RESEARCH

The effect of three basic task features on the sensitivity of acceptability judgment tasks

Paul Marty[1], Emmanuel Chemla[2] and Jon Sprouse[3]

[1] Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), DE

[2] Laboratoire de Sciences Cognitives et Psycholinguistique, Dept d'Etudes Cognitives, ENS, PSL University, EHESS, CNRS, FR

[3] Department of Linguistics, University of Connecticut, US

Corresponding author: Paul Marty (p.marty@ucl.ac.uk)

Sprouse and Almeida (2017) provide a first systematic investigation of the sensitivity of four acceptability judgment tasks. In this project, we build on these results by decomposing those tasks into three constituent task features (single versus joint presentation, number of response options, and use of response labels), and explore the consequences of those task features on the sensitivity of acceptability judgment experiments. We present 6 additional experiments (for a total of 10) designed to probe the effect of those task features on sensitivity, both independently and in combination. Our results suggest three notable conclusions: (i) there is a clear advantage to joint presentation of theoretically-related sentence types, regardless of the type of response scale used in the experiment; (ii) tasks involving a continuous slider (which have an infinite number of response options, and few labels) offer good sensitivity, while relying solely on spatial reasoning rather than numeric reasoning; and (iii) there are a number of subtle interactions among the three task features that may warrant further investigation. We discuss the potential benefits and concerns of each of these features in detail, along with the relevance of these findings for deciding how to investigate both simple and higher-order acceptability contrasts.

Keywords: acceptability judgments; experimental sensitivity; design analysis; experimental syntax; quantitative methods; linguistic methodology

1 Overview of the Study

Acceptability judgments can be prompted in numerous ways, and some experimental designs can reveal a contrast of interest with a smaller sample size than others (a.o., Schütze & Sprouse, 2014; Sprouse & Almeida, 2017) — a property we will call *sensitivity*.[1] Smaller samples can be valuable both to keep costs down, and also when researchers are working with smaller linguistic populations. More generally, the more sensitive the experimental design, the lower the risk of missing contrasts that are, or may be, of theoretical interest (Type II error), and the lower the risk of reporting an unreliable result (Type I error). Previous studies have investigated the sensitivity of acceptability judgment tasks as holistic units (a.o., Bard, Robertson, & Sorace, 1996; Cowart, 1997; Keller, 2000; Featherston, 2005, 2008; Bader & Häussler, 2010; Weskott & Fanselow, 2011; Fukuda, Goodall, Michel, & Beecher, 2012; Sprouse & Almeida, 2011, 2017; Langsford, Perfors, Hendrickson, Kennedy, & Navarro, 2018). The present study extends this line of research

---

[1] Following the suggestion of an anonymous reviewer, here we use *sensitivity* as a cover term for detecting a contrast of interest. This term is relatively neutral with respect to statistical framework since all frameworks incorporate some notion of sensitivity to a contrast of interest (as opposed to terms like *statistical power*, which may only have an interpretation within specific frameworks, like the Neyman-Pearson approach to null hypothesis testing).

by investigating how three elementary features of those tasks contribute to the sensitivity of an experiment: mode of presentation, number of response options, and use of labels.

The framework that we use to decompose tasks into features is as follows. The **mode of presentation** refers to the fact that contrasting sentences can be presented to participants one single condition at a time, or by factorial paradigms such as minimal pairs, as illustrated in (1-a) and (1-b), respectively.

(1)   **Mode of presentation**
   a.   Single presentation

   Item $x = \Big\{$ There has been a man considered sick.   No ○ Yes ○ $\Big\}$

   Item $y = \Big\{$ There has been considered a man sick.   No ○ Yes ○ $\Big\}$

   b.   Joint presentation

   Item $z = \begin{cases} \text{There has been a man considered sick.} & \text{No ○ Yes ○} \\ \text{There has been considered a man sick.} & \text{No ○ Yes ○} \end{cases}$

The **number of response options** is usually determined by the type of response scale used in the experiment. As illustrated in (2), a response scale for acceptability judgments can be reduced to a binary choice (e.g., a Yes-No scale), offer a small number of finite options (e.g., a discrete 5-point or 7-point scale) or be virtually infinite (e.g., a continuous visual scale).

(2)   **Number of response options**
   a.   Binary scale: 2 levels

         No ○   Yes ○

   b.   Likert scale: 7 levels

         least acceptable ○ ○ ○ ○ ○ ○ ○ most acceptable

   c.   Continuous scale: ∞ levels

         least acceptable [▬▬▬▬▬▬▭▭▭▭] most acceptable

The **use of labels** for specifying response options on the response scale is also usually determined by the type of task: binary scales do not leave much flexibility in this regard, while Likert and continuous scales can be more or less specified depending on whether intermediate labels are used or not, e.g. (3-a) vs. (3-b).

(3)   **Use of labels**
   a.   7-point Likert scale with endpoints only

         least acceptable ○ ○ ○ ○ ○ ○ ○ most acceptable

   b.   7-point Likert scale with endpoints and intermediate labels

         least acceptable 1 2 3 4 5 6 7 most acceptable
                           ○ ○ ○ ○ ○ ○ ○

The goal of this study is to document the single and cumulative effects of these three task features on the sensitivity of acceptability judgment experiments, and ask whether certain

features, or combinations of features, offer a sensitivity advantage over others. Building on Sprouse and Almeida's (2017) study, which we will review in the next section, we expanded on their 4 experiments to obtain a set of 10 acceptability judgment experiments minimally differing from each other along one of the three features of interest (the 4 experiments from Sprouse and Almeida (2017) and 6 new experiments). The phenomena tested were the same in all experiments and the same as those tested in Sprouse and Almeida (2017): 50 pairwise phenomena from *Linguistic Inquiry* (2001–2010) spanning the smaller half of the observable range of effect sizes in the large random sample originally tested by Sprouse, Schütze, and Almeida (2013). Following the method described in Sprouse and Almeida (2017), we compared the sensitivity of these 10 experiments by estimating and evaluating the *rate of statistical detection of acceptability rating differences* in a series of resampling simulations using the acceptability judgments for pairwise comparisons collected in each experiment. The set of simulations were conducted to cover the full range of potential sample sizes from 5 to 100 participants, and two distinct approaches to hypothesis testing, null hypothesis testing and Bayesian hypothesis testing.

First, we found that there is a sensitivity advantage in presenting contrasting conditions jointly rather than individually e.g., (1-a) vs. (1-b): presenting sentence types side by side by paradigms substantially increases the sensitivity of acceptability judgment experiments, regardless of the number of response options and the use of labels on the response scale. This benefit is found to outrank the possible benefit of asking for comparative judgments using theoretically irrelevant reference points (as in magnitude estimation), and to *partly* explain the good results of forced-choice tasks reported in the previous literature (Sprouse & Almeida, 2017). Second, we found that the number of response options and the use of labels both interact with the mode of presentation of conditions: binary and labeled graded scales (e.g., (2-a) and (3-b)) are generally more beneficial with single presentation whereas less labeled graded scales (e.g., (2-c) and (3-a)) are generally more beneficial with joint presentation. We will discuss the potential sources for these effects, the concerns researchers might have in choosing task features, and how these findings can be used to choose design options in different testing situations (in the lab, online, in the field, etc.).

In Section 2, we explain in further detail the motivations and structure of the present study, and locate its contribution in the light of previous work on the sensitivity of acceptability judgment experiments focusing on Sprouse and Almeida's (2017) study. Section 3 describes the 10 acceptability judgment experiments compared in our study together with the data analyses that were performed. Section 4 presents the results: an assessment of design sensitivity for a large range of effect sizes and sample sizes as a function of the mode of presentation of contrasting conditions, the number of response options, and the use of labels. Section 5 synthesizes our findings and Section 6 discusses their relevance for deciding how to investigate acceptability contrasts in formal and informal linguistic inquiries.

2 Motivations and Innovations of the Study

2.1 Background: Sprouse and Almeida (2017)

Sprouse and Almeida (2017, henceforth S&A) compared the sensitivity of four acceptability judgment experiments commonly used in experimental syntax, each of which involved a distinct task: a Forced-Choice (FC) task, a Magnitude Estimation (ME) task, a Yes-No (YN) task and a Likert Scale (LS) task. The general designs of these four tasks are presented and illustrated in the appendix. Table 1 provides a schematic description of these tasks as a function of the task features we aim to investigate in this study: (1) the mode of presentation of contrasting conditions, i.e. whether contrasting test sentences are presented individually or jointly, (2) the number of response options offered to participants to report their judgments, and (3) the use of labels on the response scale.

Table 1: Schematic description of the 4 acceptability judgment tasks compared in S&A as a function of the three task features investigated in this study.

Task	Experimental Item	Mode of Presentation	Number of Response Options	Use of Labels
FC	2 contrasting test sentences	joint	1 forced-choice	(not available)
YN	1 test sentence	single	2 levels	obligatory
LS	1 test sentence	single	7 levels	full labels
ME	1 reference+1 test sentence	single	∞ levels	(not available)

Each of these tasks was deployed in a separate experiment to collect acceptability judgments for 50 two-condition phenomena already known to give rise to acceptability contrasts.[2] The sensitivity of the four experiments was then empirically estimated by evaluating their *rate of statistical detection of acceptability rating differences* in a series of resampling simulations based on the acceptability judgments collected for each of the 50 pairwise contrasts tested in each experiment (see Section 3.4 for a description of the resampling procedure). For what is most relevant to our purposes, S&A's results show that FC is by far the most sensitive of the four tasks at detecting pairwise contrasts: for all effect sizes, FC requires the smallest sample sizes to produce a well-powered experiment under null hypothesis and Bayes Factor tests. Their results also indicate that ME and LS are approximately equally sensitive, albeit less sensitive than FC, and that YN is *generally* the least sensitive of the four tasks.[3]

2.2 The Dilemma

Based on S&A's results, FC appears to be an optimal choice when the nature of the research question is to ascertain the existence of an acceptability contrast between two (or more) sentence types. However, FC exhibits an important trade-off between its sensitivity advantage and the limitations imposed by its pairwise nature (see also Schütze & Sprouse, 2014; Sprouse & Almeida, 2017, for discussion). Table 2 provides a brief comparison of the four acceptability judgment tasks compared in S&A in regard of their suitability for different aspects of data collection and analysis.

FC is a categorical task (like YN) — given a pair of sentences, only one can be selected as the most acceptable. This means that FC can only indirectly provide information about the size of the difference between conditions in the form of the proportion of selections of the two sentences in the pair. This has at least two consequences. The first is that FC is less well-suited than a more gradient task (like LS or ME) for testing hypotheses that hinge on differences in effect sizes.[4] Even though we can use the proportion of selections in each

---

[2] A full list of the phenomena tested in S&A's study with example sentences along with their Cohen's *d* is available at *https://doi.org/10.5334/gjgl.236.s1*.

[3] This last observation is relative to the choice of approach to hypothesis testing. In S&A's study, the most pronounced differences in sensitivity between LS, ME vs. YN are found under null hypothesis tests; yet no such differences are observed under Bayes Factor tests. More generally, S&A note that the statistical detection rates obtained by Bayes Factors are lower than those obtained by null hypothesis tests for FC, LS and ME, while the reverse is true for YN.

[4] An anonymous reviewer points out that effect size predictions are particularly susceptible to the fact that every acceptability judgment effect requires an explicit formulation of (i) a grammatical theory, (ii) a sentence processing theory, and (iii) a theory of the task, and all three are active areas of investigation in the field. Nonetheless, effect sizes do play a role in syntactic theorizing, particularly in gradient approaches to syntax (a.o., Keller, 2000; Featherston, 2005), but also in less gradient approaches, like various claims about the partial amelioration of island effects with complex *wh*-phrases (see Szabolcsi & Lohndal, 2017, for a review). In these cases, experimenters attempt to hold as much of the variability due to sentence processing and task effects constant as possible; but as the reviewer rightly points out, it is an open question to what extent this is sucessful.

Table 2: Comparison of the acceptability judgment tasks in S&A with respect to various goals of formal experimentation. The symbols – vs. + indicate a distinction between 'suitable/well-suited' vs. 'unsuitable/less well-suited'. LS and ME appear to be more flexible for various types of analyses; YN is less well-suited, and FC is the least well suited.

	FC	YN	LS	ME
pairwise comparisons	+	+	+	+
factorial comparisons	–	+	+	+
post-hoc comparisons	–	+	+	+
effect sizes	–	–	+	+
relative acceptability	+	–	+	+
absolute acceptability	–	+	+	+

direction as a type of effect size, it is not clear what generates selections in the 'incorrect' direction in a pair. Are those errors? Is there some sort of stochastic process that leads participants to choose the other sentence in some proportion of trials? If so, what is the function that drives those errors or that stochastic process? Is it completely driven by the difference in acceptability between the two sentences, or are other factors also at play? To use the proportion of selections as a meaningful effect size, we need more research into the factors that drive selections in the opposite direction. This leads to the second consequence: because FC is less well-suited to the accurate estimation of effect sizes than other tasks, it is less well-suited to higher-order factorial designs that compare two or more effect sizes to each other (i.e., an interaction). Although it is possible to compute a differences-in-differences score for the proportion data for two pairs of sentences in an FC task, the uncertainty underlying the source of each of the differences will compound in the interaction.

FC comes with practical limitations as well. FC is ill-suited for post-hoc comparisons since the only comparisons that can be performed are those that are already constructed as pairs in the experiment itself. This limitation does not arise for the other tasks because participants report their acceptability judgments for each test sentence separately, allowing the comparison of every condition to every other condition. FC also provides less information than other tasks about where a given sentence type stands on the overall spectrum of acceptability (e.g., high, low, etc.). FC provides information about the location of one sentence relative to its pair; rating tasks like LS and ME provide information about the location of a sentence relative to (i) the participant's interpretation of the scale numbers, and (ii) the effect of the other sentences in the experiment on the participant's interpretation of the scale numbers. No task provides independent information about the location of a sentence along the scale of acceptability, but LS and ME provide more flexibility for exploring that information than FC.

In sum, because FC is specifically designed for the comparison of two (or more) conditions, it is arguably more sensitive at detecting a relative difference between these conditions than the other tasks, which only compare conditions indirectly through a response scale; but because of its very specialized design, FC is also ill-suited or less well-suited than the other tasks for a variety of analyses. Ideally, we would like to find a good compromise between the sensitivity advantage of FC and the analysis advantages of the other tasks. The goal of this study is to determine to what extent the component features of these tasks may be mixed and matched to yield a more optimal combination of advantages. In the following, we discuss our motivations for focusing on mode of presentation, number of responses, and use of labels.

2.3 Innovations of the Present Study

2.3.1 Mode of presentation

In FC, contrasting conditions are presented *jointly* so that participants always have access to a theoretically-relevant reference point when judging the acceptability of each of the test sentences (see Table 1). By contrast, in LS and YN, contrasting conditions are always presented *individually* and, in ME, the reference sentence, called the 'standard', typically does not form a theoretically-relevant contrast with the target sentence. Here, we investigate the hypothesis that part of the sensitivity advantage exhibited by FC derives from the joint presentation of contrasting conditions. Crucially, we note that this task feature can be easily implemented in most acceptability judgment tasks, with both pairwise and factorial designs.

To test this hypothesis, we will systematically compare the sensitivity of single and joint presentation across the LS and YN tasks. Concretely, we will compare the sensitivity results for the YN and LS tasks from S&A to the results of new experiments implementing joint presentation as illustrated in (4) and (5).

(4)   **YN task with joint presentation**

Item = { There has been a man considered sick.   Yes ○ No ○ }
       { There has been considered a man sick.   Yes ○ No ○ }

(5)   **LS task with joint presentation**

Item = { There has been a man considered sick.   least acceptable $\begin{smallmatrix}1&2&3&4&5&6&7\\○&○&○&○&○&○&○\end{smallmatrix}$ most acceptable }
       { There has been considered a man sick.   least acceptable $\begin{smallmatrix}1&2&3&4&5&6&7\\○&○&○&○&○&○&○\end{smallmatrix}$ most acceptable }

In the standard YN and LS tasks (see the appendix for examples), contrasting sentence types are presented individually, i.e. one at a time, and participants are asked to indicate for each sentence whether, or to what extent, this sentence sounds acceptable to them. Participants' responses to previous items would surely influence responses to following items, but the direct comparison of different items is not part of the task.[5] However, in the joint-variants of these tasks, contrasting sentence types are now presented simultaneously, and therefore participants can evaluate the acceptability of each test sentence by comparing them to each other, exactly as in the FC task.

2.4 Number of response options

S&A's results indicate that there is a benefit in offering speakers more than two response options: LS and ME both show increased sensitivity relative to YN. One might then predict that ME, with its infinite number of response options, would show increased sensitivity over LS, with its finite number of response options. This prediction is not borne out by S&A's results as LS and ME appear to be approximately equally sensitive (see Weskott & Fanselow, 2011, for similar observations). However, there may be an interfering factor in this comparison: a number of studies have suggested that participants do not (or can not) take advantage of the features of ME. It has been observed for instance that, although ME offers an infinite number of rating possibilities, participants tend to use a small set of numbers repeatedly (Featherston, 2008; Weskott & Fanselow, 2011; Sprouse, 2011a). More dramatically, Sprouse (2011a) found that participants fail to make ratio compari-

---

[5] In Magnitude Estimation tasks, one may use a 'standard' to make all judgments relative to a single reference point, which essentially serves as a pivot to compare any two sentences. Such paradigms come with other difficulties, as reviewed below.

sons of the acceptability of two sentences, contra the primary assumption of ME.[6] These limitations suggest that, despite the increased number of response options in ME, participants may be treating ME as a kind of LS task, only with a virtually infinite response scale. This line of explanation would align well with the finding that ME and LS are equally sensitive. But it also means that we cannot take S&A's results to indicate that there is no benefit to increasing the number of response options beyond 7. Instead, we need a comparison between LS and a task other than ME.

To overcome these limitations, we propose to compare YN (binary) and LS (7 points) to continuous sliders (CS henceforth). As continuous response scales, sliders allow participants to potentially rate every sentence type differently, just like magnitude estimation. Since they are not numeric, sliders also offer a more natural alternative for participants who may be less familiar with number-based psychological experiments: their use does not require any mathematical reasoning on the part of the participants, only spatial reasoning. In our study, sliders were implemented as shown in (6), and were used as response scales to construct the eponymous CS task.

(6) **Continuous Slider (CS)**
There has been a man considered sick.   least acceptable ▭▭▭▭▭▭ most acceptable

In the CS task, each test sentence was presented together with a continuous scale ranging from 'least acceptable' (leftmost) to 'most acceptable' (rightmost), i.e. the same anchors as in the LS task. Participants were asked to evaluate the acceptability of each test sentence and to give their answers along the continuous scale by using the cursor to set the right end of a blue line between the two endpoints of the scale. Consistent with our line of investigation, the CS task was tested with single presentation, as in (6), as well as with joint presentation, by presenting two contrasting sentences jointly with one slider each.

2.5 Use of labels

Labels impose predefined categories on the response options. They are also a salient aspect of a judgment task that is easy to manipulate. Yet, labelling has received little attention in the sensitivity literature.[7] The question we aim to explore is to what extent the use of fewer/more labels increases sensitivity. We note for instance that it is common practice in field research to not impose any predefined category and rather let speakers sort sentences in different categories that align with their preferred grouping. While there is an obvious practical advantage in letting speakers choose boundaries they feel comfortable with, it is an open empirical question whether the use of less specified (or less directive) response scales yields a concrete advantage in sensitivity.

To explore this question, we devised an impoverished variant of the fully labeled, 7-point Likert scale (LS_full henceforth) used in S&A featuring anchors but no intermediate labels, (7). This LS-variant, which we will refer to as LS_endpoint, was used once again

---

[6] Following Sprouse (2011a) and others, we note that the fact that the cognitive assumptions of ME do not hold for participants in acceptability judgement experiments disqualifies the idea that ME should be taken as a 'gold standard' for experimental syntax (a.o., Bard et al., 1996; Cowart, 1997; Keller, 2000). As of today, there is no evidence that ME provides a type of measurement reliability that other scaling tasks cannot achieve or delivers more meaningful data than other acceptability judgment tasks (a.o., Featherston, 2008; Weskott & Fanselow, 2011; Sprouse, 2011a).

[7] See however Križ and Chemla (2015) for a discussion of the use of 'completely true'/'completely false' rather than 'true'/'false' responses for semantic and pragmatic judgments, and the use of 'neither' rather than 'can't say' for a third option. They also report that the particular choice of labels could be more important if participants enter their responses by clicking on these labels rather than by pressing associated response keys, suggesting that in the latter case participants may more easily ignore the specific phrasing of the labels.

in two different experiments, one featuring single presentation (one test sentence with one LS_endpoint) and the other featuring joint presentation (two contrasting sentences with one LS_endpoint each).

(7)   **Likert Scale with endpoints only (LS_endpoint)**

There has been a man considered sick.   least acceptable ○ ○ ○ ○ ○ ○ most acceptable

Our motivations for including this response scale were twofold. First, by comparing LS_endpoint to LS_full, we will be able to isolate the effect of adding intermediate labels. Second, LS_endpoint also offers a more direct point of comparison than LS_full for quantifying the effect of using additional rating options: YN and CS, by definition, can only occur with endpoint labels, making LS_endpoint a minimal comparison with them in terms of the number of response options.

2.6 Summary of the tasks and research questions

Table 3 provides a description of the 10 acceptability judgment tasks compared in our study. The names of the tasks are of the form SCALE-{1,2}, where SCALE is the response scale used for the task (YN, LS_full, LS_endpoint, CS, ME, FC) and {1,2} is the mode of presentation of contrasting conditions (1 for single and 2 for joint presentation). YN-1, LS_full-1, ME-1 and FC-2 correspond to the four original tasks from S&A (previously named YN, LS, ME and FC, respectively, see Table 1).

   For the mode of presentation, we are essentially interested in knowing whether joint presentation increases sensitivity. We will address this question by comparing the single presentation tasks to their joint presentation variants (for YN, LS_full, LS_endpoint and CS). The global comparison between joint and single presentation will also enable us to settle two additional questions regarding the effect of joint presentation. First, by comparing ME-1 to the joint presentation tasks, we will be able to determine whether the benefit of theoretically-relevant joint presentation outranks the (hypothetical) benefit of the theoretically-irrelevant joint presentation inherent in ME-1, which always includes a comparison to the 'standard'. Second, by comparing FC-2 to the other joint presentation tasks, we will be able to determine whether the sensitivity advantage of the FC task is entirely rooted in the joint presentation of contrasting conditions, or alternatively results from the combination of this design feature together with the forced-choice procedure itself.

Table 3: Schematic description of the 10 acceptability judgment tasks compared in our study as a function of the three task features of interest.

Task name	Tasks from	Mode of Presentation	Number of Response Options	Use of Labels
YN-1	S&A	single	2 levels	(obligatory)
LS_full-1	S&A	single	7 levels	full labels
LS_endpoint-1	new	single	7 levels	endpoints only
CS-1	new	single	∞ levels	endpoints only
ME-1	S&A	single	∞ levels	(not available)
YN-2	new	joint	2 levels	(obligatory)
LS_full-2	new	joint	7 levels	full labels
LS_endpoint-2	new	joint	7 levels	endpoints only
CS-2	new	joint	∞ levels	endpoints only
FC-2	S&A	joint	1 forced-choice	(not available)

To assess the effect of additional rating options, we will compare YN, LS_endpoint and CS (i.e., 2 vs. 7 vs. ∞ levels) for both modes of presentation. Next, to assess the effect of additional labels, we will compare LS_full to both LS_endpoint and CS (i.e., full labels vs. endpoints only), once again for both modes of presentation. As discussed above, this second part of our study is more exploratory. To the best of our knowledge, these minimal comparisons between response scale features have never been investigated before, making it difficult, if not impossible, to state clear expectations, especially concerning the possible interactions of these features with the mode of presentation. We hope to begin to close this empirical gap.

3 Acceptability Judgment Experiments

Each of the 10 tasks was deployed in a separate experiment. Our dataset was gathered as follows. For the FC-2, ME-1, YN-1 and LS_full-1 experiments, we used the datasets already collected and analyzed by S&A. We extended S&A's datasets by conducting 6 additional experiments, one for each of the 6 new tasks described above. We used S&A's original experimental materials and analysis pipelines so that the results are directly comparable, modulo the differences in mode of presentation, number of response options, and use of labels.

3.1 Materials

The phenomena tested in all 10 experiments were the same as those originally tested in S&A's experiments: 50 two-condition phenomena from the set of 150 two-condition phenomena randomly sampled from all of the articles published in *Linguistic Inquiry* between 2001 and 2010 for the large-scale replication study by Sprouse et al. (2013). These 50 phenomena were among the 139 phenomena (out of 150) that replicated using both LS and ME tasks in Sprouse et al. (2013) in terms of both directionality (i.e., the effects were in the predicted direction) and statistical significance (i.e., the effects passed the conventional $p < .05$ criterion). We refer the readers to S&A for the description of the selection procedure and for discussion of its rationale.

The 50 phenomena were tested with each of the 10 tasks: one experiment per task, with each experiment containing 100 sentence types, one token each of the two conditions per phenomenon. Typos in the materials for three of the phenomena were already reported in S&A (for a discussion of any hypothetically possible problems with the materials see Sprouse et al., 2013). These typos were left unchanged in the 6 new experiments so as to maintain perfect parallelism in the materials used across experiments. As in S&A, these phenomena were removed from the analysis. Table 4 shows the distribution of the remaining 47 phenomena as a function of effect size. Effect sizes are grouped into four categories following the guidelines for the interpretation of *d*-values suggested in Cohen (1988, 1992a, 1992b). A full list of the phenomena, with example sentences along with

Table 4: Distribution of the 47 two-condition phenomena tested and analyzed in our study as a function of effect size, here grouped into four categories. These phenomena are those used and described in S&A's original study.

Category of effect sizes		Range of *d*-values					Phenomena
Small		0.15	≤	*d*	<	0.5	9
Medium		0.5	≤	*d*	<	0.8	11
Large		0.8	≤	*d*	<	1.1	6
Extra large		1.1	≤	*d*	≤	1.96	21
	Range:	0.15	≤	*d*	≤	1.96	Total: 47

their Cohen's *d,* is provided in the appendix of Sprouse and Almeida (2017) (available online at *https://doi.org/10.5334/gjgl.236.s1*).

The materials used in all 10 experiments were the same, and the same as those constructed for the experiments reported in S&A and Sprouse et al. (2013): there were 8 lexicalizations of each sentence type, leading to 8 lexically matched sentence sets for each of the 50 phenomena (see Sprouse et al., 2013; Sprouse & Almeida, 2017, for detail). For the single presentation experiments, the 8 lexicalizations were distributed among 8 lists using a Latin Square procedure, ensuring that participants did not see more than one sentence from each lexically-related set. Each of the lists was pseudo-randomized so that the two contrasting conditions from a single phenomenon did not appear sequentially. For the joint presentation experiments, the 8 lexicalizations were maintained in lexically-matched pairs, i.e. in minimal pairs. The pairs were distributed among 8 lists, and the order of presentation of each pair was counterbalanced across lists to minimize the effect of response biases on the results (e.g., the strategy of always choosing the first item). Two copies of each list were created, resulting in a total of 16 lists. The order of the pairs in each of the lists was randomized.

3.2 Participants

Participants in the new experiments were recruited online using the Amazon Mechanical Turk (AMT) marketplace as in S&A's experiments (see Sprouse, 2011b, for evidence of the reliability of acceptability judgment data collected using AMT), and were paid $2.50 for their time (typically less than 10 minutes). Participant selection criteria were enforced as follows. First, the AMT interface automatically restricted participation to AMT users with a US-based location. Second, we use a WorkerID-based checking procedure to prevent participants from participating more than once in our experiments. Third, we included two questions in each survey to assess language history: (1) Did you live in the US from birth until (at least) age 13? (2) Did both of your parents speak English to you during those years? It was explicitly stated that these questions were not used to determine eligibility for payment so that there was no financial incentive to lie. Participants who failed to answer 'yes' to both of these questions were excluded from the analysis. Table 5 reports

Table 5: Number of participants recruited and number of participants excluded from analyses as a function of the experiment. Each participant only ever completed one survey.

Experiment		Number of participants				
		Recruited		Excluded		Included
S&A experiments						
FC-2		144		0		144
ME-1		144		0		144
YN-1		144		5		139
LS_full-1		144		4		140
New experiments						
YN-2		131		7		124
LS_full-2		111		2		109
LS_endpoint-1		104		2		102
LS_endpoint-2		112		6		106
CS-1		113		6		107
CS-2		113		8		105
	Total:	1260	Total:	40	Total:	1220

the number of participants recruited for each experiment and the number of participants that were excluded from the analysis based on the language history questions or for obvious attempts to cheat. Based on the very low rate of exclusion observed in S&A's study, we slightly lowered the number of participants to be recruited for the six new experiments. Crucially, these sample sizes remained large enough to provide us with the ability to empirically estimate sensitivity for the same range of sample sizes in our resampling simulations.

3.3 Presentation of the experiments

Participants first saw a consent form, followed by the specific set of instructions corresponding to the task used in the experiment. To parallel S&A's experiments, no specific instructions were given regarding how to theoretically interpret the mode of presentation of sentences: participants were only told that they would be presented with single sentences or pairs of sentences. There was no explicit practice phase for sentences. However, unmarked practice items were placed as the first items of each experiment, prior to test items, to help participants decide how to use the response scale. There were 6 one-sentence practice items in the single presentation experiments, and 2 two-sentence practice items in the joint presentation experiments. These items were not included in the analysis; they were only used as a type of unannounced practice to expose participants to acceptability contrasts of different strengths before encountering test items. The test items were 100 monadic test sentences in the single presentation experiments, and 50 pairs of contrasting test sentences in the joint presentation experiments. Language history and demographic questions were included either at the beginning of the experiment or at its very end. Participants completed the experiments at their own pace.

All of the experiments were advertised on the AMT website. YN-2, CS-1 and CS-2 experiments were dynamic web-based experiments, programmed using a mix of JavaScript and HTML, and hosted on an AMT-external dedicated server. Each participant recruited through AMT was provided with a link to access the experiment, and at the end of it, was given a unique code to enter on the AMT website. The sentence items in these experiments were each presented one at a time on the screen. The other experiments were static web-based experiments, created using an HTML template, and directly hosted on the AMT-server. The sentence items in these experiments were presented all at once on the screen.

3.4 Data analysis

Following S&A, we operationalize the notion of sensitivity by estimating *the rate of statistical detection of acceptability rating differences* for each of the pairwise phenomena, in each of the experiments, at every sample size between 5 and 100 participants, by performing resampling simulations.[8] In short, these resampling simulations treat samples of participants as full populations, draw samples from each of these populations, and treat these samples as mini-populations to empirically estimate the rate of statistical detection (see S&A for discussion and refinements). To illustrate, the procedure to establish the rate of detection of a particular phenomenon for a sample of size 5 goes as follows:

---

[8] The rate of statistical detection is proportional to both effect size and sample size. First, the smaller the effect tested on a given sample, the lower the detection rate for that effect with that sample. Second, the smaller the sample for testing a given effect, the lower the detection rate for that effect. As a result, the wider the range of effect and sample sizes we consider when assessing the sensitivity of an experiment, the more informative our assessment will be. Here the choice of the range 5–100, originally proposed in S&A and persevered here for uniformity, is motivated by simple considerations of relevance: 5 is one of the lowest sample sizes that can return a significant result, while 100 is a likely upper bound for most acceptability judgment experiments.

(8)　　　*Resampling simulations: example for samples of size 5 (adapted from S&A)*

　　　a.　Draw a random sample of 5 participants (with resampling, so that participants can be drawn more than once), and run a statistical test on that sample.

　　　b.　Repeat (a) 1000 times to simulate 1000 experiments with a sample size of 5, and then calculate the proportion of simulations that resulted in a statistical test beyond the pre-established criterion for significance.

　　　c.　The proportion obtained, namely *the rate of statistical detection*, is an empirical estimate of sensitivity for samples of size 5.

Suppose that, following this procedure, 50 out of 1000 simulations for samples of size 5 resulted in a statistical test beyond significance. Then, the rate of detection of the relevant phenomenon for samples of that size will be 5%. This procedure can be then repeated for samples of size 6, 7, 8,… up to 100, and performed for all 47 phenomena, to obtain a nearly exhaustive assessment of the detection rate relationships for sample sizes and effect sizes as a function of the experiment. As sample sizes and effect sizes increase, so does the rate of statistical detection. Yet this increase might be more or less steep depending on the experiment, especially for smaller sample and effect sizes, revealing fine-grained differences in sensitivity among tasks.

　　In addition to the task, effect size, and sample size, the rate of statistical detection further depends on the choice of statistical test, and therefore on the choice of approach to hypothesis testing. To maximize the range of information we can extract from our resampling simulations, we followed S&A's strategy, and tested both a null hypothesis test and a Bayesian hypothesis test for every simulation. The statistical tests used to analyze the 6 new experiments were also the same as those used in S&A's study. This enabled us to directly compare our study to the results of S&A's resampling simulations for their original experiments without having to run new statistical tests, and then extend these previous results by running the same tests for the new experiments. In addition, these tests are relatively fast to compute, allowing us to minimize the amount of computational time that we needed to run for each approach to hypothesis testing (around 27 million simulations: 47 phenomena × 96 sample sizes × 1000 simulations × 6 new tasks).

　　For the YN-2 task, we ran repeated-measures sign-tests and Bayes Factor calculations, as S&A did for FC-2 and YN-1. For the new tasks involving graded scales (LS_full-2, LS_endpoint-1, LS_endpoint-2, CS-1 and CS-2), we ran repeated-measures t-tests and Bayes Factor calculations, as S&A did for ME-1 and LS_full-1; for these tasks, statistical tests were run on the z-score transformed ratings.[9] Bayes factors were calculated using the equation provided in Rouder, Speckman, Sun, Morey, and Iverson (2009). For the criteria of significance, we adopted the criteria proposed in S&A based on the conventions often used for publication in the cognitive sciences. For sign-tests and *t*-tests, we used a *p*-value below .05 ($p < .05$) as a criterion of significance. For Bayes Factor tests, we used a Bayes factor above 3 ($BF > 3$), i.e. the observed data is 3 times more likely under the experimental hypothesis than under the null hypothesis, which is considered 'substantial evidence' for the experimental hypothesis (Jeffreys, 1967).

　　The resampling simulations resulted in 45,120 rate of detection estimates for each approach to hypothesis testing (47 phenomena × 96 sample sizes × 10 tasks). Interpreting

---

[9] The z-score transformation is a linear transformation common to many scale-based data types. It maintains all of the relationships that exist within the data, and allows us to express each participant's responses on a standardized scale (a scale based on standard deviation units), hence removing some forms of scale bias (Featherston, 2005; Sprouse & Almeida, 2012; Sprouse et al., 2013). For a discussion of the relationships between scale-based data types (e.g., LS and ME judgment data), transformation types (e.g., z-score transformation vs. log-transformation) and statistical tests (parametric vs. non-parametric), we refer the reader to Sprouse and Almeida (2011, pp. 17–19).

the results therefore requires some amount of summarization. Following S&A, we first grouped our 47 phenomena into four categories of effect sizes (see Table 4), and then plotted the relationship between sample size and mean rate of statistical detection (our proxy measure for mean sensitivity) for each category of effect size. The resulting detection rate curves provide comprehensive visual representations of the data points (i.e., the empirical estimates of sensitivity), which we can use to ask how much the rate of statistical detection depends on different sample sizes, and thus to compare the sensitivity of our 10 tasks along the three task features of interest. As many fields of cognitive and social sciences have adopted the suggestion in Cohen (1988, 1992b, 1992a) that 80% detection is necessary for an experiment to be considered 'well-powered', we will also provide this information in our plots by indicating the sample size at which the mean rate of statistical detection first reaches 80% in our simulations.

4 Results

4.1 Mode of presentation

To assess the effect of the mode of presentation on design sensitivity, we compared the single presentation tasks to their joint presentation variants for the four response scales tested with both modes of presentation, namely YN, LS_full, LS_endpoint and CS. For these purposes, we plotted the relationship between sample size and mean detection rate for null hypothesis and Bayes Factor tests for each category of effect size (small, medium, large and extra-large effects) as a function of the mode of presentation (single vs. joint). The results are summarized in the 4×4 grid of detection rate curves in Figure 1.[10]

  We found that joint presentation offers a clear advantage in sensitivity over single presentation across-the-board: in every cell of the grid, when there is a difference between the two curves, the detection rate curve for joint presentation is systematically above the one for single presentation. In other words, for each task and each category of effect size, joint presentation yields a detection rate increase over single presentation. This increase is most pronounced for small and medium effect sizes, and least pronounced for larger effect sizes, yet still detectable, i.e. the joint presentation tasks still require the smallest sample sizes to reach 80% detection for large and extra large effects. In sum, joint presentation increases sensitivity, on both approaches to hypothesis testing and regardless of the features of the response scale.

  We also set out to address two further questions: Is joint presentation with a theoretically-relevant reference point more sensitive than joint presentation with a theoretically-irrelevant reference point (ME-1 vs. joint presentation tasks)? And is joint presentation with a response scale as sensitive as the forced-choice procedure (FC-2 vs. other joint presentation tasks)? The detection rate curves of these tasks are plotted in Figure 2. As we are interested here in evaluating the sensitivity of the group of joint presentation tasks relative to that of ME-1 and FC-2, the distinctness of their detection rate curves is intentionally minimized by using related shades of green. Differences within the set of joint presentation tasks will be analyzed in the next section when we turn to the features of the response scale.

  The first result is that ME-1 is *generally* less sensitive than the joint presentation tasks. For Bayes Factor tests, for each category of effect size, ME-1 is less sensitive (for small, large and extra-large effects) or at most as sensitive (for medium effects) as the least

---

[10] We have chosen to not include any measure of variability (credible or confidence intervals) in the plots in Figure 1 and 2 for visual clarity. This is because our estimates are based on 1000 resampling simulations, which leads to exceedingly small intervals, often barely larger than the lines in those plots. Hence, a visually distinguishable separation between two curves in those plots can confidently be taken to indicate the existence of a potentially reliable difference between those curves.
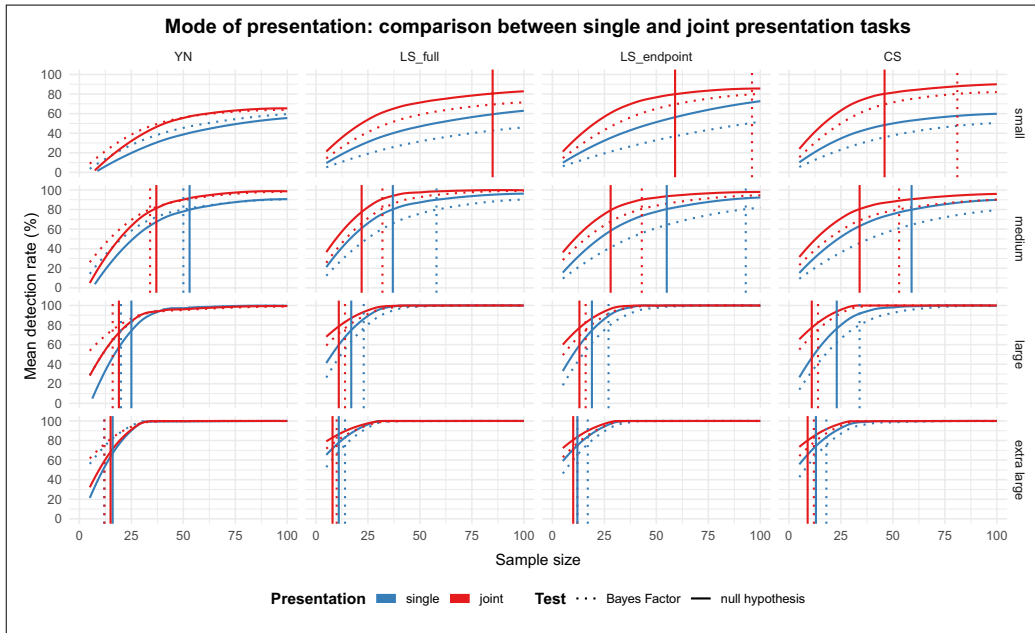
Figure 1: Detection rate curves for null hypothesis tests (solid) and Bayes Factor tests (dotted) displaying the relationship between sample size (x-axis) and estimated detection rate (y-axis) for each of the four tasks (columns) and each effect size category (rows) as a function of the mode of presentation: blue for single presentation and red for joint presentation. The vertical lines represent the sample size that first reaches 80% detection (or above) in our simulations. Cells with only one line indicate that the 80% threshold is obtained only for one mode of presentation, namely joint presentation. Cells with no line did not reach 80% detection with sample sizes less than or equal to 100. For clarity, detection rate curves were plotted by locally weighting the empirical estimates of detection (the percentage of simulations below the relevant significance threshold, here $p < .05$ and $BF > 3$, averaged over all phenomena belonging to each category) using the loess fitting procedure. Joint presentation yields a detection rate increase over single presentation across-the-board.
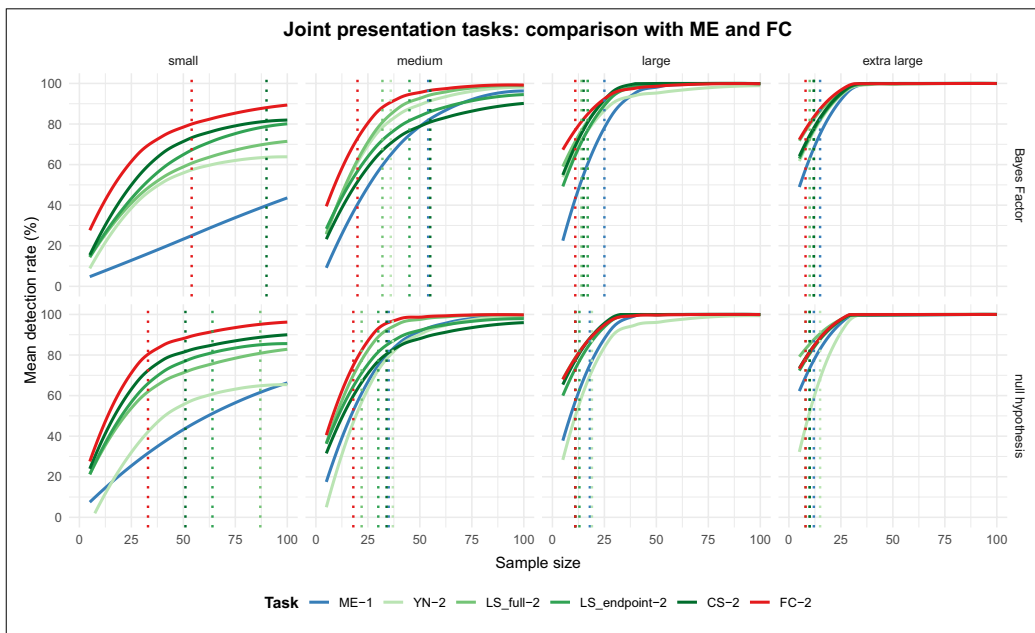
Figure 2: Detection rate curves for ME-1, YN-2, LS_full-2, LS_endpoint-2 and FC-2 displaying the relationship between sample size and estimated detection rate as a function of the category of effect size (columns) and the approach to hypothesis testing (rows): ME-1 is less or as sensitive as the least sensitive of the joint presentation tasks, while FC-2 is more or as sensitive as the most sensitive of the other joint presentation tasks.

sensitive of the joint presentation tasks in that category. Similar observations hold for null hypothesis tests with one exception: for null hypothesis tests, ME-1 receives a small sensitivity increase while YN-2 receives a small sensitivity decrease, conspiring to make them approximately equally sensitive across-the-board. The second result is that FC-2 still exhibits some advantage in sensitivity over the other joint presentation tasks, especially for small and medium effect sizes. In absolute numbers, FC-2 is basically the task that requires the smallest sample sizes to reach 80% detection for each category of effect size and under both approaches to hypothesis testing. Hence, joint presentation increases sensitivity, but this design feature is not enough by itself to explain the sensitivity advantage of the FC task: part of this advantage also stems from the forced-choice procedure itself.

Recall that all but the YN-2, CS-1 and CS-2 experiments were static web-based surveys in which all items were presented on the same webpage (see Section 3.3). Since participants could access previous items in the single-versions of these static surveys, they may have recognized that certain sentence types were similar and sometimes decided to cross-check their answers, indirectly mimicking the effect normally induced by joint presentation in their joint-counterparts. Although possible, this fall-back strategy seems unlikely in practice in the context of our experiments: the absence of lexically-matched pairs in the single presentation experiments together with the fact that each phenomenon occurred only once in each experiment conspires to make the identification of a pairwise contrast very difficult with single presentation. The present results show that, had this fall-back strategy been sometimes used, it is still more beneficial to present experimental conditions jointly rather than singly, regardless of whether experimental items are presented on the same page or on different pages.

4.2 Number of response options

To assess the effect of additional rating options, we compared the YN, LS_endpoint and CS tasks for both modes of presentation, both approaches to hypothesis testing, and each category of effect size. The results are summarized in the 4×4 grid of detection rate curves in Figure 3.

First, the detection rate curves for LS_endpoint and CS show little-to-no difference across-the-board: LS_endpoint and CS are approximately equally sensitive, regardless of the mode of presentation and choice of approach to hypothesis testing. Hence, there is no evidence of increased sensitivity of larger over smaller graded scales. Second, there are subtle interactions between the binary-graded distinction, the mode of presentation and the choice of hypothesis testing. With single presentation, the three experiments are approximately equally sensitive under null hypothesis tests, but YN has some advantage in sensitivity over LS_endpoint and CS under Bayes Factor tests. It is important to note that LS and CS use the same statistical calculations, while YN uses a different statistical calculation. This means that with results that distinguish LS/CS and YN/FC, we cannot separate out the contribution of the task from the contribution of the statistical calculation. Interestingly, with joint presentation, both LS_endpoint and CS receive a small sensitivity increase over YN for both types of hypothesis testing, suggesting that the sensitivity increase due to joint presentation is larger than the task/statistical-calculation increase for YN discussed above. This increase for LS and CS is primarily restricted to small effect sizes for Bayes Factor tests, but observable across all effect size categories for null hypothesis tests.

4.3 Use of labels

To assess the effect of additional labels, we compared LS_full with LS_endpoint and CS. The results are summarized in the 4×4 grid of detection rate curves in Figure 4. The detection rate curves show that the sensitivity of LS_full relative to LS_endpoint and CS also depends on the mode of presentation. With single presentation, LS_full has some
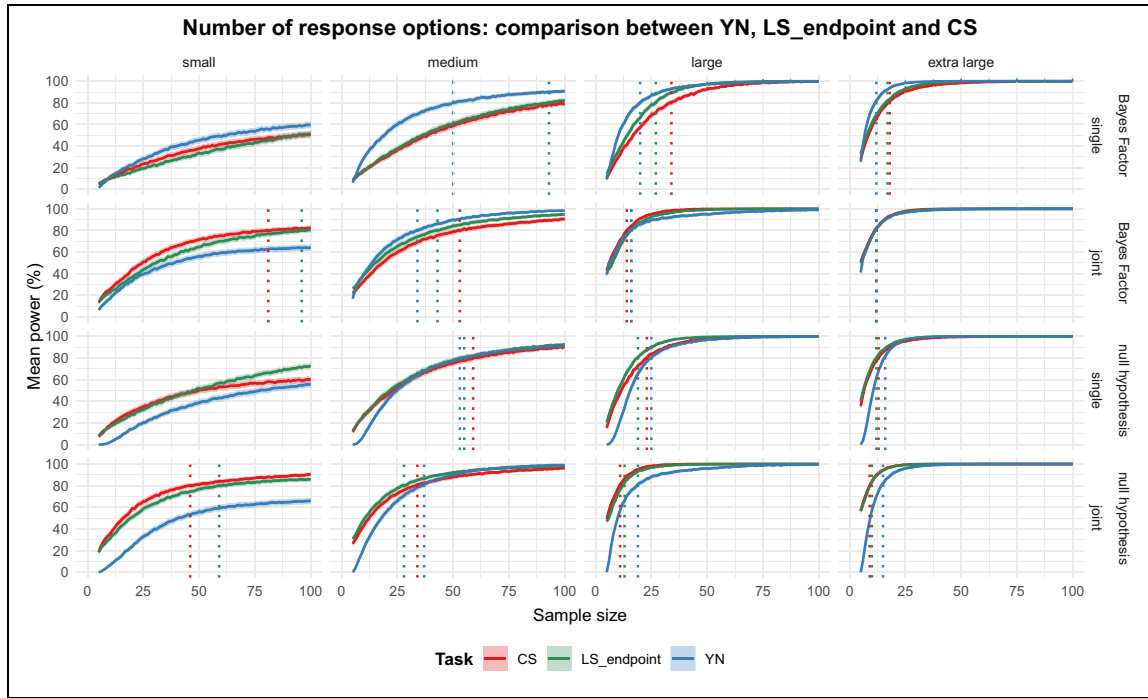
Figure 3: Detection rate curves for YN, LS_endpoint, and CS displaying the relationship between sample size and estimated detection rate for both approaches to hypothesis testing (Bayes Factor and null hypothesis) and both modes of presentation (single vs. joint) as a function of effect size category. Every estimate includes a 95% (Bayesian) credible interval based on the uninformative Jeffreys prior ($\alpha = \beta$ = 1/2), as calculated using the binom package in R. Vertical dotted lines represent the sample size that first reaches 80% detection (or above) in our simulations. The absence of such a line in a given cell indicates that the 80% detection threshold did not obtain with sample sizes less than or equal to 100. The main results are twofold. First, LS_endpoint and CS are approximately equally sensitive. Second, the sensitivity of both tasks relative to YN depends on an interaction of mode of presentation and hypothesis testing: YN is slightly more beneficial than LS_endpoint and CS with single presentation and Bayes factors, while the reverse is true for the other three possibilities: single presentation with null hypothesis testing, and joint presentation with both types of hypothesis testing.

advantage in sensitivity over LS_endpoint and CS, especially for medium effect sizes, and thus appears to be slightly more beneficial. However, with joint presentation, LS_endpoint and CS are more sensitive than LS_full at detecting small effect sizes. In sum, in a way similar to the results for number of responses discussed above, the use of less specified graded scales offers some advantage in sensitivity with joint presentation but not with single presentation.

5 Discussion

5.1 Mode of presentation

We designed our set of experiments to probe for and quantify over the benefit associated with two distinct ways of presenting contrasting conditions to participants, individually and jointly. Our study yields the following three main results:

1. Joint presentation shows increased sensitivity over single presentation.
2. Joint presentation generally shows increased sensitivity over single presentation plus a theoretically-irrelevant reference point, as in ME.
3. Forced-choice, which is inherently joint presentation, shows increased sensitivity over joint presentation plus a response scale.
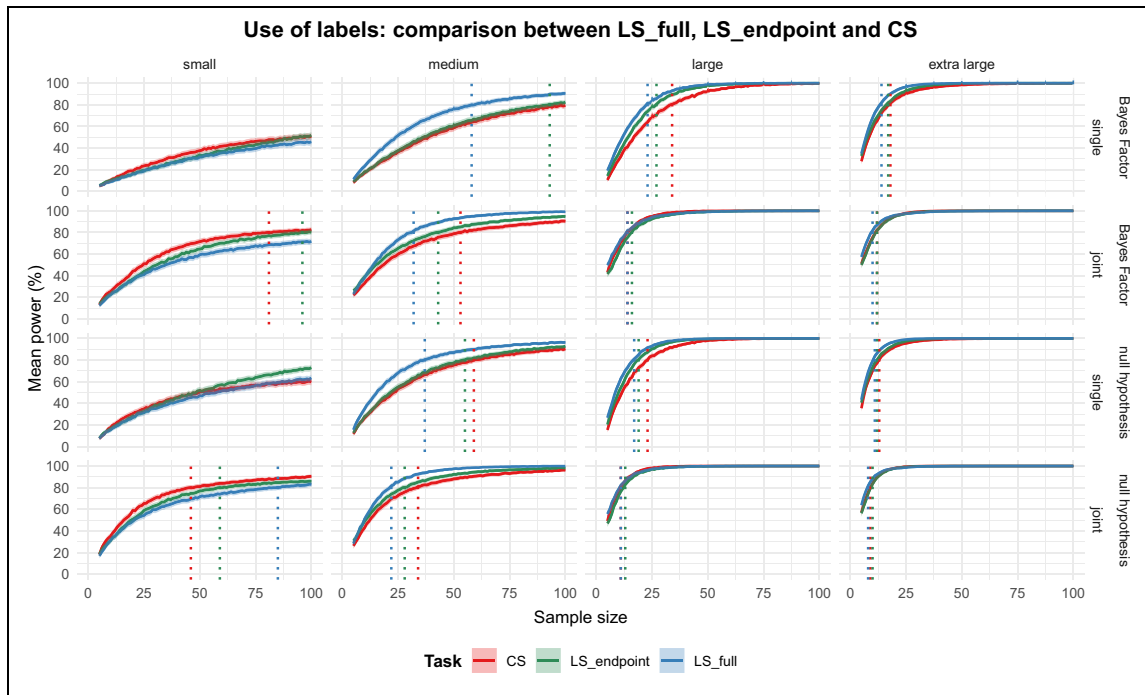
Figure 4: Detection rate curves for LS_full, LS_endpoint and CS displaying the relationship between sample size and estimated detection rate for both approaches to hypothesis testing (Bayes Factor and null hypothesis) and both modes of presentation (single vs. joint) as a function of effect size category. Every estimate includes a 95% (Bayesian) credible interval based on the uninformative Jeffreys prior ($\alpha = \beta = 1/2$), as calculated using the binom package in R. Vertical dotted lines represent the sample size that first reaches 80% detection (or above) in our simulations. The absence of such a line in a given cell indicates that the 80% detection threshold did not obtain with sample sizes less than or equal to 100. The sensitivity of LS_endpoint/CS relative to that of LS_full depends on the mode of presentation: LS_full is generally more beneficial than LS_endpoint/CS with single presentation, but slightly less beneficial than LS_endpoint/CS for small effect sizes with joint presentation.

These results establish that, when it comes to detecting pairwise contrasts, the FC task remains an optimal choice, and that part of FC's sensitivity advantage appears to stem directly from the forced-choice procedure itself. Yet these results also reveal that part of FC's advantage stems from the joint presentation of contrasting conditions. Since this task feature can be combined with any kind of response scales, researchers could choose to use this mode of presentation to potentially add additional sensitivity to other tasks, and, potentially, circumvent the shortcomings of the FC task (see Section 2.2 for discussion of those shortcomings). Though we can provide no additional empirical evidence about the relative merits and drawbacks of joint presentation, we would like to discuss in the remainder of this section some potential concerns that might arise with joint presentation, along with some thoughts about how serious these concerns might be (and perhaps some pointers to potential future projects).

One possible concern is that, unlike single presentation, joint presentation explicitly draws participants' attention to the contrasts under investigation. One may thus wonder whether joint presentation induces certain strategic effects that artificially boost the rate of statistical detection of acceptability rating differences, e.g. by leading participants to assume that all sentence types presented jointly are to be rated differently. While this question is reasonable, the benefit of joint presentation cannot be reduced to such strategic effects. First, joint presentation does not give any indication regarding the directionality of the contrasts at hand; yet what our results show is precisely that joint presentation

increases the rate of statistical detection at detecting acceptability contrasts *in the expected direction*. Second, for each category of effect size, joint presentation yields a sensitivity increase over single presentation, yet maintains the relative differences in rates of statistical detection that exist with single presentation: smaller effects are still harder to detect than medium effects, medium effects harder to detect than large effects, and so on. This observation is further substantiated in Figure 5. There, each point represents one of the 47 phenomena tested, the *x*-value represents the effect size obtained with single presentation, and the *y*-value represents the effect size obtained from joint presentation. The strong correlations ($rs > .77$) suggest that the relative ordering of effect sizes is preserved across both modes of presentation, that is, the effects which are the hardest/easiest to detect are the same across the two presentation modes. Furthermore, the slopes of the lines of best fit are all greater than 1, which suggests that joint presentation leads to larger effect sizes. In sum, larger effect sizes and greater sensitivity are obtained by offering participants the opportunity to ground their judgments on theoretically-relevant reference points.

Another possible concern could be that joint presentation might draw attention to theoretically-irrelevant differences. To our minds, this is a benefit of joint presentation, not a concern. Even though using the most sensitive mode of presentation may detect contrasts that happen upon further investigation to be driven by irrelevant factors, it also minimizes the risk of missing contrasts that are, or may be, of theoretical interest. As such, joint presentation proves useful not just for studies where all relevant linguistic factors have already been clearly identified, but also for more exploratory studies where the interpretation of the observed contrasts can be latter mediated by additional work.

A third, and related, possible concern is that joint presentation might be magnifying effect sizes, by encouraging a type of repulsion effect on the ratings of the two sentences. There is no way to eliminate this possibility: given that joint presentation yields larger effect sizes, this either means that effect sizes are magnified under joint presentation, or that effect sizes are diminished under single presentation. One option to potentially
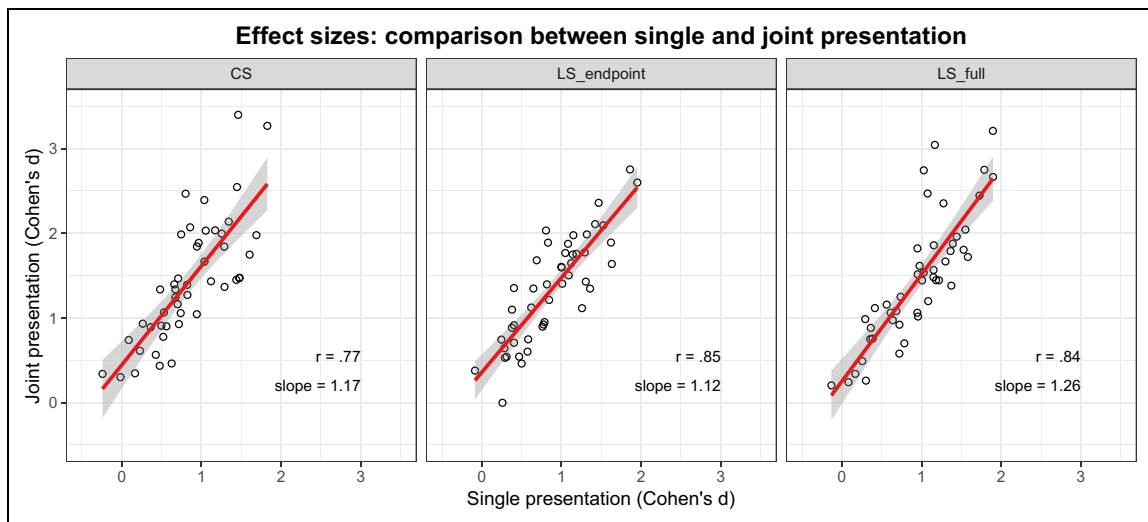

Figure 5: Effect sizes (Cohen's *d*) from the CS, LS_endpoint and LS_full experiments. Each point represents one of the 47 two-conditions phenomena from *Linguistic Inquiry* (2001–2010), the *x*-value represents the effect size obtained with single presentation, and the *y*-value represents the effect size obtained from joint presentation. The red line is the line of best fit (least squares) for the relationship between effect sizes with single and joint presentation, with 95% confidence intervals in grey. For each experiment we observe a strong correlation between effect sizes obtained with single presentation and with joint presentation (*rs* > .77), and slopes greater than 1.

minimize the repulsion effect would be to use experimental designs with three or more conditions – since acceptability is a one dimensional scale with endpoints in this task, with three conditions, the repulsion effect would at some point push one item closer to one of the others. However, to our minds, the magnifying effect of joint presentation is again a benefit, and not a concern. Joint presentation can only magnify differences that are already there. If there is no true difference, and the magnification is simply magnifying noise, then the effects should cancel out in the aggregate. Similar to magnification in other sciences, it is up to the experimenter to determine if what is magnified is theoretically meaningful or not. That said, as one anonymous reviewer noted, the YN task with joint presentation may present a case where the magnification of differences could introduce a distortion. If two sentences that would typically be given the same rating (both 'yes' or both 'no') are presented together, the magnification could lead a participant to give them different ratings, thus increasing the detection of the effect, but distorting the categorization of one of the sentences. This is thus a potential drawback to keep in mind for YN with joint presentation.

One potential advantage of joint presentation over single presentation beyond sensitivity is that it may make one of the implicit strategies adopted by participants in judgment tasks explicit, thus bringing it under the control of the experimenter. It is possible that, when some participants are asked to judge a sentence type, they attempt to verify their initial reaction by considering alternative sentence types, e.g. sentence types presented elsewhere in the experiment or simply constructed on the spot. When using single presentation, researchers have no principled way to control which implicit reference point participants rely on when evaluating a sentence type: for a given sentence, every participant can in principle consider a different reference point, perhaps ultimately resulting in implicit comparative judgments based on variable, unknown baselines. By contrast, joint presentation allows researchers to integrate this non-trivial dimension of the decision process directly into the design of the study.

5.2 Number of response options

The direct comparison of CS and LS_endpoint shows us that there is no evidence of increased sensitivity of larger graded scales (continuous sliders) over smaller graded scales (7-point Likert scale). This finding echoes S&A's previous finding that LS_full and ME are equally sensitive with single presentation, but teaches us further that this observation holds for other fine-grained scales like CS (which are not subject to the kind of practical limitations ME is subject to) independently of the mode of presentation. One limitation of these results is that they do not indicate whether there may be an increase in sensitivity for a finite number of response options that is larger than 7. We leave this question to future work.

5.3 Use of labels

Our investigation of the use of labels revealed two main results, both of which suggest an interaction between labelling and mode of presentation:

1. Graded scales with full labels, i.e. LS_full, are slightly more beneficial with single presentation, especially for medium effect sizes.
2. Continuous and graded scales with fewer labels, i.e. CS and LS_endpoint, are slightly more beneficial with joint presentation, especially for small effect sizes.

These results suggest that there is no absolute advantage or disadvantage in using more labels. However, this scale feature interacts in a subtle way with the mode of presentation

of contrasting conditions, revealing restricted benefits in certain combinations. This finding also raises an interesting question: why do graded scales with fewer labels give rise to increased sensitivity with, yet only with, joint presentation?

   One possible explanation is that this pattern of results reflects the existence of a tradeoff between the greater response flexibility offered by the CS and LS_endpoint scales, and the greater variance in responses that their flexibility tends to induce. In short, unlike LS_full, these scales do not impose any predefined categories on the many response options they offer, allowing participants to calibrate them as they wish to sort out sentences. This flexibility offers the theoretical possibility of substantially more fine-grained ratings (potential benefit), but it can also increase variance in responses (potential drawback), as evidenced for instance by the results from Weskott and Fanselow (2011) for LS and ME. With joint presentation, this drawback can easily be minimized since participants always have access to a theoretically-relevant point of comparison. In providing participants with an explicit benchmark, joint presentation can channel the evaluation process, permitting the CS and LS_endpoint scales to show their potential benefit. However, with single presentation, this drawback cannot be easily counteracted since participants may neither rely on explicit comparison points, nor provide their judgments in reference to labeled, and thus clearly identifiable, response categories. Of course, participants may attempt to monitor their judgments throughout the experiment by making up their own reference points, for instance by recollecting the sentences presented before in the experiment and by keeping track of their previous answers. These supplementation strategies are potentially cognitively demanding and therefore potentially prone to sub-optimal outputs (e.g., inaccurate recollection, irrelevant comparison points, etc.). One could argue then that, with single presentation, adding labels on the scale could play a facilitatory role in the evaluation process. As suggested by an anonymous reviewer, it could be that the labeled intermediate dots in LS_full provide a better anchor in memory than the unlabeled intermediate dots in LS_endpoint and the continuous space in CS.

5.4 A note on Bayes factors versus null hypothesis tests

By and large, the six new experiments replicate the finding from S&A that the null hypothesis tests used here (sign-tests and *t*-tests) tend to yield higher detection rates than the Bayes factor analyses (BF versions of sign-tests and *t*-tests). They also corroborate the boost that the YN task receives under a BF analysis, in comparison for instance to the LS and CS tasks (see Figure 3). As Sprouse and Almeida note, it is not clear to what extent this is an artifact of the different BF calculations used between YN tasks on the one hand, and LS and CS tasks on the other: it is possible that the BF analysis of binomial responses yields larger BFs on average than the calculation for continuous responses, which in turn could lead to greater rates of detection if the higher BFs cross the decision criterion ($BF > 3$ here) more often. We note this possibility here, and leave it for future research, as disentangling the role of the statistical analysis from the sensitivity of the task is beyond the scope of the current set of experiments.

6 Conclusion

In this paper, we compared several acceptability judgment tasks by decomposing them into three basic features (mode of presentation, number of response options, and use of labels), and then investigated how those features, independently and in combination, affect the sensitivity of the experiments. The results reveal new information about the rate of statistical detection of different kinds of acceptability judgment tasks, covering the full range of effect sizes and sample sizes that linguists are likely to encounter in their syntactic experiments. One potential consequence of these results is that they provide concrete

baselines for choosing the features of acceptability judgment experiments, as well as for considering the question of the trade-off between the sensitivity, suitability, and practicality of different acceptability judgment tasks, e.g. by identifying the cases where a slight loss of sensitivity comes at the benefit of a more suitable task.

There are two particularly notable consequences of our results. The first is that joint presentation offers increased sensitivity in non-FC tasks, without introducing any obvious drawbacks. Our results suggest that non-FC tasks with joint presentation are slightly less sensitive than FC at detecting pairwise contrasts, but they are well-suited for a broader range of purposes relevant to linguistic inquiries: they can be used beyond the study of minimal pairs to test hypotheses involving multiple conditions, to get direct information about effect sizes, to do post-hoc exploration, and much more. The second consequence is that CS and LS_endpoint tasks (which have a large number of response options, and few labels) may offer a good compromise between sensitivity and other desiderata syntacticians may have in certain research situations. CS and LS_endpoint tasks are not only relatively sensitive, but also rely only on spatial reasoning, unlike LS tasks using numerical labels, which require both numeracy and a mapping between the ordinal numbers and acceptability (either as a distance measure or as ordinal category labels) — or tasks like ME, which requires numeracy, a mapping to acceptability, and the ability to calculate ratios.

Before closing, we would like to mention some of the strands of research that our feature-based results suggest. For example, one possibility may be to explore combining the features in novel ways, such as a combined FC plus slider task, which would allow participants to report the strength of their choices (without splitting the task into two steps, i.e. a choice and a rating scale). Another possibility is to explore the consequences of joint presentation for higher-order factorial designs. Joint presentation should, in principle, be available for any design, but to our knowledge higher-order designs have not yet been tested with joint presentation. Another possibility is to explore the relevance of this feature-based approach to tasks in other areas of linguistics, such as experimental semantics. It has been suggested that there is also an advantage in using graded over binary response scales for detecting semantic ambiguities in truth-value judgement tasks: offering participants more than a simple binary choice may help reveal the existence of an interpretation that would otherwise remain hidden due to the strong preference for another, more readily accessible, interpretation (a.o., Chemla & Spector, 2011; Marty, Chemla, & Spector, 2013, 2015). Finally, an anonymous reviewer points out that there is currently not much integration between the acceptability judgment literature and broader theories of human decision-making. We hope that our practical approach here, focusing on specific design features that arise in many judgment tasks, may help facilitate future collaborations between linguists and researchers in decision-making by identifying the impacts of these features. For now, we hope that the current study and its predecessors allow for more informed experimental choices, and that future studies will continue to increase the information available to researchers as they plan their experimental studies.

Additional File

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

- **Supplementary files.** The four tasks in Sprouse and Almeida (2017). DOI: https://doi.org/10.5334/gjgl.980.s1

Acknowledgements

Division at UCL, our reviewers for Linguistic Evidence 2018 and, last but not least, our reviewers for Glossa whose comments and suggestions helped improved this article greatly. Special thanks go to Florian Pellet for his invaluable practical help.

Competing Interests

The authors have no competing interests to declare.

References

Bader, M. & Häussler, J. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46(2). 273–330. DOI: https://doi.org/10.1017/S0022226709990260

Bard, E. G., Robertson, D. & Sorace, A. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1). 32–68. DOI: https://doi.org/10.2307/416793

Chemla, E. & Spector, B. 2011. Experimental evidence for embedded scalar implicatures. *Journal of semantics* 28(3). 359–400. DOI: https://doi.org/10.1093/jos/ffq023

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences* (2nd edition). Hills-dale, NJ: Erlbaume.

Cohen, J. 1992a. A power primer. *Psychological bulletin* 112(1). 155. DOI: https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. 1992b. Statistical power analysis. *Current directions in psychological science* 1(3). 98–101. DOI: https://doi.org/10.1111/1467-8721.ep10768783

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

Featherston, S. 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, S. 2008. Thermometer judgments as linguistic evidence. In C. M. Riehl & A. Rothe (Eds.), *Was ist linguistische evidenz?* (pp. 69–90). Aachen: Shaker Verlag.

Fukuda, S., Goodall, G., Michel, D. & Beecher, H. 2012. Is magnitude estimation worth the trouble? In C. Jaehoon, A. Hogue, J. Punske, D. Tat, J. Schertz, & A. Trueman (Eds.), *Proceedings of the 29th west coast conference on formal linguistics*, 328–336. Somerville, MA: Cascadilla Press.

Jeffreys, H. 1967. *Theory of probability* (3rd ed.). Oxford, U.K.: Oxford University Press.

Keller, F. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality* (Unpublished doctoral dissertation). Edinburgh: University of Edinburgh dissertation.

Križ, M. & Chemla, E. 2015. Two methods to find truth value gaps and their application to the projection problem of homogeneity. *Natural Language Semantics* 23(3). 205–248. DOI: https://doi.org/10.1007/s11050-015-9114-z

Langsford, S., Perfors, A., Hendrickson, A. T., Kennedy, L. A. & Navarro, D. J. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a journal of general linguistics* 3(1). DOI: https://doi.org/10.5334/gjgl.396

Marty, P., Chemla, E. & Spector, B. 2013. Interpreting numerals and scalar items under memory load. *Lingua* 133. 152–163. DOI: https://doi.org/10.1016/j.lingua.2013.03.006

Marty, P., Chemla, E. & Spector, B. 2015. Phantom readings: The case of modified numerals. *Language, Cognition and Neuroscience* 30(4). 462–477. DOI: https://doi.org/10.1080/23273798.2014.931592

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. 2009. Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review* 16(2). 225–237. DOI: https://doi.org/10.3758/PBR.16.2.225

Schütze, C. T. & Sprouse, J. 2014. Judgment data. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 27–50). Cambridge: University Press Cambridge. DOI: https://doi.org/10.1017/CBO9781139013734.004

Sprouse, J. 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, J. 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, J. & Almeida, D. 2011. Power in acceptability judgment experiments and the reliability of data in syntax. *Ms., University of California, Irvine & Michigan State University*.

Sprouse, J. & Almeida, D. 2012. Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics* 48(3). 609–652. DOI: https://doi.org/10.1017/S0022226712000011

Sprouse, J. & Almeida, D. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa* 2(1). 1. DOI: https://doi.org/10.5334/gjgl.236

Sprouse, J., Schütze, C. T. & Almeida, D. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua* 134. 219–248. DOI: https://doi.org/10.1016/j.lingua.2013.07.002

Szabolcsi, A. & Lohndal, T. 2017. Strong vs. weak islands. *The Wiley Blackwell Companion to Syntax,* Second Edition, 1–51. DOI: https://doi.org/10.1002/9781118358733.wbsyncom008

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