Title

Approaching gradience in acceptability with the tools of signal detection theory

Permalink

Authors

Dillon, Brian
Wagers, Matthew

Publication Date

2019-11-19

DOI

10.31219/osf.io/apxru

Copyright Information

Dillon, B., & M. Wagers. Approaching gradience in acceptability with the tools of signal detection theory.
Preprint dated Nov 15, 2019. To appear in G. Goodall, Ed., Cambridge Handbook of Experimental Syntax.

Approaching gradience in acceptability with the tools of signal detection theory

Brian Dillon (University of Massachusetts, Amherst)
Matthew W. Wagers (University of California, Santa Cruz)

Intuitive judgments of sentence acceptability form the empirical basis of experimental syntax, and an important component of many psycholinguistic investigations (Cowart, 1997; Schutze, 1996). For this reason, one central methodological concern for experimental syntacticians is how best to collect and analyze acceptability judgment data. Experimental syntacticians continue to extend and refine the tools used to measure sentence acceptability. In this chapter we seek to contribute to this methodological expansion. We discuss the difficulties inherent in getting a quantitatively precise measurement of sentence acceptability. We suggest that the tools of signal detection theory can be applied to common acceptability judgment data. This analytical approach offers both an explicit theory of how speakers give acceptability judgments in the context of a rating task, and yields more precise measurements of sentence acceptability. The approach we outline builds on the work of previous researchers advocating similar approaches to acceptability data (e.g. Bader & Häussler, 2010; Mauner, 1995), as well as research on the magnitude estimation method for measuring acceptability judgments (Bard, Robertson & Sorace, 1996). The central goal of our chapter is to make the tools of Signal Detection Theory accessible to experimental syntacticians; to this end, we present a specimen experiment and a worked, tutorial-style analysis of acceptability judgment data using Signal Detection Theory.

1 Measuring acceptability: A brief overview

There are a variety of methods for measuring sentence acceptability in experimental contexts. These methods differ both in the response options offered to participants and the task presented to participants. For measuring the acceptability of a single sentence in isolation, binary yes-no acceptability judgments, *n*-point Likert scales, and continuous ‘thermometer’ ratings (Featherston, 2008) are all widely used. Other techniques invite participants to compare one sentence against another. For example, in two alternative forced-choice tasks (2AFC), the participant chooses the more acceptable of two sentences presented. In magnitude estimation (ME) tasks, the participant is asked to rate a target sentence relative to a baseline (or *modulus*) sentence (Bard et al., 1996). For a detailed explanation of each type of experiment, we refer the reader to Schütze & Sprouse (2014).

For all of these techniques, we might ask: how well can each acceptability measure recover true differences in acceptability between sentence tokens, or between classes of sentences? Head-to-head comparisons of the various acceptability measurements suggest that the different methods yield largely similar qualitative results for many sentence contrasts. In other words, if sentences of Type A are more acceptable than sentences of Type B as measured with one of the techniques above, other methods will generally recover this (ordinal) difference with high reliability (Bader & Häussler, 2010; Sprouse & Almeida, 2011, 2017; Weskott & Fanselow, 2011). Commonly used methods in experimental syntax have a number of desirable properties: they generally have high test-retest reliability (Langsford, Perfors, Hendrickson, Kennedy & Navarro, 2018), and in many cases yield results that are consistent with those achieved with informal methods of acceptability judgment collection (Sprouse & Almeida, 2017). In short, if the experimental syntactician is interested in establishing that there exists a simple contrast between two classes of sentences, she has a number of reliable, powerful tools available at her disposal. It is fairly straightforward for a researcher to answer the binary question *is sentence type A better than, or worse than, sentence type B?*

However, it is less straightforward to answer the more gradient sister question: *to what extent is sentence type A better than, or worse than, sentence type B?* This is because this question implies that the researcher has a reliable quantitative measure of the acceptability differences between two classes of sentence, such that she can offer a meaningful, quantitatively precise answer to this gradient question. It is widely acknowledged that this question is difficult to answer satisfactorily with existing techniques. To take one example: the quantitative differences between two sentence types on a Likert scale (or even in a binary yes/no rating context) do not have any inherent meaning; they are filtered through an individual participant's interpretation of the response scale. This renders claims based on the absolute magnitude of a difference on a Likert scale suspect, for reasons that we make precise below.

For researchers interested in the gradient acceptability question--*to what extent is sentence type A better than, or worse than, sentence type B?*--magnitude estimation has occupied a special place among other experimental techniques for measuring acceptability. In their 1996 paper, Ellen Bard and colleagues argued that linguistic acceptability should be understood in terms similar to other psychophysical judgments. That is, acceptability constitutes psychological evidence in the same way that judgments of luminosity or loudness constitute psychological evidence that can be measured and modeled. The judgment of acceptability reflects a hidden, or latent, cognitive variable that can be reliably measured, one which is the truer guide in explaining the sources of gradience. From this perspective, acceptability judgments could be measured using some of the same tools that psychophysicists had developed to quantify gradient psychological evidence in other domains. Magnitude estimation, one technique for doing this, was originally proposed by Stevens (1956) as a means of providing measurements of psychological evidence on a <u>ratio</u> scale; that is, a measurement scale with equal-sized units of measurement on which quantitative differences can be defined and directly interpreted, and for which there exists a true zero point. An example of a ratio-scale measurement is height, which has a clear zero point, licensing ratio comparisons among measurements (e.g. *Aunt Mary is twice as tall as Timmy)*. A related type of measurement is that of an *interval* scale measurement. Interval scale measurements differ from ratio scale measurements in not having a fixed zero point. An example is temperature measured in degrees Fahrenheit, which has no interpretable zero point, although it does offer equal-sized units of measurement.

The promise of a true ratio (or interval) measurement of acceptability set magnitude estimation apart from other techniques for measuring acceptability, which generally offered either *ordinal* measurements (Likert scales, yes-no ratings) or *categorical* measurements

(2AFC). Ordinal measurements allow researchers to establish a rank ordering among sentence classes, but they do not allow more precise measurements of the quantitative differences between sentences. This is because the differences between points on a Likert rating scale need not be equal-sized units of 'acceptability': there is no guarantee that the difference in 2-3 on the Likert scale is the same as the difference between 4-5 on the Likert scale. Thus, Likert ratings (and yes-no judgments) allow researchers to establish whether sentence type A is more acceptable than sentence type B with high fidelity, but they do not tell us by how much. Magnitude estimation seemed to correct that, and offer a true ratio scale judgment of acceptability. And since interval scale or ratio scale measurements of acceptability are required to provide a satisfactory answer to the gradient question, this methodological advance allowed the development and testing of theories that make fine-grained, quantitative predictions about acceptability (see, e.g., Aarts, 2007; Keller, 2000; Lau, Clark & Lappin, 2017; Sorace & Keller, 2005).

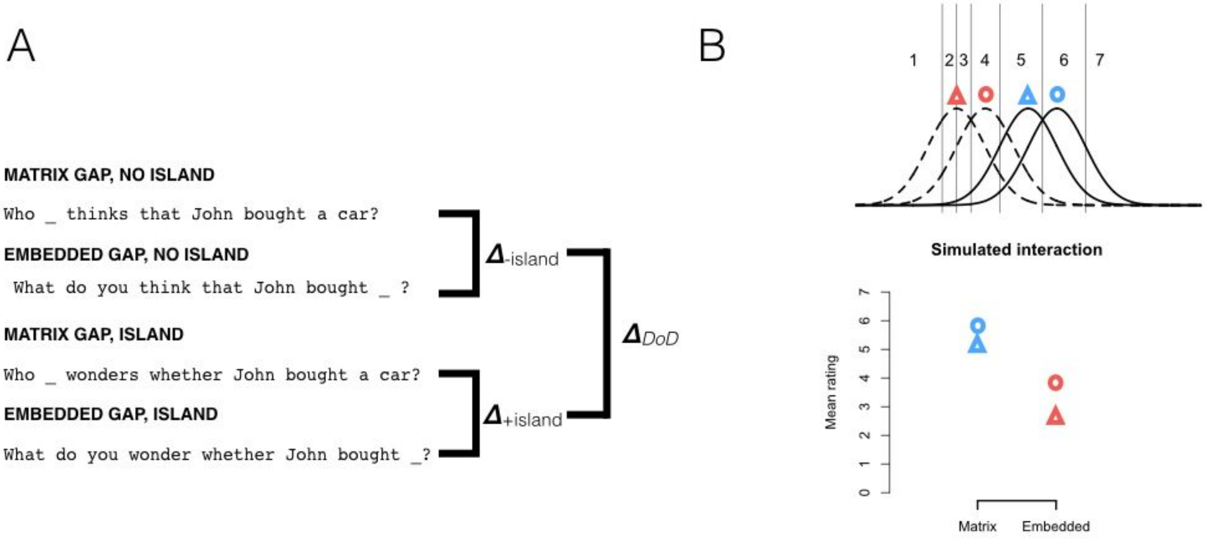
Unfortunately, it has turned out to be less straightforward to apply ME to acceptability judgments than to other psychophysical judgments of stimulus quality. Unlike judgments of brightness or loudness, say, there is no objective physical stimulus against which acceptability judgments can be compared. The 'physical axis' for brightness/loudness stimuli is important, because it allows researchers to validate the ME measurements by comparing the consistency of participant judgments against objective physical measurements of the stimulus. Since there is no physical axis for acceptability judgments, this is not possible. In its place, Bard and colleagues demonstrated that ME judgments of acceptability showed good cross-modal consistency: highly similar results were found when subjects estimated acceptability using a numerical value and when they estimated acceptability using line length. This cross-modal consistency is expected if there is a stable underlying acceptability percept that is 'read out' in the various measurement contexts; the missing physical axis in ME judgments is implied by the cross-modal consistency (Stevens, 1960; Bard et al., 1996). Despite this promising early result, several subsequent studies have raised questions about the utility of magnitude estimation as applied to acceptability judgments. Importantly, Sprouse (2011) provided evidence that participants may be unable to make true ratio judgments for acceptability (more precisely, he argued that the *commutativity assumption* of magnitude estimation does not hold). That is, participants are unable to reliably evaluate multiplicative statements like *Sentence A is twice as acceptable as Sentence B*; however, that raters can make this sort of ratio judgment is a core assumption of ME. These types of ratio judgments rely on there being a true 'zero point' on the scale being measured. It is thus unclear that there is any such a zero point for acceptability judgments (Sprouse, 2011). If Sprouse's argument is correct, then the cognitive assumptions concerning magnitude estimation do not hold, and the claim that acceptability measured in this way offers a true ratio measure of acceptability is rendered suspect. Furthermore, Langsford et al. (2018) showed that magnitude estimation has among the lowest test-retest reliability of the major experimental techniques for measuring acceptability reviewed here, although the absolute values of test-retest reliability were overall still quite high. Lastly, in a series of acceptability judgment experiments, Weskott and Fanselow (2011) showed that magnitude estimation is not more informative than Likert or binary ratings in that the proportion of variance accounted for by experimental manipulations (a measure of effect size) was not significantly different between magnitude estimation and the other tasks. In light of these results, it cannot be claimed that magnitude estimation provides a true, ratio-scale measurement of acceptability, and on a more practical level, it cannot be said that it more reliably recovers differences in acceptability between classes of sentences than other techniques of measuring acceptability (if anything, it seems to perform slightly less well than other widely used techniques: Langsford et al., 2018; Sprouse & Almeida, 2017).

2 We really do want to address the gradient question

One upshot of the preceding review is that if a researcher is interested only in establishing whether sentence class *A* is reliably more acceptable than sentence class *B*, then ordinal measurements of acceptability such as Likert scales or binary judgment measures serve the task very well. Similarly, if a researcher wishes to establish an acceptability ranking across a range of sentence classes, these same ordinal or categorical measures will suffice, and are quite easily deployed. In light of this, does experimental syntax really need to develop tools to answer the gradient question? Is there any value in getting a truly ratio, or interval measure of acceptability? We believe there is value in this, for two reasons.

2.1 Reason 1: Disconnects between statistical hypotheses and substantive hypotheses

The first reason that interval scales of measurement are of broad importance is that even if a researcher's substantive hypotheses do not critically depend on obtaining a precise interval measurement of acceptability, this same researcher's ***statistical*** hypotheses almost always do. This disconnect between what is measured and what a statistical test assumes can lead to spurious conclusions. One very common situation where this arises is testing interaction effects in crossed factorial designs. For example, Sprouse, Wagers, and Phillips (2012) constructed a 2x2 factorial design to measure the presence of island effects. This experimental design crossed the site of extract for a *wh*-dependency (a matrix gap or embedded clause gap) with the type of embedded clause (an island environment or not). They reasoned that if the penalty for embedded *wh*-movement was greater when the embedded clause was an island than when it was not, then this would constitute evidence for some additional penalty levied on extractions from islands. In other words, they expected to see a superadditive interaction of gap position and embedded clause type. Statistically, this superadditivity is realized as an interaction in an ANOVA or similar. This statistical interaction corresponds to a *difference-of-differences* score (see Fig 1A.). If the difference of differences is not zero, then the experimental factors in the 2x2 design interact, such that the effect of one factor depends on the level of the other. This use of this additive factors logic over Likert scale ratings is ubiquitous (indeed, we have routinely used this in our own work).


MATRIX GAP, NO ISLAND

Who _ thinks that John bought a car?

EMBEDDED GAP, NO ISLAND

What do you think that John bought _ ?

MATRIX GAP, ISLAND

Who _ wonders whether John bought a car?

EMBEDDED GAP, ISLAND

What do you wonder whether John bought _?

Figure 1: ***A*****:** Sprouse et al.’s factorial design. In their design, they measured the difference in acceptability between the matrix and embedded gap conditions for non-islands ($\boldsymbol{\Delta}_{\text{-island}}$), and the same acceptability difference for islands ($\boldsymbol{\Delta}_{\text{+island}}$). The difference of these two differences ($\boldsymbol{\Delta}_{\text{DoD}}$) measures the extent to which the difference is acceptability is greater for the island environments than non-island environments. ***B*****:** A hypothetical illustration of how scale bias could create a spurious interaction in the difference of differences. The top figure shows four hypothetical distributions of acceptability for the four conditions in Panel A. These distributions are defined so that the difference of differences among them is zero. The vertical lines mark the demarcation of response boundaries on a Likert scale; on the lower end of the scale, there is compression of the boundaries between rating categories. Because of this compression, averaged Likert ratings for these four conditions present a spurious superadditive interaction that is due to properties of the response scale, rather than the underlying acceptability distributions.

However, ANOVA-style interactions with Likert ratings -- even *z*-transformed ratings -- are potentially problematic. The reason is straightforward: the difference-of-differences logic, and the statistical tests used to evaluate the difference-of-differences score, critically relies on the assumption of an interval scale. That is, these tests assume that we can meaningfully compare the relative *magnitude* of the difference between pairs of conditions that sit at different points on the response scale. This assumption is not met in ordinal data, such as common Likert rating data or yes/no rating data. This opens up the possibility of misleading patterns in the data: differences in how a response scale is interpreted at the high and low end can create spurious interactions (for extensive discussion, see Loftus, 1978; Heit & Rotello, 2014; Rotello, Heit & Dubé, 2015). We suspect that most researchers using these techniques are implicitly aware of this possibility: we believe it to be widely recognized that so-called *floor* or *ceiling* effects can create spurious interactive patterns in bounded response scales (see also Liddell & Kurschke, 2018). The point we raise here is that similarly spurious interactions may obtain, for essentially similar reasons, even in the absence of obvious floor or ceiling effects.

For example, consider Figure 1, Panel B. In this hypothetical example, there is compression in the response categories at the lower end of the response scale: the distance between the edges of the 2 category is smaller on the underlying acceptability dimension that it is for the 5 category, for example. This corresponds to an experimental context where a participant is overall less willing to assign lower values on the response scale. In Figure 1B, however, the underlying pattern of acceptability is one that is entirely additive between the four hypothetical conditions: the difference of differences in the means of the four Gaussian distributions represented is zero. However, this underlying additive pattern interacts with the differences in the width of the category boundaries in the response task. As a result, we see an illusory interactive pattern in the ratings that mirrors a superadditive effect. Most importantly, if similar compression occurs across participants, then common methods of removing scale bias such as *z*-scoring will not remove it.

This simple example is intended to illustrate that in principle*,* properties of the response scale can interact with perfectly additive underlying patterns, and yield spurious interactive patterns when the resulting rating data are analyzed as if they were simple, interval measures of acceptability. This is only a hypothetical demonstration, however: It is not currently known whether this observation puts actual claims in the experimental syntax literature in peril[1], and we do not wish to imply that any single interactive pattern is in fact an artifact of scale properties. However, it is a potential danger, and failure to recognize this has mislead researchers in related fields, for decades in some cases (for example, Rotello, Heit and Dubé (2015) show that the widely-studied belief bias effect is indeed an artifact of response biases that are similar in nature to the scale compression effects we are concerned with here). It suggests that at a minimum, researchers should adopt statistical practices for analyzing interactions in judgment experiments that do not implicitly assume an interval response scale.

2.2 Reason 2: We can get more out of acceptability judgment measures

The second reason to seek an interval scale for measuring acceptability judgments is that establishing such a measurement for acceptability is a critical first step for asking more fine-grained questions about gradient effects on acceptability judgments. This arises when quantifying the relative impact various constraints have on acceptability cross-linguistically (Alexopoulou & Keller, 2007; Almeida, 2014; Häussler, Grant, Fanselow & Frazier, 2015; Kush, Lohndal & Sprouse, 2018; Sprouse, Caponigro, Greco & Cecchetto, 2016), as well as attempts to directly model the gradience in acceptability judgments (Lau, Clark & Lappin, 2017; Sprouse, Yankama, Indurkhya, Fong & Berwick, 2018; Warstadt, Singh & Bowman, 2018).

To take one example of a research question that critically turns on having an interval measurement of acceptability: Dillon, Staub, Levy and Clifton (2017) were interested in measuring the acceptability of sentences like (2):

(2) *Which flowers is the gardener planting?*

[1] In fact, we have reason to believe that the example chosen here is *not* a spurious interactive pattern: ROC Analysis of the original data in Sprouse, Wagers & Phillips (2012) shows that the interactions reported in that paper are not due to scale compression.

(2) is an object *wh*-question; the subject of the sentence is *the gardener*. Despite this, the sentence appears to be ill-formed. Dillon and colleagues proposed the plural *wh*-phrase *which flowers* interferes with the processing of the agreement on the auxiliary that immediately follows (one particular example of an agreement attraction-like effect: for recent summaries of this large research area, please consult Bock & Middleton, 2011 and Franck, 2011). One question that Dillon and colleagues asked of examples like (2) was whether the illusion of ungrammaticality was complete: when raters judged sentences like (2) to be unacceptable, were they treating it as a fully ungrammatical sentence? In other words: is the illusion of unacceptability in (2) more akin to a bistable perception of ungrammatical / grammatical agreement (a linguistic version of the Necker cube, perhaps), or do sentences like (2) simply occupy an intermediate level of acceptability, perhaps because of interference or difficulty in processing the complex agreement relationships in these examples?

A visual inspection of their data suggested a bimodal distribution of ratings for examples like (2). This in turn suggests that the answer to this question was that the perception of these sentences was essentially bimodal, with participants variably treating it as a fully acceptable or fully unacceptable sentence. Dillon and colleagues supported this observation using a formal model of the acceptability judgments given to sentences like (2). In brief, they found that the distribution of judgments in examples like (2) is well modeled by a simple mixture model according to which participants treat (2) as if it were fully ungrammatical on approximately 30% of trials, and on the remaining trials as if it were fully grammatical. A comparable model that allowed for a more gradient level of acceptability fit the data poorly.

However, Dillon et al (2017) noted that a weakness of their approach was that it assumed that the ratings on the Likert scale constituted interval measurements. If this assumption is not met, then their conclusions may not hold. For example, spurious bimodality could arise in a distribution of ratings over Likert scales if participants were for some reason unwilling or resistant to offer responses in the middle of a Likert scale (i.e. if there were scale compression in the middle of the scale). This kind of non-linearity in the mapping between underlying perception of acceptability and the response scale could create an illusion of bimodality in the responses, which in this hypothetical example would simply reflect a bias towards offering extreme responses on the response scale. If true, this would imperil these authors' conclusion about this particular linguistic illusion (we return to this effect later on in the chapter) .

2.3 The problem, The solution

The reason that Likert and binary ratings do not yield interval scale measurements essentially boils down to what we will call *response bias*: what internal criterion do participants set on their internal perception of acceptability to render a 'yes' judgment in a binary task? What internal criteria does a participant set to render a judgment of a 2 or a 3 on a Likert scale? Or to use a term more commonly used in the experimental syntax literature: what is their *scale bias*? Participants are free to interpret the rating task they are presented with as they will. This means that the threshold for labeling a sentence 'acceptable' in a binary task may well vary from one rater to another, even if they have identical underlying perceptions of a stimulus' acceptability. Similarly, what makes a sentence a '5' on a 7-point Likert rating task is likely to vary from participant to participant in ways that are idiosyncratic and not of central interest to the experimental syntactician.

In the context of a Likert rating experiment, it is common to take a *z*-score transform of each participants' ratings to address this kind of scale bias. That is, each participant's mean rating is subtracted from each response, and divided by that participant's standard deviation. This normalizes ratings across participants, but it cannot be guaranteed to result in a truly interval measure.

However, an alternative approach is to attempt to directly model the underlying acceptability values and the response thresholds. From this perspective, we take seriously the claim that sentence acceptability derives from a latent, unobserved cognitive variable that is read out or reflected in different acceptability measures; it is this underlying variable that we wish to recover and measure. As a starting point, we might take this to be a scalar (unidimensional) real value. This underlying value is mapped onto response categories in the context of an experiment when participants set an internal criterion or criteria and compare their perception of acceptability for a given sentence token to their internally determined criteria. This sort of *latent variable model* seeks to recover the underlying acceptability values by jointly modeling the underlying latent acceptability variable and the response criteria (e.g. the response bias). In one study, Langsford et al (2018) applied one widely studied latent variable model to binary judgment data: the Thurstone model. Applied to acceptability judgments, a Thurstone model seeks to recover the latent acceptability structure in a class of sentence comparisons through a series of pairwise comparisons among sentences (in this sense, it is formally similar to the ELO system for establishing chess rankings, which has also been applied to acceptability judgment data; Sprouse et al., 2018).

The Thurstone model is one example of a latent variable model for modeling judgment measures; other statistical approaches such as ordinal regression models may be interpreted in a similar fashion (Liddell & Kruschke, 2018). The great strength of the latent variable approaches is that they offer an explicit theory of how numbers get placed on the response scale in the context of a judgment experiment, and in doing so, offer a framework for inferring the underlying acceptability values from readily available judgment measures such as Likert scales or binary judgment acceptability decisions. In the following section, we discuss how to apply one relatively well-understood type of latent variable approach, Signal Detection Theory, to acceptability judgment data.

2.4 Signal Detection Theory and acceptability judgments

One broad, widely-used framework for modeling decision-making processes under uncertainty is Signal Detection Theory (SDT; Macmillan & Creelman, 2005). This framework is commonly applied to an observer's judgments of whether some stimulus is present or absent. One very common example of this is a recognition memory task, where participants are asked to decide if a stimulus is one that they previously studied (*stimulus present*) or not (*stimulus absent*). The standard SDT analytical approach categorizes response behavior in an experiment like this into *hits* (e.g. *stimulus present* responses when it is the correct response), and *false alarms* (e.g. *stimulus present* responses when it is not). SDT describes how the distribution of hits and false alarms in a detection task can be used to recover different aspects of the stimulus detection process, such as how clearly can the observer discriminate signal from noise (e.g. their *sensitivity*) and what is the threshold they use in rendering their judgments (e.g. their *bias*). Although it is most commonly applied in detection tasks, the theoretical model of the decision process implied by SDT is very broad. According to SDT, the decision process is seen as a mapping from a noisy, continuous cognitive signal onto one or more discrete response options offered to a participant in an experimental setting.

This perspective on the decision-making process has value for the experimental syntactician, because a similar decision process is plausibly at work when speakers are asked to categorize linguistic stimuli into discrete categories in a judgment task. This hypothesized decision process for both binary rating tasks and *n*-point Likert tasks is rendered graphically in Figure 2. The strength of SDT is that it allows independent estimation of the underlying cognitive signal (in our case, the latent *Acceptability* values), as well as the likelihood that a participant will respond with one response category over another (that is, response bias or scale bias).

There have been several previous attempts to apply Signal Detection Theory to acceptability judgment tasks. Mauner (1995) was one early, important application of these tools. Mauner argued that SDT was critical for analyzing grammatical judgment data given by agrammatic aphasic patients. To our knowledge, the underlying SDT theory for acceptability judgments was most explicitly developed by Bader and Häussler (2010), who proposed a model of a binary acceptability judgment task grounded in SDT. On Bader and Häussler's analysis, the process of rendering an acceptability judgment could be logically decomposed into two distinct processes: the first is the computation of a continuous acceptability value for a given sentence, and the second is a mapping from that continuous acceptability value on one of two response options in a binary rating task. Their model explained how continuous acceptability can be converted to the probability of a response in one of two categories, which in turn gave insight into why Bader and Häussler found such a tight correlation between continuous (ME) and discrete (binary Y/N judgments) measurements of acceptability in their experiments (see also Weskott & Fanselow, 2011, for a similar result). Moreover, Bader and Häussler showed that their simple two-stage decision model was able to achieve a close fit to their experimental data. That is, in their data, the proportion of 'acceptable' responses was well modeled by a continuous distribution of acceptability and a single decision criterion that mapped that continuous value of acceptability into a binary Y/N decision. In short, SDT offers a theory of how categorical responses arise in acceptability judgment experiments while being based on a fundamentally continuous, noisy signal of acceptability.

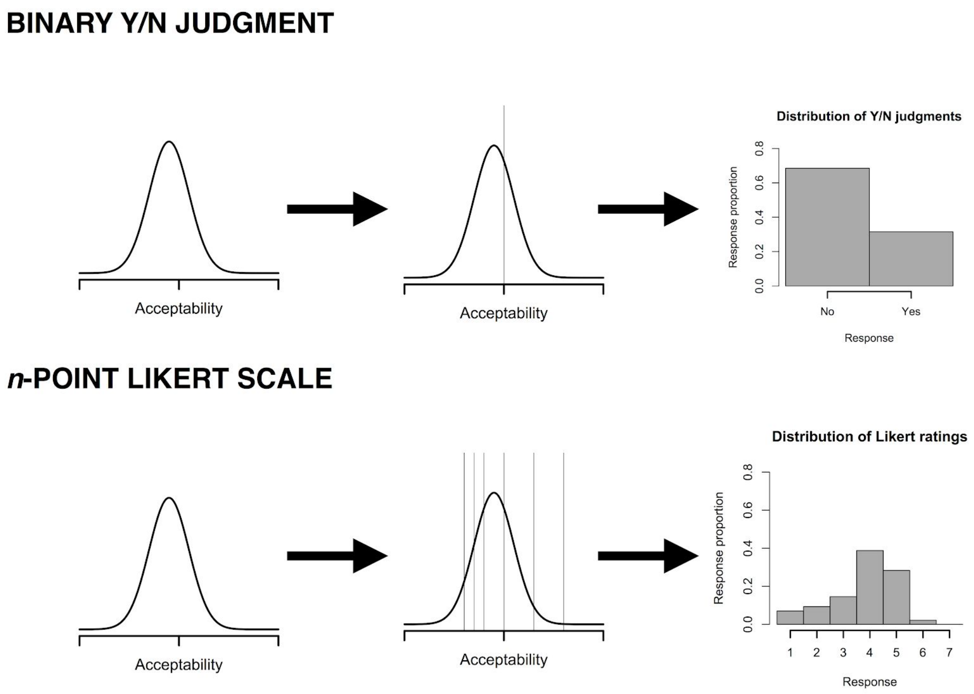


Figure 2: The process of rendering a binary or *n*-point Likert acceptability judgment from the perspective of Signal Detection Theory. The first step involves computing or determining the acceptability of a given sentence token. Decision criteria (vertical lines) are overlaid on this distribution. The second step involves mapping that value to one of the presented response options using these decision criteria to determine what response option is appropriate for a given token.

Dillon, Andrews, Rotello and Wagers (2019) further developed Bader and Häussler's Signal Detection Theoretic model of the acceptability judgment task. In their experiment, they measured speeded binary acceptability judgments, followed by a three point confidence rating task. In their analysis, they combined the rapid acceptability judgment and the graded confidence ratings into a six point scale, ranging from *very confident unacceptable* to *very confident acceptable.* On their analysis, the decision process that yields the resulting six point rating scale is fundamentally identical to Bader and Häussler's model: first, the acceptability of a sentence is computed, and second, that continuous acceptability value is mapped to one of a number of response criteria. The sole difference between the two models is the number of response thresholds necessary to model the data: a binary decision can be generalized to any arbitrary *n*-point rating scale by positing *n-1* decision criteria, which in turn partition the continuous acceptability signal into the *n* distinct, ordered response categories at the decision stage of the model.

This model of a scaled rating task implies important new routes of analysis of acceptability data, which we will discuss in detail below. In broad strokes, these analytical approaches seek to deconfound *sensitivity* from *bias.* In the present context, *sensitivity* refers to the distance between two sentence types in the underlying perceptual space; the term *sensitivity* is borrowed from classical Signal Detection Theory, where it referred to an individual's ability to discriminate signal from noise (i.e. their *sensitivity* to whether a stimulus is present). Applied to acceptability judgments, the Signal Detection Theoretic notion of sensitivity is the linguist's notion of *contrast* between a pair of sentences or sentence classes: SDT-theoretic measures of sensitivity quantify the degree of contrast between a minimal pair of sentence types, and so yields an answer to the gradient acceptability question *to what extent is sentence type A better than, or worse than, sentence type B?* Perhaps more precisely put, the analysis technique we pursue here allows the researcher to answer the question of *to what extent does sentence type A contrast with sentence type B?*

One familiar technique for computing independent measures of sensitivity and bias is simple *d'* analysis. *d'* measures the distance between the mean value of two distributions in standard deviation units. Strictly speaking*, d'* is a measure of discriminability between two classes of stimuli in the decision space that supports judgments about the stimuli. Sensitivity in the context of acceptability judgment data is most appropriately understood as the distance between the two classes of sentence in the decision space that supports the acceptability judgment. At this early stage, we remain unclear what precisely the underlying decision space that supports acceptability judgments is. On the plausible hypothesis that response behavior in an acceptability judgment task arises by mapping a unidimensional psychological value of *Acceptability* onto one of the proferred response options (Bader & Häussler, 2010), then the discriminability measure *d'* can be considered a distance measure between the location of the acceptability distributions for two classes of stimuli. While on certain assumptions, the discrimination measures can be taken to be measures of acceptability, they should not be assume to directly reflect a measurement of *grammaticality* absent a linking hypothesis about how acceptability measures relate to grammaticality. Furthermore, this interpretation of *d'* only holds if the research is willing to adopt certain assumptions about the shape of the underlying distributions of acceptability judgments. For example, the *d'* measure of discriminability assumes that the underlying distributions of acceptability are normal distributions, and that the two distributions compared have equal variance (although as we will see below, the equal variance assumption is not always justified).

However, there are other techniques for identifying independent measures of sensitivity and bias that make fewer theoretical assumptions about the data. It is possible to relax the assumption of equal variance; in this case, an appropriate measure of discriminability is d_a, a measurement of the distance between the means of the two distributions expressed in units of their root mean squared standard deviation. Other techniques can be readily applied to common experimental data. In an experimental context where there are multiple, distinct response criteria (such as a Likert scale task, or a dual task that jointly measures acceptability and confidence), it is possible to construct an empirical *receiver operating characteristic* (ROC). An ROC curve can yield a measure of sensitivity as well. In our context, this sensitivity indexd may constructed between two conditions in an acceptability judgment experiment. For a Likert scale, the empirical ROC is constructed by calculating the observed proportion of responses at the highest possible rating for both conditions; the resulting pair of values provides the x and y coordinates of the first point on the empirical ROC. The second point on the ROC is the proportion of responses in either the most acceptable rating category, or the response category just below that (e.g. 6 and 7 on a 7-point Likert scale). The subsequent points on the ROC reflect the cumulative proportion of responses at each possible response category; the final point on the empirical ROC is (1,1), reflecting the fact that all responses on a Likert scale fall into either the lowest response or above it. In the next section, we illustrate the calculation of an ROC curve in detail. At an intuitive level, the ROC visualizes how the distribution of responses in two experimental conditions differ on a point-by-point basis across the response scale. This method of comparing two conditions allows for a much more precise characterization of how two conditions may differ in acceptability that goes beyond the standard analyses of central tendency.

The SDT measures of sensitivity are directly related to other latent variable approaches to measuring acceptability. For example, *d'* is directly comparable to the inferred posterior acceptability in the Thurstone approach or the ELO approach when those measures assume equal variance between stimulus categories (Langsford et al., 2018). A similar interpretation is valid for ordinal regression (Liddell & Kruschke, 2018). However, the variance in the underlying acceptability associated with a category of sentence stimuli is not generally known. It is not at all obvious that different classes of sentence stimuli should have comparable variances; and our experience has taught us that they are often different.

At this juncture, it bears repeating that without explicitly modeling or measuring the underlying distribution of acceptability, it is very difficult to infer anything about the underlying distribution of acceptability. This distribution cannot be read directly off Likert ratings, whether they are *z*-transformed or not: what surfaces as bimodality in the Likert responses could simply be an artifact of the arbitrary placement of the response criteria, biases towards extreme responses, etc. The approach we describe here is critical for researchers interested in analyzing differences in the distribution of acceptability ratings across conditions, as the distribution of ratings on a response scale cannot be safely assumed to be a faithful reflection of the underlying acceptability distribution. The decision process can distort the underlying acceptability distribution substantially, and so there is value in explicitly modeling this aspect of the acceptability judgment task even if such a 'task model' is orthogonal to the researcher's primary theoretical interest.

The perspective developed here is a first, but critical, stepping stone. If the signal detection framework satisfactorily models acceptability judgments in Likert scale rating data, then it may offer a general framework for recovering interval scale differences in acceptability ratings between different classes of sentence. We now turn to a worked, tutorial style example of applying SDT-style data analysis to Likert rating data.

3 Tutorial: SDT & D-linking

In this section we will work an example with actual data, derived from an experiment designed to measure the acceptability of extraction dependencies and how that depends on *d*-linking (see Goodall, 2015). Our goals are two-fold. First, we seek to provide a simple proof-of-concept that the task model implied by SDT analysis provides a good approximation to response behavior in a real data set (see also Bader & Häussler, 2010). Second, we aim to give a tutorial-style introduction to the application of these techniques to facilitate the wider application of these techniques.

First we spend a little time talking through the method and design considerations. While familiar, they are not identical to a "run of the mill" syntax experiment. The topics and methods discussed in this tutorial are familiar from other areas of psychology, engineering and even radiology -- but they have not often been offered for an experimental syntax application. For reasons of space, we are unable to give a full treatment of all the various issues raised by this analysis. We recommend that the interested reader pair this section with the more complete and justified discussion of these issues in MacMillan & Creelman (1991/2005) and, in particular, their discussion of ratings experiments (Chapter 3).

3.1 Specimen Experiment

3.1.1 Method

An ROC curve is revealing about the underlying distribution of Acceptability, the latent cognitive variable of interest, because it takes relative measurements of two Acceptability distributions at multiple criterion placements. There are a number of experimental design parameters that can be used to cause participants to adopt different biases (see MacMillan & Creelman, 2005): for example, using payoffs to differentially reward correct *yes*-responses and correct *no*-responses; or using instructions that convey misleading estimates of the 'true' rate of Grammatical and Ungrammatical stimuli. But the use of a ratings experiment is perhaps the simplest means for estimating an ROC curve, and, in our experience, a quite reliable means for doing so. In our specimen experiment, we explicitly asked people to first classify the stimuli and then rate their confidence on a 3-pt scale. As a consequence, they effectively gave a rating on an 1-6 scale[2]. We did not put them under time pressure to give either judgment[3]. We implemented the experiment using Ibex on IbexFarm (Drummond, 2013).

3.1.2 Study Design Considerations

Our design must give participants a genuine opportunity to make a choice. And we need to compare the acceptability of classes of sentences that differ in a theoretically well-defined way.

[2] We suspect that much existing experimental syntax data, collected along an *n*-point Likert-style scale, could be profitably re-analyzed as a ROC curve – provided that the experiment was designed so that some conditions can be reasonably identified as sources of Hits, and others as sources of False Alarms. We have reanalyzed some of our own datasets, and find that the assumptions of an unequal variance normal-normal model are typically met (see section 3.2.2). This is an obvious area for future research.

[3] Technically the *yes*/*no* grammaticality judgment timed out after 10 sec, but we consider it virtually untimed. The mean of participant median response times was 1360 ms for grammaticality judgments, and 840 ms for confidence judgments. A version of our experiment with a 2000 ms deadline to give the grammaticality judgment yielded comparable sensitivity results. If one were interested in fine-grained modeling of RT distributions, then using the untimed task is probably preferable.

Therefore we must have a clear hypothesis about the mechanism that distinguishes two sentence classes. We adopt the hypothesis that *d*-linking improves the distinctiveness of wh-phrases -- and thus their retrievability. Therefore we should compare sentences that differ not only in whether or not the wh-phrase is lexically restricted; but crucially also in whether or not retrieval is required to interpret the sentence grammatically.

Consider the following grammatical sentence:

(3) Who do you think that the new professor is going to persuade?

What would be an appropriate control for (3)? It should elicit an opposite response in the binary judgment task, i.e., "No", and differ minimally from (3) along almost every dimension, except the theoretically relevant ones.

To recognize (3) as grammatical, the perceiver must successively (i) encode the displaced wh-*phrase,* (ii) identify the contexts in which gaps could occur, and, when they do, (iii) retrieve the filler phrase (Wagers, 2013). Therefore we constructed (4) as a control.

(4) Who thinks that the new professor is going to persuade?

(4) has a short matrix subject extraction that imposes comparable demands for (i), but effectively blocks the processes associated with (ii)-(iii). At the same time, it uses nearly the same lexical items, the same bi-clausal structure, the same argument structure, etc. It necessarily contrasts in the matrix subject. Using the local person *you* in (3) enables a reasonable comparison, since pronominal subjects are known to engender minimal additional complexity and to effectively level subject/non-subject extraction differences (Gordon, Hendrick, & Johnson, 2001).

The verb in (3-4) was selected to be obligatorily transitive: for example, practically all speakers of English require *persuade* to have a complement DP (Gahl, Jurafsky & Roland, 2004). This is important, because we want the acceptability in (3) to depend on the subprocesses of dependency comprehension succeeding. If the verb were optionally transitive (e.g., *attack*), the comprehender might assign (3) high Acceptability without engaging the processes of a theoretical interest, i.e., without finding a legitimate grammatical derivation. (5)-(6) are exactly the same as (3)-(4), but with a lexically restricted *wh*-phrase.

(5) Which donor do you think that the new professor is going to persuade?
(6) Which donor thinks that the new professor is going to persuade?

Finally, we must de-correlate grammaticality from whether or not there's a matrix subject (4,6) or embedded object (3,5) dependency. If participants implicitly learned this connection, then it would be possible to correctly classify the grammatical/ungrammatical stimuli without deeply parsing the sentences. It is in principle possible to do this by manipulating the filler sentences. We chose, instead, to do it as part of the experimental design by inserting an indefinite DP (*someone/anyone*) in the embedded object positions.

(7) Who do you think that the new professor is going to persuade anyone?
(8) Who thinks that the new professor is going to persuade anyone?
(9) Which donor do you think that the new professor is going to persuade anyone?
(10) Which donor thinks that the new professor is going to persuade anyone?

This design thus realizes 8 conditions: WhP (bare, or d-linked), Embedded VP-type (gap, or filled gap), and Grammaticality (grammatical, or ungrammatical). Table 1 repeats the full design with condition labels.

	VP	**WhP**	**Gram.**	*Sentence*
1	`Gap`	`bare`	`Gram`	Who do you think that the new professor is going to persuade?
			`Ungram`	Who thinks that the new professor is going to persuade?
2		`dlink`	`Gram`	Which donor do you think that the new professor is going to persuade?
			`Ungram`	Which donor thinks that the new professor is going to persuade?
3	`Filled Gap`	`bare`	`Gram`	Who thinks that the new professor is going to persuade anyone?
			`Ungram`	Who do you think that the new professor is going to persuade anyone?
4		`dlink`	`Gram`	Which donor thinks that the new professor is going to persuade anyone?
			`Ungram`	Which donor do you think that the new professor is going to persuade anyone?

Table 1 Example Item Set

3.2 Analysis

3.2.1 Simple sensitivity and bias

There are several possible ways to analyze the data that result from a forced-choice experiment with confidence ratings. Let us start with the simplest SDT analysis, based on *just* the binary judgment data. In Table 2 below, we've summarized the response outcomes by condition. For each condition, the empirical proportion correct (`p.c`) and its complement proportion error (`p.err`) is reported. In the final columns, `p.c` and `p.err` are annotated with a traditional SDT label: Hit (Correct *yes*), False Alarm (FA; Incorrect *yes*), Miss (Incorrect *no*), Correct Rejection (CR; Correct *no*). Numbering of the table rows indicates condition pairs to be compared ("scaled against one another"), and within the `p.c` and `p.err` columns, the values to be scaled are placed in bold.

	VP	WhP	Grammaticality	correct	error	p.c	p.err	p.c_type	p.err_type
1	gap	bare	gram	237	120	**0.664**	0.336	Hit	Miss
1	gap	bare	ungram	236	118	0.667	**0.333**	CR	FA
2	gap	dlink	gram	242	115	**0.678**	0.322	Hit	Miss
2	gap	dlink	ungram	248	110	0.693	**0.307**	CR	FA
3	fld	bare	gram	311	46	**0.871**	0.129	Hit	Miss
3	fld	bare	ungram	293	66	0.816	**0.184**	CR	FA
4	fld	dlink	gram	288	71	**0.802**	0.198	Hit	Miss
4	fld	dlink	ungram	315	42	0.882	**0.118**	CR	FA

Table 2 Binary Judgment Results

Based on the forced-choice responses alone, we can quantify participants' aggregate performance, factored into sensitivity and bias. Note that in this analysis, we are analyzing data aggregated across participants. In principle *d'* could be calculated on an individual basis, however, and that clustered data further submitted to inferential tests. Bearing this in mind, we proceed with the aggregated approach, and return to the perils and pitfalls of this in section 4.3.

The *d'* measure of sensitivity, discussed above, is the distance between two Acceptability distributions expressed in standard deviation units. To compute *d'*, therefore, we must convert the probabilities in Table 2 into standard deviation units, or *z*-scores, using the inverse cumulative normal distribution function. This is $\Phi^{-1}(\bullet)$ in standard notation, and here we use the somewhat zippier notation *z*(•) for that function.

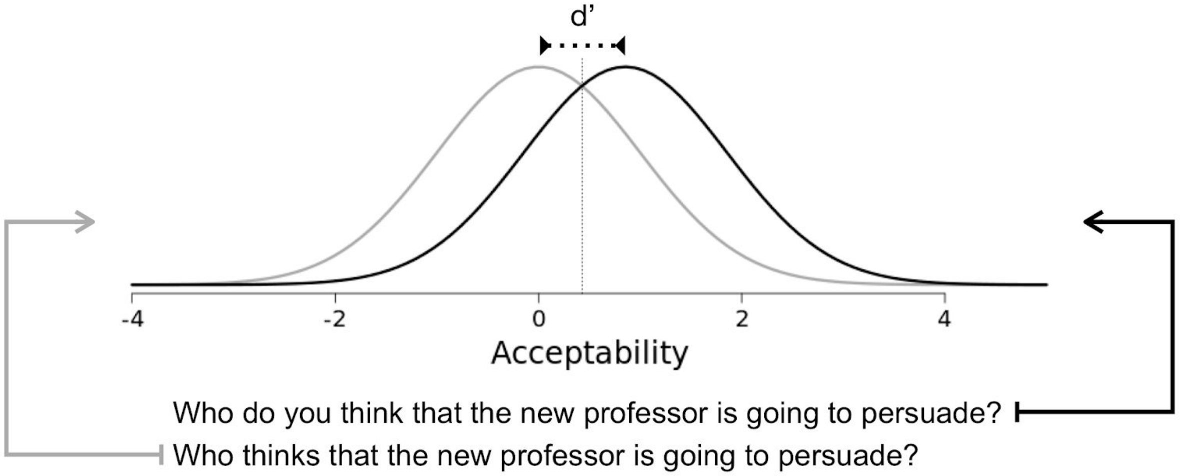
(11) d' = z(Hits) - z(FA) NB: z(p.Hits) = -z(p.Miss); z(p.FA) = -z(p.CR)

In R this distribution is implemented by the function `qnorm`, which takes a probability as its argument and returns a z-score[4]. For the *Bare* WhP conditions, with a gapped VP, we calculate d' as follows, with representative R code below.

(12)
```
# Write a function to implement d'=z(Hits)-z(FA)
sensitivity.ev <- function(Hits, FA){
      dprime <- (qnorm(Hits) - qnorm(FA))
      return(dprime)
}
> sensitivity.ev(0.664, 0.333)
[1] 0.855049
```

[4] The Normal probability distribution is defined from -∞ to +∞, and even for very large z-scores, there is an infinitesimal non-zero density. Therefore, the inverse Normal will return ±∞ for either z(0) or z(1). In group data, and even in most experimental syntax applications, this will not often be a problem because it will be rare to have perfect performance. However in individual data, it is more likely a participant will respond entirely consistently in at least one condition. In those cases, a correction must be made so that z(•) is defined. The simplest option is to add a trial to the total count, and then split it evenly between "Yes" and "No". Thus someone who gave 10 out of 10 (correct) *yes* responses, would be coded as having a corrected `p.c` of 10.5/11, or 0.955, and a correct `p.err` of 0.5/11, or 0.045. This is a form of 'smoothing' that is used when estimating probabilities off of empirical data; for a full discussion of the different approaches to correcting for extreme performance, see (Hautus, 1995).

Under the assumption of equal variance, this value implies a distribution of underlying acceptability as depicted in Figure 3: two standard normal distributions whose means are separated by 0.85 standard deviations. Observe that these two distributions overlap substantially, and errorful performance is thus guaranteed. Participants will sometimes misidentify grammatical sentences as ungrammatical, and vice versa.



Who do you think that the new professor is going to persuade?

Who thinks that the new professor is going to persuade?

Figure 3 Implied Equal-Variance Signal/Noise Distributions in Bare/VP Gap conditions

The optimal strategy to minimize errors is to set a response criterion halfway between the two peaks of the distributions -- i.e., where the two density functions intersect (indicated by the dashed line; Theodoridis & Koutroumbas, 2009). For any given value of acceptability the ratio of the heights of these two distributions (the black Grammatical distribution; the dark grey Ungrammatical distribution) define the odds that a given Acceptability value was drawn from that distribution. For higher values of acceptability, the odds in favor of the grammatical distribution grows. Where the two intersect, the odds are even -- so values above that point on the *x*-axis should elicit a Grammatical/*yes* response and values below that point should elicit an Ungrammatical/*no* response. We can characterize *bias* in the experiment as how far away from this optimum the actual criterion was set. This value, *c*, can be calculated as follows:

(13)

```
# Write a function to implement c=-[z(Hits)+z(FA)]/2
bias.ev <- function(Hits, FA){
  bias <- (qnorm(Hits)+qnorm(FA))*(-1/2)
  return(bias)
}
> bias.ev(0.664, 0.333)
[1] 0.004119758
```

In our dataset it is very close to 0, which suggests that overall bias is low in the experiment. The positive sign indicates a (slight) relative surplus of “No” responses; we will refer to positive values of *c* as ‘conservative’ response strategies, and negative values of *c* as ‘liberal’ response strategies. We can spot check this calculation by observing that Hits (0.664) in the Grammatical condition are nearly identical to Correct Rejections in the matched Ungrammatical condition (1-0.333=0.667; please see Table 2).

A positive value for *c*, indicates that the empirical response criterion is somewhat higher (more conservative) than optimal: there are more ‘no’ responses than optimal. Had there been more ‘yes’ responses than optimal, we would have expected a greater *Hit* rate in grammatical conditions, but a lower *Correct Rejection* rate in ungrammatical conditions. And correspondingly we would have obtained a negative *c*. In Table 3, the values for *d’*/sensitivity and *c*/bias are given for pairs of conditions, grammatical and ungrammatical variants, within each level of VP-type and WhP-type. Impressionistically we can see that d-linked *wh*-phrases increase participants’ sensitivity to conditions with a *Gap* in embedded object position, but not those with a Filled Gap (i.e., indefinite DP argument). There is bias shift in VP Filled Gap conditions, from more “yes” responses than optimal (negative *c*) when the WhP is bare to more “no” responses than optimal (positive *c*), when the WhP is d-linked. The bias measure *c,* in the context of our current experimental design, may be interpreted as a reflection of factors that influence the acceptability of a sentence class that are independent of the contrast of interest. For example, our observation that *c* was greater for the filled *d-link* conditions means that there is some feature of the filled-gap d-linking stimuli that caused participants to reject them at higher rates than we would have expected: participants exhibited a tendency to reject these sentences. The fact that this effect was found in *c,* our bias measure, indicates that whatever the source of this effect, it is unrelated to the process of constructing a filler-gap dependency. It may be of independent interest in its own right, but in order to isolate and identify the source of this effect, we would need to find a plausible hypothesis about its source, and construct an experimental design that would allow us to isolate and test the factors that do create this effect. At present, this effect of d-linking appears to be a general effect on this class of sentences that is unrelated to the process of filler-gap dependency completion.

```
  VP       WhP   d’     c
1 gap      bare  0.855  0.004
2 gap      dlink 0.966  0.021
3 filled   bare  2.031 -0.115
4 filled   dlink 2.034  0.168
```

Table 3 Summary of Sensitivity and Bias in the Equal Variance Analysis

The foregoing analysis is a convenient representation of our forced-choice data -- one which is arguably more digestible, as it transforms the dimensionality of the summary from 8 numbers to 4, setting bias aside. However, it makes a crucial assumption, one which we will usually find it necessary to relax. In Figure 3 the underlying Acceptability distributions have equal variance. But what if the Grammatical and Ungrammatical distributions over acceptability were not equal in variance? For the reasons explored in Section 4.2, knowing whether one class of sentences gives rise to a narrower or broader natural range of Acceptability values could be as theoretically revealing as knowing the centers of the distributions. To determine this, it will be necessary to construct a Receiver Operating Characteristic curve, or ROC curve.

3.2.2 Receiver Operating Characteristic Curve

A ROC curve describes Hits as a function of False Alarms across different degrees of bias. In our specimen experiment, we can construct this curve by grading our Y/N judgments using their confidence ratings. Consider the data just for the VP:Gap/WhP:Bare conditions, given in Table 4.

```
     VP   WhP   Grammaticality Answer.cr           No   Yes
 1   gap  bare  gram           Not confident       16    13
 2   gap  bare  gram           Somewhat confident  53    78
 3   gap  bare  gram           Very confident      51   146
 4   gap  bare  ungram         Not confident       23    15
 5   gap  bare  ungram         Somewhat confident  90    60
 6   gap  bare  ungram         Very confident     123    43
```

Table 4 Confidence Ratings in VP:Gap/WhP:Bare Conditions

To construct the ROC, let us first reshape this table into a series of tables, as illustrated in Table 5. First, starting with "Raw Counts", we order the responses along a scale from "Very Confident" *yes*-responses to "Very Confident" *no*-responses.

While we analyze data from an experiment that combined a binary judgment of acceptability with a confidence rating, this is not necessary for the analysis that follows. In fact, the ROC analysis we pursue here is one that is in principle possible for any *n*-point Likert scale data. To highlight this equivalence, the columns in Table 5 are numbered from 6 to 1 descending. This numbering is both for convenience of reference but also to reinforce the mapping onto a common Likert response scale, and show how this analysis would be applied to similar data. While the analytical tools offered here may be applied to Likert scale data without loss of generality, it is not obvious that the quantitative results we report below would replicate with judgments collected in the Likert scale method (see Wagers & Dillon, in prep).

Raw Counts

Grammatical?	YES			NO			
Confidence	Very	Smwhat	Not	Not	Smwhat	Very	Sum
	'6'	'5'	'4'	'3'	'2'	'1'	
GRAM	146	78	13	16	53	51	357
UNGRAM	43	60	15	23	90	123	354

Cumulative Counts

Grammatical?	YES			NO			
Confidence	Very	Smwhat	Not	Not	Smwhat	Very	Sum
	‘6’	‘5’	‘4’	‘3’	‘2’	‘1’	
GRAM	146	224	237	253	306	357	357
UNGRAM	43	103	118	141	231	354	354

Cumulative Proportions

Grammatical?	YES			NO			
Confidence	Very	Smwhat	Not	Not	Smwhat	Very	Sum
	‘6’	‘5’	‘4’	‘3’	‘2’	‘1’	
GRAM	0.41	0.63	0.66	0.71	0.86	1	357
UNGRAM	0.12	0.29	0.33	0.40	0.65	1	354

Table 5 Transforming Confidence Ratings to an ROC Curve. Data shown are from the VP Gap / WhP Bare conditions.

Column 6 contains the most “Very Confident” *Yes* responses to grammatical conditions. As a proportion (146/357), this represents 0.41 of all responses in that condition. Column 6 also contains the number of “Very Confident” *Yes* responses to ungrammatical conditions; as a proportion (43/354), this represents 0.12 of all responses in that condition. We can think of this pair <0.12, 0.41> as a <FA, Hit> pair representing performance achieved with maximal *no* bias -- only the highest Acceptability values would elicit a *Yes* under that bias, and even then a few ungrammatical trials fall within that range. If we move to the next most stringent *No* bias, we would include the responses under Column 5: 146+78 *Yes*-responses to grammatical conditions (=224; =0.63) and 43+60 *Yes*-responses to ungrammatical conditions (=103; =0.29). Thus our next <FA, Hit> pair is <0.29,0.63>. We continue doing this across the entire table - moving from conservative *No*-biased criteria to liberal *Yes*-biased criteria. Ultimately, we will have generated a series of 6 pairs, culminating in <1,1>. Intuitively, <1,1> is what happens when we say “Yes” to every trial: correctly capturing 100% of the grammatical conditions, but also trivially subsuming 100% of the ungrammatical conditions.

In Table 5, “Cumulative Proportions,” we’ve computed these pairs, and in Figure 4 (left column) we’ve plotted them, with False Alarms on the *x*-axis, and Hits on the *y*-axis. To complete these empirical ROCs, we added the point <0, 0>: what happens when we say *No* to every trial. Finally, we’ve gone ahead and - via the same method - computed the ROCs for all conditions in the experiment. In the right column, we’ve z-transformed each <FA, Hit> pair with the `qnorm` function. By inspecting the shape of these plots, we can already make a few first-pass conclusions about the underlying shape of the Acceptability distribution. Let us focus on the VP Gap conditions (in the top panels).

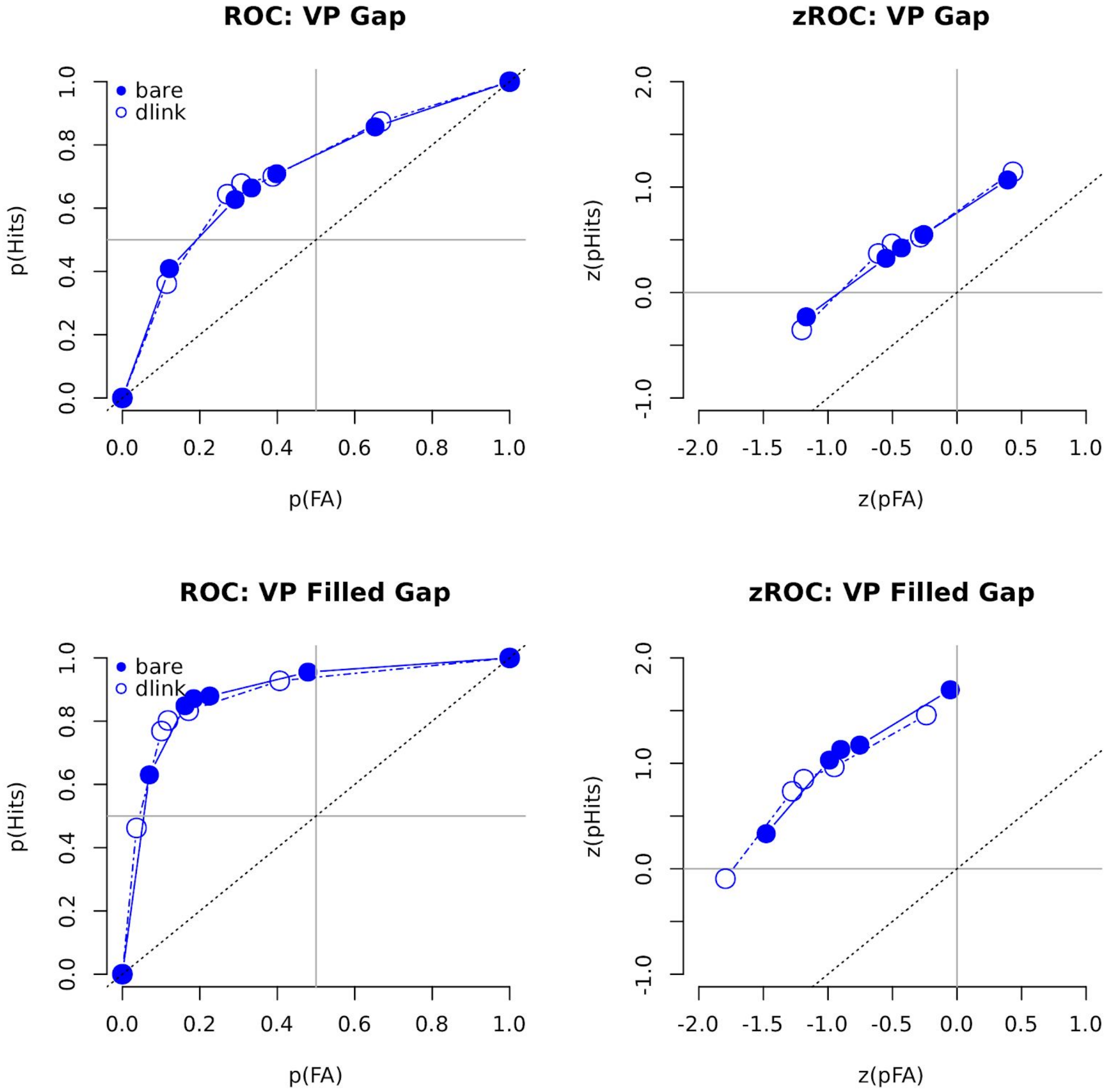


Figure 4 **ROCs, *Left*, and zROCs, *Right*, for All Comparisons**

Firstly, in both the ‘raw’ and z-transformed ROC curves, the dotted line indicates zero discrimination. But all curves representing our data, in blue, sit comfortably *above* that - our participants thus demonstrated the ability to discriminate between the grammatical and ungrammatical conditions. This is akin to a positive *d’* in the simple, equal variance calculation we performed in the previous section.

Secondly, notice that when we transformed the bowed ROC curve into z-coordinates, the resulting curve (for VP Gap) conditions is essentially a straight line. This is a good tell that the underlying distribution of acceptability is indeed normal (although this is in principle compatible with other distributions of acceptability). To a first approximation, this linear function seems consistent with the Gaussian observer model proposed by Bader & Häussler (2010). However, things look less clearly linear for the VP Filled Gap conditions, which still exhibit some curvilinearity in zROC space. As discussed below in Section 4.2, this is consistent with an underlying Acceptability distribution that is bimodal, or has some other kind of mixture distribution. The convex zROCs could also be attributed to other sources of noise in the data (Ratcliff, McKoon & Tindall, 1994).

Finally, you'll notice that the zROC line in VP Gap conditions is not quite parallel to the diagonal, and is somewhat shallow. In fact, its slope is less than 1. This is evidence that there is *more* variance in the Acceptability distribution for grammatical conditions than for ungrammatical conditions. Intuitively, the slope tells you the rate at which Hits accrue relative to False Alarms. A slope less than 1 means that Hits accrue relatively more slowly than False Alarms, a fact which implies greater variance in the signal (Grammatical) distribution. More precisely, the slope of the zROC is the ratio of the variance in the noise distribution to the variance in the signal distribution. For a slope of *s*, and a signal distribution whose variance is scaled to 1, then the variance of the noise distribution will be *s*; if instead the noise distribution's variance is scaled to 1, then the slope of the zROC is 1/*s*, and *s* is the variance in the signal distribution.

In (14) we make a simple estimate of *s* by calling Rs `lm` function, which returns the intercept and slope (in that order)[5]. For WhP:Bare/VP:Gap conditions, *s* is 0.8216.

(14)

```
> zFA <- qnorm(FA)
> zHits <- qnorm(Hits)
> lm(zHits~zFA)

Call:
lm(formula = zHits ~ zFA)

Coefficients:
(Intercept)     zFA
0.7609          0.8288
```

The fact that the variance of the underlying Acceptability distributions is unequal complicates the use of d' as a simple measure of sensitivity because the obtained sensitivity will now vary with criterion. In geometric terms, *d'* can be thought of as the distance between the zROC and the chance diagonal (zHits = zFA). When the slope of the zROC is 1, then this distance is constant across the range of zFA, and can be read directly off the intercept. But now distance to the zROC line varies along the range of zFA, and therefore d' varies. If we want to express

[5] We offer estimate only as an example. It is not generally advisable to use linear regression to find the slope; this is because the *x*- and *y*-coordinates in the ROC both constitute dependent variables, and the estimate of each is subject to uncertainty. Getting a reliable estimate of *s* involves fitting a full SDT model to the data, for example using Maximum Likelihood Estimation. To make a full UVSDT analysis accessible to researchers with a range of modeling backgrounds, Pazzaglia, Dubé and Rotello (2013) published an implementation of this using Excel's SOLVER function.

sensitivity by a single number, then we will have to take into account the fact that the underlying Acceptability distributions differ in their variance.

The measure d_a makes a kind of compromise by scaling the difference in the means of the Acceptability distributions by the root-mean-square average of their variances. Algebraically, this comes to the expression in (15), and the R calculation illustrated in (16). The value d'_2 is the y-intercept of the empirical zROC.

(15) $d_a = \sqrt{\frac{2}{1+s^2}} \cdot d_2'$

(16)
```
> emp.zROC <- lm(zHits~zFA)
> deetwo <- coef(emp.zROC)[1] # y-intercept
> s <- coef(emp.zROC)[2]      # slope
> da <- function(deetwo, s) sqrt(2/(1+s^2)) * deetwo
> da(deetwo, s)
[1] 0.8285487
```

The equation above may not be entirely intuitive at first. But, it has another guise in a proportional measure of sensitivity that is perhaps more visually comprehensible: A_z.

A_z is the area under the fitted normal ROC curve. Figure 5 demonstrates this graphically for the Bare/VP Gap conditions. The observed data points are given in blue, and the solid black line is the best-fitting curve to those points, constrained to describe the ratio between two normal distributions. The shaded area corresponds to all points <*x*,*y*> below the solid black line: here it covers 72.1% of the area, or $A_z = 0.721$. If there were no sensitivity in our experiment, such that Hits = FA, then A_z would be 0.5 -- everything below the major diagonal. If there were perfect sensitivity, the shading would fill the entire plot, and A_z would equal 1.

More generally, the area under the ROC curve is one important index of sensitivity. If one assumes that the underlying distributions that generate the ROC are normal, then A_z equals the area under the curve (AUC). However, it is also possible to calculate the area under the ROC curve without making this assumption about the parametric shape of the underlying distributions; in this case, the area under an empirical ROC can be calculated by simply using the 'trapezoid' method, that is, successively summing the areas of trapezoids that connect the points in the empirical ROC (Melo, 2013). It remains to be seen whether, in general, the assumption of normal acceptability distributions yields a good fit to acceptability judgment data. In our experience, however, fitted normal-normal ROCs often yield a very good fit to empirical ROCs (indeed; this can be seen in Figure 5: the empirical points lie quite close to the fitted curve).

A_z can be converted to d_a and vice-versa. The equation and code-snippet in (17) shows how to do this in both directions. We used the R library *pROC* (Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, & Müller, 2011) to fit the normal-normal ROC curve in Figure 5, to compute A_z and to create the plot. In Section 4, we return to some recommendations about software and procedures.

(17) a. $d_a = \Phi^{-1}(A_z) \cdot \sqrt{2}$

```
> Az2Da <- function(Az) qnorm(Az)*sqrt(2)
> Az2Da(0.721)
[1] 0.8284672
```

b. $A_z = \Phi \cdot (d_a / \sqrt{2})$

```
> Da2Az <- function(Da) pnorm(Da/sqrt(2))
> Da2Az(0.8284672)
[1]  0.721
```

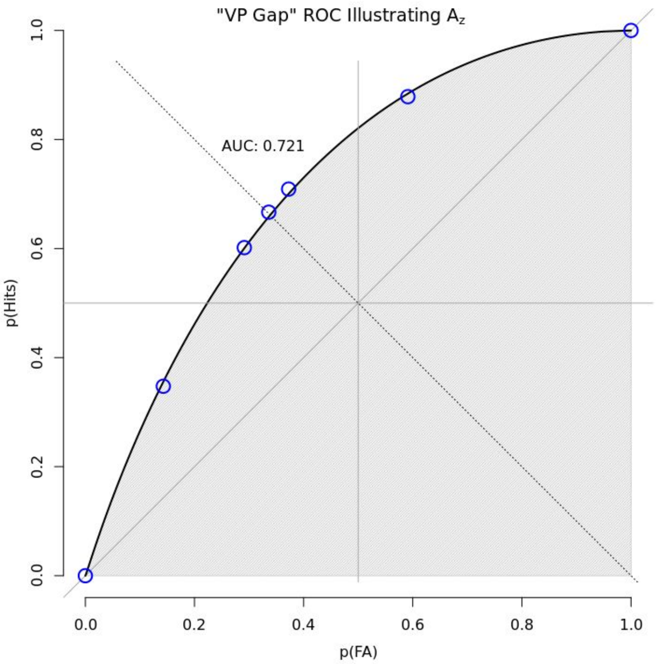


Figure 5 ROC for VP Gap/WhP Bare Condition illustrating A_z

Table 6 summarizes the 4 crucial comparisons in our experiment, reporting both d_a and A_z. Finally we report *s*, the ratio of the variance in the noise/ungrammatical distribution to the signal/grammatical distribution, as estimated from the zROC line. Figure 6 plots the implied Acceptability distributions.

VP	WhP	d_a	A_z	s
1 gap	bare	0.83	0.72	0.83
2 gap	dlink	0.86	0.73	0.89
3 filled	bare	2.0	0.91	0.92
4 filled	dlink	1.9	0.91	0.95

Table 6 Summary of Sensitivity and Variance in Unequal Variance Analysis

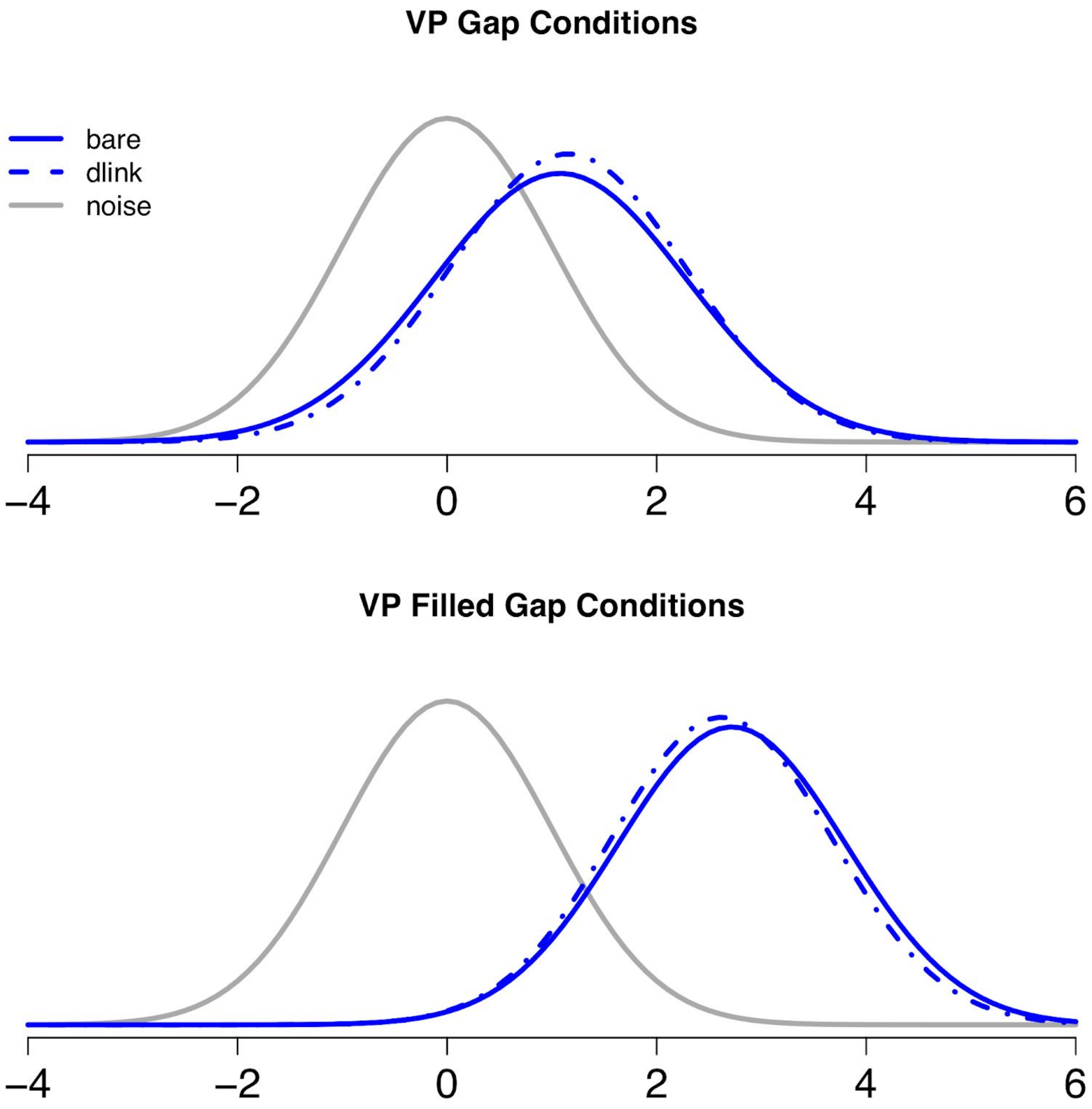


Figure 6 Implied Unequal-Variance Signal/Noise Distributions

3.3 Interpreting the Results

Constructing the ROC curve shows that, in our experiment, d-linking had, a best, a modest impact on sensitivity (and a non-significant one; see section 4.1). It did not greatly shift the distribution of Acceptability. There was a modest effect on *s*, suggesting relatively less variance in the grammatical distribution when the WhP was d-linked compared to when it was bare[6]. Participants were considerably more sensitive in VP Filled Gap conditions. We hypothesize that this is because all of the information required to discriminate between grammatical and ungrammatical conditions accumulates before retrieval is ever necessary. The presence of the 'filled gap' is enough to classify the ungrammatical sentences as ungrammatical. In other words, in these sentences, retrieval may never happen (cf. Wagers & Phillips, 2014). Interestingly here d-linking modestly reduces sensitivity, which suggest that the distinctiveness, or goodness-of-fit, of the restricted *wh*-phrase makes it a tempting, if ungrammatical, lure to integrate into the filled gap site.

4 Further Considerations, Caveats and Future Research Opportunities

In this chapter, we have pointed out that many standard methods of measuring acceptability judgments fall short as true interval measurements of acceptability, a situation that interferes with the statistical analysis of acceptability judgment data, and which presents a problem for researchers interested in drawing inferences from the distribution and gradience in their acceptability judgment data. Following Bader & Häussler (2010), we proposed that Signal Detection Theory offers a set of analytical tools that can resolve these issues. We presented a specimen experiment aimed at investigating the processing of filler-gap dependencies using acceptability judgment measurements, and a worked Signal Detection Theoretic analysis of these data. We discussed simple indices of sensitivity (such as *d'*) and ROC analysis as ways of quantifying the contrast in acceptability between two types of sentences.

For reasons of space, this is as far as this chapter can take us. However, there are many open, unresolved issues; researchers interested in applying SDT to their own judgment data will face these issues in practice. Here we briefly mention a few.

4.1 Why SDT?

A plausible reaction to our proposal might be: *why bother with all of this, just to analyze acceptability judgment data?* The SDT approach we offer here might strike the reader as a solution in search of a problem. Indeed, we have not offered a clear example of a spurious claim in the literature that results from not taking into account the analytical considerations we raise here. It remains an open question whether there are such examples to be found, and our aim here is not to find one such an example to hold up as a demonstration.

Instead, we would like to suggest that the SDT approach we advocate here is worthy of consideration in its own right for several reasons. First, and perhaps foremost, is the simple observation that most work in experimental syntax implicitly, yet incorrectly, assumes that data from Likert ratings offers an interval scale. This assumption underlies the simple tabulation of

[6] Our results are not straightforwardly compatible with the finding of Goodall (2015) that d-linking improves acceptability even in non-island dependencies. That paper does note the existence of prior studies which also found limited effects of d-linking in non-island dependencies. But we do not read too much into this apparent non-replication, given the substantial differences between our two studies. We leave it as an area for future investigation.

descriptive statistics like mean rating value and standard error, and the interval scale assumption is required for inferential statistical tests (most pressingly for those involving interactions). Despite this, it is widely recognized that ordinal Likert rating data does generally yield interval measurements of acceptability (cf. Cowart, 1997). Happily, interval-scale measurements are precisely what SDT's measures of discriminability and bias offer. And as we have shown here, the SDT analysis can be carried out using Likert rating data. In our view, this motivates the use of SDT for the analysis of those rating data.

Second, it seems to us that the theoretical perspective implied by the SDT model--a Gaussian observer who maps a noisy, unidimensional acceptability value onto one of a handful of discrete response options--is a useful theoretical perspective on the acceptability judgment process. It offers a precise model for how acceptability judgment responses are given in the context of a judgment task, and it makes explicit the key parameters of the acceptability decision task (acceptability versus scale usage). Should it prove to be a valid model of the acceptability judgment task, this model will allow researchers to ask more precise questions about acceptability judgment data.

Last, there is a broader reason why we think the SDT perspective is useful for experimental syntax: it suggests a useful experimental design heuristic. In our specimen experiment, we expanded the experimental design of Goodall (2015) to include ungrammatical and grammatical variants of each condition. Part of this motivation was analytical: SDT analysis requires the analyst to sample both from the 'signal' and the 'noise' distributions, which the latter offering an unacceptable baseline that in turn allows the analyst to quantify response bias or scale usage for a given structural configuration. But these baselines also served to make a more diagnostic experimental design. Specifically, they allowed us to potentially distinguish the impact of *d*-linking on the retrieval of a filler phrase (i.e. the hypothesis we were testing) from other nuisance factors that might have contributed to differential ratings for *d*-linked and bare *wh*-phrases. In the context of our SDT analysis, these nuisance factors were captured in our 'response bias' measures, which reflected baseline differences in the ratings between *d*-linked and bare-*wh* phrases. For example, we speculated that raters might offer more generous ratings for *d*-linked sentences overall, simply because they contain more lexical content. This would amount to a 'response bias' in favor of *d*-linked sentences, creating higher ratings for d-linked sentences for reasons independent of the factors of interest (e.g. memory retrieval processes): including structurally matched, unacceptable controls allowed us to diagnose this. However, we hasten to add that this issue is not limited to d-linking phenomena we study here: it can be very difficult to establish exactly what features of a sentence contribute to a judgment of its acceptability. The SDT analysis and concomitant design heuristics may prove useful to experimental syntacticians in developing experimental designs that can help reduce some of this interpretive uncertainty.

4.2 Inference and software recommendations

In our tutorial we focused on computing descriptive indices of sensitivity and bias from experimentally collected data. Statistical inference from a sample data with ROCs is less straightforward. For both descriptive fitting of ROC curves, as well as statistical inference, we recommend the *pROC* package (Robin et al., 2011). This package implements a number of common statistical tests for ROC data in *R,* and allows for straightforward fitting of ROC curves using model syntax familiar from other *R* packages (see Robin et al., 2014, for a complete introduction to *pROC*). We illustrate this syntax in (18). The basic *roc* function estimates an empirical ROC; here, we are fitting the ROC in Figure 5, to evaluate sensitivity in the bare *wh*-phrase conditions with a gapped VP. The basic call to *roc* fits an empirical ROC, and returns a measure of sensitivity (AUC) that is calculated using the trapezoid method on the empirical ROC. `Grammaticality` is a binary factor with two levels, `gram` and `ungram`, which contribute our hit and false alarm rights respectively. `ordinal` is our six-point response variable defined above, ranging from 6 (very confident 'Yes' responses) to 1 (very confident 'No' responses).

(18)
```
> roc.gap.bare <- roc(Grammaticality ~ ordinal,  data = bare.gap.data)
> roc.gap.bare
Call:
       roc.formula(formula = Grammaticality ~ ordinal, data =  bare.gap.data)

       Data:  ordinal  in  357  controls  (Grammaticality  gram)  >  354  cases
(Grammaticality ungram).
Area under the curve: 0.7063
```

The AUC for the empirical ROC is .7063; again, this is calculated assuming no particular functional form for the underlying distributions. However, a normal-normal ROC can be fit by setting `smooth` to `TRUE`:

(19)
```
> roc.gap.bare <- roc(Grammaticality ~ ordinal, smooth=T, data = bare.gap.data)
> roc.gap.bare
Call:
       roc.formula(formula = Grammaticality ~ ordinal, data =  bare.gap.data)

Data: ordinal in 357 controls (Grammaticality gram) > 354 cases (Grammaticality
ungram).
Smoothing: binormal
Area under the curve: 0.721
```

Here, the normal-normal fitted AUC is equivalent to A_z above. One benefit of using *pROC* is that once an *roc* object has been fitted, there are several methods that can be applied to the fitted curve to calculate inferential statistics. For example, confidence intervals over ROC statistics can be calculated with *ci*(), which will yield stratified bootstrapped confidence intervals.

In addition to confidence intervals, *roc.test()* implements a statistical test for statistically significant differences in area under the curve (AUC) or partial AUC for a pair of ROC curves. The partial area under the ROC curve, pAUC, is the AUC computed for some portion of the ROC curve (Ma, Bandos, Rockette & Gur, 2013). One problem that arises in the context of statistical inference for paired ROC curves is the problem of correlated observations across the curves: for a discussion of how the techniques discussed here address this issue, see Hanley and McNeil (1983), and Robin et al. (2014).

One straightforward and flexible method of testing for a reliable difference between two ROC curves relies on a bootstrap procedure: for *B* bootstrap replicates, the data are resampled (with replacement), and for each replicate, the statistic of interest is calculated (Efron & Tibshirani, 1994). For a comparison of two ROC curves, the normalized difference between the two resulting AUCs is well approximated by a normal distribution; thus, a *Z* statistic can be calculated by bootstrap, which then yields a *p*-value for a statistical test of the hypothesis that the AUCs of the two curves differ (Robin et al., 2014). Alternative approaches to testing differences between the AUC for two correlated ROCs is that of Delong, Delong and Clarke-Pearson (1988), which uses an approximation of the variance-covariance matrix between correlated ROCs, and Venkantraman (2000), which uses a permutation test applied directly to the ROC curves themselves, rather than to the AUC.

A bootstrap test in *pROC* applied to two fitted ROC curves using the code in (20). Here, we fit two ROCs to the data, and submit them to a simple test of significance using *roc.test():*

(20)

```
> roc.test(roc.gap.bare, roc.filledgap.bare, method = 'b’, are.paired=T)
Bootstrap test for two ROC curves

data:  roc.gap.bare and roc.filledgap.bare
D = -8.4598, boot.n = 2000, boot.stratified = 1, p-value < 2.2e-16
alternative hypothesis: true difference in AUC is not equal to 0
sample estimates:
Smoothed AUC of roc1 Smoothed AUC of roc2
     0.7210194            0.9122634
```

The bootstrap test yields a test statistic *D,* which is the normalized difference in the AUCs between the two ROCs across bootstrap replicates (Robin et al., 2014; see Hanley & McNeil, 1983 for how this value is calculated for paired ROCs); this test statistic is compared to a standard normal distribution to derive a *p*-value. By default, *roc.test()* will yield a two-sided test of the alternative hypothesis that the true difference in the AUC between the two ROC curves is not 0. The value of the test statistic in this particular example is large; correspondingly, we can reject the null hypothesis that the AUC for the gap and filled-gap ROCs are the same. This test suggests that the observation we made above--that raters showed greater sensitivity to grammaticality in filled-gap over gap-less structures--is likely to generalize beyond the present data set.

Some caveats to the preceding are in order. The structure of a typical experiment in experimental syntax or psycholinguistics is considerably more complex than that assumed by the out-of-the-box tests in *pROC*. The typical psycholinguistic experiment is typically a repeated-measures design with multiple random grouping factors; most commonly, these grouping factors are participant and stimulus item (it has been standard practice in psycholinguistics to treat stimulus as a random factor since Clark, 1973). For the last decade or so, the complex hierarchical structure to the data of the typical psycholinguistic experiment has driven rapid advances in the statistical treatment of these datasets (e.g. Baayen, Davidson &

Bates, 2008; Jaeger, 2008). The out of the box ROC tests ignore this important correlational structure in the data. These violated assumptions as a result may result in an inflated Type I error rate for tests applied to these ROC curves. To sidestep this issue, Dillon et al. (2019) resampled at the level of individual participants in a bootstrap test for differences in the key statistics they report; this is in essence a by-participants analysis, with no guarantee that the effect will generalize to other experimental items. At present, we know of no straightforward implementation of this participant-level bootstrap procedure that can be deployed out of the box; however, interested readers may consult Dillon et al. (2019), and download associated code at https://osf.io/sd3hu/. See the Appendix for a regression-based alternative that allows (limited) mixed-effects modeling.

4.4 Variance in underlying distributions

In this chapter we have focused on the role that ROC analysis can play in distinguishing sensitivity from bias in acceptability judgment experiments. However, ROC analysis also allows researchers to evaluate the relative variance in two stimulus categories: above, we did this by estimating the slope of the zROC. For normal-normal ROC curves, the slope of the *z*ROC is the ratio of the noise distribution's variance to that of signal distribution; in the example above, we set the signal distribution's variance to 1. Thus *s* represented the noise distribution's variance and the slope of the zROC alike. Because the slope of the zROC reflects the ratio of the variances of the two distributions, the empirical (*z*)ROC is informative not just about the distance between the underlying acceptability distributions, but also the relative variance in those distributions.

In other areas where ROC analysis is applied, a difference in slope has proven to be theoretically meaningful. To take one prominent example, this pattern of unequal variances is widely observed in the recognition memory literature (Ratcliff, Sheu & Gronlund, 1992). This indicates there is greater variance in the distribution of memory strength for studied items over unstudied lure items; this increased variance might arise, for example, if the distribution of memory strengths for studied items included a mixture of successfully encoded items and items that were not successfully encoded (Ratcliff et al., 1992; de Carlo, 2002). The shape of the ROC can also in principle reflect more unusual distributions of acceptability. For example, if one underlying distribution is bimodal or other type of mixture distribution, then the resulting ROCs can exhibit curvilinearity (see de Carlo, 2002, for extended discussion).

The shape of the ROC may be of theoretical interest to the extent that it illuminates differences in how two distributions of acceptability judgments differ. To illustrate this, let us return to the bimodal distribution of acceptability judgments in Dillon et al (2017). (21a) is the critical configuration that Dillon and colleagues wanted to investigate; recall that the central empirical question was whether the distribution of ratings associated with these examples was bimodal or unimodal. (21b) is a matched ungrammatical control from Dillon et al's study that can be used for the purposes of SDT scaling.

(21) a. Which flowers is the gardener planting?

b. Which flowers is the gardeners planting?

The aggregated distribution of ratings is in the leftmost panel of Figure 7. It can be seen that there is a pronounced bimodality for (21a). Earlier, we raised the concern that the conclusions Dillon et al drew on the basis of this were a potential artifact of how the scale is used. However, consider this worry in light of the empirical ROC and ZROCs presented in the middle and rightmost panels of Figure 7. Here, it can be seen that there is a pronounced curvilinearity in the ROC that is characteristic of a bimodal or mixture distribution (de Carlo, 2002). The ROC analysis is consistent with the conclusion reached on the basis of reasoning about the distribution of raw Likert ratings: the curvilinear (z)ROC is exactly what is expected if the underlying distribution of acceptability values is bimodal. In other words, the (z)ROC suggests that the underlying acceptability distribution for sentences like *which flowers is the gardener planting* is bimodal, just as is the Likert ratings.

Although in this instance, the conclusion licensed by the raw data and the ROC analysis align, this is not guaranteed. The broader point is that claims based on apparent distributions in Likert data are not watertight; ROC analysis can help secure empirical conclusions based on the distribution of rating data.

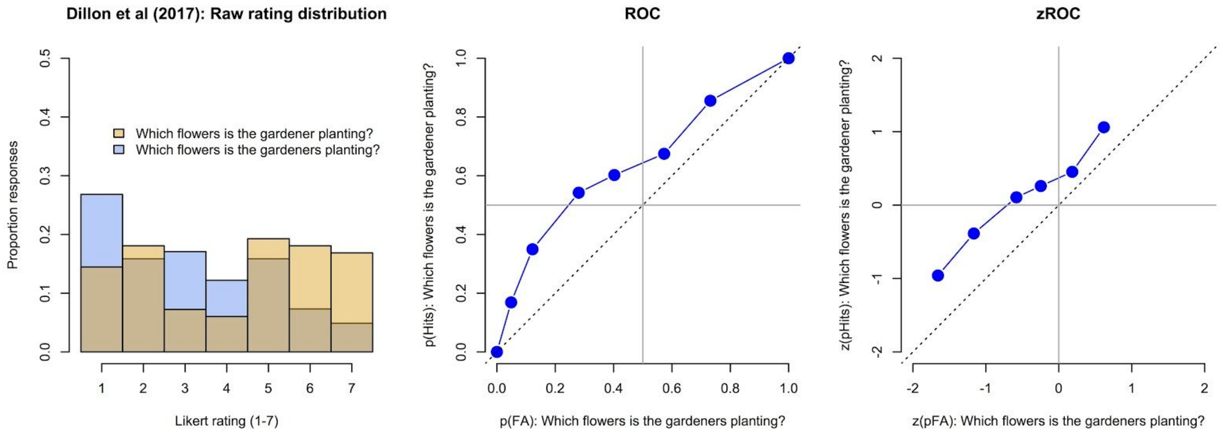


Figure 7 **Diagnosing Bimodality with ROC Analysis:** A reanalysis of the data from Dillon et al. (2017). The ROC and zROC suggest that the bimodality in the rating data does not reflect a simple bias towards extreme responses in the task.

4.5 Individual and group level analysis

Throughout, we have aggregated our data across participants. This step merits careful scrutiny: it is not guaranteed, in general, that the pattern seen in the aggregated data characterizes the data pattern seen for any one individual. For this reason, the pattern seen in the aggregate should be approached cautiously. Likewise, the variability seen in the aggregated data may stem from within-participant variation or between-participant variation. In analyzing the group-level ROCs, we make the implicit assumption that the participants share the same sensitivity, criterion locations, and slope of the zROC.

This assumption is almost certainly too strong. However, aggregating data across participants is often necessary, as not many experiments yield a sufficient number of data points per condition per subject to allow reliable estimates of the points of an empirical ROC at an individual level. For practicality reasons, then, most experimental syntacticians will be analyzing group-level measures of sensitivity. What is the effect of this aggregation on those measure? MacMillan and Kaplan (1985) show that the sensitivity estimates from aggregated data are biased estimates of the true sensitivity, but that the bias is slight: the effect of aggregation slightly underestimates the true sensitivity for a given comparison (see also MacMillan, Rotello & Miller, 2004). MacMillan et al (2004) note that the pooling of data across participants is preferred when the number of observations per individual is low (as is often the case in acceptability judgment experiments); they argue that the biasing effects of pooling data are likely to be modest, and that the estimates of sensitivity are *less* biased when based on aggregated data than when based on averaging individual-level estimates of sensitivity, when the amount of data for any one individual is small (Hautus, 1997).

Still, it seems fair to say that further research is necessary to understand exactly how group-level ROC statistics reflect individual-level variation in experiments with the structure of a typical within-subjects and within-items design used in experimental syntax. An alternative approach to the issue is suggested by researchers deploying similar techniques, such as the speed-accuracy trade-off (SAT) method (e.g. McElree, 2000; McElree, Foraker & Dyer, 2003). Like ROC analysis, SAT analysis requires stable estimates of Hit rates and False Alarm rates at multiple points, and in multiple conditions; also, like ROC analysis, group-level analysis may yield a data pattern that doesn't characterize the performance of any one individual participant (Liu & Smith, 2009). For this reason, SAT experiments typically involve collecting much more data at the individual level than is common in experimental syntax; most published SAT studies involve more than 30 observations per participant for each hit or false alarm rate estimate. Although this would allow more robust estimates of the ROC curve at an individual level basis, there are other obvious practical difficulties that travel along with this approach; it may require multiple testing sessions to gather sufficient data from a single individual, and with so many repeated observations, a participant may adapt to the structures being tested over the course of the experiment.

4.6 Conclusions

In this chapter, we have sketched how Signal Detection Theory can be applied to acceptability judgment data. We have argued that latent variable models such as SDT hold substantial promise for experimental syntacticians by offering a way of precisely answering the quantitative question in experimental syntax: *to what extent is sentence type A better than, or worse than, sentence type B?* We offered a worked, tutorial style analysis of a sample data set to show how ROC analysis can be applied to a data set that has a similar structure to Likert rating tasks and binary forced-choice acceptability judgments in experimental syntax.

In closing, we noted several challenges that arise in the context of the analytical approach pursued here. In particular, there are unresolved issues concerning statistical inference using ROCs, and it remains unclear exactly to what extent aggregating data across participants will distort estimates of sensitivity and bias in acceptability judgment experiments. A bit further afield, we note that the paradigm we used here--binary judgments with a secondary confidence rating--may yield results that are different from the Likert scale ratings that are more commonly deployed in the experimental syntax literature. Head-to-head comparisons of these methods of collecting judgment data would be valuable. Despite these challenges, we hope to have communicated our enthusiasm for this approach, and the promises we see for its application to acceptability judgment data.

5 Appendix: cumulative ordinal regression models and SDT

An unequal variance signal detection theory model can be estimated as a particular parameterization of a cumulative ordinal regression model (DeCarlo, 1998). Ordinal regression models are most appropriate for ordered response categories, including the responses produced from Likert-type trials. Here we'd like to briefly spell-out the connection to SDT, for two reasons. Firstly, to show you how an SDT analysis can be conducted in the more familiar setting of regression. But also, to urge consideration of the broad class of ordinal regression models, with or without an SDT interpretation, as a superior alternative to linear regression or ANOVA-style approaches that treat acceptability ratings as numbers instead of response categories. Ordinal regression is not yet widely used in the analysis of acceptability judgment studies, but it can overcome many of the pitfalls we mentioned in Section 2.1. These are discussed in much greater detail by Liddell & Kruschke (2018) and an excellent practical overview and tutorial is given by Bürkner & Vuorre (2019) using R and the Bayesian *brms* package. DeCarlo (2003) presents an SDT-as-ordinal-regression tutorial using SPSS. Below we use the R library *ordinal* (Christensen, 2019), which we like for its user-friendly syntax and its capability for estimating (some) mixed-effects cumulative models as well.

The ordinal regression model derives ordered response categories from the classification of latent variables. The cumulative model, in particular, assumes a latent continuous distribution which is partitioned by a set of thresholds, analogous to the presentation in Figure 2 (bottom panel). The placement of the thresholds determines the probabilities of selecting each response category. Thus the set of thresholds are like the set of multiple response criteria assumed in a ROC curve analysis. Predictor variables in the regression shift these thresholds, analogous to a sensitivity parameter like the *d* family of parameters. Finally a scale, or variance, parameter can shrink or widen the latent distribution, just like the *s* parameter in a ROC analysis.

It's easiest to work an example. We will illustrate a basic ordinal regression model with our WhP:Bare/VP:Gap trials from Table 5. In Table 7 below, we show 10 trials from our dataset. As before, we've mapped the set of <Judgment, Confidence> pairs onto a <1-6> scale, called *rating*. What matters here is that we've created an ordered factor. (Once again, it's worth stressing that this decision is not tantamount to assuming complete task equivalence between a Likert-judgment and SDT confidence rating task; the degree of overlap between these tasks remains an open empirical question).

```
   Item Condition Answer.gj          Answer.cr Grammaticality Rating
```

1	11	a	Yes	Not confident	gram	4
2	22	b	No	Very confident	ungram	1
3	32	b	No	Very confident	ungram	1
4	23	a	No	Very confident	gram	1
5	3	b	Yes	Somewhat confident	ungram	5
6	8	b	Yes	Very confident	ungram	6
7	1	a	No	Very confident	gram	1
8	12	a	No	Somewhat confident	gram	2
9	7	a	Yes	Very confident	gram	6
10	24	a	No	Very confident	gram	1

Table 7 Sample trials prepared for ordinal regression

In (22) we define the Ungram level of Grammaticality as the reference level (thus it is coded as a 0 in the regression; Gram is thus 1). In (23) we call the function *clm* and in Figure 8, we visualize its interpretation. Like other `lm`-style functions, its primary argument is *formula*, which defines the regression equation. Here the dependent variable, on the left-hand side of ~, is `rating` and the fixed effect of `Grammaticality` is on the right-hand side. The *link* argument species the distribution family to which the underlying latent variable belongs; and here we specify the use of the probit link, i.e., $\Phi^{-1}(\bullet)$ or z(•). By default, *clm* uses a logit link, and other density functions are possible.

(22)
```
> bare.gap.data$Grammaticality <-
       relevel(bare.gap.data$Grammaticality, ref="ungram")
```

(23)
```
> evsdt.clm <- clm(formula = rating~Grammaticality,
       link = "probit",
       data = bare.gap.data)
> summary(evsdt.clm)
formula: rating ~ Grammaticality
data:    bare.gap.data

 link   threshold nobs logLik   AIC     niter max.grad cond.H
 probit flexible  711  -1105.38 2222.75 6(2)  8.88e-09 1.4e+02

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
Grammaticality1  0.82054    0.08288     9.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
1|2 -0.34089    0.06344  -5.373
2|3  0.25510    0.06248   4.083
3|4  0.40541    0.06316   6.419
4|5  0.51382    0.06387   8.045
5|6  1.09124    0.07040  15.501
```

First, we focus on the table of *Threshold coefficients,* or θ*,* indicated in bold. These should be interpreted in standard deviates, because our latent variable is the Standard distribution $\mathcal{N}(0,1)$. They define the cumulative response probabilities for *ungram* responses. For example, the cumulative probability of responding with the first ordered category (1 ~ "Ungrammatical/Very confident") is $\Phi(\theta_{1|2})$. Figure 8, Panel A, shows this graphically and (24) shows how to calculate the modeled cumulative response probabilities for the ungrammatical condition.

(24)
```
     > pnorm(evsdt.clm$Theta)
        1|2   2|3   3|4   4|5   5|6
[1,] 0.367 0.601 0.657 0.696 0.862
```

<INSERT FIGURE 8 HERE; ONLINE COPY>

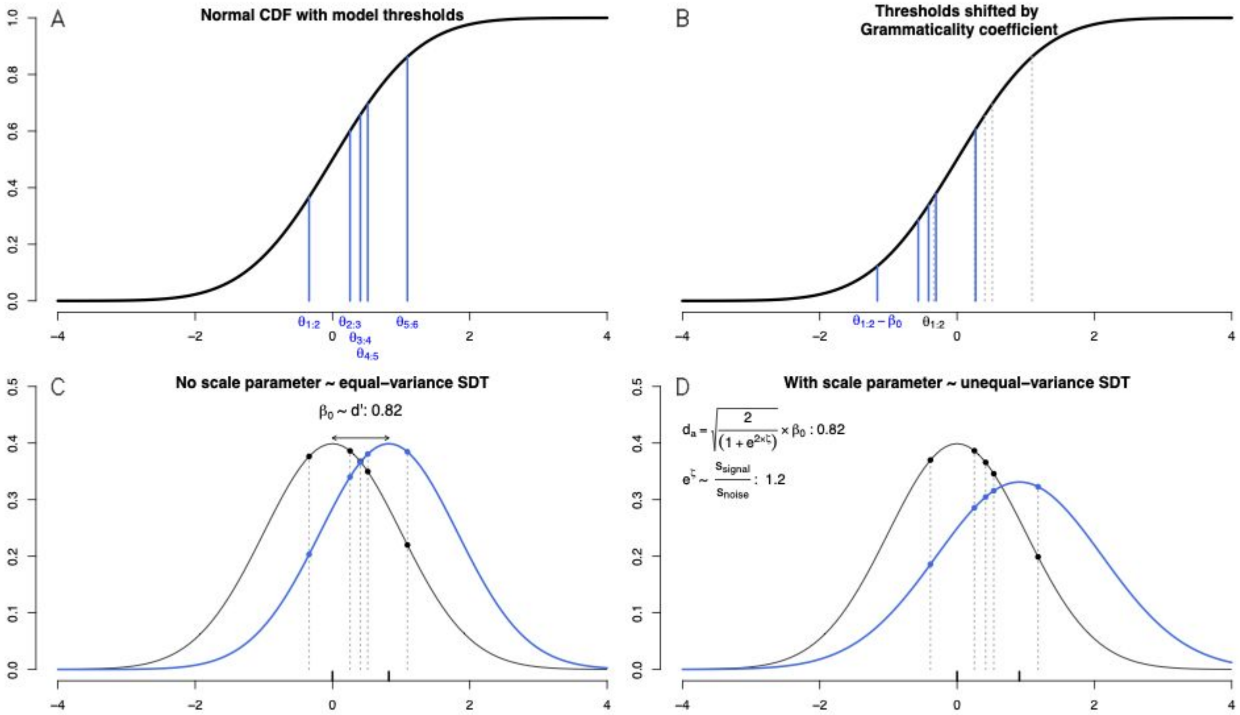


Figure 8 Relating cumulative ordinal regression models to SDT analysis

The difference between *gram* and *ungram* responses is modeled as a shift in the thresholds by the Grammaticality coefficient β_{GRAM}. Figure 8, Panel B, shows this graphically and (25) recovers the modeled values.

(25)
```
     > pnorm(evsdt.clm$Theta - evsdt.clm$beta)
        1|2   2|3   3|4  4|5   5|6
[1,] 0.123 0.286 0.339 0.38 0.607
```

Instead of conceptualizing the thresholds as moving by β_{GRAM} to the left, we could also imagine the density function is moving by β_{GRAM} to the right, which Figure 8, Panel C, illustrates. This makes it clearer that (23) is essentially just another representation of an equal-variance SDT model. Here β_{GRAM} may be interpreted as *d'*.

Overall the values we recovered in (24)-(25) do compare pretty well with the actual observations, as (26) shows, but it's not perfect: in particular, for Grammatical sentences, we've very slightly underestimated the low ratings and overestimated the high ones.

(26)

```
> bare.gap.data %$%
       table(Grammaticality, rating) %>%
       prop.table(1) %>%
       apply(1,cumsum) %>% t
              rating
Grammaticality     1     2     3     4     5 6
        ungram 0.347 0.602 0.667 0.709 0.879 1
        gram   0.143 0.291 0.336 0.373 0.591 1
```

This shouldn't be a surprise, however, since we discovered in the ROC curve analysis (Section 3.2) that the variance in the underlying signal and noise distributions were not equal. We can account for this in an ordinal regression model by adding a scale parameter. (27) updates the model call in (23) with such a parameter.

(27)

```
>uvsdt.clm <- clm(formula = rating~Grammaticality, scale = ~Grammaticality,
       link = "probit",
       data = bare.gap.data)
>summary(uvsdt.clm)
formula: rating ~ Grammaticality
scale:   ~Grammaticality
data:    bare.gap.data

 link   threshold nobs logLik   AIC     niter max.grad cond.H
 probit flexible  711  -1103.22 2220.44 9(2)  1.02e-07 1.3e+02

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
Grammaticalitygram   0.9076     0.1019   8.906   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

log-scale coefficients:
                   Estimate Std. Error z value Pr(>|z|)
Grammaticalitygram  0.18598    0.08978   2.071   0.0383 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
1|2 -0.38815    0.06789  -5.718
```

```
2|3  0.25211    0.06329   3.983
3|4  0.41595    0.06425   6.474
4|5  0.53483    0.06562   8.150
5|6  1.17987    0.08417  14.018
```

The scale parameter, estimated above as ~ 0.186, is reported on the log scale; (28) returns it to the probit scale (it's called `zeta` in *ordinal*'s CLM data structure).

```
(28)    > uvsdt.clm$zeta %>% exp
Grammaticalitygram
          1.204396
```

We can interpret this value as follows: the latent variable that supports judgments in this experiment has standard deviation approximately 1.2 as great in the grammatical condition, compared to the ungrammatical condition. This corresponds to 1/*s* (or ~ 0.83) in our ROC analysis. Figure 8, Panel D, visualizes this CLM. (29) shows how the modeled values can be directly computed; they compare much more favorably to the actual data, than those generated by the model without a scale parameter.

```
(29)    > rbind(pnorm(uvsdt.clm$Theta, sd = 1),
                pnorm(uvsdt.clm$Theta - uvsdt.clm$beta, sd = exp(uvsdt.clm$zeta)))
       1|2   2|3   3|4   4|5   5|6
[1,] 0.349 0.600 0.661 0.704 0.881
[2,] 0.141 0.293 0.342 0.378 0.589
```

We can close the circle on this demonstration by fully modeling the gap conditions in our d-linking data, with the call given in (30). As a reminder, the variable *WhP* indicates whether the sentences included a *d*-linked *wh*-phrase; and the first two lines center the predictors so that Grammaticality:gram and WhP:dlink are represented as the positive values of their respective contrasts.

```
(30)    > contrasts(all.data$Grammaticality) <- -contr.sum(2)/2
        > contrasts(all.data$WhP) <- -contr.sum(2)/2
        clm(rating~Grammaticality*WhP,
              scale=~Grammaticality*WhP,
              data=gap.data, family="probit") %>% summary
formula: rating ~ Grammaticality * WhP
scale:   ~Grammaticality * WhP
data:    subset(all.data, VComp == "gap")

 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  1426 -2208.33 4438.65 9(2)  5.92e-13 1.9e+02

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
Grammaticality1        1.4002     0.1010   13.86   <2e-16 ***
WhP1                  -0.0551     0.0974   -0.57     0.57
```

```
Grammaticality1:WhP1  -0.1169     0.1948   -0.60     0.55
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

log-scale coefficients:
                     Estimate Std. Error z value Pr(>|z|)
Grammaticality1        0.1526     0.0664    2.30    0.022 *
WhP1                  -0.1083     0.0643   -1.68    0.092 .
Grammaticality1:WhP1  -0.0802     0.1286   -0.62    0.533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
1|2  -1.3140     0.0650  -20.21
2|3  -0.2777     0.0576   -4.82
3|4  -0.0387     0.0575   -0.67
4|5   0.1314     0.0578    2.27
5|6   1.1907     0.0656   18.16
```

This ordinal regression result confirms our ROC analysis: grammatical sentences were associated with significantly higher ratings and were associated with greater variance in the underlying decision variable. In this experiment, we found no significant effect of D-linking on the location of the decision variable; however the negative WhP scale coefficient indicates that sentences with d-linked WhP phrases were associated with relatively less variance in the underlying decision variable[7].

Finally we note it is possible to incorporate random-effects structure into a “mixed” ordinal logistic regression. In *ordinal/*`clmm` this is limited to the location parameter (~ d’). For a more powerful set of options in the Bayesian framework, we point the reader to the *brms* package via the Bürkner & Vuorre (2019) tutorial.

[7] This result, marginally significant at p < .10, suggests a clear path for future research: that *d*-linking can affect acceptability judgments not (only) by changing the location of the latent “Acceptability” variable, but by narrowing its distribution. This is potentially compatible with an interpretation of the d-linking effect according to which wh-dependencies involving lexically restricted wh-phrases are more likely to be correctly parsed.

6 References

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

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412.

Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics*, 46(2), 273-330.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 32-68.

Bock, K., & Middleton, E. L. (2011). Reaching agreement. *Natural Language & Linguistic Theory*, *29*(4), 1033-1069.

Bürkner, P. C., & Vuorre, M. (2019). Ordinal regression models in psychology: a tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77-101.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Jouffrnal of verbal learning and verbal behavior*, *12*(4), 335-359.

Cowart, W. (1997). *Experimental syntax*. Sage.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: theoretical developments with applications to recognition memory. *Psychological review*, *109*(4), 710.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

Dillon, B., Staub, A., Levy, J., & Clifton Jr, C. (2017). Which noun phrases is the verb supposed to agree with?: Object agreement in American English. *Language*, *93*(1), 65-96.

Dillon, B., Andrews, C., Rotello, C. M., & Wagers, M. (2019). A new argument for co-active parses during language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(7), 1271.

Drummond, A. (2013). Ibex farm. *Online server: http://spellout. net/ibexfarm*.

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological review*, *117*(3), 831.

Efron, B. and Tibshirani, R. (1993) An Introduction to the Bootstrap. Chapman & Hall.

Featherston, S. (2009). Relax, lean back, and be a linguist. Zeitschrift für Sprachwissenschaft, 28(1), 127-32.

Franck, J. (2011). Reaching agreement as a core syntactic process. *Natural Language & Linguistic Theory*, *29*(4), 1071-1086.

Fukuda, S., Goodall, G., Michel, D., & Beecher, H. (2012). Is Magnitude Estimation worth the trouble. In Proceedings of the 29th West Coast Conference on formal linguistics (pp. 328-336). Somerville, MA: Cascadilla Proceedings Project.

Gahl, S., Jurafsky, D., & Roland, D. (2004). Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 432-443.

Goodall, G. (2015). The D-linking effect on extraction from islands and non-islands. *Frontiers in psychology*, *5*, 1493.Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of experimental psychology: learning, memory, and cognition*, *27*(6), 1411.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*(3), 839-843.

Häussler, J., Grant, M., Fanselow, G., & Frazier, L. (2015). Superiority in English and German: Cross‐Language Grammatical Differences?. *Syntax*, *18*(3), 235-265.

Hautus, M. J. (1997). Calculating estimates of sensitivity from group data: Pooled versus averaged estimators. *Behavior Research Methods, Instruments, & Computers*, *29*(4), 556-562.

Heit, E., & Rotello, C. M. (2014). Traditional difference-score analyses of reasoning are flawed. *Cognition*, *131*(1), 75-91.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*(4), 434-446.

Kush, D., Lohndal, T., & Sprouse, J. (2018). Investigating variation in island effects. *Natural language & linguistic theory*, *36*(3), 743-779.

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).

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202-1241.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79, 328-348.

Liu, C. C., & Smith, P. L. (2009). Comparing time-accuracy curves: Beyond goodness-of-fit measures. *Psychonomic Bulletin & Review*, *16*(1), 190-203.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*(3), 312-319.

Ma, H., Bandos, A. I., Rockette, H. E., & Gur, D. (2013). On use of partial area under the ROC curve for evaluation of diagnostic performance. *Statistics in medicine*, *32*(20), 3449-3458.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychological bulletin*, *98*(1), 185.

Macmillan, N. A., & Creelman, C. D. (2005). Detection Theory: A User's Guide Lawrence Erlbaum Associates. *New York*, 73.

Macmillan, N. A., Rotello, C. M., & Miller, J. O. (2004). The sampling distributions of Gaussian ROC statistics. *Perception & Psychophysics*, *66*(3), 406-421.

Mauner, G. (1995). Examining the empirical and linguistic bases of current theories of agrammatism. *Brain and Language*, *50*(3), 339-368.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of psycholinguistic research*, *29*(2), 111-123.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67-91.

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, *139*(6), 1173.

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological review*, *99*(3), 518.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 763.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), 77.

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*(4), 944-954.

Schutze, C. T. (1996). The empirical base of linguistics. *Grammaticality Judgments and Linguistic Methodology. Chicago: The University of Chicago*.

Schütze, C. T., & Sprouse, J. (2014). Judgment data. *Research methods in linguistics*, 27.

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, 115(11), 1497-1524.Sprouse, Jon. 2011. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87(2): 274-288

Sprouse, Jon & Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: a journal of general linguistics*, 2(1), 14.

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

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

Sprouse, 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., Yankama, B., Indurkhya, S., Fong, S., & Berwick, R. C. (2018). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review*, *35*(3), 575-599.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *The American journal of psychology*, *69*(1), 1-25.

Venkatraman ES: A Permutation Test to Compare Receiver Operating Characteristic Curves. Biometrics 2000, 56: 1134–1138

Wagers, M.. (2013). Memory mechanisms for wh-dependency formation and their implications for islandhood. *Experimental syntax and island effects*, 161-85.

Wagers, M. W., & Phillips, C. (2014). Going the distance: memory and control processes in active dependency construction. *The Quarterly Journal of Experimental Psychology*, *67*(7), 1274-1304.

Warstadt, A., Singh, A., & Bowman, S. R. (2018). Neural Network Acceptability Judgments. *arXiv preprint arXiv:1805.12471*.

Weskott, T., & Fanselow, G. (2011). On the informativity of different measures of linguistic acceptability. *Language*, 249-273.