1

Sentence Acceptability Experiments: What, How, and Why

Grant Goodall

Sentence acceptability experiments have become increasingly common since Cowart (1997) first presented a detailed method for carrying them out, but there is still relatively little clarity among syntacticians about what goes into a well-designed experiment, how to perform one and interpret the results, and why one might want to do this in the first place. This chapter addresses these concerns, by providing recommendations and perspective on how experimental approaches to acceptability can be understood and put to use. Section 1.1 discusses the notion of acceptability in general and the role that it has played in linguistic research. Section 1.2 enters into the details of experimental design, giving an overview of the best practices in acceptability experiments that have emerged from the last two decades of research. Section 1.3 explores the variety of factors, both grammatical and extra-grammatical, that acceptability experiments seem to be able to detect, while Section 1.4 addresses the question of why one might undertake the effort of conducting acceptability experiments.

1.1 Acceptability

Any description of a language inevitably includes a description of what is possible, e.g., a listing of the phonemes, the allowable syllable structures, the preferred word order, etc. These are essentially descriptions of what is "acceptable" in the language, and such descriptions form the core of grammatical research in all traditions. By characterizing what is acceptable, such descriptions also make implicit claims about what is not acceptable. More explicit claims of unacceptability were occasionally included

I am grateful to the members of the Experimental Syntax Lab at UC San Diego and to research assistants Chengrui Zheng and Noah Hermansen for their valuable comments, discussion, and assistance in the preparation of this chapter.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

in classical, Renaissance, and American structuralist grammars (see Householder 1973; Myers 2017), but with the advent of generative grammar in the mid-twentieth century, and its emphasis on explicitness in grammatical description, the distinction between acceptability and unacceptability became much more important. This distinction played a crucial role in Chomsky (1957), for instance, and was much more extensively discussed in Chomsky (1965).

The focus on acceptability vs. unacceptability also led to a more nuanced understanding of what these concepts mean. The most widespread view, proposed originally in Chomsky (1965), is that (un)acceptability may be influenced by a variety of factors, of which (un)grammaticality is only one. Under this view, then, "acceptability" and "grammaticality" are not synonyms, in that a sentence that is well-formed according to principles of the grammar, for instance, may turn out to be unacceptable due to parsing difficulties, etc. Both concepts are presumably gradient (i.e. acceptability and grammaticality are both "a matter of degree," in Chomsky's (1965) terms), but only acceptability is perceived directly. Grammaticality, like the other factors that contribute to acceptability, can only be inferred based on the evidence available.

Given this distinction between acceptability and grammaticality, it should be clear why one performs "acceptability experiments," but not "grammaticality experiments." Expressions such as "grammaticality judgments" are traditional, but as has often been pointed out, they are a misnomer (Schütze 2016; Myers 2017) and appear to be declining in frequency relative to "acceptability judgments" (Myers 2009).

Acceptability is assumed to be a percept that occurs when a speaker encounters a linguistic stimulus, and in an experiment, the speaker is typically asked to report on this percept (Schütze & Sprouse 2014). For example, speakers have a percept in response to *Girl the boy the saw* (presumably different from the percept they would have with *The girl saw the boy*) and can report on it. Referring to how this percept and the following report come to be has always been problematic, however. The term "introspection" is traditionally used, but this brings to mind introspectionist psychology and the idea that experiment participants can report on their internal cognitive mechanisms. Since no one assumes that the primary mechanisms underlying linguistic behavior are accessible to consciousness, this term can be misleading. Similarly, the term "intuition," rightly or wrongly, can give the impression that the process is capricious or unempirical. "Judgment" may avoid these unwanted implications, but it carries one of its own: it suggests that the process involves protracted and conscious deliberation on whether the sentence is acceptable or not, whereas in practice, the process appears to be virtually instantaneous.

The terms "introspection," "intuition," and "judgment," then, might all seem inadequate in one way or another, although all three are commonly used. As Schütze (2016) points out, terms like "sensation" or "reaction"

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

come much closer to capturing the true nature of acceptability as a percept. Neither is currently in regular use for this purpose, but it is helpful to keep them in mind when designing and interpreting the results from acceptability experiments, because they allow us to think much more clearly about what these experiments are actually measuring. Acceptability is obviously different from other "sensations" that one might want to measure, such as pain or thirst, but there are also important similarities. Like pain, for instance, acceptability is a percept that one feels subjectively, without necessarily being aware of the causes or mechanisms behind it, and there is a clear sense of fine-grained gradience: a sentence may feel slightly more acceptable or less acceptable than another, just as pain can subtly increase or decrease. In addition, the most reliable way to measure either acceptability or pain is to rely on what the individual reports.[1]

Among the various experimental methods that are used in linguistics, acceptability is often categorized as being both "behavioral," meaning that participants' overt response to the stimulus is what is being measured, and "offline," meaning that the participants' response is untimed and comes once the stimulus is complete (Garrod 2006). Acceptability thus contrasts with methods that are neurolinguistic, where brain responses are measured directly, and/or "online," where participants' responses are timed and come while the stimulus is in progress, such as in self-paced reading or eye-tracking (see Chapters 23, 24, and 27). These distinctions are valuable, but the offline vs. online contrast in particular should not be exaggerated. In standard acceptability experiments, participants typically read the stimulus sentence, give their response, and move to the next stimulus within 5 seconds, suggesting that participants are registering an immediate sensation without much conscious deliberation.[2] The main difference, then, between online methods and acceptability seems to be when the response/measurement take place (while the sentence is in progress vs. after it is complete), rather than the extent to which the response might be influenced by conscious thought (see Phillips & Wagers 2007; Lewis & Phillips 2015).

1.2 Best Practices in Acceptability Experiments

At a bare minimum, an experiment that attempts to measure acceptability will need to present a stimulus sentence to a participant and give that

[1] In fact, pain is standardly measured in terms of a Visual Analog Scale (VAS; Carlsson 1983), which bears many similarities to the types of scales used for measuring acceptability, and the effectiveness of treatments is determined by calculating whether there is a statistically significant difference in VAS between the treatment group and a control group, in a way analogous to acceptability studies (though unlike acceptability studies, pain studies also deal with the question of whether a given statistically significant difference in VAS is clinically meaningful (e.g. Forouzanfar et al. 2003; Dworkin et al. 2008)).

[2] This time estimate is based on our observations of the amount of time that most participants need to finish an entire acceptability experiment.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

participant a way to express the sensation that arises in response. Beyond that, though, several difficult questions about design and procedure must be addressed, and here I will present suggestions for doing this. I will follow the basic guidelines from Cowart (1997), supplemented with lessons learned in the subsequent decades of experimental practice.

1.2.1 Factorial Design

In general, presenting a single sentence in isolation to participants will be of little use. Most sentences will be neither 100 percent nor 0 percent acceptable, so their intermediate status will only be comprehensible in relation to other sentences. In addition, it is only the comparison to other sentences that can tell us what contributes to an increase or decrease in acceptability for a given sentence. For example, participants presented with a sentence like (1) will feel some degree of discomfort.

(1) Who do you think that _ will hire Mary?

Without anything else to compare it to, we won't know whether this level of unacceptability is a lot or a little, and in any event, we won't know what to attribute the unacceptability to. To remedy this, we can compare acceptability in (1) and (2).

(2) Who do you think _ will hire Mary?

(2) is exactly like (1) except that in (2), there is no overt complementizer. This is the well-known *that*-trace phenomenon (see Chapter 10, as well as Perlmutter (1971) and Pesetsky (2017)), so we expect that (2) will be of much higher acceptability than (1). Given these two data points, we can conclude that omission of *that* ameliorates extraction of an embedded subject, but we don't know how much of this effect might be due to the omission of *that* alone, which might affect acceptability even without extraction of the subject. To test for this, we can include control conditions to give us a baseline indication of the effect of *that* on acceptability. For example, we could consider the counterparts to (1) and (2) but with extraction of the object, as in (3) and (4).

(3) Who do you think that Mary will hire _ ?

(4) Who do you think Mary will hire _ ?

We could also use yes/no questions or declarative statements for this purpose; the crucial part is to get a baseline measure of the presence/absence of *that* independently of extraction of the embedded subject. The choice of which conditions count as controls is partly a matter of perspective: if we are testing subject vs. object extraction, such as (1) and (3), then (2) and (4) act as controls.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

In our example so far, then, we are manipulating two factors, each with two levels: *that* (presence vs. absence) and extraction site (subject vs. object). This gives us the four conditions (1)–(4), as in Table 1.1.

This "factorial design" of our stimuli (Fisher 1935) allows us to measure each of the two factors on their own, but more importantly, it allows us to detect any interactions between them, i.e. cases where the effect of one factor differs depending on the level of the other factor. In cases like this, for example, researchers have generally found that the effect of *that* is much larger when the extraction site is the embedded subject than when it is the embedded object, and the effect of the extraction site is much larger when *that* is present than when it is not (see Chapter 10 for more details).

A factorial design bears many important similarities to minimal pairs/sets as they are traditionally used in linguistics, and the use of the full set of four conditions in (1)–(4) to demonstrate the existence of an effect will be familiar even to syntacticians without backgrounds in experimentation. As with a minimal pair/set, the conditions in a factorial design are lexically matched to the extent possible, so that any differences found can be attributed to the factors being tested, rather than to random lexical differences across conditions. More so than with a traditional minimal pair/set, though, a factorial design allows one to examine the interaction of two factors with great care and precision. The robustness of the *that*-trace effect may obscure this point here, but it is not hard to imagine how in principle (and often in practice), the differences among conditions could be small enough that only a factorial experiment would be able to detect an interaction.

The example in Table 1.1 is known as a "2 × 2" design, because each of the two factors has two levels. It is of course possible to have more levels (e.g. 2 × 3) or more factors (e.g. 2 × 2 × 2), resulting in more conditions. This is sometimes justified, but it must be weighed against the disadvantages that come with a more complex design. Going beyond 2 × 2 can make lexically matched stimuli more difficult to construct, the statistical analysis more cumbersome, and the results more difficult to interpret. In addition, the more conditions in the design, the more stimuli in the experiment, so the amount of time that people can participate in the experiment without undue fatigue also sets an upper limit on the

Table 1.1 *A factorial design for measuring the* that-*trace effect*

		that	
		+	–
Extraction Site	**subject**	(1) Who do you think that _ will hire Mary?	(2) Who do you think _ will hire Mary?
	object	(3) Who do you think that Mary will hire _?	(4) Who do you think Mary will hire _?

complexity of the factorial design. With some other experimental techniques (e.g. self-paced reading), the relative difficulty of getting statistically significant contrasts can also discourage researchers from designing experiments with more than a 2 × 2 design. This is less of an issue in acceptability studies, where experiments finding significant contrasts among four to six conditions are not uncommon, but still, it should be kept in mind. Overall, then, 2 × 2 is generally the default design, with more complex designs recommended only when needed to test the hypothesis at hand and when the potential disadvantages have been carefully considered.

1.2.2 Lexicalizations

Having lexically matched stimuli as in Table 1.1 is an important first step in designing an experiment, since as we have seen, only true minimal sets like these allow us to know that differences in acceptability are due to differences in the factors. Running an experiment with only these four stimuli would not be a good idea, however. First, we would not know whether the results obtained were specific to these particular lexical items or were due to something more general about the structure. Since we are typically interested in the latter, it is advisable to abstract away from the effects of individual items by testing many different lexical versions of a set like Table 1.1, thus allowing any deeper structural effect to emerge. Second, presenting experiment participants with four such similar sentences could cause a number of problems. It could mean, for instance, that their experience of reading the first sentence, in which every word is new, will be different from their experience of reading the other sentences, in which the words are already familiar, that the ratings for one sentence could affect the ratings of the other, and that the monotony of the sentences could result in declining levels of attention among the participants as the experiment proceeds. Any of these consequences would be undesirable, since they could affect the results in unpredictable ways.

Given these considerations, we would ideally want each participant to see the four conditions of Table 1.1 in such a way that each condition is lexicalized differently. To make this possible, we need to construct four separate "lexicalization sets" of our factorial design; Table 1.1 already constitutes one such set. At a minimum, there must be as many lexicalization sets as there are conditions in the design. These sets need to be constructed by hand, so the potential for error is great (with potentially very bad consequences for the experiment). To avoid such errors, it is best to take one of the conditions from the factorial design and break the sample sentence into components in a table or spreadsheet (Cowart (1997)). Additional stimuli can then be created either by copying the component into the entire column, in cases where the component will

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

Table 1.2 *A table of components for Condition 1*

Who	do you	think	that	will hire	Mary	?
Who	do they	suppose	that	might contact	Tom	?
Who	did he	decide	that	should interview	Sophia	?
Who	did she	imagine	that	could photograph	Ryan	?

be found in all lexicalizations of that condition, or by populating the other cells in that column with similar lexical items, in cases where each lexicalization for that component will be different. An example is shown in Table 1.2, where the first, fourth, and seventh columns are invariable, while the others change.

The other conditions can now be constructed simply by copying the table and then manipulating it. Deleting the fourth column, for instance, will create Condition 2, while Condition 3 can be constructed by reversing the order of the fifth and sixth columns. When this process is done, the table or spreadsheet can be converted to ordinary text. By following this procedure, a full set of 16 stimuli (4 lexicalization sets of 4 conditions each) can be created with only a minimal possibility of error.

1.2.3 Counterbalancing

Let us suppose that at this point, for each of the conditions in Table 1.1, we have a set of four lexicalizations as in Table 1.2. For the reasons given earlier, we do not want any participant to see a given lexicalization in more than one condition. We also do not want different lexicalizations of a given condition to be seen by different numbers of participants, since we don’t want certain stimuli to be overrepresented in the results. These considerations are aspects of “counterbalancing,” and the usual way to ensure counterbalancing of the stimuli is to perform a “Latin square” procedure (using either a spreadsheet or a script). This refers to distributing the stimuli into lists such that each list contains only one representative of each lexicalization and the same number of stimuli for each condition, as illustrated in Table 1.3, without any repetition of stimuli across lists.

Each list in Table 1.3 corresponds to what an individual participant would see in the experiment. In this example, each participant will see one example of each condition. If we want participants to see more than that, as we often do (see Section 1.2.6 below), the same principles and procedures apply, but we will need to create additional stimuli in order to achieve our counterbalancing goals. For example, if we want each participant to see 5 examples of each condition, then we need to create 20 lexicalization sets (5 examples × 4 conditions), or put differently, 20 rows in the equivalent of Table 1.2.

Table 1.3 *Counterbalanced lists of experimental stimuli using a Latin square design*

List 1	List 2	List 3	List 4
Who do you think that will hire Mary?	Who do they suppose that might contact Tom?	Who did he decide that should interview Sophia?	Who did she imagine that could photograph Ryan?
Who did she imagine could photograph Ryan?	Who do you think will hire Mary?	Who do they suppose might contact Tom?	Who did he decide should interview Sophia?
Who did he decide that Sophia should interview?	Who did she imagine that Ryan could photograph?	Who do you think that Mary will hire?	Who do they suppose that Tom might contact?
Who do they suppose Tom might contact?	Who did he decide Sophia should interview?	Who did she imagine Ryan could photograph?	Who do you think Mary will hire?

1.2.4 Order

Presenting the lists in Table 1.3 as they are introduces a confound: Condition 1 is always first, Condition 2 is always second, etc. This could influence the results in undesirable ways, so the obvious solution is to randomize the order in each list so that each participant sees the conditions in different orders. The lists in Table 1.3 are short enough that this could easily be done manually, but in most experiments, this is not practical. A common solution using a spreadsheet is to insert a column to the left of each list, insert a random number in each cell (using the RAND function in Excel, for instance), and then sort each list by the random number column. If there are multiple tokens of each condition in a list, then randomization will sometimes result in two such tokens being adjacent. Some researchers prefer to avoid this by performing "pseudo-randomization," in which such cases are manually separated after the regular randomization process is complete.

Counterbalancing for order is also advisable. This can be done by re-randomizing the lists in Table 1.3 or by reversing the randomized order already obtained. Either method will produce a new set of four lists.

1.2.5 Fillers

Even with experimental stimuli fully counterbalanced and randomized (or pseudo-randomized), participants will still see lists of sentences that are all structurally very similar. This could result in participants becoming desensitized to the distinctions we are interested in or beginning to speculate as to the structure being investigated, all of which could affect their judgments in unknown ways. "Filler" items (i.e. stimuli that are not part of the factorial design of the experiment) help avoid this outcome, in that they "cleanse the palate" between experimental items, while also disguising the purpose of the experiment.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

Table 1.4 *Planning table for fillers assuming a factorial design as in Table 1.1, 6 tokens per condition, and a 2:1 filler–experimental ratio*

Acceptability	Experimental items	Filler items	Total
High	18	6	24
Intermediate	0	24	24
Low	6	18	24
Total	24	48	72

For filler items to serve their purpose, there should be at least a 1:1 ratio of fillers to experimental items. Having more fillers is better, though this has to be balanced against the need for the overall list to be of reasonable length. The filler–experimental ratio that we choose will tell us how many fillers need to be constructed, but knowing what type they should be requires doing an inventory of the anticipated acceptability levels of the experimental items. The reason for this is that the fillers should help create a list of stimuli that is roughly balanced in terms of acceptability, so that participants will use the full range of the response scale provided. Table 1.4 provides an example of such an inventory for an experiment with a factorial design as in Table 1.1. In that design, we can anticipate that Condition 1 will be of relatively low acceptability, but the other three will be relatively high. Table 1.4 assumes that we have chosen to have participants see 6 tokens of each condition, with a 2:1 filler–experimental ratio.

As Table 1.4 shows, fillers can be used to correct the imbalance in acceptability among the experimentals, so that the overall list of stimuli is distributed well across the full range of acceptability.

It would be counterproductive for participants to become aware of the filler–experimental distinction, so to the extent possible, most of the fillers should be superficially similar to the experimental items. Table 1.5 gives examples of possible fillers at three rough levels of acceptability for an experiment as in Table 1.1.

Having fillers of extremely high or extremely low acceptability can serve another purpose: They can guard against ceiling or floor effects with the experimental stimuli. Both of the “high” examples in Table 1.5, for example, would probably be of higher acceptability than the high-acceptability experimentals in Table 1.1, and conversely for the “low” examples in relation to the low-acceptability experimental. In addition, these fillers at the extremes of acceptability can be used to detect participants who are not attending to the task (see Section 1.2.8).

Another function of fillers is to act as practice items at the beginning of the experiment. These need not (and should not) be explicitly marked as such, but having the first 3–5 stimuli be fillers allows participants to

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

Table 1.5 *Example fillers for an experiment with a factorial design as in Table 1.1*

High	Who says that it will rain tomorrow? What is the name of your dentist?
Intermediate	Who do you wonder whether anyone will approach? What do you believe that the man who you saw in the park ate?
Low	Who do thinks you that the birds are singing? What will they say about a cars are park on the street?

become familiar with the task and the response scale before they start reacting to experimental items.

It is sometimes convenient to join two or more small experiments as sub-experiments of a larger experiment. In this instance, the experimentals from one sub-experiment can act as fillers for another, but because of counterbalancing in each sub-experiment, the fillers will not be uniform for all participants. Even in this case, though, an analysis as in Table 1.4 should be done and some true fillers created so that gross imbalances in acceptability across the stimuli are avoided.

1.2.6 Presentation of Stimuli to Participants

If stimuli are counterbalanced as in Section 1.2.3, the number of conditions in the factorial design will determine the minimum number of lists to be created. In a 2 × 3 design, for instance, there will need to be 6 lists (or a multiple of 6) in order for counterbalancing to be preserved. The number of lists, in turn, determines the ideal number of participants. With 6 lists, for instance, having 6 participants (or again, a multiple of 6) makes it possible to have the same number of participants for each list and thus maintain full counterbalancing. Small deviations from this ideal are unlikely to have a large effect on the results, but major deviations (e.g. having many more participants for one list than for others) should be avoided, since the resulting lack of counterbalancing could plausibly affect the outcome of the experiment.

Apart from counterbalancing, statistical power should also be taken into consideration when deciding on the number of experimental participants. Increasing the number of participants increases the statistical power, as does increasing the number of tokens of each condition that participants see. Sprouse and Almeida (2017) give some useful guidelines on this (see also Cowart 1997), but as a rough rule of thumb, two or three dozen participants are generally sufficient to yield useful results with most experimental designs. It is also often possible to have at least four or five tokens per condition without the overall experiment becoming too long.

In practice, presenting the stimuli in written form is by far the most common option. Having participants listen to the stimuli instead might

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

seem preferable, but the amount of work required to record stimuli uniformly across all levels of acceptability discourages most researchers from doing this. Nonetheless, recorded stimuli have been used in some cases where oral presentation was particularly crucial (see Polinsky et al. 2013; Ritchart, Goodall, & Garellek 2016; Sedarous & Namboodiripad 2020).

In addition to the experiment itself, a consent form and a language background questionnaire are often also presented to participants. When possible, it is preferable to present the questionnaire after the experiment, to avoid the possibility that it might affect participant responses.

1.2.7 Response Method

As discussed in Section 1.1, acceptability is a kind of sensation experienced by speakers of the language after they have heard a linguistic stimulus, so if we are to measure this experimentally, we must provide the stimuli that will induce these sensations and a way for participants to record them. There are many such response methods (see Chapter 2 for detailed discussion), but a fixed numerical scale is very widely used, because it is easy to implement for researchers and easy to understand for participants (see also the detailed comparison between this method and others in Langsford et al. (2018)).

A fixed numerical scale, such as one going from 1 to 7, is technically an ordinal scale: there is a rank order among the elements (i.e. the numerals) on the scale (Stevens 1946). To interpret the results, however, it is helpful to be able to treat the scale as an interval scale, in which the difference between any two adjacent elements on the scale is the same. The use of numerals for the points on the scale implies this, but does not guarantee it. For whatever reason, participants might treat the difference between 1 and 2, for instance, as being smaller than the difference between 6 and 7, and in that case, we no longer have an interval scale. Fortunately, there are a number of steps the researcher can take to discourage participants from conceiving of the scale in this way. First, the numerals should be spread out evenly on the scale so that the amount of physical space between any two adjacent elements is always the same. Second, the extremes of the scale should be labeled (e.g. "good" and "bad"), but the intermediate points on the scale should not be. Such intermediate labels (e.g. "fair," "not very good," etc.) virtually ensure that participants will not treat the scale as an interval scale and may even discourage them from treating it as ordinal, thus making the results very difficult to interpret. Third, the numerals should all be similar in appearance and status so that participants do not perceive a "break" at any point along the scale. For this reason, scales including both single-digit and double-digit numbers, or both positive and negative numbers, should be avoided. Fourth, the scale should consist of an odd number of numerals, so that participants don't introduce an artificial break and divide them, consciously or unconsciously, into "good" and "bad" ones.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

Scales that range from 1 to 5 or from 1 to 7 make it easy to satisfy the above constraints, so they are very frequently used. Because of the way that humans interact with left-to-right arrays and mentally represent numerical scales (see Natale et al. 1983; Davidson 1992; Dehaene et al. 1993; Zorzi et al. 2002; Harvey et al. 2013), it is best for the lowest number (i.e. 1) to correspond to lowest acceptability and be on the left, while the highest number (i.e. 5 or 7) should correspond to highest acceptability and be on the right. In other words, the numbers should be arranged in ascending order from left to right and from least acceptable to most acceptable (at least for languages in which that is the customary direction of text).

Participants of course need to be told what will happen and what they need to do in the experiment, but it appears that the exact form of these instructions does not crucially affect the results (Cowart (1997), though see Beltrama and Xiang (2016) for results perhaps suggesting otherwise). It is probably best to keep the instructions very short, as in (5), and place filler items representing a range of acceptability as the first several stimuli.

(5) Sample instructions: Rate each sentence on a scale from 1 (bad) to 7 (good) based on how it sounds to you. Use only your sense as a speaker of the language. There are no correct or incorrect answers and you should not try to analyze the sentence.

The initial filler items give participants a better idea of what is being asked of them and should allow them to feel comfortable with the task by the time they get to any experimental items.

1.2.8 Analyzing the Results

Before the participants' numerical responses are analyzed, it is worth employing a screening procedure to eliminate any participants who appear not to have completed the task appropriately, which could occur because they do not actually speak the language being investigated, because they unintentionally inverted the "good" and "bad" labels on the scale, or because they chose not to follow the instructions for some reason (see Chapter 4 for further discussion). At its most basic level, screening can be done by visually inspecting the results to detect participants who gave the same rating to all sentences, but more involved procedures are worthwhile in most cases. There are a variety of ways to do this, but the basic idea is to look for participants who have many anomalous responses to the filler items. One could, for instance, select filler items that are especially good or especially bad (and for which one would thus expect responses on one side or another of the rating scale) and eliminate participants who exceed a certain threshold for "errors" on these items (where an "error" might be a rating of 3 or below on a seven-point scale for a sentence that is known to be highly

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

acceptable, for instance). It is also possible to compute the group mean and standard deviation for each filler, note cases where a participant's ratings are more than two standard deviations away from the mean, and then eliminate participants who go beyond a certain number of such cases. If participants are eliminated in this way, they may need to be replaced by new participants if counterbalancing as in Section 1.2.6 is to be preserved.

Such screening procedures should be used cautiously and judiciously, of course, so as not to eliminate legitimate results, but they can be a useful way to remove some of the noise that inevitably arises in experiments of this type. Another potential source of noise stems from the fact that individual participants may choose to use the numerical scale for their responses in very different ways. One participant, for instance, might concentrate most responses around the center of the scale, while another might use the extremes much more readily, with the result that a response of "7" on a seven-point scale could mean different things for the two participants. Because of this type of variation, it is advisable to standardize the results by transforming them to *z*-scores. Each *z*-score shows how many standard deviations the participant's response is above or below that participant's mean for all items. Two participants might both have a *z*-score of 0 for a particular item, for instance, meaning that their responses for that item were the same as their overall mean, even though their raw ratings for that item were different. Any statistical analysis of the results should then be based on the *z*-scores (though doing a parallel analysis using the raw scores is also possible).

A basic analysis of the results using descriptive statistics should include the mean, standard deviation, and response distribution for each condition in the experimental design. Visualization of the results usually includes at least a bar or line graph showing the means and error bars for all conditions. Line graphs make interactions (or a lack of interaction) more perspicuous and avoid giving the appearance that there is an absolute lower bound to acceptability (which there will not be if *z*-scores are being used). In addition to descriptive statistics, an analysis in terms of inferential statistics needs to be done in order to test for statistical significance. This has traditionally been done using a repeated-measures ANOVA, but is now more commonly done with a linear mixed-effects model.

1.3 What Acceptability Includes

Following an experimental procedure as outlined in Section 1.2 will very regularly result in statistically significant contrasts among conditions. In fact, it is not uncommon to be able to detect significant differences among three or four conditions, signifying three or four distinct levels of

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

acceptability, in effect. Acceptability studies are more fine-grained in this sense than many other experimental methods for the study of syntax, such as self-paced reading or event-related brain potential (ERP) (see Chapters 23 and 24), where such multiple contrasts are unusual. Interpreting the contrasts that the experiment detects is not always a simple matter, however, since acceptability seems to be sensitive to a number of factors, including ones that the experimenter did not intend to investigate. In this section, we survey some of the main factors that are known to influence acceptability.

1.3.1 Grammaticality

Not surprisingly, acceptability experiments are sensitive to grammaticality, in that sentences that are generable by the postulated grammatical model are generally more acceptable than those that are not. Studies such as Sprouse and Almeida (2012) and Sprouse, Schütze, and Almeida (2013) have found that where linguists have traditionally claimed a difference in grammaticality, experimental work finds a difference in acceptability in over 90 percent of the cases. There are exceptions to this, and these are worth investigating, but as a general rule of thumb, one can expect that if there is a well-founded claim of a difference in grammaticality between two conditions, this difference will be detectable in an acceptability experiment.

Cowart (1997) shows this convincingly across an interesting range of phenomena. In terms of *wh*-extraction out of NPs, for instance, he finds (6b) to be significantly less acceptable than (6a), an effect that has been claimed to stem from grammatical constraints on extraction (Chomsky 1977; Fiengo & Higginbotham 1981). (6c) does not show this degradation, again in line with what grammatical accounts predict.

(6) a. Who did the Duchess sell [a portrait [of __]] ?
b. Who did the Duchess sell [Max's portrait [of __]] ?
c. Who did the Duchess sell [Max's portrait] [to __] ?

For the four conditions in our earlier discussion of the *that*-trace effect as in (7) (repeated from (1)–(4) above), Cowart found that the contrasts predicted by grammatical accounts of the effect are easily detected in acceptability experiments.

(7) a. Who do you think that _ will hire Mary?
b. Who do you think _ will hire Mary?
c. Who do you think that Mary will hire _ ?
d. Who do you think Mary will hire _ ?

Specifically, (7a) is significantly less acceptable than any of the other conditions. Finally, with respect to the behavior of anaphors, Cowart

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

finds that Binding Theory effects show up as significant contrasts in acceptability experiments (see Chapter 11 for further discussion). In (8)–(10), the anaphor has a local antecedent in the (a) examples, in accord with Principle A, but a remote antecedent in the (b) examples, in violation of Principle A.

(8) a. Cathy's parents require that Paul support himself.
b. Paul requires that Cathy's parents support himself.

(9) a. Cathy's parents require that Paul support himself and the child.
b. Paul requires that Cathy's parents support himself and the child.

(10) a. Cathy's parents require that Paul support both himself and the child.
b. Paul requires that Cathy's parents support both himself and the child.

In all three cases, Cowart shows experimentally that (a) is significantly more acceptable than (b), as predicted.

It may turn out that some or all of the contrasts in (6)–(10) are not ultimately due to the grammar per se, but are to be accounted for in other ways (see Phillips & Wagers (2007) for general discussion). Nonetheless, these cases show that the types of contrasts that syntacticians are typically interested in generally yield very robust distinctions when examined within an acceptability experiment. Beyond this, however, most acceptability experiments also uncover distinctions that are clearly not tied to standard notions of grammaticality, in that significant degradations are revealed experimentally that do not correspond to any expected declines in grammaticality.

This fact highlights the importance of the distinction between "acceptability" and "grammaticality" (see Section 1.1), and in Sections 1.3.2–1.3.4, we survey those factors beyond grammaticality that are thought to play a role in acceptability (see also Chapter 5 for more detailed discussion).

1.3.2 Presence of a Dependency

One of the most solid, well-replicated findings in the experimental literature is that the presence of a *wh*-dependency results in a sharp decline in acceptability relative to controls without such a dependency. In the experiments from Cowart (1997) sketched above, for instance, all of the sentences in (6) are significantly less acceptable than the control in (11).

(11) Why did the Duchess sell a portrait of Max?

This is a striking finding because of the sentences in (6), (a) and (c) seem unobjectionable when presented in isolation and are standardly taken to be fully grammatical, on a par with (11). The difference, though, is that (6a)

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

and (c) involve a clear dependency between a *wh*-filler and a gap, whereas (11) does not (either because there is no dependency or because there is no subcategorized gap).[3]

Similar results are seen in Sprouse, Wagers, and Phillips (2012) and much subsequent work, where sentences with a *wh*-dependency are significantly less acceptable than matched controls where there is only a vacuous *wh*-dependency. (12a) is an example of such a control, in that there is only a vacuous dependency, while (12b) is the type of sentence that shows significant degradation, apparently because of the non-vacuous dependency in this case.

(12) a. Who __ thinks that John bought a car?
b. What do you think that John bought __ ?

Both (12a) and (12b) are textbook examples of grammatical sentences, so the difference observed experimentally does not seem to be attributable to a difference in grammaticality.

Filler–gap dependencies such as those just seen have been studied extensively in the experimental literature on acceptability, and the results have been remarkably consistent: the presence of a non-vacuous dependency leads to significant degradation. Whether other types of syntactic dependencies, such as binding, control, and A-dependencies (e.g. in passives, unaccusatives, etc.), show similar effects remains largely unexplored. Recent work by Dayoung Kim (Kim & Goodall 2018) found no significant difference in acceptability between simple active and passive sentences as in (13).

(13) a. The director hugged the actress in the theater.
b. The actress was hugged by the director in the theater.

In many analyses, there is a non-vacuous dependency in (13b) between *the actress* and the object position in the clause, but this putative relation does not seem to lead to the type of degradation seen with *wh*-dependencies (though Kim does find an effect for the dependency when it spans more than one clause). Overall, though, much work remains to be done on the possible effects on acceptability of syntactic dependencies other than traditional filler–gap dependencies.

1.3.3 Length of Dependency

In addition to the simple presence of a *wh*-dependency, the length of the dependency also has an effect on acceptability, with greater distance (in terms of syntactic structure) resulting in a larger degradation. This was

[3] Hofmeister, Culicover, and Winkler (2015) provide another very clear example of how the simple presence of a *wh*-dependency results in significant degradation, as do Omaki et al. (2020) with regard to the presence of a scrambling dependency in Japanese. See also Namboodiripad (2017).

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

first seen in cases where the dependency crosses an NP boundary. In the experiment from Cowart (1997) mentioned above, a significant difference is found between sentences like (14a), where an NP boundary is crossed, and (14b), where it is not.

(14) a. Who did the Duchess sell [a portrait [of __]] ? (= (6a))
b. Who did the Duchess sell [a portrait] [to __] ?

This results in significantly lower acceptability for (14a) relative to (14b).

Crossing clause boundaries seems to have a similar effect. Work by Bethany Keffala (reported in Keffala (2011); Keffala & Goodall (2011, 2013); Goodall (2017)), for instance, shows that placing the gap inside an embedded clause, such as in (15b) relative to (15a), significantly degrades acceptability, as seen in Figure 1.1 (results reported on an eleven-point scale).

(15) a. These are the potatoes [that we prepared __].
b. These are the potatoes [that we realized [that the chef prepared __]].
c. These are the potatoes [that we inquired [how the chef prepared __]].
d. These are the potatoes [that we spoke to the chef [that prepared __]].

(15c) and (15d) represent island violations (a *wh*-island and a relative clause island, respectively), so their low level of acceptability is expected, but the long extraction in (15b) occupies an intermediate position: less acceptable than a simple case of extraction as in (15a), but more acceptable than island violations as in (15c) and (15d). The degradation in (15b) could also be related to the fact that it simply has more words than (15a), but given the size of the effect, it seems likely that the length of the dependency is also playing a major role. Alexopoulou and Keller (2007) present similar results based on *wh*-questions, rather than relative clauses.

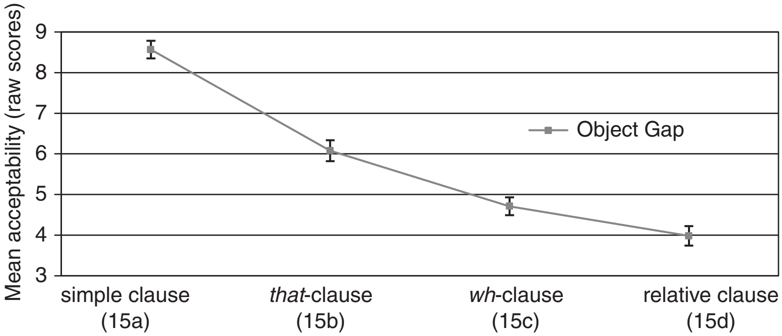

Figure 1.1 Effect of dependency length on relative clauses (Keffala 2011)

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

Whether acceptability is sensitive to more fine-grained distinctions in dependency length is less clear. In cases where the number of NP and/or clause boundaries remains constant, for instance, one might expect that all other things being equal, dependencies with the gap in object position would be less acceptable than those with the gap in subject position. This expectation is not confirmed, however, in a number of studies of extraction from embedded clauses in English. In the experiment in Cowart (1997) mentioned above, for example, no significant difference is found between embedded subject and object gaps in *wh*-questions such as those in (16) (see also Fukuda et al. 2012).

(16) a. Who do you think [_ will hire Mary] ? (= (7b))
b. Who do you think [Mary will hire _] ? (= (7d))

1.3.4 Frequency

There are two types of frequency that might be relevant to acceptability: the relative frequency of the lexical items used and the relative frequency of the syntactic structure (see also Chapter 25). With regard to the first, it seems intuitively likely that sentences with high-frequency words, as in (17a), would be more acceptable than matched sentences with one or more low-frequency words, as in (17b).

(17) a. The girl found the box.
b. The seamstress found the satchel.

It also seems intuitively likely that such effects would only emerge with very extreme contrasts in frequency, as in the examples in (17). To my knowledge, however, intuitions like these have not been tested experimentally in a systematic way (but see Sag, Hofmeister, & Snider 2007). Since most acceptability experiments in syntax use lexically matched sets of stimuli, the issue of differences in lexical frequency does not typically arise.

Experimental studies of the effects of particular lexical items on the acceptability of syntactic structures do exist, of course (see, e.g., Fukuda 2012, 2017), though they are often based on the semantic properties of the lexical items in question, rather than their frequency. Frequency-based analyses also exist, however, such as in Bresnan and Ford (2010), where it is shown that speakers' acceptance of a given verb in one or another of the two patterns in the dative alternation (e.g. *gave the book to the girl* vs. *gave the girl the book*) correlates with the relative frequency of that pattern with the verb. With regard to the effect of the frequency of the structure, abstracting away from particular lexical items, and how this might affect acceptability, much less is known. On the face of it, however, it seems unlikely that frequency and acceptability will always correlate. Active clauses, for instance, are much more frequent than passive clauses in

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

corpora, but it is not obvious that active clauses are more acceptable than passive clauses. Indeed, as discussed above for (13), there seems to be no significant difference between the two when this is tested experimentally. Frequency and acceptability thus seem to be largely independent measures. Some generalizations about the relationship between the two can be made, however, as Bermel and Knittl (2012) point out that if a structure is of high frequency, it will also be of high acceptability (but not vice versa), and that if a structure is of low acceptability, it will also be of low frequency (but not vice versa). Crucially, the reverse of each of these conditional statements is clearly not true. That is, if a structure is of high acceptability, we cannot conclude that it will also be of high frequency, and if a structure is of low frequency, we cannot conclude that it will also be of low acceptability.

1.4 Motivations for Doing Acceptability Experiments

In the previous sections, we have seen the many steps involved in creating a well-designed acceptability experiment and the reasons why the researcher should carry these steps out with appropriate care. We have also seen that sentence acceptability experiments are very sensitive not only to the grammaticality of the stimuli, as one would naturally assume, but also to factors that a syntactician might not initially feel the need to measure experimentally, such as the presence/absence of a dependency or the length of that dependency. Given all of this, one might reasonably conclude that doing acceptability experiments is a lot of trouble and that the payoff is not that great, in that the results are hard to interpret because of the uncertainty as to what is causing the level of acceptability that is measured.

In this section, we will see that this discouraging view of acceptability experiments, though based on fact, is not the whole story, and that there are a number of ways that researchers can use acceptability experiments to address important syntactic questions with a level of precision and insight that is not possible with other methods.

1.4.1 Testing Claims of (Un)acceptability

Perhaps the most obvious way that acceptability experiments can be useful is by adjudicating cases where more traditional methods of determining acceptability have yielded unclear or controversial results. One area where experimental work of this type has played an especially important role is in Superiority, the phenomenon in which an object *wh*-phrase may not be fronted when the subject is also a *wh*-phrase, as in (18) (see Chapter 5).

(18) a. **Who** bought **what**?
b. ***What** did **who** buy __ ?

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

It has sometimes been claimed that German does not show this effect (e.g. Grewendorf 1988), but Featherston (2005) uses the results of an acceptability experiment to argue that in fact it does, despite initial appearances. This is thus an instance where the experimental procedure is able to detect a contrast (i.e. between the equivalents of (18a) and (18b) in German) that more traditional methodologies had suggested might not exist.

There have also been examples going in the other direction, where traditional work claims that there is a contrast, but experimental work suggests that there is not. An example of this is Fedorenko and Gibson (2010), who examine the claim that Superiority violations like (18b) become more acceptable when a third *wh*-phrase is added, as in (19) (Bolinger 1978; Kayne 1984)

(19) **What** did **who** buy __ **where**?

Fedorenko and Gibson present the results of an experiment suggesting that there is no difference in acceptability between (18b) and (19).

Another such example concerns the claim that *wh*-extraction from within an embedded subject, known to result in very low acceptability, shows significant amelioration when the containing subject is itself a *wh*-phrase that has been fronted within its clause (Torrego 1985; Lasnik & Saito 1992; Kayne 1984). For instance, (20c) is claimed to be more acceptable than (20a) or (20b).

(20) a. [Which animal] will [several movies about __] be shown to the visitors?
b. [Which animal] do you wonder whether [several movies about __] will be shown to the visitors?
c. [Which animal] do you wonder [how many movies about __] will be shown to the visitors?

Some researchers have argued that this claim is not correct (e.g. Gallego 2009; Müller 2010), so I tested it by means of an acceptability experiment in which 48 participants saw four tokens of each of the conditions in (20), in addition to 9 other conditions and 57 filler items (Goodall 2015). Lists of stimuli were fully counterbalanced and pseudo-randomized. Participants rated each stimulus on a seven-point scale and the results (converted to *z*-scores) are presented in Figure 1.2.

The results show that the expected amelioration for extraction out of a fronted *wh*-subject, as in (20c), does not occur, suggesting that the initial claim about these structures was incorrect.

All of the above examples show that acceptability experiments can be an excellent supplement to (or even substitute for) acceptability judgments gathered in more traditional ways. None of this should suggest, however, that acceptability experiments are simply the handmaid of traditional methods, since even if the ultimate goal is merely to determine how

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

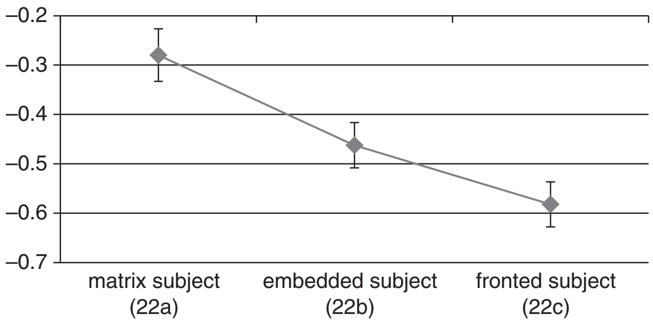

Figure 1.2 *Wh*-extraction from three types of subjects

acceptable a given structure is, formal experiments can provide a type of evidence that more traditional methods are in principle incapable of doing.

One reason that experiments can do this is that they seem to be able to measure contrasts with equal precision in sentences of both high and low acceptability. A good example of this comes from an experimental study of the difference between bare *wh*-phrases (e.g. *who*, *what*, etc.) and more lexically elaborated *wh*-phrases (e.g. *what boy*, *which of the apples*, etc.), often referred to as D-linked *wh*-phrases (Goodall 2015). The experiment examines this difference in the case of extraction from embedded clauses, including both ordinary complement clauses, as in (21), and island clauses, as in (22) and (23).

(21) a. **What** do you believe that he might buy __ ?
b. **Which of the cars** do you believe that he might buy __ ?

(22) a. **What** do you believe the claim that he might buy __ ?
b. **Which of the cars** do you believe the claim that he might buy __ ?

(23) a. **What** do you wonder who might buy __ ?
b. **Which of the cars** do you wonder who might buy __ ?

It has long been noted that D-linking causes amelioration of island violations, so the (b) sentences in (22) and (23) are expected to be more acceptable than their counterparts in (a). The experimental results show this expected effect, but they also show an effect of approximately the same size in non-island stimuli such as (21). This is surprising because no such effect in non-islands had previously been noticed and it is not something that most standard accounts of D-linking would lead us to expect. This finding is thus of interest for our understanding of the larger phenomenon, but for present purposes, what matters more is the fact that this finding could probably not have been obtained by gathering acceptability judgments in the traditional way. For most speakers, the contrasts in (22)

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

and (23) seem easy to articulate (e.g. "(a) sounds relatively bad and (b) sounds relatively good"), but this is much less true in (21), where both (a) and (b) sound good. Even if speakers could articulate the difference between (21a) and (21b), it is not clear that they could judge whether the amelioration there is of the same size as it is in (22) and (23), as the experimental results suggest.

The ability to measure "degrees of goodness," in addition to "degrees of badness," is thus one of the important attributes of acceptability experiments that is not shared by more traditional methods of ascertaining acceptability. Another concerns the fact that acceptability experiments typically make use of explicitly formulated factorial designs. Traditional work often uses minimal sets, which are created in a spirit very much in line with factorial designs, but they are usually less explicit, and the relative lack of precision in the measurements means that they can't be as informative as a factorial design. To see how useful factorial designs can be, consider the experimental study of the freezing effect in extraposition in Hofmeister, Culicover, and Winkler (2015). It has been known for many years that extraposed phrases become "frozen" to any subextraction from within them (i.e. they become islands, in effect). Hofmeister et al. explore this by constructing stimuli with both Extraction (+ and −) and Extraposition (+ and −) as factors, as shown in (24).

(24) a. Kenneth revealed that he overheard a nasty remark [about the President] earlier.
[–Extraction, –Extraposition]
b. Kenneth revealed which President he overheard a nasty remark [about _] earlier.
[+Extraction, –Extraposition]
c. Kenneth revealed that he overheard a nasty remark earlier [about the President].
[–Extraction, +Extraposition]
d. Kenneth revealed which President he overheard a nasty remark earlier [about _].
[+Extraction, +Extraposition]

In many ways, the results obtained are what one would expect: there is a significant degradation associated with Extraction (e.g. (24b) is less acceptable than (24a)) and with Extraposition (e.g. (24c) is less acceptable than (24a)). Neither of these declines in acceptability is surprising given what we saw in Section 1.3.2 about the presence of dependencies. Also not surprising is the fact that condition (24d) is the least acceptable of all, since this is the "freezing" case. What is surprising is that condition (24d) does not show "superadditivity." That is, the level of unacceptability of (24d) can be accounted for by simply adding the degradation due to Extraction

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

and the degradation to Extraposition, without the need for a separate freezing principle.[4] This study thus shows the utility of carefully setting up a factorial design and treating that design seriously. It is only by measuring the acceptability of each condition and examining the relationships among those four measurements that we can conclude that the low acceptability of (24d) falls out from the same factors that cause the degradation in (24b) and (24c).

There are many cases like (24d) in the syntax literature. That is, there are many cases where the low acceptability of a structure is noted, but a careful factorial design is not constructed and the possibility that the low acceptability might stem from the additive effect of other factors is not considered. To take just one example, consider the phenomenon of partial *wh*-movement ("simple partial *wh*-movement," in the terminology of Fanselow (2017)) found in a variety of languages around the world. In this phenomenon, *wh*-phrases may move to their scope position, move to a position lower than their scope position, or remain *in situ*. These possibilities are illustrated in (25) for Indonesian (Saddy 1991).

(25) a. **Siapa** yang Bill tahu Tom cintai __ ? (Full *wh*-movement)
who FOC Bill knows Tom loves
b. Bill tahu **siapa** yang Tom cintai __ ? (Partial *wh*-movement)
Bill knows who FOC Tom loves
c. Bill tahu Tom men-cintai **siapa**? (*Wh-in situ*)
Bill knows Tom loves who
'Who does Bill know that Tom loves?'

Full *wh*-movement in Indonesian obeys standard island constraints, as one might expect, and *wh-in situ* does not. Partial *wh*-movement also obeys island constraints, both with respect to the overt movement that occurs and with respect to the distance between the *wh*-phrase and its scope position (Saddy 1991; Cole & Hermon 1998, 2000). (26) shows representative sentences for the three types of *wh*-questions with a negative island.

(26) a. ***Apa** yang Tom tidak harap Mary beli __? (Full *wh*-movement)
what FOC Tom not expect Mary bought
b. *Tom tidak meng-harap **apa** yang Mary beli __?
(Partial *wh*-movement)
Tom not expect what FOC Mary bought
c. Tom tidak meng-harap Mary mem-beli **apa**? (*Wh-in situ*)
Tom not expect Mary bought what
'What did Tom not expect that Mary would buy?'

[4] In fact, (24d) is slightly more acceptable than expected given this addition. See Hofmeister et al. (2015) for details.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

Table 1.6 *A factorial design for measuring the island effect with partial* wh-*movement*

		Question type	
		in situ	**partial movement**
Island structure	–	(25c)	(25b)
(negation)	+	(26c)	(26b)

The crucial and perhaps most surprising fact here is the unacceptability of (26b), i.e. the sensitivity of the *wh*-phrase to an island structure through which it has not moved overtly. Viewing this data point in isolation, however, it is difficult to know whether there is truly an island effect here or whether the low acceptability of (26b) is due to the additive effect of the negative structure and the partial *wh*-movement. This issue could be addressed with a factorial design as in Table 1.6 (with items lexically matched appropriately).

If the unacceptability of (26b) can be reduced to the sum of the degradation induced by partial movement itself (seen by comparing (25c) and (25b)) and the degradation induced by the more complex island structure (seen by comparing (25c) and (26c)), then we could conclude that the island effect for partial movement is in a sense illusory; there is no separate island constraint. This is the type of conclusion, for instance, that Hofmeister et al. (2015) reach with respect to the freezing properties of extraposition. If, on the other hand, the unacceptability of (26b) is significantly greater than the sum of the partial movement effect and the island structure effect, then we could conclude that partial movement truly is sensitive to island constraints even when the *wh*-phrase has not moved overtly through the relevant structure. This is the type of conclusion, for instance, that Sprouse, Wagers, and Phillips (2012) reach with regard to a number of classical island effects in English (see Chapter 9). It is also the conclusion that has been adopted in the literature for partial *wh*-movement, but without an explicit design as in Table 1.6 and a precise means of measuring acceptability, it is hard to be sure that this conclusion is correct.

Determining the correct answer here and knowing whether the apparent island sensitivity of partial movement is truly due to an island constraint or is simply what happens when two relatively complex properties coexist in a single sentence is crucial to our understanding of extraction phenomena and locality in syntax, so a lot hinges on the outcome of an experiment such as that sketched in Table 1.6. The same can be said for many phenomena and issues in the syntax literature, and in this sense, formal acceptability experiments can provide important insights that more traditional techniques cannot.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

1.4.2 Cross-linguistic Comparisons

Another important benefit of experimental approaches to acceptability comes from the ability to make precise and well-grounded cross-linguistic comparisons (see Chapter 7 for more extensive discussion of this point). Examples of a given structure being possible in one language but not in another are very common in the syntactic literature and uncovering and accounting for such cases is the syntactician's stock-in-trade. It is also common, however, to see claims that a given structure is possible (or impossible) in both languages under discussion, but that it is nonetheless more acceptable in one of the languages than in the other. Such observations have played an important role in the literature on phenomena such as parasitic gaps, weak islands, resumptive pronouns, and extraction from complement clauses, but it has sometimes been difficult to characterize these cross-linguistic differences, much less to account for them.

Acceptability experiments clearly have a lot to offer in this domain, because they allow us to quantify particular syntactic effects and then make comparisons across languages based on that. For example, if we were examining languages that seem to have the *that*-trace effect, we could conduct experiments along the lines of what was described in Section 1.2 above for each of the languages, calculate an effect size for each language based on the results (e.g. using DD-scores as in Sprouse, Wagers, and Phillips (2012)), and then compare these effect sizes across the languages. This kind of quantitative description is of course not an analysis, but it could plausibly be the first step toward one and to a deeper understanding of what lies behind this type of cross-linguistic variation. Alexopoulou and Keller (2007), Sprouse et al. (2011), Sprouse et al. (2016), and Chacón (2015) are all examples of this type of work.

1.4.3 Comparing Populations

Acceptability experiments can also be extremely useful when comparing groups of speakers, rather than individuals. Traditional fieldwork and similar techniques can only detect extremely robust between-group differences, but the fine-grained measurements, large participant pools, and statistical analyses of experiments make it possible to detect much more subtle differences that might otherwise go unnoticed. One traditional area where this is helpful is in looking for linguistic differences across regions. If one wants to know, for instance, whether the *that*-trace effect is valid and similarly robust across a wide geographical area, acceptability experiments offer a straightforward way to probe this, as Cowart (1997) and Chacón (2015) have in fact done (see Chapters 7 and 10). Conducting large-scale experiments across regions poses interesting challenges, such as the need to ensure that the experimental procedure is the same for all participants and that the lexical items used in the stimuli are equally

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

comprehensible in all regions, but these challenges can be managed and the extra work is justified by the results, which typically cannot be obtained by other means. Guajardo and Goodall (2019) is an example of this type of work.

Another dimension along which groups of speakers are often compared concerns the speakers' acquisition history or general language background (see also Chapter 14). One might want to compare native speakers to non-native speakers, for instance, or those whose exposure to the language began at age 3 to those for whom it began at age 8. Here too, acceptability experiments allow the kind of fine-grained analysis in which subtle differences among such groups can be detected (or ruled out). Kim and Goodall (2016), for instance, perform a series of acceptability experiments testing island phenomena in Korean using two groups of participants: native speakers of Korean residing in Korea and heritage speakers of Korean who were born in the US (or moved to the US before age 7) and who reside in the US. The results showed an intricate pattern of behavior across different island types, but this pattern was remarkably similar across the two populations, despite the very different linguistic environments that the two groups were exposed to in childhood. Kim and Goodall use these results to argue that the island effects being examined are largely immune to environmental influences.

General cognitive measures have also been a common tool for comparing speakers in relation to their performance in acceptability experiments. In perhaps the most well-known example of this, Sprouse, Wagers, and Phillips (2012) conduct both an acceptability experiment and two measures of working memory capacity on a large group of participants. The acceptability experiment measures speakers' sensitivity to four types of island effects, but no correlation is found between speakers' island sensitivity and the measures of their working memory capacity. The authors use this lack of correlation to suggest that island effects are not ultimately due to limitations on processing capacity. Michel (2014) uses a wider battery of working memory measures but finds a similar lack of correlation with regard to acceptability. It would not be warranted to conclude that general cognitive measures are unrelated to acceptability, however, since Hofmeister, Staum Casasanto, and Sag (2014) find a clear correlation between reading span scores and acceptability, but only for sentences where low acceptability is attributable to processing costs (such as when arguments are relatively distant from their heads). There is little doubt that the relation between acceptability and other cognitive measures will remain a lively topic of research for many years (see Chapters 4, 9, and 24).

There are many other ways that one can distinguish among individuals or groups and that one can imagine might have an effect on their acceptability. Many traditional forms of categorization, for instance, such as gender or familial left-handedness (e.g. Bever et al. 1989), are known to

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

have neurological correlates, and it is conceivable that these could affect one or more of the components that go into acceptability judgments (as described in Section 1.3). More broadly, known neuroanatomical differences across individuals or genetic markers of various types could also be examined for correlations with individuals' behavior in acceptability experiments. I am not aware of studies of this type, but given that acceptability experiments are relatively easy and quick to conduct, yet also yield very fine-grained results, it is probably only a matter of time before such work begins to be done.

1.5 Conclusion

Constructing, running, and analyzing a sentence acceptability experiment requires a number of techniques and concepts that are not traditionally part of most syntacticians' training, yet the benefits of learning how to do so make it worth the effort in many cases. As we have seen here, formal experiments allow us to measure acceptability with a degree of precision and confidence that is not possible with other methods, but they also give us a type of data that would not be accessible to us otherwise. They allow us to measure contrasts across the full range of acceptability, for instance, including among sentences that are presumably fully grammatical, and they allow us to determine whether a given effect might be the additive result of smaller effects or might require positing something additional. They also allow us to make well-defined comparisons across languages and across speaker populations that differ by region, age of acquisition, gender, or other factors.

Sentence acceptability experiments are often compared to more traditional methods of collecting acceptability judgments, such as fieldwork or reliance on one's own judgments. Much discussion in the literature has focused on the question of which is better, with some arguing that traditional methods are sufficient for most syntactic research and that experiments are only needed in exceptional cases, and others arguing that traditional methods should be abandoned and that experiments should be used in almost all cases (Phillips 2009; Culicover & Jackendoff 2010; Gibson & Fedorenko 2010, 2013; Gibson, Piantadosi, & Fedorenko 2013). Given what we know now about acceptability experiments, however, and especially what we know about the factors that these experiments are sensitive to (see Section 1.3) and the ways that these experiments can be put to use (see Section 1.4), this debate in the literature seems ill conceived. Traditional methods of measuring acceptability have been extraordinarily productive, both in documenting facts about understudied languages and uncovering new facts about languages that have already been studied extensively. It is not clear that experimental methods can take over all of these functions, but it is clear, as we have seen, that experimental methods

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

can do many things that more traditional techniques cannot. Just as no one would argue that we need to choose between self-paced reading and ERP techniques for the study of syntax, similarly no one should argue that we need to choose between traditional methods and acceptability experiments. Each can do things that the other cannot and they both have important roles to play in syntactic research.

References

Alexopoulou, T. & Keller, F. (2007). Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language*, 83(1), 110–160.

Beltrama, A. & Xiang, M. (2016). Unacceptable but comprehensible: The facilitation effect of resumptive pronouns. *Glossa: A Journal of General Linguistics*, 1(1), 29. DOI:10.5334/gjgl.24

Bermel, N. & Knittl, L. (2012). Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory*, 8(2), 241–275.

Bever, T. G., Carrithers, C., Cowart, W., & Townsend, D. J. (1989). Language processing and familial handedness. In A. M. Galaburda, ed., *From Reading to Neurons*. Cambridge, MA: MIT Press, pp. 331–360.

Bolinger, D. (1978). Asking more than one thing at a time. In H. Hiz, ed., *Questions*. Dordrecht: D. Reidel.

Bresnan, J. & Ford, M. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*, 86(1), 168–213.

Carlsson, A. M. (1983). Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale. *Pain*, 16(1), 87–101.

Chacón, D. A. (2015). Comparative psychosyntax. Doctoral dissertation, University of Maryland.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1977). On *Wh*-movement. In P. Culicover, A. Akmajian, & T. Wasow, eds., *Formal Syntax*. New York: Academic Press, pp. 71–132.

Cole, P. & Hermon, G. (1998). The typology of *wh*-movement and *wh*-questions in Malay. *Syntax*, 1(3), 221–258.

Cole, P. & Hermon, G. (2000). Partial *Wh*-Movement: Evidence from Malay. In U. Lutz, G. Müller, & A. Von Stechow, eds., *Wh-scope Marking*. Amsterdam: John Benjamins, pp. 101–130.

Cowart, W. (1997). *Experimental Syntax*. New York: Sage.

Culicover, P. W. & Jackendoff, R. (2010). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences*, 14(6), 234–235.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

Davidson, R. J. (1992). Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition*, 20, 125–151.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, 122(3), 371.

Dworkin, R. H., Turk, D. C., Wyrwich, K. W., Beaton, D., Cleeland, C. S., Farrar, J. T., & Brandenburg, N. (2008). Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *Journal of Pain*, 9(2), 105–121.

Fanselow, G. (2017). Partial *wh*-movement. In M. Everaert & H. C. Van Riemsdijk, eds., *The Wiley Blackwell Companion to Syntax*, 2nd edn. New York: John Wiley & Sons.

Featherston, S. (2005). Universals and grammaticality: *Wh*-constraints in German and English. *Linguistics*, 43(4), 667–711.

Fedorenko, E. & Gibson, E. (2010). Adding a third *wh*-phrase does not increase the acceptability of object-initial multiple-*wh*-questions. *Syntax*, 13(3), 183–195.

Fiengo, R. & Higginbotham, J. (1981). Opacity in NP. *Linguistic Analysis*, 7, 347–373.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, 98(1), 39–82.

Forouzanfar, T., Weber, W. E., Kemler, M., & van Kleef, M. (2003). What is a meaningful pain reduction in patients with complex regional pain syndrome type 1? *Clinical Journal of Pain*, 19(5), 281–285.

Fukuda, S. (2012). Aspectual verbs as functional heads: evidence from Japanese aspectual verbs. *Natural Language & Linguistic Theory*, 30(4), 965–1026.

Fukuda, S. (2017). Split intransitivity in Japanese is syntactic: Evidence for the Unaccusative Hypothesis from sentence acceptability and truth value judgment experiments. *Glossa: A Journal of General Linguistics*, 2(1), 28. DOI:10.5334/gjgl.268

Gallego, Á. (2009). On freezing effects. *Iberia: An International Journal of Theoretical Linguistics*, 1(1), 33–51.

Garrod, S. (2006). Psycholinguistic research methods. In K. Brown, ed., *Encyclopedia of Language and Linguistics*. Amsterdam: Elsevier, pp. 251–257.

Gibson, E. & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, 14(6), 233.

Gibson, E. & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1–2), 88–124.

Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes*, 28(3), 229–240.

Goodall, G. (2015). The D-linking effect on extraction from islands and non-islands. *Frontiers in Psychology: Language Sciences*, 5, 1493. DOI: 10.3389/fpsyg.2014.01493

Goodall, G. (2017). Referentiality and resumption in *wh*-dependencies. In J. Ostrove, R. Kramer, & J. Sabbagh, eds., *Asking the Right Questions: Essays in Honor of Sandra Chun*, pp. 65–80. eScholarship, University of California. http://escholarship.org/uc/item/8255v8sc

Grewendorf, G. (1988). *Aspekte der deutschen Syntax*. Tübingen: Narr.

Guajardo, G. & Goodall, G. (2019). On the status of Concordantia Temporum in Spanish: An experimental approach. *Glossa: A Journal of General Linguistics*, 4(1), 116. DOI:10.5334/gjgl.749

Harvey, B. M., Klein, B. P., Petridou, N., & Dumoulin, S. O. (2013) Topographic representation of numerosity in the human parietal cortex. *Science*, 341(6150), 1123–1126.

Hofmeister, P., Culicover, P., & Winkler, S. (2015). Effects of processing on the acceptability of "frozen" extraposed constituents. *Syntax*, 18(4), 464–483.

Hofmeister, P., Staum Casasanto, L., & Sag, I. A. (2014). Processing effects in linguistic judgment data: (Super-)additivity and reading span scores. *Language and Cognition*, 6(1), 111–145.

Householder, F. W. (1973). On arguments from asterisks. *Foundations of Language*, 10(3), 365–376.

Kayne, R. (1983). Connectedness. *Linguistic Inquiry*, 14, 223–249.

Kayne, R. S. (1984). *Connectedness and Binary Branching*. Dordrecht: Foris.

Keffala, B. (2011). Resumption and gaps in English relative clauses: Relative acceptability creates an illusion of "saving." In C. Cathcart et al., eds., *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*. Berkeley, CA: University of California, Berkeley Linguistics Society, pp. 140–154.

Keffala, B. & Goodall, G. (2011). Do resumptive pronouns ever rescue illicit gaps in English? Poster presented at the 24th Annual CUNY Conference on Human Sentence Processing, Stanford, California.

Keffala, B. & Goodall, G. (2013). On processing difficulty and the acceptability of resumptive pronouns. Paper presented at Linguistic Evidence – Berlin Special, Humboldt-Universität, Berlin.

Kim, B. & Goodall, G. (2016). Islands and non-islands in native and heritage Korean. *Frontiers in Psychology: Language Sciences*, 7. DOI:10.3389/fpsyg.2016.00134

Kim, D. & Goodall, G. (2018). Complexity effects in A- and A'-dependencies. Poster presented at 31st CUNY Conference on Human Sentence Processing, UC Davis.

Langsford, S., Perfors, A., Hendrickson, A. T., Kennedy, L. A., & Navarro, D. J. (2018). Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: A Journal of General Linguistics*, 3(1), 37. DOI: 10.5334/gjgl.396

Lasnik, H. & Saito, M. (1992). *Move Alpha: Conditions on Its Application and Output*. Cambridge, MA: MIT Press.

Lewis, S. & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, 44(1), 27–46.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

Mahowald, K., Graff, P., Hartman, J., & Gibson, E. (2016). SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language*, 92(3), 619–635.

Michel, D. (2014). Individual cognitive measures and working memory accounts of syntactic island phenomena. Doctoral dissertation, University of Calfiornia, San Diego.

Müller, G. (2010). On deriving CED effects from the PIC. *Linguistic Inquiry*, 41 (1), 35–82.

Myers, J. (2009). Syntactic judgment experiments. *Language & Linguistics Compass*, 3(1), 406–423.

Myers, J. (2017). Acceptability judgments. In M. Aronoff, ed., *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press. DOI: 10.1093/acrefore/9780199384655.013.333

Namboodiripad, S. (2017). An experimental approach to variation and variability in constituent order. Doctoral dissertation, University of Calfiornia, San Diego.

Natale, M., Gur, R. E., & Gur, R. C. (1983). Hemispheric asymmetries in processing emotional expressions. *Neuropsychologia*, 19, 609–613.

Omaki, A., Fukuda, S., Nakao, C., & Polinsky, M. (2020). Subextraction in Japanese and subject–object symmetry. *Natural Language & Linguistic Theory*, 38, 627–669.

Perlmutter, David M. (1971). *Deep and Surface Structure Constraints in Syntax*. New York: Holt, Rinehart, and Winston.

Pesetsky, D. (2017). Complementizer-trace effects. In M. Everaert & H. C. Van Riemsdijk, eds., *The Wiley Blackwell Companion to Syntax*. New York: John Wiley & Sons.

Phillips, C. (2009). Should we impeach armchair linguists? *Japanese/Korean Linguistics*, 17, 49–64.

Phillips, C. & Wagers, M. (2007). Relating structure and time in linguistics and psycholinguistics. In P. Levelt & A. Caramazza, eds., *The Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press, pp. 739–756.

Polinsky, M., Clemens, L. E., Morgan, A. M., Xiang, M., & Heestand, D. (2013). Resumption in English. In J. Sprouse & N. Hornstein, eds., *Experimental Syntax and Island Effects*. Cambridge: Cambridge University Press.

Ritchart, A., Goodall, G., & Garellek, M. (2016). Prosody and the *that*-trace effect: An experimental study. In K. Kim et al., eds., *Proceedings of the 33rd West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project, pp. 320–328.

Saddy, D. (1991). *Wh*-scope mechanisms in Bahasa Indonesia. *MIT Working Papers in Linguistics*, 15, 183–218.

Sag, I., Hofmeister, P., & Snider, N. (2007). Processing complexity in subjacency violations: the complex noun phrase constraint. In *Proceedings of the 43rd Annual Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, pp. 215–229.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press

Schütze, C. T. (2016). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Berlin: Language Science Press.

Schütze, C. & J. Sprouse (2014). Judgment data. In R. Podesva & D. Sharma, eds., *Research Methods in Linguistics*. Cambridge: Cambridge University Press, pp. 27–51.

Sedarous, Y. & Namboodiripad, S. (2020). Using audio stimuli in acceptability judgment experiments. *Language and Linguistics Compass*, 14:e12377. DOI: 10.1111/lnc3.12377

Sprouse, J. & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, 48(3), 609–652.

Sprouse, J. & Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: A Journal of General Linguistics*, 2(1). DOI:10.5334/gjgl.236

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, 134, 219–248.

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language*, 88(1), 82–123.

Sprouse, J., Caponigro, I., Greco, C., & Cecchetto, C. (2016). Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory*, 34(1), 307–344.

Sprouse, J., Fukuda, S., Ono, H., & Kluender, R. (2011). Reverse island effects and the backward search for a licensor in multiple *wh*-questions. *Syntax*, 14(2), 179–203.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103 (2684), 677–680.

Torrego, E. (1985). On empty categories in nominals. Unpublished manuscript, University of Massachusetts, Boston.

Zorzi, M., Priftis, K., & Umiltà, C. (2002). Brain damage: Neglect disrupts the mental number line. *Nature*, 417(6885), 138–139.

https://doi.org/10.1017/9781108569620.002 Published online by Cambridge University Press