear Optimality Theory as a Model of Gradience in Grammar. In Gisbert Fanselow, Caroline Fery, Ralph Vogel, and Matthias Schlesewsky, eds., Gradience in Grammar: Generative Perspectives, 270–287. Oxford: Oxford University Press.

Kim, Boyoung, and Grant Goodall. 2016. Islands and non-islands in native and heritage Korean. Frontiers in Psychology 7: 134.

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

Kush, Dave, Terje Lohndal, and Jon Sprouse. 2018. Natural Language and Linguistic Theory 36: 743-779.

Langendoen, D. Terence, Nancy Kalish-Landon, and John Dore. 1973. Dative questions: A study in the relation of acceptability to grammaticality of an English sentence type. *Cognition* 2: 451-478.

Langsford, Steven, Amy Perfors, Andrew T. Hendrickson, Lauren A. Kennedy, and Danielle J. Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a journal of general linguistics* 3: 37.

Lau, Jey. H., Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A probabilistic view of linguistic knowledge. Cognitive Science 41(5). 1201-1241

Linzen, Tal, and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics* 3: 100.

Luka, Barbara J., 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: 436-459.

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

Marantz, Alec. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* 22: 429-445.

Marcus, Gary F. 2001. The algebraic mind: Integrating connectionism and cognitive science. MIT press.

Matchin, William, Jon Sprouse, and Greg Hickok. 2014. A structural distance effect for backward anaphora in Broca's area: an fMRI study. *Brain and Language* 138: 1-11.

Michel, Dan. 2014. Individual cognitive measures and working memory accounts of syntactic island phenomena. Doctoral dissertation. University of California, San Diego.

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

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

Phillips, Colin. 2006. The real-time status of island phenomena. Language 82: 795-823.

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

Raaijmakers, Jeroen GW, Joseph MC Schrijnemakers, and Frans Gremmen. 1999. How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. Journal of Memory and language 41: 416-426.

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

Smolensky, Paul, and Géraldine Legendre. 2006. The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1. MIT press.

Sorace, Antonella. 2000. Gradients in auxiliary selection with intransitive verbs. *Language* 76: 859-890.

Spencer, Nancy Jane. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of psycholinguistic research* 2: 83-98.

Sprouse, Jon. 2009. Revisiting Satiation. *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. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48: 609-652.

Sprouse, Jon and Diogo Almeida. 2013. The empirical status of data in syntax: A reply to Gibson and Fedorenko. Language and Cognitive Processes 28: 222-228.

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

Sprouse, Jon, Ivano Caponigro, Ciro Greco, & Carlo Cecchetto. 2016. Experimental syntax and the variation of island effects in English and Italian. Natural Language and Linguistic Theory 34: 307-344

Sprouse, Jon, Shin Fukuda, Hajime Ono, & Robert Kluender. 2011. Reverse island effects and the backward search for a licensor in multiple wh-questions. *Syntax* 14(2):179-203.

Sprouse, Jon, Carson T. Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. Lingua 134: 219-248.

Sprouse, Jon, Beracah Yankama, Sagar Indurkhya, Sandiway Fong, & Robert C. Berwick. (in press). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. The Linguistic Review 35: 575-599.

Stabler, Edward P. 1997. Derivational minimalism. In Christian Retoré, editor, Logical Aspects of Computational Linguistics. Lecture Notes in Computer Science 1328, 68-95, NY: Springer-Verlag.

Stepanov, Arthur, Manca Mušič, and Penka Stateva. 2018. Two (non-)islands in Slovenian: A study in experimental syntax. Linguistics 56: 435-476.

Stevens, Stanley Smith. 1956. The direct estimation of sensory magnitudes: Loudness. The American journal of psychology 6: 1-25.

Stowe, Laurie A. 1986. Parsing WH-constructions: Evidence for on-line gap location. *Language and cognitive processes* 1: 227-245.

Sturt, Patrick. 2003. The time-course of the application of binding constraints in reference resolution. Journal of Memory and Language 48: 542–562.

Szabolcsci, Anna, and Frans Zwarts. 1993. Weak islands and an algebraic semantics for scope taking. Natural Language Semantics 1: 235-284.

Tucker, Matthew, Ali Idrissi, Jon Sprouse, & Diogo Almeida. 2019. Resumption ameliorates different islands differentially: Acceptability data from Modern Standard Arabic. *Perspectives on Arabic Linguistics 30,*. Edited by Matthew Tucker.

Van Gompel, Roger P. G., & Liversedge, Simon P. 2003. The influence of morphological information on cataphoric pronoun assignment. Journal of Experimental Psychology: Learning, Memory, and Cognition 29: 128–139.

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural Network Acceptability Judgments. arXiv preprint arXiv:1805.12471.

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

Weskott, Thomas, and Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language*: 249-273.

Wike, Edward L., and James D. Church. 1976. Comments on Clark's "The Language-as-Fixed-Effect Fallacy". Journal of Memory and Language 15: 249-255.

Yoshida, Masaya, Nina Kazanina, Leticia Pablos & Patrick Sturt. 2014. On the origin of islands. Language, Cognition and Neuroscience 29: 761-770.

Annotated bibliography for acceptability judgment methods

For readers looking to learn how to design and analyze acceptability judgment experiments, I have created a course that covers design and analysis from start to finish. It includes (i) course slides covering experimental design and statistical analysis, (ii) R scripts for data wrangling, statistical analysis, and generating publication-quality figures, and (iii) the materials and data from a real experiment on island effects that can be used to demonstrate the workflow from start

to finish. I will endeavor to keep the most up-to-date version linked on my professional website at all times. I have placed the current link below.

> Sprouse, Jon. 2019. Online course materials for methods in experimental syntax. https://sprouse.uconn.edu/courses/experimental-syntax/

Every experimentalist must find a set of tools for processing data and generating publication-quality figures. There are several computing language that can serve this purpose, such as Python, Matlab, and R. I personally prefer R because it is free, open-source, and specifically designed for data analysis and visualization. The scripts included in the course above are designed to provide a basic introduction to R that is focused on the tasks that arise in acceptability judgment experiments. For readers who wish to go beyond these scripts, there are any number of free resources online for learning R. A good place to start is the free online book written by the creators of RStudio and the 'tidyverse' package:

> Grolemund, Garrett, and Hadley Wickham. 2019. R for Data Science. https://r4ds.had.co.nz/

Statistics is a living, and ever-evolving, field of study in its own right. As such, statistical analysis is probably one of the more difficult aspects of experimental methods to learn. The course above presents an introduction to statistics that is specifically targeted to syntacticians who want to use rating tasks for judgment experiments. For readers interested in going beyond that introduction, this textbook by Fields, Miles, and Fields is a good first introduction to a wide range of analyses in frequentist statistics. It also uses R for analysis and visualization.

> Field, Andy, Jeremy Miles, and Zoë Field. 2012. *Discovering statistics using R*. Sage publications.

For readers who want to truly understand the theory underlying frequentist statistics, this textbook by Maxwell and Delaney is a terrific next step. It is an advanced text, so I would not necessarily recommend it as a first textbook on statistics. That said, I do recommend that any readers who plan to make frequentist statistics a major component of their analysis pipeline think about adding it to their library.

> Maxwell, Scott E., and Harold D. Delaney. 2003. *Designing experiments and analyzing data: A model comparison perspective*. Routledge.

Bayesian statistics and frequentist statistics differ in their philosophical approach, and therefore provide different types of information. For readers interested in exploring Bayesian statistics in their research, Kruschke's textbook is probably the best place to start. It provides a comprehensive introduction to Bayesian statistics, as well as code to implement Bayesian analyses for all of the most common experimental designs.

> Kruschke, John. 2014. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press.

Bayes factors provide valuable information in their own right (the odds ratio of the probability of the evidence under the experimental and null hypotheses), and also provide a lower cost

introduction to Bayesian analysis than full-fledged posterior models. For readers interested in Bayes factors, the BayesFactor package in R is a comprehensive solution. The manual for the package includes examples for the most common experimental designs, and links to articles in the primary literature.

Morey, Richard D. 2019. Using the 'BayesFactor' package.
https://richarddmorey.github.io/BayesFactor/