RESEARCH

Quantifying sentence acceptability measures: Reliability, bias, and variability

Steven Langsford[1], Amy Perfors[2], Andrew T. Hendrickson[3], Lauren A. Kennedy[1] and Danielle J. Navarro[4]

[1] The University of Adelaide, School of psychology, AU

[2] University of Melbourne, School of psychological sciences, AU

[3] Tilburg University, Department of cognitive science and artificial intelligence, NL

[4] The University of New South Wales, School of psychology, AU

Corresponding author: Steven Langsford (steven.langsford@adelaide.edu.au)

Understanding and measuring sentence acceptability is of fundamental importance for linguists, but although many measures for doing so have been developed, relatively little is known about some of their psychometric properties. In this paper we evaluate within- and between-participant test-retest reliability on a wide range of measures of sentence acceptability. Doing so allows us to estimate how much of the variability within each measure is due to factors including participant-level individual differences, sample size, response styles, and item effects. The measures examined include Likert scales, two versions of forced-choice judgments, magnitude estimation, and a novel measure based on Thurstonian approaches in psychophysics. We reproduce previous findings of high between-participant reliability within and across measures, and extend these results to a generally high reliability within individual items and individual people. Our results indicate that Likert scales and the Thurstonian approach produce the most stable and reliable acceptability measures and do so with smaller sample sizes than the other measures. Moreover, their agreement with each other suggests that the limitation of a discrete Likert scale does not impose a significant degree of structure on the resulting acceptability judgments.

1 Introduction

Acceptability judgments have formed a large part of the study of language since at least Chomsky (1965). They are one of many sources of evidence, alongside corpus linguistics (Sampson 2007), psychological experiments (Noveck & Reboul 2008), and neuroscience techniques (Shalom & Poeppel 2007), that each offer distinct and complementary information about language (Arppe & Järvikivi 2007). One major factor in the popularity of acceptability judgments is the way they allow theories to be tested against artificial constructions that passive observation would rarely or never provide (Schütze 1996). For instance, acceptability judgments can differentiate between constructions that are ungrammatical and those that are rare or missing but still grammatical.

Acceptability judgments come in a number of possible forms, each with their own advantages and disadvantages. The main differences between different forms are in the kind of response required from the participant. People can be offered a discrete rating scale, a real-valued scale, or be asked to make a relative comparison between items. The choice of what response options to offer is critical in two important respects: it determines the statistical tests available to researchers, and it may also significantly influence

people's interpretation of the task. For these reasons, the characteristics of different kinds of acceptability measures are well studied. We know that acceptability judgment data are influenced by details such as the selection of participants (Dąbrowska 2010), sample size (Mahowald et al. 2016), task structure (Featherston 2008), participant engagement (Häussler & Juzek 2016), and data processing decisions (Juzek 2015).

Most of the existing literature focuses on the question of to what extent acceptability judgment data can be used to adjudicate about individual phenomena or effects of linguistic interest (e.g., by presenting pairs of sentences that capture a specific contrast relevant to a particular theoretical claim). However, one might be interested in evaluating the range of acceptability measures along other dimensions as well. To what extent do acceptability judgments from different elicitation tasks support claims about larger-scale generalizations across many different sentences and phenomena? To what extent do different measures of acceptability agree with each other about specific items or sentences? To what extent is each measure robust to differences within individuals at different time points? This paper focuses on exploring these questions.

In the work presented here, we attempt to quantify the extent to which acceptability judgment data from a variety of different elicitation tasks supports different kinds of claims: claims about the global structure of acceptability across a large set of diverse sentences, claims based on the magnitude of acceptability differences, and claims made at the level of single items or sentences. We accomplish this by quantifying the relative contribution of multiple factors – individual participant differences, sample size, task structure, and response-style mitigation in data processing – to the empirical reliability of acceptability scores over specific items (rather than over specific effects) for different measures. We chose to focus on reliability because reliability places a ceiling on how appropriate acceptability judgments are as a test of linguistic theories. If acceptability judgments for some measure or in the presence of some factor are not reliable, we should be cautious about relying on them. Moreover, understanding what factors influence the reliability of a measure can be informative about exactly what that measure reflects.

Our approach aims to differentiate between possible sources of bias and variance. It is currently unclear what proportion of the variability seen in acceptability judgment data is due to lapses of attention, idiolect differences between participants, differences in interpretation of acceptability scales, or interference from simultaneously presented items.

A standard response to the diversity of potential sources of variability is to give them all equal status as *noise independent of the linguistic effect* and ask what can be concluded about true linguistic effects (focused on specific phenomena) in the presence of this noise, regardless of its source. An extensive literature explores this question, looking at the chance of identifying an effect where none exists (Sprouse & Almeida 2011; Sprouse et al. 2013), the chance of failing to identify an effect that is truly present (Sprouse and Almeida 2017a), and differences in sensitivity of different measures compared on a particular known effect (Weskott & Fanselow 2011). The consensus of such studies is that acceptability judgments are highly reliable across replications (Sprouse & Almeida 2017b).

As this literature shows, differentiation between different sources of bias and variance is not strictly necessary in order to test specific linguistic **effects**, which are the primary currency of linguistic research. Many measures of sentence acceptability have good psychometric properties when they are used for such a purpose (e.g., testing whether a set of sentences licensed under some linguistic theory have different acceptability than a set of sentences that are not licensed). If such differentiation is not necessary, why are we attempting to do so here?

The first reason is that such differentiation is important if we want to use acceptability judgments to explore questions that are not focused on hypothesis testing about specific

linguistic effects. For instance, it is quite possible that the nature of the elicitation task may impose structure on the *overall distribution* of acceptability scores across multiple kinds of sentences. Thus, understanding to what extent different tasks do this is important for investigations of the global structure of acceptability in language. Such investigations would include issues like the extent to which clustering structure may be apparent in acceptability judgments (Hofmeister et al. 2013; Lau et al. 2016), whether there are dialect or language differences in global acceptability structure, or whether low acceptability sentences show greater variability than high acceptability ones. Indeed, global acceptability judgments (if they are reliable) may even provide a means to differentiate between dialects or evaluate the knowledge or fluency of individual speakers.

The second reason we are interested in distinguishing between different sources of variability is the expectation that some of these sources fall under an experimenter's control and can be minimized. Different elicitation tasks may vary in their vulnerability to particular sources of variability, which affects their relative quality as scientific instruments. In general, a task that is more difficult might be expected to incur greater variability due to distraction or mistaken responding. Tasks with a small number of unambiguous response options, such as forced choice tasks, may be less vulnerable to response style variability than tasks with flexible free response options that are open to differences of interpretation, such as magnitude estimation. Conversely, forced choice tasks may be more vulnerable to item neighborhood effects, with sentences potentially processed differently in the context of a contrast rather than in isolation.

How much do these tasks vary and how large are these different sources of variation? Our goal is to provide a quantitative answer to this question.

The many possible sources of bias and variability cannot be completely disentangled, since they are generally all present in some unknown degree in every response. We give quantitative bounds for the distinct contribution of certain sources of variability in two different ways.

First, we contrast between and within participant test-retest reliability. Between-participant test-retest reliability is an important metric of measure quality in its own right, since no strong conclusions can be drawn from the results of a measure if it is liable to give different answers to the same question on different occasions (Kline 2013; Porte 2013; Brandt et al. 2014; DeVellis 2016). While distinct in the way it avoids appealing to a ground truth, between-participant test-retest reliability is closely related to error-rate reliability, if the underlying truth is considered stable over the time scales involved. As such, it is widely reported in existing work on the reliability of acceptability judgment data (Sprouse et al. 2013; Sprouse & Almeida 2017a). However test-retest reliability within the same participant can offer additional information, especially when contrasted with between-participant reliability. This contrast, which has no analogue in error rates, is informative about the composition of the variability: variability inherent to the construct itself and random noise due to inattention or other error can be expected in both, while individual differences in response style and subjective acceptability only contribute to the variability of between-participant replications. As a result, between-participant replications are expected to be less reliable, and the size of the reliability gap quantifies the combined impact of these particular sources of variability. Even further decomposition into the source of this within/between reliability gap is possible as well. For instance, the variability due to response style differences can be estimated by examining the effect of data pre-processing steps (e.g., z-transformation of scores) known to mitigate this particular source of variability.

Second, we contrast these within and between participant test-retest reliability results for measures based on different tasks. The tasks differ primarily in the kind of response

options offered, which could potentially impose structure on results. For example, asking people to give responses on a discrete Likert scale might force them to collapse distinct acceptabilities onto one response if there are too few options or encourage them to make spurious distinctions if there are too many (Schütze 1996; Carifio & Perla Likert 2008; Schütze 2011). The comparisons involved in forced choice judgments could also direct people's attention to specific syntactic details, particularly when the two sentences are related, as is typical of a well-controlled test pair. This might lead to different acceptability ratings than if each sentence was considered in isolation (Cornips & Poletto 2005). Contrasts between measures are therefore useful both in identifying the best-performing measures (Sprouse et al. 2013; Sprouse & Almeida 2017a) and to test the degree of agreement between them (Schütze 2011; Weskott & Fanselow 2011; Sprouse & Almeida 2012).

However, from the perspective of decomposing sources of bias and variance, distinct tasks may also be differently vulnerable to different sources of variability. As a result, we may be able to use them to cross-check against each other's potential biases.

The structure of this paper is as follows. We first give a detailed introduction to the measures considered in this paper, the processing steps and statistical tests associated with each, and the series of experiments within which we collect the data. When reporting the results our primary focus is on test-retest reliability; it is first evaluated in terms of raw score correlation of all sentences in a dataset, then in terms of the decisions yielded by each measure on particular contrasts of interest. For each of these we compare within and between participant reliability and examine the impact of sample size. We conclude by examining the mutual agreement between the measures, with reference to expert judgments in the published literature. In the discussion, we address limitations of this work, consider recommendations for researchers interested in measuring sentence acceptability, and discuss future directions.

1.1 The measures

Early work on the reliability of formal measures was prompted by concerns about the practice of "armchair linguistics", which considered phrases or sentences as the primary unit of evidence on which linguistic theories were built, taking for granted that the acceptability status of these sentences would be immediately obvious to a native speaker. With reference to previously discredited introspective approaches in psychology (Danziger 1980), critics pointed out that the intuitions of a linguist about a sentence they constructed themselves to demonstrate a particular point of syntax might not be the same as those of the broader language community (Spencer 1973; Schütze 1996; Wasow & Arnold 2005; Dąbrowska 2010).

Proponents of informal approaches argued in response that linguists were mainly concerned with phenomena that gave very large effect sizes, making multiple opinions on a particular acceptability difference redundant (Phillips & Lasnik 2003; Featherston 2009; Phillips 2009). This approach defended the legitimacy of the large literature built on such informal tests, but left open the question of how to decide what counts as an obvious case (Linzen & Oseki 2015).

Recent systematic work comparing expert and naive judgments has largely supported the argument that the majority of claims published in the linguistics literature are consistent with the results of formal tests against the judgments of large numbers of naive native speakers (Culbertson & Gross 2009; Sprouse & Almeida 2012; Sprouse et al. 2013). However the same program of research has shown that even for contrasts with large effect sizes, formal tests offer more information than informal ones. As well as giving an objective measure of whether a test sentence is more or less acceptable than a control to a language community, a formal test can

also give an indication of the size of the difference, and the relative acceptability of both sentences on a global acceptability scale (Sprouse and Schütze 2017). It has also been argued that as a result of much productive work on large effects, smaller effects have become increasingly important to further progress (Gibson & Fedorenko 2013; Gibson et al. 2013).

One potential drawback of formal methods is their higher cost in time and participant-hours. However, as Myers (2009) points out, more representative samples and quantitative replicability need not be prohibitively expensive or complicated. Moreover, cost depends in part on the measurement task as well as the question being asked. For instance, many fewer judgments are required for a forced-choice task on an "obvious" effect (Mahowald et al. 2016) than for answering finer-grained questions about statistical power or sensitivity (Sprouse & Almeida 2017a).

Our goal in this paper was to evaluate all of the most commonly used formal measures of sentence acceptability, as well as variants on them, in order to isolate and expose the impact of task-specific assumptions. The primary distinction between existing measures is whether they ask participants to give each sentence a rating on a scale of some sort (a rating task) or make a choice between two sentences (a choice task). The two rating tasks we consider are LIKERT scales and Magnitude Estimation (ME), while the two choice tasks involve either deciding between two related sentences (TARGET PAIRS) or two random sentences (RANDOM PAIRS). This yields four separate tasks, but for two we separately evaluate alternative statistical methods for transforming the raw results, giving six distinct measures. One task for which we consider multiple analyses is magnitude estimation, where scores can be log transformed (ME(LOG)) or both log and z-transformed (ME(z-SCORE)). The other is the judgments involving random sentence pairs, which can either be used directly or input into a THURSTONE model based on a standard measurement approach in psychophysics.

The six measures, ME(z-SCORE), ME(LOG), LIKERT, THURSTONE, TARGET PAIRS, and RANDOM PAIRS are described in more detail in the Method section. One reason for this choice of tasks is to reflect current practice: LIKERT, TARGCHOICE, and ME are probably the most common instruments for eliciting acceptability judgments (Podesva & Sharma 2014). However another consideration is their diversity of assumptions. In particular, LIKERT and ME each supply a particular rating scale, while the choice tasks do not. A key contribution of this paper is the presentation of the THURSTONE model, which allows comparisons between these perspectives by inferring scale structure from choice data (Thurstone 1927). The THURSTONE model is capable of representing a wide range of latent acceptability structures: the degree of consistency between the structure inferred from choice task data and rating task data gives an indication of the extent to which the researcher-supplied scales impose structure on people's responses.

1.2 Measure evaluation

In this paper we systematically investigate three criteria for evaluating each of the six measures: test-retest reliability, agreement, and robustness to sample size.

Measure agreement is an important check of validity for diverse measures claiming to reflect the same underlying construct. Here we are also interested in the vulnerability of different measures to different sources of noise, with the goal of allowing researchers to minimize the variability in results that are due to controllable properties of the elicitation task rather than the linguistic construct of interest. Although robustness to sample size is not directly related to the decomposition of measure variability and bias that is the main focus of this paper, we include it as important information for readers interested in the implications of this work for study design.

Test-retest **reliability** can be defined at various levels from responses (when repeating questions within-participants) to items (an aggregation of many responses) to effects (which aggregate over many theoretically-related items). Here we are primarily concerned with the item level, for several reasons. First, effect-level reliability is already well studied. Second, including only one item per effect (as we do) allows us to maximize variability across items and thus creates a much more stronger test of each *measure*. If a measure is highly reliable even across an extremely varied sentence set, this is more informative than finding that it is reliable along a more narrow set of stimuli. Finally, item-level reliability is not itself well-studied, yet is theoretically important: if people's judgments about specific items are reliable for a given measure, a much wider range of theoretical claims about language are open to study with this data type.

The assessment of reliability depends in part on the nature of the hypothesis being tested. Some researchers might be particularly interested in a *decision* problem: determining whether people make different judgments for two different sentences or kinds of sentences. Others might be interested in an *estimation* problem, being able to accurately position sentences relative to each other on an acceptability scale. In this paper we evaluate reliability using both kinds of assessment. For a decision problem, we rely on statistical significance testing of the difference between acceptability scores produced by a particular measure for the two sentences. This allows us to precisely characterize our uncertainty in the estimate of the difference for each pair of sentences, and compare that degree of uncertainty across measures in a principled way. For estimation problems, we calculate correlations between scores from different time periods or people. Reliability at this level of detail is relevant to claims about the overall structure of acceptability, for example whether or not it exhibits strong clustering (Sprouse 2007).

A secondary factor we focus on is **sensitivity to sample size**. We do this by systematically repeating our reliability analyses with the judgments derived from different sample sizes of participants and comparing this to the results from the full sample. This is directly useful in estimating the sample size required for a target level of reliability in studies using these measures. It also gives an indication of how efficiently these measures are able to extract information from responses; this is useful because different methods might take different numbers of trials to produce reliable answers (Li et al. 2016; Sprouse & Almeida 2017a).

Our final factor of interest is the **agreement** between measures. This is of interest not only because substantial agreement suggests that the measures reflect genuine acceptability judgments rather than superficial measure-specific behavior, but also because such agreement provides converging evidence about the nature of those judgments. Cross-measure agreement is better studied than reliability (Schütze 2011; Weskott & Fanselow 2011; Sprouse & Almeida 2012), but still has not been investigated within the full array of measures we consider. It is therefore valuable as a replication and extension of previous work.

2 Method

2.1 Sentences

In order for the comparisons to be fair, all of the measures are evaluated on the same set of sentences. Sprouse et al. (2013) selected these sentences from a subset of English data points published in *Linguistic Inquiry* between 2001 and 2010.

Sprouse et al. (2013) subdivide these sentences into 148 distinct linguistic phenomena, roughly corresponding to 150 distinct sources (with two instances where different sources discussed the same construction). Each linguistic phenomenon was then represented by

multiple items (eight instances). Since our focus is not on the content of any particular linguistic claim, we selected one matched pair of acceptable/unacceptable items at random from the 150 distinct sources to create a set of 300 sentences. This decision limits our ability to make claims about the status of any particular phenomenon, since each is represented by a single item. However, our focus is on the reliability and variability inherent to specific *measures,* and for this the diversity of sentences is a significant advantage: it is important to evaluate them over the full range of sentence acceptability levels and effect sizes. In addition, we can also estimate the variability inherent to individual items. The full list of sentences is in Appendix A.

2.2 Measures

Our reliability and sample size analyses involve comparing the six different measures of sentence acceptability described above. When analyzing agreement, we additionally include informal expert judgments from the published literature (Informal). The procedures for deriving scores and significance tests for each measure are given below, followed by the details of data collection. Table 1 summarizes this information.

2.2.1 Informal

The Informal measure captures the binary judgments presented in the *Linguistic Inquiry* journal for each of the sentences in question. For each of the 150 pairs, one sentence was judged to be acceptable and one was unacceptable (as noted with a judgment diacritic like * or ? in the journal). We include this measure because of the intense interest in comparing informal and formal methods (Featherston 2007; Munro et al. 2010; Myers 2012; Sprouse & Almeida 2012; Gibson & Fedorenko 2013; Sprouse et al. 2013), although our main focus in this paper is on evaluating the test-retest reliability and mutual consistency of the formal methods.

One important caveat for the interpretation of the comparison with Informal results in this paper is the fact that each phenomenon is represented here by a single example sentence, rather than the multiple items as is the usual practice for formal studies (Myers 2009). For our purposes (i.e., investigating item-level reliability and especially the extent to which acceptability judgment data supports tests of global structure), this feature of the item set is an advantage: to the extent that different instances of the same phenomenon have similar acceptability, using one item per phenomenon gives the maximum variability over the item set and maximum coverage over the acceptability space. However

Table 1: Method summary: we examined four different tasks, two choice tasks and two rating tasks, analyzing two of these in two different ways for a total of six distinct measures. For each of these measures, we evaluate the set of acceptability scores for all sentences (which supports comparisons using Pearson correlations) as well as decisions made on pairs of sentences (which allows us to focus on targeted contrasts between two particular sentences in a hypothesis-testing framework).

Task	Measure	Sentence Score	Hypothesis test
Targeted contrasts	Target pairs	Proportion endorsements	Difference of proportions
Random contrasts	Random pairs	Proportion endorsements	Difference of proportions
Random contrasts	Thurstone	Mean posterior acceptability	Credible interval
Magnitude estimation	ME (log)	Mean of log responses	t-test
Magnitude estimation	ME (z-score)	Mean of z-transformed log responses	t-test
Likert rating	Likert	Mean of z-transformed ratings	t-test

it also means that there is some risk that any specific phenomenon in question will be represented by an atypical example.

Since typically only a single instance of an INFORMAL judgment is available for any particular item, test-retest reliability does not apply, so they are assessed only in terms of cross-measure agreement.

2.2.2 Likert

In a typical LIKERT task, each sentence is presented with a series of possible acceptability rating options. This task is widely used in the psychological literature (Likert 1932; Hartley 2014) and is generally considered fairly intuitive. LIKERT scales are one of the most widely-used formal measures of linguistic acceptability (Schütze & Sprouse 2014) and have been shown to substantially agree with informal judgments (Sprouse et al. 2013), with experts and non-linguists coming to largely the same conclusions (Culbertson and Gross 2009). In our experiments, following Sprouse et al. (2013) and Mahowald et al. (2016), we aggregated LIKERT scores by first converting each individual participants' responses to z-scores. The acceptability score for each sentence was thus the average of all z-scores associated with it. This normalization scheme mitigates the impact of individual differences in response style.

2.2.3 Magnitude estimation

ME tasks were developed to estimate the relative magnitude of differences between items by supplying interval data (Bard et al. 1996). In this procedure, adapted from psychophysics (Stevens 1956), participants are given an initial reference item to calibrate their judgments and then asked to compare other items by assigning them any positive real number. Although unable to provide true ratio data as initially claimed (Sprouse 2008; Weskott & Fanselow 2008; Sprouse 2011a), ME is still commonly used (Cowart 1997; Keller 2003; Featherston 2005; Murphy et al. 2006; Johnson 2011; Schütze 2011; Erlewine & Kotek 2016). Interpreted as a linear scaling task rather than a direct recording of people's mental representations, it is distinct from other measures in the extreme freedom it gives for arbitrarily precise responses, although whether that extra variability actually encodes information about linguistic effects has been questioned (Weskott & Fanselow 2011). ME has been shown to agree with other forms of acceptability judgment (Keller & Asudeh 2001; Weskott & Fanselow 2009).

The typical aggregation scheme for ME data in linguistics, following Bard et al. (1996), is to average the log of the raw scores associated with each item (Sorace 2010; Weskott & Fanselow 2011). Originally, this was because the log transform is natural for ratio data, which is the form requested in the instructions to participants. Recent work has shown that participants are in general unable to produce responses conforming to the properties of true ratio scales (Sprouse 2011a), and it may in fact be impossible to do so since acceptability does not have a clearly defined zero point. We adopt the log transformation here primarily because it has historically been a standard approach for reducing the impact of the outliers typical of ME data. Moreover, other possibilities, such as trimming the data or Windsorizing, would remove information.

In order to evaluate the role played by response style differences, we additionally investigate the impact of also applying a z-transformation, which is sometimes recommended for ME scores for that purpose (Featherston 2005; Sprouse & Almeida 2011; Fukuda et al. 2012). The z-transformation mitigates response style differences in two ways. First, participant ratings are scored relative to their mean rating (which compensates for individual differences in which part of the scale people use) and distances are expressed in standard deviation units (which compensates for individual differences in the range of the scale that they use).

By contrasting the test-retest reliability of ME data both with and without the z-transform applied, it is possible to see how effective it is in mitigating response style differences. Specifically, the z-transformation is predicted to improve reliability in the between-participant replication to a much greater extent than the within-participant replication, since response style differences are a between-participant source of variability. To the extent that it is effective, this contrast gives an indication of the degree to which variation in ME scores can be attributed to variability in people's usage of the scale.[1]

2.2.4 Target pairs

The TARGET PAIRS judgment task asks people to select the more acceptable sentence of two candidates specifically chosen to isolate a particular contrast of theoretical interest. This is perhaps the simplest measure. By focusing only on the differences that are of theoretical interest, this measure increases the statistical power for determining the differences within those targeted pairs, but sacrifices the ability to compare pairs to one another. The TARGET PAIRS comparison is widely used (Rosenbach 2003; Myers 2009) and has been shown to substantially agree with informal judgments (Sprouse & Almeida 2011; Sprouse et al. 2013).

We consider acceptability scores in the TARGET PAIRS task to be the proportion of times the preferred option was chosen, without distinguishing between responses indicating equal acceptability or the alternative option. Unlike the other aggregate scores, this measure does not capture global structure, since decisions regarding each pair are isolated by design. The primary outcome of interest for this measure is the outcome of the significance test of the estimated proportion ($\hat{P}$) with respect to the number of judgments (N). This was calculated by determining if the 95% confidence interval around the estimated proportion included random guessing (0.5). If the interval did not include 0.5, the null hypothesis that people did not prefer one sentence over the other was rejected. The standard formula for calculating a confidence interval around a proportion was used:

$\hat{p} \pm Z_{crit} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$ where $Z_{crit} = 1.96$.

2.2.5 Thurstone

The THURSTONE measure, which has a long history in psychophysics (Thurstone 1927; Roberts et al. 1999; Fabrigar & Paik 2007), is used to make inferences about the subjective perception of stimuli based on forced-choice comparison data.

The basic idea is to ask people to make acceptability judgments about a random subset of pairs drawn from a large set of stimuli (for us, this corresponds to asking people to give forced-choice judgments on two sentences sampled at random from the full set of 300 sentences). It is important that the pairs are random rather than the theoretically-motivated pairs as in the TARGET PAIRS task because comparing each sentence to many others imposes strong statistical constraints on the set of possible orderings of all of the sentences (Thurstone 1927). Distances on the inferred acceptability scale are given meaning by the model's mapping between acceptability differences and probability of endorsing a particular response. The observed responses constrain the plausible outcomes for responses in unobserved comparisons, assuming transitivity of acceptability. As a result, only a small subset of all possible pairs is necessary to make inferences about the acceptability of all of the sentences relative to each other. Technical details for the THURSTONE measure are described thoroughly in Appendix B.

---

[1] We also ran all of these analyses with raw judgments (no transformations at all), judgments receiving only the z-transform (rather than z and log), or judgments that were converted to ranks. None had superior reliability than LIKERT or THURSTONE, and judgments that did not incorporate some way of taming outliers did not produce meaningful results.

The THURSTONE model represents the acceptability of each sentence as a distribution over an inferred acceptability scale. To derive an overall acceptability score from this measure, we simply take the mean of each distribution as representing the acceptability score for that sentence. For a decision rule corresponding to the significance test in other measures, we constructed a credible interval over the difference between sentences. The distribution of credible differences was generated by repeatedly sampling from each posterior and taking the difference of those samples. The result was considered inconclusive if the range between 0.025 and 0.975 quantiles included 0, otherwise the observed difference was considered significant.

The THURSTONE model has a well-established record of performance in other domains that require inferring latent acceptability orderings, such as product preferences in marketing research (O'Mahony et al. 2003; Ennis 2016). It is also a prominent tool in the *wisdom of crowds* literature, where it is used to define a meaningful consensus aggregating over individual judgments that cannot be simply be averaged together (Miller et al. 2009; Selker et al. 2017). Previous work on experimental syntax methodology has identified forced choice tasks as a particularly sensitive and reliable method of eliciting acceptability judgments (Sprouse et al. 2013; Schütze & Sprouse 2014), while noting that they are restricted in the way they give limited ordinal information about only the particular sentences involved in the contrast at hand. The THURSTONE method retains main benefits of this task type, which are the simple unambiguous response options and the way individual items can target arbitrarily small acceptability differences, while also aggregating information over all responses to derive a real-valued acceptability score that is directly comparable over all items. By providing real-valued data on a psychologically meaningful scale (Nosofsky 1992; Borg & Groenen 2005), THURSTONE modeling draws on much of the same motivation that originally drove the adoption of ME (Schütze 2011). By shifting the responsibility for quantifying acceptability from participants to a measurement model, it avoids problems associated with the difficulty people have using the number line in the requested way (Sprouse 2008; 2011a).

2.2.6 Random pairs

The THURSTONE model requires choice task data over random pairs rather than the theoretically related pairs that are usually compared in a choice task. This means that we have a dataset – the raw scores on RANDOM PAIRS – which provides a baseline against which to compare the THURSTONE and the TARGET PAIRS measures. Analyzing the RANDOM PAIRS measure may be helpful in both determining how much of the performance of the THURSTONE measure depends on the model, as well as in quantifying the impact the choice of contrast sentence has in the TARGET PAIRS task.

As in the TARGET PAIRS measure, the proportion of trials in which a sentence was endorsed over the alternative or the both-equal option was taken as its overall acceptability score. Unlike TARGET PAIRS, this is an estimate of global acceptability across the whole set of sentences considered, albeit a noisy measure that depends on the randomly sampled set of alternative sentences each sentence appeared with. The significance test applied was the same test of proportion equality applied to TARGET PAIRS.

2.3 General procedure

To examine within and between participant reliability, three data sets were needed, an INITIAL reference set, followed by a WITHIN PARTICIPANTS replication and a BETWEEN PARTICIPANTS replication. Participants involved in the BETWEEN PARTICIPANTS replication gave the series of acceptability judgments used in the INITIAL dataset. They then performed a short distractor task designed to interfere with their ability to remember

their answers to particular items, after which they repeated the same set of acceptability judgments (in a different random order) to create the BETWEEN PARTICIPANTS data set. A second group of participants was recruited to supply the BETWEEN PARTICIPANTS data set: the same procedure was used, except that these participants did not see the distractor task or give a second set of judgments.

In order to keep the time commitment per participant under approximately 30 minutes, we divided the four tasks into two groups that were presented to the same set of participants, with the RANDOM PAIRS task grouped with the LIKERT rating task, and the TARGET PAIRS task grouped with the ME task. With these groupings each participant saw one choice task and one rating task, which minimized possible fatigue due to always making the same type of judgment or interference between similar task types. Furthermore, requiring participants to complete more than one task increases the time and attention expended between responses to identical items. This decreases the chance that responses reflect an explicit memory of the first judgment for the WITHIN PARTICIPANTS replication.

In the first set of trials (the first half of the experiment for the INITIAL/WITHIN PARTICIPANTS group, the entirety of the study for the BETWEEN PARTICIPANTS group) participants saw two blocks (one rating task and one choice task of 42 trials each. The order of tasks within a block was randomized for each participant, and the order of items within each task was randomized on each presentation of a block. Each block contributed 40 trials to the data analysis. The additional two questions were attention checks designed to have a clear correct answer, used only to exclude participants whose incorrect responses indicated either inattention or misunderstanding of the task (see Appendix A). Participants involved in the BETWEEN PARTICIPANTS study completed at this point, while those involved in the WITHIN PARTICIPANTS study then did the distractor task, followed by a repetition of the exact same trials, with the same task order as the initial presentation but a re-drawn random order of items within each task. No sentences were repeated in different items for any one participant. Each participant thus saw only a random subset of the 300 sentences, but across participants all sentences were seen a similar number of times.

The distractor task was based on a change blindness demonstration (Simons & Rensink 2005). We chose it because it is non-linguistic and known to be a very attention-grabbing task (Rensink et al. 1997). During it, people were shown two images that were identical except for one difficult-to-identify discrepancy: for instance, one showed a city street in which the window of one of the buildings was present in one image and absent in the other. The images were presented sequentially and repeatedly for 800ms each with an 800 ms white mask in between. Participants were asked to identify the discrepancy and click on it. Once they had done so or thirty seconds had elapsed, they were shown another pair of alternating images. There were six such images. Because the point of this task was just to provide a break between the acceptability judgment tasks, performance was not analyzed.

In all conditions participants saw the same general set of instructions, shown below:

> This study will ask you some questions about the acceptability of sentences. There's no objective standard for what makes a phrase feel "more acceptable", but we're confident that you'll know it when you see it. Some phrases are natural while others are clumsy or just plain wrong, and we expect you'll find it pretty easy to judge how acceptable a phrase is, even across very different topics. There are two different types of question. Some of the questions will ask you to give a sentence an acceptability rating. Others will ask you to compare two sentences and say which one is more acceptable.

All participants were asked to answer two multiple choice questions to make sure they understood the instructions (see Appendix A) before beginning the experiment. Those who did not answer both questions correctly were returned to the instructions page and could not begin until both were answered correctly.
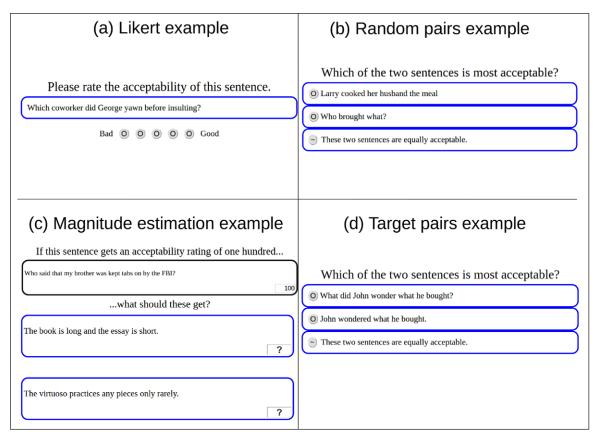
2.4 Task-specific procedures

2.4.1 Random pairs blocks

In these blocks, people were presented with three vertically arranged options. Each was surrounded by a blue border under the title "Which of the two sentences is most acceptable?" The first two options were sentences randomly drawn from the full pool of 300. The third option read "These two sentences are equally acceptable." Participants clicked on a sentence to choose it, as shown in Figure 1(b). As in the LIKERT blocks, a progress marker indicating the item and block number was displayed, and no feedback was given.

2.4.2 Target pairs blocks

These trials were exactly the same as the RANDOM PAIRS trials in the other version of the experiment. The only difference is that the sentences were both in the pair of theoretical interest rather than randomly selected from the entire set; an example is shown in Figure 1(d).

Figure 1: Example trials for each of the four question types. In one version of the experiment ((a) and (b)), participants saw blocks of sentences presented in random order in a LIKERT task and a choice task in which the sentences were randomly drawn from the entire sentence pool (RANDOM PAIRS). In the other version ((c) and (d)), the blocks were either in a magnitude estimation (ME) or typical choice task in which the sentence pairs were theoretically motivated (TARGET PAIRS). For each measure, the associated panel reflects the appearance of a typical trial.

2.4.3 Likert blocks

In these blocks, on each trial people saw a single sentence surrounded by a blue border under the title "Please rate the acceptability of this sentence." Under the sentence was a row of five unmarked buttons labeled "Bad" at the far left and "Good" on the far right, as shown in Figure 1(a). Below this was a progress marker giving the trial and block number. Clicking any of the response buttons disabled them for 500 ms and displayed the next sentence to be judged. No feedback was given.

2.4.4 Magnitude estimation blocks

In these blocks, people saw six pages of seven sentences each. On each page the top of the screen contained a fixed title banner that remained in position when the page was scrolled. It consisted of some reminder instructions ("If this sentence gets an acceptability rating of one hundred … [reference sentence] … what should these get?"). The reference sentence, following Sprouse (2011b), was "Who said that my brother was kept tabs on by the FBI?"), and was surrounded by a black border that also contained a non-editable text box in the lower right corner that was pre-filled with the reference value 100. This was followed by the test items, which were surrounded by a blue border and contained an editable text box in the lower right corner initially containing a question mark. An example is shown in Figure 1(c).

Test items were arranged vertically with seven to a page with approximately two or three test sentences visible at once on the screen and the remaining sentences visible by scrolling. Each set of seven sentences was followed by progress marker and a next button which presented a fresh set of seven sentences, with no option to return to a previously rated set. Input was restricted to positive numbers, and no feedback was given, other than a prompt to give positive number ratings in order to continue if an unparsable or empty input was detected when the next button was clicked.

In order to ensure that people understood the ME task, before they rated any sentences each participant practiced the task on line lengths. They were required to give ratings for six different test lines (relative to a reference line length of 100). There were five test lines presented in random order, with lengths ranging between $\times 0.01$ and $\times 2.5$ of the reference line. Although the exact lengths of test lines were randomized to avoid encouraging participants to only use round numbers, there was one example each of very short (length ∼ 25% of the reference line), short (∼75%), roughly equal (∼125%), long (∼175%), and very long (∼225%) lines.[2] During these practice trials there was feedback on every response, and people did not continue to the next trial until their estimates were within 30 of the correct answer. Participants successfully completing this practice were considered to have understood the process of ME.

2.4.5 Participants

There were four rounds of recruitment to cover the two pairs of tasks (Likert/Target pairs and ME/Target pairs) in two presentation formats (a two-session format giving

---

[2] The ME specific instructions were:

> Some of the questions will ask you to compare the acceptability of each sentence to a standard reference sentence and tell us the result as a number. The standard reference sentence always has an acceptability rating of 100. A sentence that is twice as good should get a rating that is twice as large, a sentence that is half as good should get a rating that is half as large, and so on. Any positive number is a valid rating, please do try to use a wide range of numbers. More detailed responses carry more information about how acceptable you feel the sentences are, and that's really what we're interested in. Having said that, you don't need to spend a lot of time doing a deep analysis of every little detail, we're much more interested in your first impressions.

INITIAL and WITHIN PARTICIPANTS data, and a single-session format giving BETWEEN PARTICIPANTS data).

Two-session LIKERT **and** RANDOM PAIRS 150 adults were recruited via Amazon Mechanical Turk. Participants were paid US$3.00 for an average of 33 minutes work. They ranged in age from 20 to 65 (mean: 34.6) and 81 of them (55%) were male. Fifteen people were excluded from the analysis: three had non-compatible browsers so their data failed to save, one reported being a non-native English speaker, and 11 gave at least one incorrect response to the attention check questions. Of the 135 remaining participants, 133 were from the US and two were from India. Three reported speaking additional languages other than English but all 135 included participants reported being English native speakers.

Two-session ME **and** TARGET PAIRS 160 adults were recruited via Amazon Mechanical Turk. Participants were paid US$4.00 for an average of 38 minutes work. They ranged in age from 19 to 66 (mean: 34.0) and 91 of them (57%) were male. Twenty-five people were excluded from the analysis: one reported being a non-native English speaker, two were found to have participated in the previous round, and 22 gave at least one incorrect response to the attention check questions. Of the 135 remaining participants, 132 were from the US, with one each from India, Chile, and Ireland. One reported speaking an additional language other than English but all 135 included participants reported being English native speakers.

Single-session LIKERT **and** RANDOM PAIRS 150 adults were recruited via Amazon Mechanical Turk. Participants were paid US$1.60 for an average of 21 minutes work. They ranged in age from 22 to 69 (mean: 34.5) and 93 of them (62%) were male. Twenty-three people were excluded from the analysis: two had participated in a previous round, four reported being non-native English speakers, and 17 gave at least one incorrect response to the attention check questions. Of the 127 remaining participants, 125 were from the US, one was from Dominica, and one was from India. Two reported speaking additional languages other than English but all 127 included participants reported being English native speakers.

Single-session ME **and** TARGET PAIRS 151 adults were recruited via Amazon Mechanical Turk. Participants were paid US$3.00 for an average of 31 minutes work. They ranged in age from 18 to 70 (mean: 34.8) and 89 of them (59%) were male. Fourteen people were excluded from the analysis: four reported being non-native English speakers, and 10 gave at least one incorrect response to the attention check questions. Of the 137 remaining participants, 135 were from the US with one participant from Canada and one from the United Kingdom. Three reported speaking additional languages other than English but all 137 included participants reported being English native speakers.
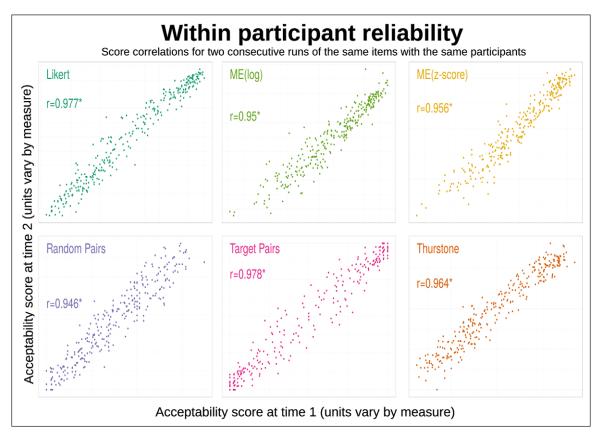
3 Results

We begin by examining the test-retest reliability of the scores derived from each measure. For these analyses, we use the Pearson correlation between scores drawn from the relevant data sets: INITIAL and WITHIN PARTICIPANTS for **within participant reliability** or INITIAL and BETWEEN PARTICIPANTS for **between participant reliability**. Reliability at this level of detail may be required to test claims involving comparisons over more than two items, such as whether or not acceptability exhibits strong clustering, or claims expressed in terms of the degree of difference between items rather than the binary presence or absence of a difference (Sorace & Keller 2005; Gibson et al. 2013).
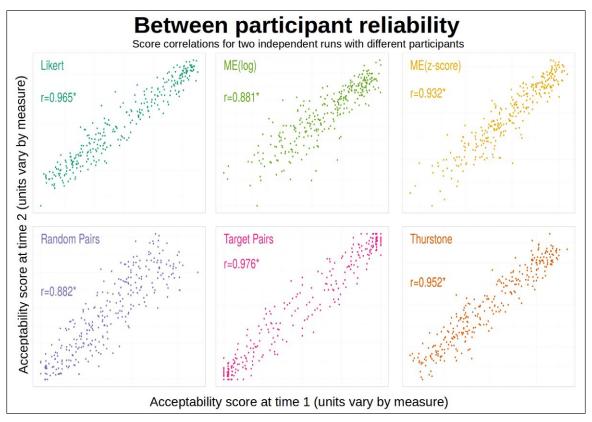
3.1 Global measures

3.1.1 Reliability

We quantify the global reliability of a measure across different data sets using the Pearson correlation between acceptability estimates. Correlations between scores obtained between scores based on the INITIAL dataset and those based on the WITHIN PARTICIPANTS replication data are shown in Figure 2, with the score based on INITIAL responses on the x-axis and scores based on WITHIN PARTICIPANTS replication on the y-axis. The strong linear relationships obtained show that all measures were highly reliable. Test-retest correlations were large and statistically significant for every measure. LIKERT scores and TARGET PAIRS were the most reliable measures.

Correlations between scores obtained in INITIAL data and BETWEEN PARTICIPANTS data are shown in Figure 3, with scores obtained from the INITIAL data on the x-axis and scores obtained from the BETWEEN PARTICIPANTS replication on the y-axis. As in the WITHIN PARTICIPANTS case, all measures were highly reliable, with all correlations large and statistically significant. However, each correlation is somewhat lower than the within-participant counterpart. This extra variation must be driven by those factors unique to the BETWEEN PARTICIPANTS case: either individual differences among the participants in the two participant pools or item effects due to the re-drawing of the items shown to participants (within-participants tests used identical items each time).

Figure 2: Within-participant reliability measured by correlations between sentence acceptability rankings. All of the formal measures aggregate responses into an acceptability score for each sentence. For each, the x-axis reflects the score using that measure in the INITIAL data, while the y-axis reflects the score from that measure in the WITHIN PARTICIPANTS data. The r values indicate Pearson's correlation coefficient, and the stars (*) indicate significance at $p < 0.001$. All measures are both highly linear and highly significant, suggesting that all these measures have good within-participant reliability.
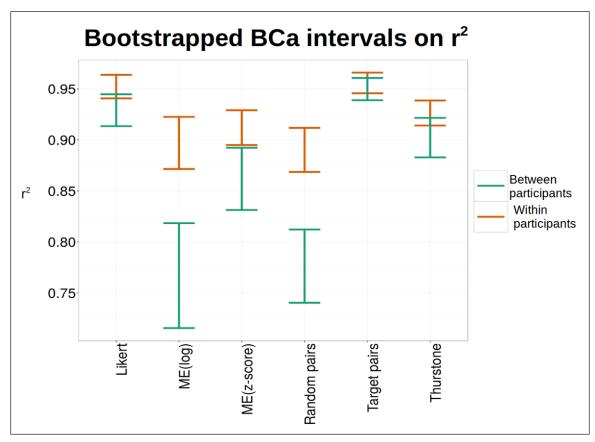
Figure 3: Between-participant reliability measured by correlations between sentence acceptability rankings. All of the formal measures aggregate responses into an acceptability score for each sentence. For each, the x-axis reflects the ranking derived using that measure in the INITIAL data. The y-axis reflects the score from that measure in the BETWEEN PARTICIPANTS data. The r values indicate Pearson's correlation coefficient, and the stars (*) indicate significance at $p < 0.001$. Between-participant reliability is naturally lower than within-participant reliability for all measures, but the relationship between scores derived from the two data sets are still linear and highly significant. All these measures show good between-participant reliability.

Given that all the measures seem to be relatively reliable, we next seek to test whether the relative differences in reliability can be considered significant. One way to test the significance of the differences observed between these correlations is to bootstrap 95% intervals around them. We used the R package *boot* (Davison & Hinkley 1997) to generate adjusted bootstrap percentile intervals (BCa) around the r^2 estimate of variance explained in re-test scores given only scores from the INITIAL data set, assuming linearity.

Results are shown in Figure 4. There is a significant difference in correlations between within-participant and between-participant reliability for the ME and RANDOM PAIRS measures but not the other ones.

The relatively large drop in correlation when moving to from within to between-participant data for the ME scores is most likely driven by individual differences in response styles, as shown by the large reduction in this gap when mitigating response style differences using ME(z-score). In contrast, there is limited scope for response style differences in the RANDOM PAIRS task, so the large drop in reliability when moving from within to between-participants data is likely to reflect the fact that new items were drawn, which gave each sentence a different set of comparison sentences.

What is most noteworthy about these effects is the fact that the LIKERT and THURSTONE scores do not appear to suffer from them. Despite LIKERT ratings being vulnerable in theory to response style differences, these results suggest they do not appear to be a major source of variation in practice. Although the THURSTONE acceptability estimates

Figure 4: Comparing reliability correlations with bootstrapped r^2. The correlations indicating score reliability were compared across scores and across between/within contrasts by bootstrapping a 95% BCa interval. These intervals are of the r^2 for the linear model predicting scores in the second data set (WITHIN PARTICIPANTS or BETWEEN PARTICIPANTS) from scores derived from the INITIAL dataset, and are based on 1000 samples. The results show that ME scores and RANDOM PAIRS scores are significantly impacted by participant and item effects where the other scores are not. TARGET PAIRS is the single most reliable measure. Of the measures allowing global comparisons, the most reliable is LIKERT.

are based on exactly the same responses the RANDOM PAIRS endorsement proportions derive from, they do not show strong item effects, which is a testament to the robust nature of the THURSTONE approach.

3.1.2 Sample size dependence

The analyses so far yield estimates of between- and within-participant reliability of global sentence acceptability judgments for each measure, but all involve our full sample of judgments derived from all included participants. Although even this quantity of judgments is relatively cheap and straightforward using platforms such as Amazon Mechanical Turk, it is important to understand how robust reliability is when sample sizes are lower. By repeatedly dropping some subset of participants at random from the full sample and re-running all analyses on the retained participants only, we obtained estimates for the number of participants required for a given level of reliability up to the level achieved in the full sample. These required-sample-size estimates are directly useful for researchers planning future studies, and also give an indication of how efficiently each measure extracts information from its input. All measures can be expected to asymptote to some maximum level of reliability given the underlying variability of responses, with more efficient measures approaching this maximum more quickly.

We explored sample size by performing a sub-sampling procedure in which only a subset of participants were drawn without replacement[3] from the total population (of around 150) at sample sizes ranging from 30 to 120 in increments of 10. Only the subset of judgments was used to derive the reliability estimates. We carried out 30 repetitions of the sub-sampling procedure at each sample size and averaged them to estimate the reliability measures at that sample size. Although this smooths out variation associated with the random choice of participants retained, it does not fully reflect the variability expected at each sample size because the repetitions cannot be totally independent. Especially when the sub-sample is a large proportion of the full sample, there is extensive overlap in the data retained across iterations. Sub-sampling was also constrained to only allow samples where every item appeared at least once so that an acceptability score was always computable for each sentence and the targeted comparisons were guaranteed to be feasible.

As Figure 5 shows, reliability decreased for every measure with decreasing sample size, but less reliable measures also showed larger decreases and the drops were higher for between-participant reliability. Reassuringly, the relative ordering of measures did not change and most became reasonably close to their performance on the full dataset at samples between 50 and 100 people. These results also suggest that the most reliable measures are most efficient, as they approach their maximum reliability more quickly in the number of responses.
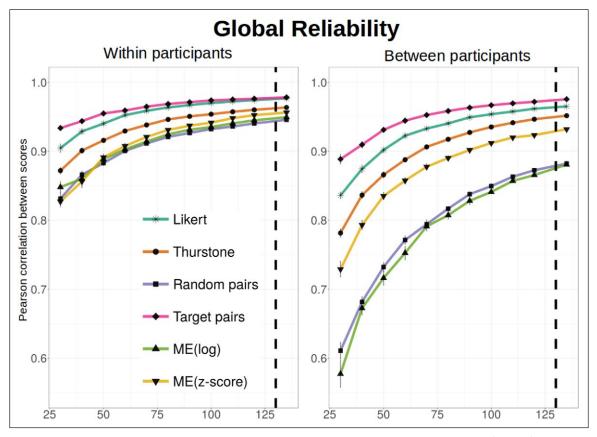
3.1.3 Discussion

Overall, all of the measures have high test-retest reliability, especially LIKERT, THURSTONE, and TARGET PAIRS; the most reliable judgments are obtained by TARGET PAIRS. This task is unusual in not offering acceptability scores that are comparable across all sentences: of the measures that do offer global comparisons, LIKERT scores are most reliable. RANDOM PAIRS and the two ME scores were the least reliable. Contrasting the within and between-participant r^2 values suggests one possible reason: these scores are particularly vulnerable to individual response style or item effects. Of these two possibilities, individual differences in response style is likely to be the major contributing factor for ME, as shown by the way the z-transformation improves reliability and reduces the gap between within and between participant reliability. The RANDOM PAIRS measure is more likely to be showing item effects. There is little scope for response style differences in a choice task, but the measure is clearly sensitive to the changing identity of the alternative choices, which were re-drawn for the new participants.

In principle, LIKERT scales are also vulnerable to response style differences, and since the THURSTONE scores are based on the same input as random-pairs they are exposed to the same item effects. However both measures include protection against these influences: z-transformation in the case of LIKERT scores and the inferred latent scale for the THURSTONE scores. These results suggest that in practice these protections are effective.

Examining subsets of participants shows that the relative reliability of the different measure types does not change with sample size, and that the most reliable measures were also the least impacted by the number of participants. For the sentences considered here, reliability scores approached their maximum values at approximately 100 participants, which with 40 trials per participant and 300 items corresponds to an average of 13.3 trials per item. The degree of variability in responses might be expected to vary with

---

[3] Other work (Sprouse & Almeida 2017a) draws samples with replacement for similar analyses, but unlike our work, their items were organized into lists, preserving an even distribution across people. Because our participants all rated different sets of sentences, sampling people multiple times greatly distorts the distribution of items within the dataset in a way that they would never be distorted had that been the target sample size.

Figure 5: Global reliability measured by Pearson correlation. Pearson correlations (y-axis) between acceptability estimates based on different data sets were used to quantify the reliability of each measure over different sample sizes (x-axis). Results to the right of the dotted vertical bar are based on the full sample, results to the left are averages of 30 samples of the designated size drawn from the full sample. For all measures, smaller sample sizes are less reliable, but gains in reliability from increasing sample size become progressively smaller. **Within-participant reliability** (left panel) shows variability in estimates based on responses from the same people to the same items, and can be interpreted as variability deriving from the difficulty of the task and the inherently probabilistic nature of people's responses. **Between-participant reliability** (right panel) is subject to the same sources of noise plus individual differences and variability introduced when re-drawing the items presented, so contrasting within and between participant reliability indicates the vulnerability of each measure to these extra factors. These results show that RANDOM PAIRS and ME scores are particularly vulnerable to participant and item differences, and that TARGET PAIRS and LIKERT ratings are consistently most reliable.

the particular sentences used, so this relation between reliability and number of trials per item holds only to the extent that the sentences considered here represent a typical range of acceptability for research targets.

3.2 Decision measures

Global reliability is useful when testing claims applicable to diverse collections of sentences, but some hypotheses are most naturally tested with targeted contrasts between particular pairs of sentences. Does each measure yield the same *decision* about which item of a pair is more acceptable? This sort of targeted comparison can expose changes in acceptability due to a particular syntactic manipulation while controlling for other factors like length, plausibility, and complexity. The global scores discussed above do allow pair-wise contrasts based simply on the difference between two acceptability scores, but for targeted contrasts researchers would typically conduct a measure-specific significance test instead. These are preferable because they take full advantage of a researcher's

knowledge of the test structure to appropriately characterize the variability associated with the acceptability estimates, which in turn offers control of the long-run Type 1 error rate.

Of course, if a researcher's primary goal was to evaluate a particular theoretical claim, they would present participants with multiple item pairs that all instantiate the syntactic manipulation of interest, rather than the one-item-per-effect that we have evaluated here. It is nevertheless interesting for us to evaluate the decision reliability of *items* as we do here, for several reasons. First, if items are highly reliable across tests or people, that is both noteworthy and highly reassuring about whether *effects* might also be reliable. Second, looking at item-level decision reliability is still informative about the overall reliability of each measure, and can tell us about the sources of variability within each measure.

We thus investigate the reliability of each of the measures with regard to the decisions a researcher would draw based on a significance test for a contrast of interest. The particular significance tests we used differ for each measure as described in the *Measures* section in Table 1: some involve two-tailed independent sample t-tests, while for others the structure of the data requires more complex analyses. We consider only the 150 targeted contrasts used in the targeted pairs task, reflecting the particular linguistic phenomena under investigation in the original *Linguistic Inquiry* articles. Since we are interested in the reliability of decisions rather than the content of any particular decision, we did not control for multiple comparisons in any of these tests, mimicking the situation that would obtain if each contrast was being studied independently. As in the previous analysis, we contrast within and between-participant reliability.

We examine three outcomes relevant to the test-retest reliability and sensitivity of such decisions: the number of *inconsistent* decisions across time points, the number of those inconsistencies which involve decision *reversals*, and the *null decision count*. Each individual decision admits three possible outcomes: option A is more acceptable, option B is more acceptable, or the null hypothesis of no difference cannot be rejected. For each measure, we evaluate the number of decisions (of 150 pairs) which were **inconsistent** (i.e., at one time option A was selected by the measure but at the other time either option B or the null was). An inconsistent measure indicates that an error of some sort (either Type 1 or Type 2) was made at some point, but it in most cases it is impossible to determine what kind of error it was. Flipping from option A to option B, which we call a **reversal** indicates a Type 1 error, and is quite rare: no measure produces a reversal on the full dataset. Flipping from a null to a non-null result could be either a Type 1 error (if the non-null result was incorrect) or a Type 2 error (if the null result was incorrect).

An indication of the sensitivity of a measure is given by the number of **null decisions** (i.e., the measure was unable to reject the null hypothesis in the INITIAL data set): it would be possible for a measure to never produce an inconsistent decision, but only because it was unable to ever reject the null hypothesis, which would not be a very interesting measure.

The raw numbers of inconsistent and null decisions are shown in Table 2. No significant differences in the number of inconsistent measures was found either within participants ($\chi^2(5) = 10.933$, $p = 0.0527$) or between participants ($\chi^2(5) = 8.0842$, $p = 0.15$), but there were significant differences in the number of null decisions ($\chi^2(5) = 37.35$, $p < 0.001$). In particular, TARGET PAIRS had notably fewer null responses: it was a more sensitive measure. There were no reversals across data sets for any measure in the full sample, but some did occur at smaller sample sizes.

Table 2: Decision reliability measured by agreement on targeted contrasts. The reliability of each measure was quantified based on the number of decisions, out of a 150 total, that suggested different conclusions at different time points. All inconsistent decisions here were significant at one time point and null in the other: sign reversals appeared only at smaller sample sizes. To test if high reliability was based on insufficient power resulting in consistent null decisions, the total number of null results is also shown. There was no significant difference in the number of inconsistent decisions across measures for within- or between- participant datasets, however there was a significant difference in the number of null decisions, with the TARGET PAIRS measure showing the fewest null decisions.

	Likert	**ME (z-score)**	**ME (log)**	**Thurstone**	**Random pairs**	**Target pairs**
Inconsistent decisions						
Within	11	15	18	12	26	13
Between	17	21	28	18	27	15
Null decisions						
Initial	24	35	44	31	49	11

3.2.1 Sample size dependence

As in the previous analysis, we examined the impact of sample size on decision reliability by analyzing 30 random subsets of the full sample for each increment of 10 participants between 30 and 120. Results are shown in Figure 6. As expected, within-participant reliability is generally higher than between-participant reliability. TARGET PAIRS is highly reliable and less sensitive to sample size, although at sample sizes of 75 and above it enjoys no particular advantage over the most reliable global scores, LIKERT and THURSTONE. A similar sample size analysis for the null results is shown in Figure 7, whose left panel indicates the number of sentence pairs for which each measure concluded the evidence was insufficient to reject the null in the INITIAL data set.

Overall, then, TARGET PAIRS appears to be both highly reliable and extremely sensitive, yielding relatively few inconsistent decisions (Figure 6) even with very low numbers of null decisions. That said, the main drawback of the measure is evident upon comparing the effect of sample size on the number of decision **reversals** across measures: the number of decisions on which the measure indicates one option was significantly more acceptable at one time but the same measure indicates that the *other* option was significantly more acceptable at the other time. Reversals are one kind of inconsistency, but are singled out here because they reflect a larger and more consequential difference than inconsistencies that involve being unable to reject the null in one sample but not another. As the right panel of Figure 7 indicates, only TARGET PAIRS yields decision reversals. Especially at small sample sizes, it may show a statistically significant preference for one item of a pair only to prefer the *other* item with more data. Thus, its sensitivity comes at a cost – being more likely to completely flip in the direction of a statistically significant judgment.

3.2.2 Discussion

The relative reliability of the targeted contrast decisions derived from each measure is consistent with the ordering observed in the score correlation analysis, with TARGET PAIRS and LIKERT the most reliable measures. The consistency of decisions is however distinct from the consistency of the underlying scores because of the way it depends on the threshold for rejecting the null. When the null hypothesis is true, the number of agreements across replications is completely determined by the alpha level of the test, but in the presence of a real effect it also depends on the sensitivity of the measure and the true effect size. In this study the effect sizes are constant across measures, so differences
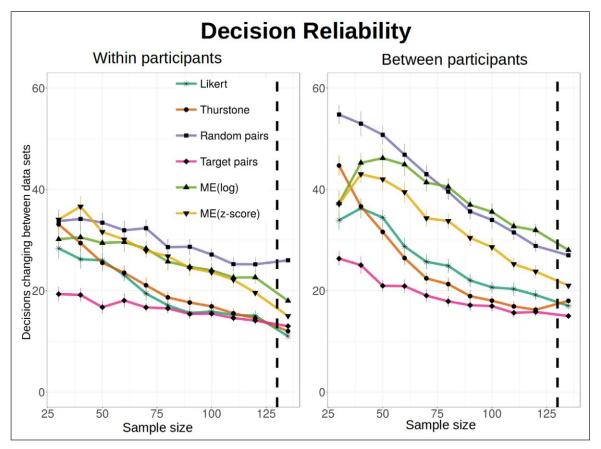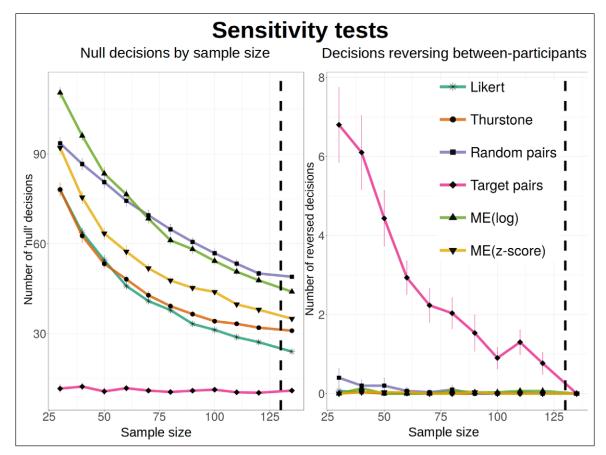
Figure 6: Decision reliability measured by agreement on targeted contrasts. The reliability of each measure was quantified as the number of decisions resulting in inconsistent outcomes (y-axis) tested across various sample sizes (x-axis). Outcomes were considered inconsistent if a contrast was considered null at one time point but significant at another, or the two results were significant in opposite directions: either case guarantees an error, although the first scenario is ambiguous as to error type. Results to the right of the dotted vertical bar are based on the full sample, results to the left are averages of 30 samples of the designated size drawn from the full sample. For all measures, smaller sample sizes are less reliable. Within-participant reliability (left panel) shows variability in estimates based on responses from the same people to the same items, and can be interpreted as variability deriving from the difficulty of the task and the inherently probabilistic nature of people's responses. Between-participant reliability (right panel) is subject to the same sources of noise plus individual differences and variability introduced when re-drawing the selection of items presented. Contrasting within and between participant reliability indicates the vulnerability of each measure to these extra factors. Between-participant reliability also gives a direct measure of how well acceptability scores might be expected to replicate. These results show that RANDOM PAIRS and ME scores are particularly vulnerable to participant and item differences, and that THURSTONE and LIKERT scores are consistently reliable and approach the reliability of the TARGET PAIRS method.

between measures reflect their sensitivity. A measure's sensitivity ultimately depends on the information content of responses, and the extent to which information is lost by the process of aggregating responses to produce an acceptability score. Unlike the alpha level, this is not a property of the decision rule and can only be estimated empirically. It is determined by the allowable range of variability in responses and the extent to which observed variability is systematic. An ideal measure would be high on both, but the two properties conflict in the sentence acceptability context to the extent that increasing the flexibility of response options makes the task more difficult. The different tasks considered here represent different trade-offs in intuitive ease-of-use (helping participants respond systematically) and expressiveness (widening the range of response options). As previous authors

Figure 7: Sensitivity tests by sample size. The number of non-significant differences declared by each decision rule (left panel) and the number of times significant effects appear to reverse between samples (right panel). Each measure evaluated the same 150 contrasts between target-pairs of interest. Results to the right of the dotted vertical bar are based on the full sample, results to the left are averages of 30 samples of the designated sample size drawn from the full sample. TARGET PAIRS stands out as rejecting the null most often, and at a similar rate across a range of sample sizes. Taken together, these results highlight the unique properties of the TARGET PAIRS measure, which almost always arrives at a decision even at small sample sizes, but with decisions that may not be stable under repeated measurement. Decision reversals are shown for the between-participants test, the within-participants analogue shows the same qualitative pattern with roughly half the number of decision reversals at each sample size for TARGET PAIRS. The other measures are less sensitive in the sense of producing more null decisions, but more conservative in the way repeated samples only ever disagree on the magnitude of an effect, never its sign. Among these conservative measures, LIKERT ratings are the most likely to detect acceptability differences.

have noted (Weskott & Fanselow 2008; Fukuda et al. 2012), the greater expressiveness of ME's free responses appears to be offset by increases in unsystematic variation. As in the score correlations discussed above, contrasting between and within participant reliability suggests that this extra noise is introduced by individual participants' idiosyncratic use of the scale, and can be mitigated by z-transformation. LIKERT appears to be effective in the compromise it achieves between allowing variability in responding and constraining unsystematic variation.

TARGET PAIRS was found to achieve a very high power on the limited contrasts it considers, in that it arrived at fewer null decisions and was relatively insensitive to sample size. The risk of producing significant results in the wrong direction, as shown by the decision reversals, is a consequence of this high power along with the fact that controlling Type 1 error rates does not entail controlling error magnitudes. The design of the decision rule

used allows for the possibility of significant findings in the wrong direction so long as the rate of such outcomes obeys the specified limits (Gelman & Tuerlinckx 2000; Cumming and Maillardet 2006).

The practice of testing several instances of any one phenomenon of interest (Schütze 2011) provides protection against these potential sign errors, since measurement error will be randomly distributed across individual items. The cost is inflation of the item set size, which may be a relatively minor burden compared to increasing the number of participants to the levels required for comparable power with a more conservative measure such as LIKERT or THURSTONE.

Although item effects may be driving the higher level of sign errors for TARGET PAIRS, it is unlikely that they are responsible for the relatively greater number of null decisions yielded by our LIKERT measure than were reported in other work (Sprouse et al. 2013; Häussler et al. 2016; Mahowald et al. 2016). On closer examination, it probably emerges because our dataset involved fewer responses per effect. For instance, the Sprouse et al. (2013) results are based on 12 or 13 responses per item, with 8 items per phenomenon giving roughly 100 responses per effect. Our data involve approximately 20 responses per item (with some variability due to the random draw of items), but since there is only one item per phenomenon this is also 20 responses for each effect. With the smaller number of responses there are naturally more null decisions. Despite our much smaller Ns per item, the fact that we still had relatively few null decisions is overall reassuring, especially in light of the fact that a researcher investigating a specific effect would test multiple items.
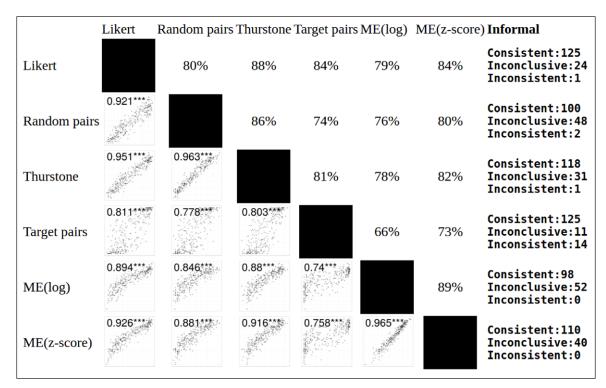
3.3 Agreement between measures

Having investigated the within and between-participant reliability of all of our measures, a natural question is whether they give the same answers as *each other*. Indeed, the question of how well different measures compare to informal judgments (as published in scholarly journals or textbooks) is much of the focus of related research in the literature (Featherston 2007; Munro et al. 2010; Myers 2012; Sprouse & Almeida 2012; Gibson & Fedorenko 2013; Sprouse et al. 2013).

Here we build on that by also incorporating comparisons to the INFORMAL measure of acceptability, though that is not our main focus. Instead, we explore a much wider set of comparison measures between all of the formal measures. We examine both global structure and decision agreement, using the same tests of agreement between measures as the reliability analyses, but for agreement between measures rather than across data sets.

Figure 8 summarizes these results, presenting both correlations between acceptability estimates and the percentage of decisions (out of 150 total) where the measures arrived at the same conclusions. The scores used in this analysis were from the INITIAL dataset for LIKERT and TARGET PAIRS measures, and from the BETWEEN PARTICIPANTS data for THURSTONE, RANDOM PAIRS, and ME scores. This avoids comparing unrelated measures on data derived from the same participants, potentially inflating their agreement. Related measures (the two versions of ME, or the two analyses of the random pairs choice task) are derived from the same data sets, so any disagreements between them are consequences of the different analysis only.

3.3.1 Discussion

There was substantial agreement between measures. Measures based on the same responses (THURSTONE/RANDOM PAIRS and ME(LOG)/ME(z-score)) were highly correlated ($r \approx 0.96$). Between measures based on different responses, correlations ranged between 0.74 and 0.95, with the highest agreement appearing between the LIKERT and THURSTONE measures. This highest level of agreement is comparable to the reliability

	Likert	Random pairs	Thurstone	Target pairs	ME(log)	ME(z-score)	Informal
Likert	■	80%	88%	84%	79%	84%	Consistent:125 Inconclusive:24 Inconsistent:1
Random pairs	0.921***	■	86%	74%	76%	80%	Consistent:100 Inconclusive:48 Inconsistent:2
Thurstone	0.951***	0.963***	■	81%	78%	82%	Consistent:118 Inconclusive:31 Inconsistent:1
Target pairs	0.811***	0.778***	0.803***	■	66%	73%	Consistent:125 Inconclusive:11 Inconsistent:14
ME(log)	0.894***	0.846***	0.88***	0.74***	■	89%	Consistent:98 Inconclusive:52 Inconsistent:0
ME(z-score)	0.926***	0.881***	0.916***	0.758***	0.965***	■	Consistent:110 Inconclusive:40 Inconsistent:0

Figure 8: Agreement between different measures. The lower left of this table presents Pearson correlation as a measure of global agreement between scores. The upper right presents the percentage of targeted contrasts for which different measures arrived at the same conclusion (A is more acceptable, B is more acceptable, or the evidence is inconclusive). The exception to this format is the informal expert judgment measure of acceptability, which presents only the number of decisions where informal judgments were consistent with the formal measure, the number that were inconsistent, and the number of decisions where the formal measure found the evidence inconclusive. Overall inter-measure agreement is strong. Among the formal measures, the highest agreement obtains between the LIKERT and THURSTONE scores. Agreement with the informal judgments is also high, with TARGET PAIRS notable as the only measure indicating an appreciable number of contrary conclusions.

between participants for these measures, indicating that switching from one measure to the other introduces no more variation than using the same measure twice, despite substantial differences in the presentation of the task.

All measures are largely consistent with the informal judgments. Almost all differences were observed in the expected direction, although on average 23% of these differences were deemed too small to be considered statistically significant. The measures were also found to be highly consistent with existing quantitative studies of these items, with the Likert acceptability scores reported here correlating at $r = .94$ with those obtained by Sprouse et al. (2013). Although item-level agreement is high, the rate of null decisions is higher here than previously reported for these measures in Sprouse et al. (2013). This is primarily because the decisions reported here are based on more responses per item but many fewer responses per effect, due to the way each effect was represented by a single sentence pair.

Although these design decisions limit conclusions about any particular effect, because they were constant over measures, contrasts can still be drawn between the measures. In particular, this comparison highlights a striking difference between TARGET PAIRS and the other measures, as it identified 14 contrasts in the opposite of the predicted direction, a disagreement rate of 9.3%. There are several possible explanations for these inconsistencies. It is possible that they represent measurement errors, although the sub-sampling

analysis suggests that it is unlikely that re-measurement at the sample size considered here would ever be expected to reverse more than one decision. More likely is that in most cases these are genuine effects counter to the predicted direction, but they are extremely small and TARGET PAIRS is the only measure with enough power to identify them. This interpretation is supported by the fact that most of these contrary decisions by TARGET PAIRS are identified as null results by the other measures.

Measurement error seems unlikely for cases where two different measures agree with each other and informal judgments, so there may also be instances where contrasting highly similar sentences directs people's attention to features of the sentences that are less salient when the two items are presented separately. One example of such an effect is the pair of sentences *There are leaves burnt* and *There are leaves green*. This pair of sentences, constructed by Sprouse et al. (2013), follows a pattern designed to demonstrate that (English) passives have event arguments (Basilico 2003). The option endorsed in the INITIAL TARGET PAIRS data is *There are leaves green*.[4] One possible explanation for this counterintuitive result is that it arises from people recognizing that the sentences are identical except for the words *burnt* and *green*, then responding to a strong association between *green* and *leaves*. The interpretation that the choice is somehow induced by the particular contrast of these two sentences is supported by the way the measures that do not present the two sentences together agree that *There are leaves burnt* is the more acceptable option. The practice of using multiple items targeting each phenomenon under study would be an effective defense against this kind of potentially misleading result, since it depends heavily on the *green/leaves* association, which would be unlikely to have an analogue in other sentences targeting this particular passive construction. Across all three data sets collected, only 19 contrasts are involved in a decision result conflicting with expert judgment at any point. These are presented in Appendix A.

4 Summary and conclusions

Our main focus in this work is the test-retest reliability survey of the most common tasks used to measure sentence acceptability. All tasks considered here showed high reliability, with even the least reliable measure, TARGET PAIRS, producing large positive correlations across re-test data sets. By contrasting within-participant reliability with between-participant reliability on the same sentences with the same measures, we estimated what proportion of the variability observed can be attributed to factors unique to the between-participant replication. In all cases between-participant reliability was lower, and this reliability drop was particularly pronounced for ME and RANDOM PAIRS, suggesting these measures are particularly vulnerable to variability across people or how items are paired together. The TARGET PAIRS and LIKERT ratings showed not only the highest within-participant reliability but also had the least amount of decrease in reliability when comparing between- to within-participant correlations. This pattern is a hallmark of well-calibrated measurement instruments.

Secondly, we ask to what extent acceptability estimates depend on the particular assumptions of each measurement tool, and whether the conclusions a researcher would reach would change based on the measurement task they used. Here we find high consistency between measures, including near-uniform agreement with expert judgment. The least accurate global score (RANDOM PAIRS) was still highly correlated ($r \approx .9$) with the most accurate global score (LIKERT). Where disagreements occurred between the measures,

---

[4] This result re-appears in the WITHIN PARTICIPANTS replication, but not the BETWEEN PARTICIPANTS replication, which is null.

it was usually in the magnitude rather than the direction of the difference, with the less reliable scores more likely to not reject the null for closely matched pairs.

This overall consistency is striking given the structural differences between these tasks, especially between the LIKERT and THURSTONE tasks. Both these measurement tasks incorporate strong assumptions, and in different domains have not always agreed with each other (Roberts et al. 1999; Drasgow et al. 2010).

Specifically, the assumptions made by the LIKERT task center around people's interpretation of the scale, which may impose structure on responses (Schütze 1996; Carifio & Perla 2008) or be vulnerable to differences in response style (Lee et al. 2002; Johnson 2003). The THURSTONE measure avoids these issues by removing the researcher-supplied scale and forcing a discrete choice, but instead assumes transitivity of acceptability, which is known to be violated in similar preference-ranking tasks (Tversky 1969). Such violations have been observed in sentence acceptability judgments (Danks & Glucksberg 1970; Hindle & Ivan 1975).

A core contribution of this paper is that these measures provide converging evidence in the domain of sentence acceptability: theoretically motivated concerns about the restrictions a fixed LIKERT response scale imposes on participants turn out not to matter in practice, with the scale-free THURSTONE measure based on choice task data arriving at essentially identical acceptability estimates. Although the LIKERT and THURSTONE acceptability scores agree, LIKERT scores are marginally more reliable and have the advantage of more easily accepting additional sentences into an existing set of comparisons.

Despite the close agreement between measures, TARGET PAIRS stands out as having noteworthy decision reliability. It showed the highest power, yielding very few null results, but as a result was also the only measure vulnerable to complete reversals of a significant decision. This pattern is characteristic of high-powered tests, where significant differences observed under high-noise/low information conditions tend to entail exaggerated estimates of effect size (Loken and Gelman 2017). While TARGET PAIRS is the highest performing measure in terms of test-retest consistency, and maintains this performance at small sample sizes, the relatively few errors it produces at low sample sizes can be of a qualitatively different and potentially much more misleading kind. Relatedly, the TARGET PAIRS measure had by far the highest disagreement with the informal expert ratings of any measure, endorsing the informally dispreferred sentence on 14 of the items (9.3%) while the other measures endorsed at most two. When using the TARGET PAIRS measure it is critical for researchers to include multiple pairs of target sentences within the same construct to increase decision reliability.

We find that ME tasks produce acceptability scores that are consistent with the other measures but somewhat less reliable. Contrasting the within and between participant test-retest reliability shows that this greater variability is likely to be due to variation in participant response styles, which appears as noise in the final measure. This source of variability can be mitigated somewhat by processing the scores using a transformation sensitive to response style, such as the z-transform. However, this is less effective than offering restricted responses in the task itself, as the LIKERT and THURSTONE measures do. In general, although ME measures performed overall better than we expected them to, they were still consistently inferior to most of the alternatives.

Although we expect these results to be indicative of the relative test-retest reliability of these measures, the particular reliability results we observed can be expected to depend to some extent on factors such as the specific sentences and the number of trials per participant, which were controlled across measures to ensure the comparisons were fair. For the rating tasks, reliability can be expected to be a function of the number of trials per item, so the analysis over participant sample sizes gives some indication of how reliability

might be expected to change with different sentence set sizes. The situation is less clear for the THURSTONE and RANDOM PAIRS measures, which may be sensitive to the diversity of contrasts presented as well as the average number of presentations per sentence. By choosing to hold the set of sentences constant we ensured that each measure was tested on the same range of effect sizes, but this does limit the generalizability of our reliability results. However, we believe these 150 sentences are representative of the kinds of sentences commonly used for sentence grammaticality judgments.

Although individually these measures make a range of assumptions that could be considered strong limitations, the high agreement between them suggests that these measure-specific assumptions do not have a strong impact on acceptability judgments. We find that if multiple items targeting the same contrast are used, none of the methods considered here have an appreciable chance of giving a strongly misleading result (although there are differences in efficiency, with ME measures requiring more trials for any given level of reliability).

While we find that the most common measurement tasks are all reasonably effective, the LIKERT task performed especially well. In addition to achieving relatively high test-retest reliability, our results also suggest that the LIKERT measure admits a stronger interpretation of sentence acceptability scores than is usually attributed to it. Our findings suggest that the interpretation of LIKERT data need not be constrained by concerns that the limited response scale may impose structure on the data, or that the subjective distance between response options is unknown and may vary between people. The structure suggested by the LIKERT data is in high agreement with the structure suggested by the THURSTONE measure. Since the latter is both agnostic about the underlying structure of acceptability and capable of recovering various clustered or gradient but non-linear distributions of acceptability, this high agreement suggests that the nature of the LIKERT scale is not significantly shaping the structure of acceptability judgments it yields. The minimal difference between within-participant test-retest reliability and between-participant test-retest reliability suggests that the z-transformation offers effective protection against potential differences in the interpretation of the scale.

One interesting aspect of our results hinges on the fact that our dataset involved only one item per effect. This was intentional since it thus made the item set maximally variable and offered a stronger test of each measure. Our results indicating that many of these measures can reliably reflect *global* acceptability, rather than just effect-level acceptability, is gratifying and reassuring. It is also interesting that our *item*-level reliability is so high, differing from other work measuring effect-level reliability primarily in yielding slightly higher numbers of null decisions at lower sample sizes (Sprouse et al. 2013; Häussler et al. 2016; Mahowald et al. 2016). Aside from this, we found item-level reliability that was nearly as good as effect-level reliability incorporating multiple items. Taken together with the high item-level variability observed around effects in other work (Sprouse et al. 2013), this may suggest that people are surprisingly consistent on specific items but that the effect-level phenomena within any given item can at least sometimes be obscured by lexical choices or other superficial differences between sentences.

In terms of design recommendations for researchers interested in efficiently obtaining results that replicate with high confidence, we replicate previous results pertaining to the reliability of effects defined as ordinal relationships between sentence classes and extend them to include recommendations for ensuring the reliability of distances between individual items. We reproduce here both the general finding that acceptability judgments are highly reliable in between-participant replications (Sprouse & Almeida 2012; Sprouse et al. 2013), and also more detailed claims such as the high power of TARGET PAIRS (Schütze & Sprouse 2014; Sprouse & Almeida 2017), the lack of extra information in the extra

variability of ME ratings (Weskott & Fanselow 2011), and the qualitative relationship between decision reliability and sample size (Mahowald et al. 2016). We further show that these reliability results extend to estimation analyses, with a high correlation in the acceptability scores assigned by different tasks to different sentences.

Overall, our work demonstrates that formal acceptability results are even more informative than previously realized. They agree substantially with each other (as well as informal measures) across the global structure of acceptability, not just individual targeted sentence pairs. Moreover, the best-performing measures (like LIKERT and THURSTONE) appear not to impose substantial structure of their own onto the pattern of acceptability responses. This licenses us to use acceptability judgments to address a wider variety of questions that we have previously been able – from identifying dialectical or language differences (or possibly even individual fluency) using acceptability judgments, to investigating the global structure of grammatical knowledge (e.g., is it all-or-none or multi-dimensional?). Not all of these questions may pan out, but it is exciting to think that the formal tools we have developed for evaluating targeted sentence pairs may have something to say about them as well.

Additional Files

The additional files for this article can be found as follows:

- **Appendix A.** Corpus details. DOI: https://doi.org/10.5334/gjgl.396.s1
- **Appendix B.** Details of the THURSTONE model. DOI: https://doi.org/10.5334/gjgl.396.s1

Acknowledgements

Competing Interests

The authors have no competing interests to declare.

References

Arppe, Antti & Juhani Järvikivi. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159. DOI: https://doi.org/10.1515/CLLT.2007.009

Bard, Ellen Gurman, Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language,* 32–68. DOI: https://doi.org/10.2307/416793

Basilico, David. 2003. The topic of small clauses. *Linguistic Inquiry* 34(1). 1–35. DOI: https://doi.org/10.1162/002438903763255913

Borg, Ingwer & Patrick J. F. Groenen. 2005. *Modern multidimensional scaling: Theory and applications*. Berlin: Springer Science & Business Media.

Brandt, Mark J., Hans IJzerman, Ap Dijksterhuis, Frank J. Farach, Jason Geller, Roger Giner-Sorolla, James A. Grange, Marco Perugini, Jeffrey R. Spies & Anna Van't Veer. 2014. The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology* 50. 217–224. DOI: https://doi.org/10.1016/j.jesp.2013.10.005

Carifio, James & Rocco Perla. 2008. Resolving the 50-year debate around using and misusing Likert scales. *Medical Education* 42(12). 1150–1152. DOI: https://doi.org/10.1111/j.1365-2923.2008.03172.x

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Cornips, Leonie & Cecilia Poletto. 2005. On standardising syntactic elicitation techniques (part 1). *Lingua* 115(7). 939–957. DOI: https://doi.org/10.1016/j.lingua.2003.11.004

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

Culbertson, Jennifer & Steven Gross. 2009. Are linguists better subjects? *The British Journal for the Philosophy of Science* 60(4). 721–736. DOI: https://doi.org/10.1093/bjps/axp032

Cumming, Geoff & Robert Maillardet. 2006. Confidence intervals and replication: Where will the next mean fall? *Psychological Methods* 11(3). 217. DOI: https://doi.org/10.1037/1082-989X.11.3.217

Dąbrowska, Ewa. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27(1). 1–23. DOI: https://doi.org/10.1515/tlir.2010.001

Danks, Joseph H. & Sam Glucksberg. 1970. Psychological scaling of linguistic properties. *Language and Speech* 13(2). 118–138. DOI: https://doi.org/10.1177/002383097001300204

Danziger, Kurt. 1980. The history of introspection reconsidered. *Journal of the History of the Behavioral Sciences* 16(3). 241–262. DOI: https://doi.org/10.1002/1520-6696(198007)16:3<241::AID-JHBS2300160306>3.0.CO;2-O

Davison, Anthony C. & David V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

DeVellis, Robert F. 2016. *Scale development: Theory and applications* 26. Thousand Oaks, CA: Sage publications.

Drasgow, Fritz, Oleksandr S. Chernyshenko & Stephen Stark. 2010. 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology* 3(4). 465–476. DOI: https://doi.org/10.1111/j.1754-9434.2010.01273.x

Ennis, Daniel M. 2016. *Thurstonian Models: Categorical Decision Making in the Presence of Noise*. Richmond, VA: The Institute for Perception.

Erlewine, Michael Yoshitaka & Hadas Kotek. 2016. A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory* 34(2). 481–495. DOI: https://doi.org/10.1007/s11049-015-9305-9

Fabrigar, Leandre R. & Jung-Eun Shelly Paik. 2007. Thurstone Scales. In Neil Salkind (ed.), *Encyclopedia of Measurement and Statistics*, 1003–1005. SAGE publications.

Featherston, Sam. 2005. Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua* 115(11). 1525–1550. DOI: https://doi.org/10.1016/j.lingua.2004.07.003

Featherston, Sam. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33(3). 269–318. DOI: https://doi.org/10.1515/TL.2007.020

Featherston, Sam. 2008. Thermometer judgments as linguistic evidence. *Was ist linguistische Evidenz*, 69–89.

Featherston, Sam. 2009. Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft* 28(1). 127–132. DOI: https://doi.org/10.1515/ZFSW.2009.014

Fukuda, Shin, Grant Goodall, Dan Michel & Henry Beecher. 2012. Is Magnitude Estimation worth the trouble? In Jaehoon Choi, E. Allan Hogue, Jeffrey Punske, Deniz Tat, Jessamyn Schertz & Alex Trueman (eds.), *Proceedings of the 29th West Coast Conference on Formal Linguistics*, 328–336.

Gelman, Andrew & Francis Tuerlinckx. 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* 15(3). 373–390. DOI: https://doi.org/10.1007/s001800000040

Gibson, Edward & Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28(1–2). 88–124. DOI: https://doi.org/10.1080/01690965.2010.515080

Gibson, Edward, Steven T. Piantadosi & Evelina Fedorenko. 2013. Quantitative methods in syntax/semantics research: A response to sprouse and almeida (2013). *Language and Cognitive Processes* 28(3). 229–240. DOI: https://doi.org/10.1080/01690965.2012.704385

Hartley, James. 2014. Some thoughts on Likert-type scales. *International Journal of Clinical and Health Psychology* 14(1). 83–86. DOI: https://doi.org/10.1016/S1697-2600(14)70040-7

Häussler, Jana & Thomas Juzek. 2016. Detecting and discouraging noncooperative behavior in online experiments using an acceptability judgment task. In Hanna Christ, Daniel Klenovšak, Lukas Sönning & Valentin Werner (eds.), *A blend of MaLT: Selected contributions from the methods and linguistic theories symposium 2015*, 15. 73–100.

Häussler, Jana, Thomas Juzek & Tom Wasow. 2016. Unsupervised prediction of acceptability judgements. In Patrick Farrell (ed.), *To be grammatical or not to be grammatical – is that the question*. Annual Meeting of the Linguistic Society of America.

Hindle, Donald & Sag Ivan. 1975. Some more on anymore. In Ralph Fasold & Roger Shuy (eds.), *Analyzing variation in language: Papers from the second colloquium on new ways of analyzing variation*, 89–110.

Hofmeister, Philip, T. Florian Jaeger, Inbal Arnon, Ivan A. Sag & Neal Snider. 2013. The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes* 28(1–2). 48–87. DOI: https://doi.org/10.1080/01690965.2011.572401

Johnson, Keith. 2011. *Quantitative methods in linguistics*. Manchester, MI: John Wiley & Sons.

Johnson, Timothy R. 2003. On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika* 68(4). 563–583. DOI: https://doi.org/10.1007/BF02295612

Juzek, Thomas. 2015. *Acceptability judgement tasks and grammatical theory*. Oxford: University of Oxford Phd thesis.

Keller, Frank. 2003. A psychophysical law for linguistic judgments. In Richard Alterman & David Kirsh (eds.), *Proceedings of the 25th annual conference of the cognitive science society*, 652–657.

Keller, Frank & Ash Asudeh. 2001. Constraints on linguistic co-reference: Structural vs. pragmatic factors. In Johanna Moore & Keith Stenning (eds.), *Proceedings of the 23rd annual conference of the cognitive science society*, 483–488.

Kline, Paul. 2013. *Handbook of psychological testing*. Routledge.

Lau, Jey Han, Alexander Clark & Shalom Lappin. 2016. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*.

Lee, Jerry, Patricia Jones, Yoshimitsu Mineyama & Xinwei Esther Zhang. 2002. Cultural differences in responses to a Likert scale. *Research in Nursing & Health* 25(4). 295–306. DOI: https://doi.org/10.1002/nur.10041

Li, Linjie, Vicente Malave, Amanda Song & Angela Yu. 2016. Extracting human face similarity judgments: Pairs or triplets? *Journal of Vision* 16(12). 719–719. DOI: https://doi.org/10.1167/16.12.719

Likert, Rensis. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 140. 44–60.

Linzen, Tal & Yohei Oseki. 2015. *The reliability of acceptability judgments across languages.* New York, MS: New York University Press.

Loken, Eric & Andrew Gelman. 2017. Measurement error and the replication crisis. *Science* 355(6325). 584–585. DOI: https://doi.org/10.1126/science.aal3618

Mahowald, Kyle, Peter Graff, Jeremy Hartman & Edward Gibson. 2016. SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language* 92(3). 619–635. DOI: https://doi.org/10.1353/lan.2016.0052

Miller, Brent, Pernille Hemmer, Mark Steyvers & Michael D. Lee. 2009. The wisdom of crowds in rank ordering problems. In Andrew Howes, David Peebles & Richard Cooper (eds.), *9th International conference on cognitive modeling*, 86–91. Manchester: ICCM.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen & Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In Jon Weese (ed.), *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, 122–130.

Murphy, Brian, Carl Vogel & Conny Opitz. 2006. Cross-linguistic empirical analysis of constraints on passive. In *Presentation to the symposium on interdisciplinary themes in cognitive language research.* Helsinki: Finnish Cognitive Linguistics Association.

Myers, James. 2009. The design and analysis of small-scale syntactic judgment experiments. *Lingua* 119(3). 425–444. DOI: https://doi.org/10.1016/j.lingua.2008.09.003

Myers, James. 2012. Testing adjunct and conjunct island constraints in Chinese. *Language and Linguistics* 13(3). 437.

Nosofsky, Robert M. 1992. Similarity scaling and cognitive process models. *Annual Review of Psychology* 43(1). 25–53. DOI: https://doi.org/10.1146/annurev.ps.43.020192.000325

Noveck, Ira & Anne Reboul. 2008. Experimental pragmatics: A Gricean turn in the study of language. *Trends in Cognitive Sciences* 12(11). 425–431. DOI: https://doi.org/10.1016/j.tics.2008.07.009

O'Mahony, Michael et al. 2003. Discrimination testing: A few ideas, old and new. *Food Quality and Preference* 14(2). 157–164. DOI: https://doi.org/10.1016/S0950-3293(02)00109-X

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

Phillips, Colin & Howard Lasnik. 2003. Linguistics and empirical evidence: Reply to edelman and christiansen. *Trends in Cognitive Sciences* 7(2). 61–62. DOI: https://doi.org/10.1016/S1364-6613(02)00046-3

Podesva, Robert & Devyani Sharma. 2014. *Research methods in linguistics.* Cambridge: Cambridge University Press.

Porte, Graeme. 2013. Who needs replication? *CALICO Journal* 30(1). 10–15. DOI: https://doi.org/10.11139/cj.30.1.10-15

Rensink, Ronald, Kevin O'Regan & James Clark. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science* 8(5). 368–373. DOI: https://doi.org/10.1111/j.1467-9280.1997.tb00427.x

Roberts, James, James Laughlin & Douglas Wedell. 1999. Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement* 59(2). 211–233. DOI: https://doi.org/10.1177/00131649921969811

Rosenbach, Anette. 2003. Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in english. *Topics in English Linguistics* 43. 379–412. DOI: https://doi.org/10.1515/9783110900019.379

Sampson, Geoffrey. 2007. Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory* 3(1). 1–32. DOI: https://doi.org/10.1515/CLLT.2007.001

Schütze, Carson. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago, IL: University of Chicago Press.

Schütze, Carson. 2011. Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(2). 206–221. DOI: https://doi.org/10.1002/wcs.102

Schütze, Carson & Jon Sprouse. 2014. Judgment data. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, chap. 3. 27–51. Cambridge: Cambridge University Press.

Selker, Ravi, Michael D. Lee & Ravi Iyer. 2017. Thurstonian cognitive models for aggregating top-n lists. *Decision* 4(2). 87. DOI: https://doi.org/10.1037/dec0000056

Shalom, Dorit Ben & David Poeppel. 2007. Functional anatomic models of language: Assembling the pieces. *The Neuroscientist.*

Simons, Daniel & Ronald Rensink. 2005. Change blindness: Past, present, and future. *Trends in Cognitive Sciences* 9(1). 16–20. DOI: https://doi.org/10.1111/j.0963-7214.2005.00332.x

Sorace, Antonella. 2010. Using magnitude estimation in developmental linguistic research. In Elma Blom & Sharon Unsworth (eds.), *Experimental methods in language acquisition research*, 57–72. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/lllt.27.05sor

Sorace, Antonella & Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115(11). 1497–1524. DOI: https://doi.org/10.1016/j.lingua.2004.07.002

Spencer, Nancy Jane. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2(2). 83–98. DOI: https://doi.org/10.1007/BF01067203

Sprouse, Jon. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1. 123–134.

Sprouse, Jon. 2008. Magnitude estimation and the non-linearity of acceptability judgments. In Natasha Abner & Jason Bishop (eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics*, 397–403. Somerville, MA: Cascadilla Press.

Sprouse, Jon. 2011a. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87(2). 274–288. DOI: https://doi.org/10.1353/lan.2011.0028

Sprouse, Jon. 2011b. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1). 155–167. DOI: https://doi.org/10.3758/s13428-010-0039-7

Sprouse, Jon & Carson Schütze. 2017. Grammar and the use of data. In Bas Aarts, Jill Bowie & Gergana Popova (eds.), *The Oxford handbook of English grammar*, chap. 3. Oxford University Press.

Sprouse, Jon, Carson 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. DOI: https://doi.org/10.1016/j.lingua.2013.07.002

Sprouse, Jon & Diogo Almeida. 2011. *Power in acceptability judgment experiments and the reliability of data in syntax*. Irvine CA & Lansing MI: University of California & Michigan State University Master's thesis.

Sprouse, Jon & Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48(03). 609–652. DOI: https://doi.org/10.1017/S0022226712000011

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

Sprouse, Jon & Diogo Almeida. 2017b. Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences* 40. DOI: https://doi.org/10.1017/S0140525X17000590

Stevens, Stanley Smith. 1956. The direct estimation of sensory magnitudes: Loudness. *The American Journal of Psychology*, 1–25. DOI: https://doi.org/10.2307/1418112

Thurstone, Louis. 1927. A law of comparative judgment. *Psychological Review* 34(4). 273. DOI: https://doi.org/10.1037/h0070288

Tversky, Amos. 1969. Intransitivity of preferences. *Psychological Review* 76(1). 31. DOI: https://doi.org/10.1037/h0026750

Wasow, Thomas & Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115(11). 1481–1496. DOI: https://doi.org/10.1016/j.lingua.2004.07.001

Weskott, Thomas & Gisbert Fanselow. 2008. Variance and informativity in different measures of linguistic acceptability. In Natasha Abner & Jason Bishop (eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics*, 431–439. Somerville, MA: Cascadilla Press.

Weskott, Thomas & Gisbert Fanselow. 2009. Scaling issues in the measurement of linguistic acceptability. *The Fruits of Empirical Linguistics* 1. 229–245. DOI: https://doi.org/10.1515/9783110216141.229

Weskott, Thomas & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87(2). 249–273. DOI: https://doi.org/10.1353/lan.2011.0041