The role of experimental syntax in an integrated cognitive science of language.

Jon Sprouse, Department of Cognitive Sciences, University of California, Irvine
Diogo Almeida, Department of Linguistics and Languages, Michigan State University

1 Introduction

Acceptability judgments form the primary empirical foundation for generative syntactic theories (Chomsky 1965, Schütze 1996). As such, the methodology of acceptability judgment collection has been a topic of research since the earliest days of generative syntax (e.g., Hill 1961, Spencer 1973). However, the past fifteen years have seen a dramatic increase in the number of articles devoted to the topic. It seems clear that the recent increase in interest in methodological issues is related to advances in technology that have made it easier than ever to construct, deploy, and analyze formal acceptability judgment experiments, which following Cowart (1997) have come to be called experimental syntax (a practice that we will follow in this chapter). The question at the center of this literature is deceptively simple: How can formal acceptability judgment experiments help achieve the goals of generative syntacticians? As we will see in this chapter, answering this question is surprisingly complex. A comprehensive answer to this question requires (at least) three components: (1) an explicit formulation of the goals of generative syntax, (2) an enumeration of the potential obstacles to those goals, and (3) an empirically-driven evaluation of the ability of formal experiments to eliminate those obstacles. In this chapter we will present a comprehensive review of the recent acceptability judgment literature with respect to these three components in an attempt to provide (our version of) an answer to the question of how formal judgment experiments can help generative syntactic theory.

2 The goals and obstacles of generative syntactic theory

Our starting assumption is that all cognitively-oriented language researchers share the goal of constructing a theory of language that integrates at all three of Marr’s famous levels of analysis: the computational level, the algorithmic level, and the implementational level (Marr 1982, see also discussion in Phillips 1996, Phillips and Lewis 2010, Kluender 1991, Frazier 1978, Embick and Poeppel 2005, and many others). Marr (1982) used cash registers as an illustrative example to define these three levels for information processing devices (such as the human brain). The computational level of the theory is a description of the properties of the problem that must be solved by the device, as well as the operations that the device must perform, that abstracts away from the exigencies of actually solving the problem in practice. For a cash register, the computational level description is the theory of addition, with properties such as commutativity and associativity, abstracting away from the precise algorithms that are necessary to carry-out addition. For the sentence level phenomena of language, syntactic theories are computational level descriptions, as they describe the properties of the final syntactic structures that must be built, as well as the properties of the structure building operations that are required to build them, but abstract away from the requirements of real-time sentence processing. The algorithmic level of the theory is a description of the actual operations that must be deployed to solve the problem (i.e., an algorithm). For a cash register this could be the base-10 addition algorithm that we learned in school: start from the right, and “carry over the ones.” For language, parsing theories
are algorithmic level theories, as they describe the specific parsing operations that must be deployed during real-time sentence processing, including the strategies that dictate the deployment of those operations, and the ways in which parsing resources constrain the operation of the parser. Finally, the implementational level of the theory is a description of how the processes стратегий ресурсы are implemented in the hardware of the device. For a cash register, there are several hardware options that can influence this level (e.g., spinning drums versus electronic processors). However, for (a cognitive approach to) language, there is only one set of hardware, the human brain. Neurolinguistic theories, which seek to identify the cortical networks involved in various linguistic computations, are a first step toward implementational level descriptions (Embick and Poeppel 2005, Sprouse and Lau 2012).

There are at least two major obstacles to the construction of an integrated theory of language. The first is the black box problem: there is no method to directly measure cognitive mechanisms. What this means in practice is that researchers must (i) identify observable data types (behavior, electrophysiological responses, hemodynamic responses), and (ii) identify linking hypotheses that license (empirically valid) inferences from the observable data to the unobservable cognitive mechanisms. The black box problem affects all three levels of the theory; however, in this chapter we will focus on the data and linking hypotheses underlying syntactic theory (see Sprouse and Lau 2012 for a description of the data types and linking hypotheses at the algorithmic and implementational levels). Crucially, the black box problem presents a framework for investigating the empirical contribution of experimental syntax by focusing the discussion on the following questions: To what extent is the data underlying current incarnations of syntactic theory sound? And, What types of inferences are licensed by the linking hypothesis between acceptability judgment data and syntactic theory?

Whereas the first major obstacle to the construction of an integrated theory of language, the black box problem, presented a framework for investigating the empirical contribution of experimental syntax, the second major obstacle presents a framework for understanding the historical and sociological context of recent investigations of experimental syntax. Even a cursory glance at the experimental syntax literature suggests that significantly more attention has been devoted to the question of the soundness of the data underlying syntactic theory than to the question of what inferences are licensed by the linking hypothesis between data and theory. As we will see in this chapter, we believe that this has been a distraction for the field, as there appears to be no evidence that the existing data is faulty, and growing evidence that traditional informal collection methods are appropriate for the majority of phenomena of interest to syntacticians. We believe that this distraction can be (at least partially) traced to the second major obstacle to the construction of an integrated theory of language: the difficulty in establishing linking hypotheses between the levels (computational, algorithmic, implementational) of the theory (see also Phillips 1996, Townsend and Bever 2001, Ferreira 2005).

Establishing a level-level linking hypothesis, for example between syntactic theories (computational) and parsing theories (algorithmic level), requires the resolution of at least two complex theoretical issues. The first is to determine exactly how much of the sentence processing system should be captured by the syntactic theory; in other words, a line must be drawn that separates the aspects of the system that will be abstracted away from, and the aspects of the system that will be part of the syntactic theory. The second is to determine exactly what the linking hypothesis will be between the mechanisms in the syntactic theory and the mechanisms in the parsing theory. One early attempt at an integration of syntactic and parsing
Theories was the Derivational Theory of Complexity (DTC) (Miller 1962, McMahon 1963, Miller and McKean 1964, Gough 1965, 1966; for reviews see Fodor, Bever, and Garrett 1974, Berwick and Weinberg 1983, Pritchett and Whitman 1993, Phillips 1996, Townsend and Bever 2001). The DTC assumed an early version of transformational syntactic theory that contained structure building operations (e.g., transformations), but abstracted away from other aspects of sentence processing such as meaning, parsing strategies, probabilistic information, etc. The DTC also assumed an isomorphic linking hypothesis between structure building operations in the syntactic theory and parsing operations in the parsing theory. Under this view, for every transformation that was necessary for a given sentence in the syntactic theory, there was a complementary process in the parsing theory to ‘un-do’ the transformation during sentence comprehension. In this way, the DTC predicted that behavioral responses that tracked parsing difficulty (such as reaction times) would be directly affected by the number of transformations that were necessary to derive a given sentence in the syntactic theory, as each transformation would trigger complementary processes during sentence comprehension. As is well known, this prediction did not hold for many types of complex sentences.

The failure of the DTC as a linking hypothesis between syntactic and parsing theories continues to shape the interaction of syntacticians and psycholinguists, as there is some truth to the observation that each side of the computational/algorithmic divined a different lesson from the failure. Though it is clear that the failure of the DTC was likely due to problems with all three components (the syntactic theory, the parsing theory, and the isomorphic linking hypothesis between the two; see Phillips 1996, Townsend and Bever 2001, Phillips and Lewis 2010), syntacticians tend to be more suspicious of the veracity of parsing theories, and psycholinguists tend to be more suspicious of the veracity of syntactic theories. This latter suspicion, coupled with a long tradition of formal experimentation in psycholinguistics, may be the cause of the increased attention given to the soundness of acceptability judgment data, as unsound data would obviously lead to unsound theories (Edelman and Christiansen 2003, Ferreira 2005, Gibson and Fedorenko 2010a,b). However, as will become clear in the next section, the soundness of acceptability judgment data does not appear to be a true impediment to an integrated theory (see also Phillips and Lasnik 2003, Phillips 2009, and Culicover and Jackendoff 2010); instead, the real impediment seems to be the complexity of the problem, as the space of possible syntactic theories, the space of possible parsing theories, and the space of possible linking hypotheses that can account for the data that we do have are still all relatively large.

3 To what extent are the acceptability judgments underlying syntactic theory sound?

Perhaps the most obvious target of criticism for researchers who are skeptical of syntactic theory is whether the judgments reported in any given paper can be trusted to be a true reflection of the acceptability of the sentences in question. We will call this the veracity of judgment data. Establishing the veracity of judgment data is no easy task: the fundamental problem is that, unlike the properties of physical objects, there is no device that can objectively measure the properties of cognitive objects. Instead, cognitive scientists must rely on behavioral experiments to indirectly establish the quantity or quality of the cognitive objects in question. In the case of acceptability, the behavioral experiments in question actually ask the participants to report their judgment of acceptability; however, it should be clear that this report of acceptability is not necessarily the ‘true’ acceptability response generated by the cognitive system of language. The process of establishing the veracity of judgment data is actually the process of establishing
confidence that the reported values of acceptability accurately reflect the ‘true’ acceptability response (see also Featherston 2007, Myers 2009a). The question then is how can experimental syntax help establish confidence in the acceptability judgments reported in the syntactic literature.

3.1 Direct comparisons of informally collected and formally collected judgments

Currently, there appears to be two approaches to using experimental syntax to establish confidence in the data underlying syntactic theory. The first approach simply assumes that formal acceptability judgment experiments provide superior data to the informal acceptability judgment experiments that have traditionally been used in the field. Under this assumption, all one has to do to establish confidence in the data underlying syntactic theory is run (or re-run) formal experiments to double check the data. Any discrepancy between the informal results and the formal results is evidence for the unreliability of the informal data (as the formal data is assumed to be more accurate). This is the tack has been taken recently by Gibson and Fedorenko 2010b, who report comparisons of informally and formally collected judgments for 7 sentence types: a pairwise comparison of two sentence types from Gibson 1991, a pairwise comparison of two sentence types from Kayne 1983 (first tested in Clifton et al. 2006), and a one-way comparison of three sentence types from Chomsky (1986). The informally reported judgments for these three comparisons suggest that there are true differences between the sentence types; however, Gibson and Fedorenko (2010b) find no significant differences for the first two comparisons, and a significant difference contrary the informally reported results for the third comparison. This, they argue, proves that it is at least logically possible that the informal methods that have characterized data collection in syntactic theory have led to unsound theorizing.

Sprouse and Almeida (submitted) adopted this approach in an effort to determine how different the data underlying syntactic theory would be if formal experiments were used to establish a representative set of data points that form the foundation of (generative) syntactic theory. They tested 469 data points from an introductory syntax textbook (Adger 2003) in formal experiments using 440 naïve participants, the magnitude estimation (Stevens 1957, Bard et al. 1996) and yes-no tasks, and three different types of statistical analyses (traditional frequentist tests, linear mixed effects models (Baayen et al. 2008), and Bayes factor analyses (Rouder et al. 2009)). The results of that study suggest that the maximum replication failure rate between the informal judgments for those 469 data points and the formal judgments for those 469 data points is 2%. Sprouse and Almeida (submitted) conclude that, though it is logically possible that formally collected judgment data would be substantially different from informally collected judgment data, at least for the 469 data points in Adger’s (2003) introductory syntax textbook, there would be very little difference in the shape or empirical coverage of the theory (see also Featherston 2009 and Phillips 2009 for similar conclusions).

Though the direct comparison of informally collected and formally collected judgments provides an efficient assessment of how likely the data underlying syntactic theory would be to change under the widespread adoption of experimental syntax techniques, it should be clear at this point that the assumption that formally collected judgments are closer to the ‘truth’ is just that, an assumption. A difference between the two methodologies cannot a priori establish which (if any) of the results is more accurate. Perhaps one of most interesting aspects of experimental syntax is that it provides a set of tools, and a body of knowledge, that can be used to tease apart
the potential differences between informally and formally collected judgments, and ultimately determine to what extent the methodologies result in meaningful empirical differences.

3.2 A comparison of informal and formal collection techniques

We follow Marantz (2005) in assuming that the standard approach to informal judgment experiments consists of the following steps: First, the syntactician constructs a set of conditions to minimally contrast the relevant structural property. These conditions are carefully constructed to rule out known syntactic nuisance variables, the set of which has been incrementally accumulated by prior linguistic research. Next, the syntactician constructs a set of sentences for each condition in an attempt to rule out any lexically driven extraneous factors (such as sentence plausibility or glaring word frequency imbalances) that may impact acceptability. Finally, the syntactician asks (verbally or by e-mail) 5-10 linguists or graduate students to rate the relative acceptability of the conditions. The task usually consists of a direct comparison between a target and control condition (i.e., a forced choice task), or an absolute acceptable/unacceptable (yes-no) categorization. The results are then summarized in a journal article using a diacritic placed to the left of an archetypical example of each sentence type: * indicates that most subjects found most of the sentences of the condition very unacceptable, ? indicates that most subjects found most of the sentences of the condition to lie in the middle of the spectrum, and no diacritic indicates that most subjects found the sentences of the condition acceptable\(^1\). This methodology can be compared to formal judgment experiments, which generally consist of the following steps: First, the syntactician constructs a set of conditions to minimally contrast the relevant structural property. Next, the syntactician constructs a set of sentences for each condition (usually 8 or more) in an attempt to rule out any lexically driven extraneous factors (such as sentence plausibility or word frequency) that may impact acceptability. Then the syntactician recruits a sample of 20 or more naïve (non-linguist) participants to perform an acceptability judgment task. The task is typically a numerical rating task that requires unrelated distracter items in a full survey format, such as magnitude estimation or Likert scales. Finally the results are summarized using standard descriptive statistics (mean, standard deviation) and analyzed using standard inferential statistical tests (t-test, ANOVA, linear mixed effects modeling), as opposed to informal experiments, which are typically analyzed with a non-numerical descriptive summary only (diacritics).

3.3 Naïve versus expert participants

Perhaps one of the most contentious aspects of informal judgment experiments is the use of professional linguists as participants. Several critics of informal experiments have suggested that this introduces the logical possibility of cognitive bias on the part of the participants: as professional linguists, the participants will likely be aware of the theoretical consequences of their judgments, and this awareness may impact the judgments that they ultimately report.

\(^1\) Unfortunately, the informal nature of this diacritic notation often blurs the important distinction between an empirical observation (e.g., * = judged unacceptable or less acceptable than the relevant control condition), and a theoretical claim (e.g., * = ungrammatical), and although the meaning intended by the authors is often transparent in context, that is not always so (Devitt 2010, Gross and Culbertson 2011).
Supporters of informal experiments counter this possibility with two logical arguments. First, acceptability judgment experiments are easily replicable, as they require no special equipment. This means that any given data point can be replicated on the spot: audiences at conferences, reviewers of articles, and even the readership of journals can quickly and easily check the reported judgments for accuracy, and thus ferret out any influence of cognitive bias. Second, the theoretical awareness of professional linguists may provide a sort of expert knowledge that increases the reliability, and possibly the sensitivity, of linguists’ judgments over non-linguists’ judgments (Newmeyer 1983, 2007, as well as Fanselow 2007, Grewendorf 2007, and Haider 2007 for possible examples in German, and Devitt 2006, 2010, Culbertson and Gross 2009, Gross and Culbertson 2011 for a discussion of what could be meant by ‘expert knowledge’). The empirical question then is whether there is any evidence of cognitive bias in the judgments of professional linguists, and whether there is any evidence that professional linguists provide more sensitive judgments than naïve participants.

To our knowledge, the only existing study of cognitive bias was conducted by Dabrowska (2010). She compared generative linguists, functional linguists and naïve participants in a rating study that included Complex NP islands (What did John make the claim that Mary bought?). One plausible prediction of the cognitive bias hypothesis is that generative linguists would rate the examples of island violations lower than the functional linguists because island constraints are a core part of the generative theory of syntax, but have been argued by several researchers to be an epiphenomenon of language use (e.g., Kuno 1973, Deane 1991, Kluender & Kutas 1993, Goldberg 2007). In other words, generative linguists have a motivation to confirm the reliability of Complex NP islands, whereas functional linguists have a motivation to disconfirm the reliability of Complex NP islands. Dabrowska (2010) actually found that generative linguists’ ratings were higher than the functional linguists’ and naïve participants ratings, which, contrary to the cognitive bias hypothesis, suggests that the generative linguists were actually biased against their own theoretical interests.

In a similar vein, there have not been many studies of the relative sensitivity of using linguists versus non-linguists as experimental subjects. Given the fact that some phenomena have been shown to produce disparate results in informal and formal experiments, it would be interesting to see whether this could be due to a difference in relative power to detect a predicted difference between conditions. For example, Gibson (1991) reports the following contrast based on informal experiments involving professional linguists:

(1)  a. *The man that the woman that the dog bit likes eats fish.  
      b. ?I saw the man that the woman that the dog bit likes.

---

2 Dabrowska (2010) interpreted this to be the result of different frequencies of exposure to specific kinds of unacceptable sentences. Since the training of generative linguists include reading scores of textbooks’ and articles’ examples of specific ungrammatical sentences, this could lead to a higher familiarity with them, which might lower generative linguists’ sensitivity to the unacceptability of these sentences. It is important to note, however, that this kind of explanation systematically predicts the opposite of the cognitive bias hypothesis for the phenomena studied by syntacticians: judgments of generative linguists are going to be systematically less sensitive to the predicted contrasts than the judgments of naïve participants.
Gibson and Fedorenko (2010b) report that subsequent formal experiments using numerical scales failed to corroborate the reported difference. One strategy for investigating the sensitivity of linguists’ judgments would be to attempt to increase the statistical power of the formal experiment, and see if the contrast in (1) appears. We asked 98 naïve participants to choose which of the two sentences in (1) sounded more natural in a binary forced-choice task, counterbalancing for the order of presentation. 87 out of the 98 participants chose 1b ($p = 0.0000000000000017$ by sign test), which suggests that the informal experiments conducted by the professional linguists may have had a sensitivity that is greater than the sensitivity in the formal experiments mentioned in Gibson and Fedorenko (2010b). Culbertson and Gross (2009) take a slightly different tack in investigating this question. They compared the numerical judgments of linguists, non-linguists with some experience participating in behavioral experiments, and non-linguists with no experience participating in behavioral experiments. They found a difference in the numerical ratings between participants with and without experience participating in behavioral experiments, but not between linguists and non-linguists. This result suggests that it is not the theoretical knowledge of linguists that provides the sensitivity per se, but rather previous experience with the unnatural demands of behavioral experiments.

3.4 Small versus large samples

It is well known that informal judgment experiments tend to be conducted on relatively few participants (almost always fewer than 10), whereas formal judgment experiments tend to use samples of 20 or more. Whether this difference in sample size is relevant for the reliability of the results of the experiments is an empirical question that can only be answered relative to the sentence types under investigation. To put this in concrete terms, we analyzed the relationship between sample size and the probability of detecting a significant difference for two phenomena that are central to syntactic theory: Whether-island effects and Complex NP island effects.

(2) Whether island
   (a) What do you think [CP that John bought __]? 
   (b) *What do you wonder [CP whether John bought __]?

(3) Complex NP island
   (a) What did you claim [CP that John bought __]? 
   (b) *What did you make [NP the claim [CP that John bought __]]?

We chose these two island effects because they are undeniably central to linguistic theory, as they have generated dozens of theoretical articles over the years, and are featured in nearly every syntactic textbook. They are also typical of complex syntactic violations, in that naïve participants cannot easily identify the “error” (Crain and Fodor 1987). Crucially, several researchers have also claimed that there is significant variation in the acceptability of island effects, suggesting that the prevailing syntactic analyses are incorrect (e.g., Kuno 1973, Grimshaw 1986, Deane 1991, Hofmeister and Sag 2010, but cf. Sprouse, Wagers, & Phillips submitted). Moreover, the versions of the island effects chosen here are generally considered to be the most variable, and therefore the most likely to demonstrate significant variability (cf. the less variable versions: WH-islands and Relative Clause islands).
We ran resampling simulations on the data from Sprouse et al. (submitted), which is a data set of responses from 173 naïve participants who rated 4 tokens of each condition using the magnitude estimation task, in order to assess the relationship between sample size and probability of detecting a significant result. As an example of how a resampling simulation works, suppose we wanted to assess the probability of a significant result for an experiment with a sample size of 5. The simulation algorithm would be:

1. Draw a random sample of 5 participants (allowing participants to be potentially drawn more than once -- this is called “sampling with replacement”)
2. Run a set of statistical tests on the sample. For the results reported here, we used simple paired $t$-tests.
3. Repeat this procedure 1000 times, and compute the proportion of significant results at $p < .05$.

This gives us an estimate of the probability of detecting a significant effect with 5 participants.

We then repeat this procedure for every other possible sample size from 6 to 173 to derive estimates for every sample size. Figure 1 reports the results of these simulations for Whether islands and Complex NP islands based on the Sprouse et al. (submitted) data.

Figure 1: Resampling simulation results for island effects

The $x$-axis is sample size, and the left-side $y$-axis is proportion of results at $p < .05$. The solid black line reports proportion significant in the expected direction. The dashed black line reports the proportion significant in the opposite direction. The solid grey line reports the difference between condition means in $z$-units using the right-side $y$-axis. The inset provides a zoomed in view of the proportion significant in the correct direction for sample sizes 5-25.

As Figure 1 makes clear, with a sample size of 5, the detection rate is already above 80%, which is generally considered a well-powered experiment (Cohen 1965). Reaching the extraordinary level of 100% detection requires only 9 participants for Whether islands and 12 participants for NP islands.
On the one hand, these results suggest that it is possible that the sample sizes used in informal experiments are appropriate for the phenomena that are central to syntactic theory (even “variable” phenomena such as these two island effects). However, this is not to say that all acceptability judgment effects are robust at small sample sizes. For example, the Sprouse et al. (submitted) data set also includes phenomena that are generally studied in real-time sentence processing experiments in order to investigate the behavior of the human sentence parser. Two such phenomena are the center embedding effect from Frazier (1985), attributed to Janet Fodor, which holds that doubly center-embedded sentences can be made more acceptable by deleting the second VP as in (4a-b), and the agreement attraction effect from Bock and Miller (1991), which holds that ungrammatical plural number marking on verbs can be made more acceptable by including a (non-subject) plural NP between the subject and verb as in (5a-b):

4. Center Embedding effect (Frazier 1985, Gibson and Thomas 1999)
   a. *The ancient manuscript that the graduate student who the new card catalog had confused a great deal was studying in the library was missing a page.
   b. ?The ancient manuscript that the graduate student who the new card catalog had confused a great deal was missing a page.

   a. *The slogan on the poster unsurprisingly were designed to get attention
   b. ?The slogan on the posters unsurprisingly were designed to get attention

The results of the resampling simulations for these two phenomena are presented in Figure 2.

Figure 2: Resampling simulations for two phenomena from the sentence processing literature.

The x-axis is sample size, and the left-side y-axis is proportion of results at p < .05. The solid black line reports proportion significant in the expected direction. The dashed black line reports the proportion significant in the opposite direction. The solid grey line reports the difference between condition means in z-units using the right-side y-axis.
There is a stark contrast between the island phenomena in Figure 1 and the “processing” phenomena in Figure 2. The effect size for the island phenomena (solid grey line in the figures) is between .6 and .8 z-units, and 80% detectability is reached with a sample size of only 5 participants. The effect size for the “processing” phenomena is between .1 and .2 z-units, and 80% detectability is only reached with sample sizes above 30 for agreement attraction, and above 78 for the center embedding effect. The question then is to what extent the phenomena studied in the syntactic literature are more like the island effects, and thus require relatively small samples to ensure detectability, and to what extent the phenomena are more like the “processing” effects, and thus require large samples to ensure detectability. Though we tried to choose test cases that would be considered typical of syntactic phenomena (though still relatively variable), it is clear that more research is necessary to expand these findings to the thousands of sentence types that form the basis of syntactic theory. Nonetheless, these findings may suggest a potential explanation for the disciplinary divide between the majority of critics of informal experiments (mostly researchers working outside generative linguistics), and that the majority of supporters of informal experiments (mostly generative syntacticians): it may be the case that the phenomena of interest to the two groups have substantially different behavior with small samples, such that the critics are mistaking a property of the phenomena for a property of the methodology itself.

3.5 The different acceptability judgment tasks

There are at least four distinct acceptability judgment tasks in the experimental syntax literature: the forced-choice task (FC), the yes-no task (YN), the Likert scale task (LS), and the magnitude estimation task (ME). A cursory review of the experimental syntax literature reveals an interesting correlation: FC and YN tasks tend to be employed in informal experiments, whereas LS and ME tasks tend to be employed in formal experiments. This correlation has led to (an often unstated) perception that LS and ME are superior tasks, which is in many ways parallel to the assumption that formal experiments are inherently superior to informal experiments. However, as we will see in this section, the costs and benefits of each task can only be evaluated with respect to the research questions that the task is intended to answer. The YN and FC tasks were primarily designed to answer qualitative questions, whereas the LS and ME tasks were primarily designed to answer quantitative questions. The choice of task depends on which type of question the researcher is interested in answering, as each task comes with its own unique set of benefits and limitations.

In a forced-choice task, participants are presented with two (or more) sentences, and instructed to choose the sentence that is most (or least) acceptable. In this way, FC is explicitly designed to qualitatively compare two (or more) conditions, and directly answer the qualitative question Is there a difference between these conditions? There are two explicit benefits to FC tasks. First, FC tasks are relatively easy to deploy, since each trial in an FC task is an isolated experiment unto itself. In other words, participants do not need to see any sentences other than the two (or more) being directly compared in order to complete the trial accurately. This means that FC tasks do not require filler sentences (e.g., to fill out the response scale). FC tasks also do not require multiple lexicalizations of each condition, as minimal pairs can be presented simultaneously without risk that participants will overlook the structural difference due to the identical lexicalization. In fact, the only survey procedure that is critical to the FC task is to counterbalance the order of presentation within each trial in order to avoid response biases (e.g.,
some participants may choose the first sentence simply because it is first). The ease of
deployment of FC tasks may explain its popularity within informal experiments.

The second benefit of FC tasks is sensitivity. FC tasks are the only task explicitly
designed for the comparison of two (or more) conditions; the other tasks compare conditions
indirectly through a response scale (either yes-no, or a numerical scale). This sensitivity
difference can be seen in a direct comparison of FC to ME. For example, we compared the ME
results for Whether islands, CNPC islands, Center Embedding, and Agreement Attraction (176
participants) from Sprouse et al. (submitted) to the results of an FC experiment (106 participants)
conducted using Amazon Mechanical Turk (Sprouse 2011a). We ran resampling simulations to
answer the question What is the likelihood of single participant reporting an effect in the
predicted direction using this task? Table 1 reports the results of the resampling simulations.

Table 1: Resampling simulations comparing the likelihood of detecting a result in the single trial
between ME and FC.

<table>
<thead>
<tr>
<th></th>
<th>Magnitude Estimation</th>
<th>Forced Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size of differences (in raw scores)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Whether Island</td>
<td>77%</td>
<td>83%</td>
</tr>
<tr>
<td>Complex NP Island</td>
<td>72%</td>
<td>75%</td>
</tr>
<tr>
<td>Center Embedding</td>
<td>39%</td>
<td>47%</td>
</tr>
<tr>
<td>Agreement Attraction</td>
<td>50%</td>
<td>55%</td>
</tr>
</tbody>
</table>

As Table 1 demonstrates, FC outperforms ME at detecting a qualitative difference between
conditions even when the margin of difference is reduced to the absolute possible minimum (=1;
i.e., a rating of 166 is considered higher than a rating of 165). The primary limitation of FC is
that it can only indirectly provide information about the size of the difference between conditions
in the form of the proportion of responses (see also Myers 2009b). Therefore, if the nature of the
research question is to simply ascertain the existence of a predicted acceptability contrast, the FC
task seems to be the optimal choice. However, if the research question is quantitative in nature
(i.e., about the magnitude of an acceptability contrast), it may be better to use one of the
numerical tasks (LS or ME).

In the Yes-No task, participants are presented with one sentence at a time, and instructed
to judge the sentence as a member of one of two categories: acceptable/grammatical/yes or
unacceptable/ungrammatical/no. The YN task is similar to the FC task in that it is primarily a
qualitative task; however, the two also present substantial differences. The YN task is designed
to answer the question does this sentence belong to the yes-category or the no-category? In this
way the YN task probes the relationship between a single sentence and the two categories
presented to the participant (rather than the relationship between two sentences as in the FC
task). From the point of view of syntactic theory, this sounds like a direct test of grammaticality;
however, it should be noted that i) it is not clear whether all subjects use the same category
boundary between yes-no, and ii) it is not clear whether the yes-no boundary in any given subject
maps to the theoretically relevant grammatical-ungrammatical boundary. Therefore it is
important to interpret the category names in YN tasks cautiously, as it is not clear what they truly
represent to the participants. The primary advantage of the YN task is that it is quick to
deploy, as it does not require filler sentences or complex instructions (Myers 2009b). In fact,
several researchers have demonstrated that the YN task can be used to compare the relative difference between conditions (at least for the sentence types tested) by computing the proportion of yes-responses for each condition (Myers 2009b, Bader and Häussler 2010). The primary disadvantage of the YN task is that it is likely less sensitive than the FC task at detecting qualitative differences between two conditions (because the difference is always relative to the category boundary) and likely less sensitive than the LS and ME tasks at establishing numerical estimates of the quantitative difference between conditions (because the difference is indirectly computed through proportions).

In a Likert scale task, participants are given a numerical scale, usually consisting of an odd number of points such as 1-5 or 1-7 with the endpoints defined as acceptable or unacceptable, and asked to rate each sentence along the scale. One of the primary benefits of LS is that it is both numerical and intuitive. This means that LS can be used to answer questions about the size of a difference between conditions by leveraging inferential statistical tests such as ANOVA and linear mixed-effects modeling. Recent research has suggested that LS has at least as much statistical power as ME, and in fact may lead to lower variance in the responses given the limited number of response points in an LS task (Weskott and Fanselow 2011). The primary limitations of LS are all related to the use of the numerical scale. For example, the scale itself suggests that the intervals between points are uniform: the interval between 1 and 2 is one unit, the interval between 2 and 3 is one unit, etc. However, because participants can only use the limited number of response points (i.e., there is no 3.5 on the scale), it is impossible to ensure that the intervals are truly uniform. This problem is compounded when aggregating across participants in a sample. In practice, this risk can be minimized by including anchoring examples at the beginning of the experiment to establish the points along the scale, and by including filler sentences in the experiment to ensure that each point along the scale is used an equal number of times. Furthermore, participant’s responses can be z-score transformed prior to analysis to eliminate some additional forms of bias such as scale contraction (e.g., only using points 3-5 in a 1-7 scale) or scale skew (e.g., only using the high end of the scale).

In the magnitude estimation task, participants are given a reference sentence and told that the acceptability of the reference sentence is a specific numerical value (e.g., 100). They are then asked to rate additional sentences as a proportion of the value of the reference sentence. For example, a sentence that is twice as acceptable as the reference sentences would be rated 200. ME was developed by Stevens (1957) explicitly to overcome the problem of potentially non-uniform, and therefore non-meaningful, intervals in the LS task (in the domain of psychophysics). In the ME task, the reference sentence acts as a unit of measure for all of the other sentences in the experiment. In this way, the intervals between sentences can be expressed as proportions of the reference sentence (the unit of measure). This offers the possibility of substantially more accurate ratings (Bard et al. 1996, Cowart 1997, Keller 2000, Featherston 2005a,b) than the LS task. In addition, the response scale in ME is the entire positive number line, which means that participants can report a potentially infinite number of levels of acceptability (Bard et al. 1996, Keller 2000), as opposed to the finite number in the LS task (see also Featherston 2008 for a novel task using the positive number line). Finally, because the ME task is a measurement task involving the reference sentence and a single sentence to be measured, each trial in an ME experiment is encapsulated similar to the trials in an FC experiment. This means that in principle ME does not require full survey procedures such as fillers and multiple lexicalizations, as each measurement should be logically independent of the other sentences in the experiment.
Given the purported benefits of ME, it is perhaps unsurprising that over the past 15 years ME has assumed the role of a “gold standard” among the acceptability judgment tasks. However, a series of recent studies of the ME task have called into question many of those benefits. For example, one of the primary assumptions of the ME task is that participants truly use the reference sentence as a unit of measure. In order for this to be true, participants must be able to make a ratio comparison of two sentences (e.g., sentence B is 1.5 times the magnitude of sentence A). Adapting a series of techniques developed in the psychophysics literature (Narens 1996, Luce 2002), Sprouse (2011b) tested this assumption directly, and found that participants could not make ratio comparisons of the acceptability of two sentences. This failure of the primary assumption of the ME task to hold suggests that participants may be treating the ME task as a type of LS task, only with an open and infinite response scale. This result accords well with recent results from Weskott and Fanselow (2011) and Bader and Häussler (2010) that suggest that the results of ME experiments are no more informative than the results of LS (and even YN) experiments, at least for the phenomena that they tested. The burgeoning consensus among these researchers is that the true value of ME lies in (i) the increased number of levels of acceptability that participants can report, though this might come at the cost of higher variance (Weskott and Fanselow 2011), and (ii) the sociological impact of using a task that is considered more ‘sophisticated’ than LS.

3.6 The interpretation of variation across participants

Another issue raised by the experimental syntax literature is how to interpret the variability in acceptability judgments across participants. To make this discussion concrete, imagine that a researcher is interested in the difference between two sentence types, as is typical in acceptability judgment experiments. In general statistics terminology, this difference is the treatment effect. When investigating the treatment effect, the research can ask if the sample as a whole shows a treatment effect by comparing the mean response of the sample for each condition. If the two means are different enough, traditional statistical significance testing (SST) will report that there is a significant treatment effect for the sample. However, finding a statistically significant treatment effect for the sample does not mean that every participant demonstrated the treatment effect. In practice, given sufficient statistical power, very few participants need to show the treatment effect in order for the sample to show a significant treatment effect. A recurring question in the experimental syntax literature is what to make of this variability. If 100% of the participants show the treatment effect, it is pretty clear that the effect is a robust fact for all of the members of the sample. However, what if 75% show the effect, and 25% do not? What if only 25% show the effect, and 75% do not?

There seem to be three different approaches to this question in the literature:

1. Since measurement involves noise, only the central tendency of the sample matters, and it is expected that not every participant or every item in the sample show the treatment effect.

2. If a large enough proportion of participants do not show the predicted treatment effect, this might be evidence for a different dialect.
3. Given a strong theory-data linking hypothesis that ungrammatical sentences should overwhelmingly judged to be unacceptable, a large enough proportion of participants that fail to show the predicted treatment effect, or that judge supposedly ungrammatical sentences no worse than awkward will be taken as evidence that the theoretical prediction is disconfirmed.

The first approach assumes that the participants who do not show the treatment effect are simply being influenced by random noise. This is the default assumption in most domains of cognitive science, as it is assumed that all behavioral responses are the result of a combination of the experimentally manipulated behavior, and various sources of random noise (sometimes called unsystematic variation). Under this approach, it only matters whether the sample as a whole shows the treatment effect: if SST reveals a treatment effect in the sample, then there is a real treatment effect. This is by far the most common approach in the experimental syntax literature, as many of the best practices of experimental syntax, including the use of SST, have been directly adapted from the experimental fields of cognitive science such as psycholinguistics and psychophysics. A second approach is to investigate whether participants who do not show the treatment effect are actually from a different population than the participants who do show the effect. In most domains of cognitive science, the population of interest is all humans; in linguistics, the population of interest is all speakers of a given language. It is always a logical possibility that the participants who do not show an effect have a different grammar than the speakers who do show the effect (see also den Dikken et al. 2007). A third approach is to assume that only manipulations that yield a treatment effect in (almost) 100% of participants are real. While this is certainly a strong criterion to impose on experimental results, it is not without a certain logic. It is common in the syntactic literature to talk about possible sentences and impossible sentences. If one truly believes that a given sentence is impossible in a certain language, then one could also conclude that no amount of random noise should be enough to cause participants to rate that sentence as acceptable (see also Hoji 2010). This approach nonetheless makes several assumptions: (i) that acceptability judgments directly reflect the grammaticality of the sentences, without contamination from other cognitive systems, (ii) that fatigue and distraction do not affect judgments, (iii) that the crucial analysis is categorical (sentence A is acceptable or unacceptable), as opposed to relative differences (A is better or worse than B), and (iv) that the empirical domain of syntactic theory is only the difference between possible and impossible sentences.

Because of the domain-specific issues related to syntactic theory (e.g., language variation, possible/impossible sentences), it is critical to keep these three approaches to variation in mind when interpreting the results of formal acceptability judgment experiments. Failure to do so can lead to substantially different interpretations of the data. For example, Langendoen et al. (1973) investigated the claim made by Fillmore (1965) and others that the first object of a double-object construction cannot be questioned:

(6) *Who did you buy a hat? 
   (cf. What did you buy Mary?)

Langendoen et al. performed an answer completion task to test this claim formally. They asked 109 students to answer questions like (7) with a complete sentence:
Their hypothesis was that if questions like (6) are indeed unacceptable, then the answers to (7) should *consistently* place the answer NP at the end of the sentence *(I showed the woman my daughter)*. Langendoen et al. reported two findings: that one-fifth of the participants responded with the NP in the first object position *(I showed my daughter the woman)* and these participants were all from the metropolitan New York City area. Langendoen et al. (1973) considered following approach two, i.e., concluding that there are two dialects at work in the sample: speakers from NYC, who can question first object, and everyone else, who cannot. Their favored conclusion, however, was more nuanced. Noticing the theoretical difficulty in incorporating the necessary restrictions in the grammar of English to explicitly rule out questions from the first object of a double object construction (a point previously raised by Jackendoff & Culicover, 1971), Langendoen et al. (1973) proposed that these constructions are in fact licensed by the grammar of English. The difference between the population that finds them acceptable and the population that does not, they argue, is due to a different parsing strategy employed by the two groups. Taking a different perspective, Wasow and Arnold (2005) and Gibson and Fedorenko (2010a,b) have interpreted Langendoen et al (1973)’s result under approach three, and concluded that Fillmore’s (1965) original claim is incorrect: it is, in fact, possible to question first objects. Of course, it is also possible to assume approach one: the one-fifth of participants who created first object answers did so because of random noise in the experiment. This means that 87/109 participants responded in accordance with Fillmore’s (1965) claim. A one-tailed sign test yields *p* = .0000000018 – a highly significant result. What should be obvious here is that the problem is not with the data itself, since no experimental result disputed the fact that, by and large, speakers of English found questions constructed from the first object of a double object construction to be unacceptable. The problem is with the interpretation of what these results might mean for the theory of grammar: Langendoen et al. (1973)’s favored interpretation was motivated first and foremost by theory-internal considerations, while Wasow and Arnold (2005)’s and Gibson and Fedorenko (2010a,b)’s conclusions were dictated by their data-theory linking hypothesis.

A similar situation arises with Wasow and Arnold’s (2005) test of a claim from Chomsky (1955/1975) that the complexity of a noun phrase strongly determines the position of that noun phrase within a verb-particle construction. Chomsky’s claim is twofold. First, he claims that the most natural place for multi-word NPs is after the particle, therefore both (a) and (b) below should be more acceptable than both (c) and (d). Second, he claims that complex NPs (relative clauses) are less acceptable than simple NPs when they occur between the verb and particle, therefore (d) should be less acceptable than (c).

(8)  
a. The children took in all our instructions. [3.4 out of 4]  
b. The children took in everything we said. [3.3 out of 4]  
c. The children took all our instructions in. [2.8 out of 4]  
d. The children took everything we said in. [1.8 out of 4]

Wasow and Arnold (2005) ran a formal rating experiment, the results of which are in square brackets in (8). According to approach one, which assumes that only a difference in means is necessary to verify a claim, the formal results match Chomsky’s informal results perfectly: there is a significant interaction between particle position and NP type (*p* < .001). However, Wasow and Arnold (2005, p. 1491) interpret the results as problematic because “17% of the responses to
such sentences [d.] were scores of 3 or 4.” It seems that Wasow and Arnold (2005) were assuming approach three, which requires that sentences be judged unacceptable close to 100% of the time if we are to accept their status as unacceptable.

It is crucial for the language community to be explicit about their assumptions regarding the variability of acceptability judgments moving forward. As we have already seen, several high profile criticisms of informal experiments rest upon the assumption that there should be little or no variability among participants (Wasow and Arnold 2005, Gibson and Fedorenko 2010a,b), but it is important to notice that when this very strong assumption between the relationship between the theory and the data is relaxed (for instance, to allow for things like sampling error), the exact same set of results can be seen as providing strong evidence for the opposite conclusion (see also Labov (1996) for a similar discussion of the wanna contraction, and Raaijmakers (2003) for a similar discussion of the interpretation of variation across participants in the sentence processing literature).

3.7 The veracity of the effect versus the importance of the effect

One final issue that should be mentioned is the critical difference between the veracity of the effect, and the importance of the effect. As we have seen, experimental syntax provides several tools to establish the veracity of effects: large sample sizes and SST can help quantify the confidence that there is a true difference between conditions. However, there is no experimental methodology that can tell the researcher whether a given effect is important (i.e., whether the theory should attempt to explain the effect). It is somewhat common to read discussions of subtle (small) effects in the literature followed by the suggestion that formal experiments may better address the issue. In reality, an effect that is small in an informal experiment will also be small in a formal experiment. This can be seen in the grey lines in Figures 1 and 2. The grey lines represent the size of the difference between the mean ratings of the conditions (in z-units). Note that the lines are horizontal, indicating that the size of the effect does not change as more participants are added to the experiment: the large effects in Figure 1 are large with 5 participants, and remain large with 175; The small effects in Figure 2 are small with 5 participants, and remain small with 175. This demonstrates the fact that large, formal experiments can increase confidence in the existence of a small effect, but it can’t change the size of the effect (see also Phillips 2009). It is up to the researcher to decide if a small effect is relevant for a particularly theory and how to best account for it.

3.8 Conclusion

In this section we have seen that experimental syntax techniques provide a useful toolkit for exploring precisely which properties of formal and informal experiments increase our confidence in the veracity of the results (i.e., whether there is indeed a true differences between conditions). Though it is relatively common to assume that formal experiments provide ‘better’ results than informal results, the current state of the field suggests that many of the perceived benefits of formal experiments ultimately disappear under closer empirical scrutiny. This raises the very real possibility that the problem facing acceptability judgment data is a sociological one, not an empirical one: researchers who are accustomed to formal experiments are disinclined to have confidence in the results of informal experiments, regardless of whether the informal
experiments are empirically appropriate for the research questions that they are intended to address.

4 What types of inferences are licensed by the linking hypothesis between acceptability judgments and syntactic theory?

The majority of experimental syntax studies have focused on the veracity of the data underlying syntactic theory. As the previous section made clear, we believe that this has been a (necessary) distraction: there appears to be no evidence that the existing data is faulty, and growing evidence that the informal methods are appropriate for the majority of phenomena of interest to syntacticians. However, there is reason to believe that experimental syntax techniques also provide new tools to investigate the inferences licensed by the linking hypothesis between acceptability judgments and syntactic theory. In this section we will review two ways in which experimental syntax has added to our understanding of the nature of syntactic theory: (i) testing reductionist claims about the correct locus of acceptability judgment effects (so-called “processing” explanations), and (ii) examining the complex theoretical issues surrounding the interpretation of continuous acceptability judgments.

4.1 Reductionist approaches to acceptability judgment effects

The fundamental component of the linking hypothesis between judgment data and syntactic theories is the assumption that manipulations of the structural properties of a sentence will lead to modulations of acceptability. Regular readers of the syntactic literature are aware that it is relatively common for syntacticians to establish the structural nature of acceptability differences; by holding non-syntactic factors constant (semantics/plausibility, phonetics/phonology, morphology/lexical properties), syntacticians can be relatively certain that it is the structural manipulation that is driving the effect. However, because acceptability judgments are the result of successful sentence processing, and because the operation of the parser is (by definition) intricately tied to structural properties of sentences, there is always a possibility that acceptability effects may be driven by properties of the parsing system rather than grammaticality per se (cf. the conclusion reached by Langendoen et al. 1973 for the questions based on the double dative construction mentioned in the previous section). The question then is how experimental syntax techniques can help tease apart acceptability differences due to grammaticality effects, and acceptability differences due to properties of the parsing system.

The first step in teasing apart this ambiguity is to be clear about what is meant when one suggests that acceptability differences are driven by properties of the parsing system. These types of accounts are sometimes called “processing explanations” to contrast with “syntactic explanations,” but as several researchers have remarked, this label is less than ideal (Phillips 2011, Sprouse et al. submitted). The problem with this label is that, by definition, structural manipulations will result in different behavior by the parser. This is precisely what we want: the theory of syntax (a computational theory of structural properties of sentences) should be closely related (by a level-level linking hypothesis) to the theory of parsing (an algorithmic theory of syntactic structure building). Viewed from this perspective, every “syntactic explanation” is a “processing explanation”, as the syntactic theory is a form of abstraction or idealization of the structure-building component of the parser. Because of this tight relationship between the syntactic theory and the structure-building component of the parser, the syntactic properties of a
sentence will necessarily affect the behavior of the syntactic structure-building component of the parser. This means that in order for the so-called “processing explanations” to be distinct from “syntactic explanations”, the “processing explanation” must not be related to the syntactic structure-building component of the parser.

To clarify the content of these types of questions, Phillips (2011) and Sprouse et al. (submitted) suggest the term reductionist instead of “processing explanation”. They argue that the logic of these types of explanations is clearly reductionist, in that the acceptability effect is argued to be reducible to non-structure-building components of the parser, such as parsing strategies or parsing resource capacity. Under a reductionist approach, the relationship between the acceptability effect and the structural manipulation that is normally assumed to be driven by the syntactic system is actually epiphenomenal; the true causal nexus lies between the extra-syntactic factors, such as parsing strategies or parsing resources, and the acceptability effect. The second order correlation between acceptability and the structural manipulation arises because structural manipulations necessarily affect parsing strategies or parsing resource allocations. In this way, the complexity of the syntactic system is reduced in favor of extra-syntactic components. To the extent that these extra-syntactic components are independently necessary, reductionist explanations may be preferred to syntactic explanations according to theory building metrics such as Occam’s razor.

As a concrete example, take the two island effects that we have discussed previously: Whether islands and Complex NP islands (see also Alexopoulou and Keller 2007, Sprouse 2008, and Sprouse et al. 2011 for other examples of the parsing system affecting acceptability judgments). The standard analysis within the syntactic literature is that these island effects are rated unacceptable by native speakers because there is a syntactic constraint, such as the Subjacency Condition, that rules these structures out as ungrammatical. However, several researchers have proposed alternative explanations that do not involve syntactic constraints at all, but rather potentially independently motivated properties of the parsing system such as working memory (Kluender and Kutas 1993, Kluender 1998, 2004, Hofmeister and Sag 2010), attention (Deane 1991), and focus (Erteshik-Shir 1973, Goldberg 2007). For example, Kluender and Kutas (1993) argue that island violations such as (2b) and (3b) are in fact grammatical structures, but that the unacceptability reported by speakers is in fact the result of a combination of two relatively resource-intensive processes that are necessary to successfully parse the sentences. These two processes require more resources than are available to the parsing system, and therefore cause the parsing system to fail to successfully parse the sentences, resulting in the perception of unacceptability.

Kluender and Kutas (1993, see also Kluender 1998, 2004) are very explicit about the two processes that they believe are the cause of the unacceptability, and about the resources in question. They argue that the first process is the maintenance of a displaced wh-word in working memory during the processing of the sentence between the wh-word and the downstream gap site. The second process is the construction of the island structure itself, which as can be seen in (2b) and (3b) involves a CP clause that is in some ways more complex than CPs headed by that. Kluender and Kutas (1993) argue that each of these processes requires a certain amount of working memory resources to be deployed. Although each of these processes can be deployed in isolation, when deployed simultaneously, the combined resource requirements are greater than the pool of available resources. At this point it should be clear that this reductionist theory defines island effects as a psychological interaction of two (sets of) parsing processes that occurs because the processes rely upon a single pool of resources. Sprouse et al. (submitted) suggest that
this psychological interaction can be translated into a *statistical interaction* between two factors, each with two levels: **LENGTH** (short, long) and **STRUCTURE** (non-island, island) (see also Myers 2009b for a discussion of the use of factorial designs in syntax). The factor **LENGTH** manipulates the length of the wh-dependency at two levels: within a bi-clausal constituent question, a short dependency is created by extraction of the matrix subject; a long dependency is created by extraction of the object argument in the embedded clause. The factor **STRUCTURE** refers to the **STRUCTURE** of the embedded clause.

Table 2: Independent manipulation of dependency length and island structures

<table>
<thead>
<tr>
<th>LENGTH</th>
<th>STRUCTURE</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>short</td>
<td>non-island</td>
<td>Who __ thinks that John bought a car?</td>
</tr>
<tr>
<td>long</td>
<td>non-island</td>
<td>What do you think that John bought __?</td>
</tr>
<tr>
<td>short</td>
<td>island</td>
<td>Who __ wonders whether John bought a car?</td>
</tr>
<tr>
<td>long</td>
<td>island</td>
<td>What do you wonder whether John bought __?</td>
</tr>
</tbody>
</table>

Defining island effects in this way has several advantages. First, it allows us to isolate the effect of each of the individual factors on continuous acceptability ratings. For example, the effect of processing long-distance wh-dependencies can be seen by comparing the short, non-island condition to the long, non-island condition, and the effect of processing island structures can be seen by comparing the short, non-island condition to the short, island condition. Second, it allows us to quantify the statistical interaction of the two factors. If there were no statistical interaction between the two factors (i.e., if the two sets of processes impact acceptability ratings independently), we would expect a graph like that in Figure 3a. Figure 3a is an example of simple linear additivity between each factor in which the cost of each process leads to a decrement in acceptability ratings, and in which each cost sums linearly with respect to the short/non-island condition. This linear additivity in decrements leads to two parallel lines. However, if there were an interaction between the two factors, we would expect a graph like that in Figure 3b: super-additivity when the **long** and **island** levels of the two factors are combined, leading to non-parallel lines. (The hypothetical ratings in Figure 4 are displayed in terms of standardized z-scores, which can be derived from any approximately continuous rating measure, such as Likert scales or magnitude estimation.)
Figure 3: Example results for main effects and interaction

a. Main effects only

b. Interaction

Figure 3b is in fact the pattern that is consistently observed when the factors LENGTH and STRUCTURE are independently manipulated in acceptability experiments, although there is variation in the size of the effect of island structures alone (Kluender and Kutas 1993, Sprouse 2007a, Sprouse et al. submitted). In this way, island effects can be defined as a statistical interaction between two structural factors, exemplified by a super-additive decrease in acceptability in the long, island condition.

Sprouse et al. (submitted) used this interaction-based definition to test the role of working memory resources in the acceptability of island effects. They argued that the Kluender and Kutas (1993) analysis would predict an inverse relationship between working memory capacity in individuals and the strength of the super-additive interaction because the super-additive interaction (as opposed to linear additivity) arises due to insufficient working memory capacity. Therefore one would expect participants with higher working memory capacity to have smaller interactions, and participants with lower working memory capacity to have larger interactions. Sprouse et al. (submitted) tested over 300 participants with two different working memory tasks and two different acceptability judgment tasks (LS and ME), and found no relationship between working memory capacity and the size of the super-additive interaction. From these results, they conclude that it is unlikely that working memory capacity is driving the unacceptability of island effects. In this way, a combination of factorial designs, numerical acceptability judgment tasks, and parsing resource tests (working memory tests, etc.) can be used to investigate the probability of reductionist versus grammatical explanations for acceptability effects, and help establish the data-theory linking hypothesis between acceptability judgments and syntactic theory.

Though this is just one example of using experimental syntax to investigate a reductionist explanation for acceptability effects, it does suggest a general methodology for future studies. The first step is to identify a set of (non-structure-building) parsing-related factors that are hypothesized to drive the effect. The simplest possible reductionist explanation is one in which the acceptability effect of each of the factors will sum (linearly) to the full effect (e.g., the lack of interaction in Figure 3a). If the factors do not sum (linearly) to the full effect (e.g., the super-additive interaction in Figure 3b), then there must be an explanation for the super-additivity. In the case of island effects, the super-additive interaction was explained by the assumption that the two processes draw on the same limited pool of working memory resources. The explanation for
the super-additivity can then be tested by searching for correlations between the strength of the interaction and the relevant parsing properties (e.g., working memory capacity).

4.2 Gradient acceptability and the nature of syntactic theory

Although it is possible to categorize sentences as either acceptable or unacceptable in qualitative tasks such as the YN task, the results of numerical judgment tasks such as ME have suggested that acceptability is better described by a continuous scale, with sentence types taking values at any point along the scale. The fact that acceptability is a continuous measure has led several researchers to investigate to what extent the nature of the grammar itself may be continuous (Keller 2000, Sorace and Keller 2005, Featherston 2005b, Bresnan 2007). It is not uncommon to encounter those who believe continuous acceptability necessitates a continuous (or gradient) syntactic system. However, there is no necessary link between the nature of acceptability and the nature of the syntactic system. The question in fact hinges on (at least) three complex theoretical issues: (i) What is the relationship between acceptability judgment data and syntactic theories? (ii) What is the correct level of abstraction for a computational theory? (iii) What is the continuous syntactic property that could give rise to continuous acceptability?

Up to this point, the discussion of the relationship between acceptability judgment data and syntactic theories has been one of inference from data to theory: when certain extra-syntactic factors (such as parsing strategies, parsing resources, and issues related to acceptability judgment tasks themselves) are held constant, modulations in acceptability judgments are interpreted (via a linking hypothesis) as evidence about the properties of the syntactic system. However, it is also possible to reverse the direction of the relationship and investigate how well the syntactic theory predicts the acceptability judgment data. While this approach is more common in the computational modeling literature, where the term generative is used to refer to models that can be used to generate the observable data, it seems clear that the question of how best to account for the continuous nature of acceptability judgments is a predictive/generative question. This can be seen in the structure of a common argument that syntactic theories should be gradient: (i) acceptability is continuous, (ii) categorical grammars predict categorical acceptability, (iii) gradient grammars predict continuous acceptability, (iv) therefore the grammar must be gradient. Crucially, this style of argument assumes that the syntactic theory is the correct locus for the mechanisms that lead to continuous acceptability. It is clear that this is not a logical necessity: all of the extra-syntactic components of the language faculty that must be controlled in order to make careful inference from acceptability data to syntactic theory (such as parsing strategies, parsing resources, and the conscious mechanisms that underlie judgment tasks) may be contributing to the gradience in the acceptability judgment data, and therefore are potential sources of the mechanisms that generate continuous acceptability. In short, once one adopts a predictive/generative approach to modeling acceptability judgment data, the question is to what extent should the continuous mechanisms be part of the syntactic theory, and to what extent should the continuous mechanisms be part of the extra-syntactic components of the language faculty.

In order to determine to what extent the syntactic theory should predict continuous acceptability, we must be explicit about what a syntactic theory is a theory of, and which aspects of that theory can give rise to continuous acceptability. In section one, we presented a view of syntactic theory as a computational level description of a part of the human language faculty. In other words, syntactic theory is a formulation of the properties of the syntactic structure building
mechanisms of the human parser that abstracts away from parsing strategies, parsing resources, and other issues that are specific to the real-time implementation of parsing algorithms. From this perspective, there are only two options for including gradient mechanisms within the syntactic theory. The first option is to actually abstract away form the algorithmic level less, such that one or more of the gradient mechanisms of the algorithmic level actually exists in the computational level description. Bresnan (2007) has advocated a probabilistic approach to the syntax of the dative construction in English that may be an example of this sort of ‘weaker’ abstraction, as the syntactic theory appears to include information about the probability of the dative construction under various morphosyntactic, semantic, and information-structure environments – information that has previously been of primary interest to algorithmic level sentence processing theories (see also Bresnan and Hay 2008, Bresnan and Ford 2010).

The second option is to include an additional property in the syntactic theory that can capture gradience. Once again, this is a complex ontological issue, as any property that is included in the syntactic (computational) theory must be mapped to the actual language faculty as it is implemented in the human brain (i.e., there is a mentalistic commitment). This leads to a stark contrast between syntactic theories in which grammaticality is a purely theoretical construct, and syntactic theories in which grammaticality is a mentalistic construct. In the former, grammaticality is not a property that is available to the mental system, but rather a label that theoreticians can apply to sentences rather than using the non-technical terms ‘possible/impossible’ (e.g., Chomsky 1957). In the latter, grammaticality is a property that is available to the mental system in some form, such as when different syntactic constraints are assumed to lead to different levels of unacceptability (e.g., Huang 1982, Chomsky 1986), or when structures are assumed to be marked as ungrammatical (e.g., the star feature in Chomsky 1972). As Keller (2000) demonstrates for Optimality Theory, and Featherston (2005b) demonstrates for the Decathlon model, syntactic theories that assume that grammaticality is a mentalistic construct can be used to directly predict continuous acceptability without altering the level of abstraction in the computational/algorithmic divide; however, the cost is assuming that grammaticality is a mentalistic construct rather than simply a theoretical one.

On the one hand, it is clear that experimental syntax techniques, especially numerical tasks like ME and LS, yield a new form of continuous acceptability data that provide researchers with the opportunity to reverse the normal direction of data-theory inference, and construct predictive/generative syntactic theories. On the other hand, the discussion in this subsection suggests that the interesting questions raised by this approach are not data-driven questions. In other words, the data enable this line of questioning, but the data don’t determine the answer (see also Sprouse 2007b). The questions raised in this section (such as: What is the right level of abstraction for a computational theory? and Should syntactic theories include a gradient, mentalistic property called grammaticality?) are theoretical questions. And like all theoretical questions, they can only be answered through careful comparison of the empirical adequacy of competing theories.

5 Conclusion

Our goal in this chapter was to review the role of experimental syntax in the construction of an integrated cognitive science of language. While this role is undoubtedly still evolving, it seems clear that experimental syntax is well positioned to make substantial contributions to two questions that are central to the integration of syntactic theories with parsing theories: Is the data
underlying existing syntactic theories sound? and, What types of inference are licensed by the linking hypothesis between acceptability judgments and syntactic theories? Although both questions are scientifically relevant to the theory, the current state of evidence suggests that questions about the veracity of existing judgment data may have been a (historically driven) distraction: there appears to be no evidence that the existing data is faulty, and growing evidence that the informal methods are appropriate for the majority of phenomena of interest to syntacticians. This suggests that the contribution of experimental syntax in the coming years will be as a tool for investigating what the acceptability judgment data reveals about the nature of syntactic theory. We have seen two examples of this approach in this chapter: the question of reductionist approaches to complex syntactic phenomena (e.g., island effects), and the question of gradient approaches to syntactic theory. Undoubtedly there are more questions waiting to be discovered as the field progresses toward an integrated theory of language.
References


