A preliminary comparison of the acceptability and coreference judgment tasks across the core binding phenomena in English

Pasha Koval
New York University Abu Dhabi
pasha.koval@nyu.edu

Jon Sprouse
New York University Abu Dhabi
jon.sprouse@nyu.edu

Abstract This paper offers a detailed comparison of two binding tasks, the acceptability judgment task and the coreference judgment task, in their ability to detect violations of Conditions A, B, and C in English. Based on the results of four experiments, we provide a comprehensive analysis of the validity, performance, and statistical power of the two tasks based on the results of four experiments. The results suggest that the coreference judgment task produces more consistent response distributions across the entire scale and across binding phenomena of different magnitudes. In turn, the acceptability judgment task produces smaller effect sizes across all binding phenomena, requires larger sample sizes, and may be more taxing for participants, which together result in a worse overall performance.

Keywords: experimental syntax; binding; acceptability judgment task; coreference judgment task; methodology; Condition A; Condition B; Condition C; reconstruction

1 Introduction

This study compares two binding tasks, the acceptability judgment task and the coreference judgment task, in their ability to detect violations of Conditions A, B, and C in English. The extensive use of both tasks in linguistic research warrants a comparative study that holds the promise of providing valuable new insights. The three main points of comparison are the validity, performance, and statistical power of the two tasks. Our findings endorse several recommendations that can be used to plan new experimental studies and evaluate the results of existing ones. In addition, they contribute to the discussion about the use of traditional linguistic methodology in large-scale formal experiments.

Previous research on experimental syntax methodology has focused mainly on tasks that compare grammatical and ungrammatical sentences (Sprouse & Almeida 2012a; Sprouse et al. 2013; Sprouse & Almeida 2017; Linzen & Oseki 2018; Marty et al. 2020). Binding, although ubiquitous in syntactic argumentation, typically relies on different readings of an otherwise grammatical sentence and therefore was necessarily excluded from these studies. For example, Binding Theory Conditions (Chomsky 1981) control the coreferential interpretation by either enforcing it (Condition A) or blocking it (Conditions B and C) in a given syntactic configuration. Therefore, to test these experimentally, we need a task that can access and then assess the status of the coreferential interpretation.

Theoretical syntactic literature (Chomsky 1981; Lasnik 1989) uses a two-step algorithm that can be implemented in the acceptability judgment task.[1] When following this

[1] Large-scale experimental binding studies using the acceptability judgment task are rare. One example is found in Temme & Verhoeven (2017).

algorithm, a participant evaluating a sentence in (1) must first identify the coreferential reading in (1a) by isolating it from the non-coreferential reading in (1b) and then assess the acceptability of (1a). Note that while Condition C renders the coreferential reading in (1a) unacceptable, the non-coreferential reading in (1b) remains fully acceptable and available to the participant. Thus, the first, *metalinguistic* step of the algorithm during which the two readings are isolated from each other is an indispensable part of the task.

(1) He paid for Timothy.

a. *He$_i$ paid for Timothy$_i$. coreferential reading

b. He$_j$ paid for Timothy$_i$. non-coreferential reading

In contrast, the experimental literature typically uses the coreference judgment task (Gordon & Hendrick 1997; Kazanina et al. 2007). In this task, participants see a sentence with two NPs marked in some way and then indicate on a scale whether these NPs can refer to the same person or must refer to different people. The modal statements "can refer to the same person" and "must refer to different people" describe the relationship between coreferential and non-coreferential readings. The former suggests that both readings are available, while the latter indicates that the coreferential reading is impossible, as in (1). By placing these statements at opposite ends of the same scale, the experimenter can infer the status of the coreferential interpretation without asking the participants to engage in metalinguistic reasoning.[2]

Both acceptability and coreference judgment tasks come with their own unique sets of challenges and limitations. As mentioned above, the acceptability judgment task requires participants to use metalinguistic reasoning to focus on one reading while blocking the others. The skill of metalinguistic reasoning is usually taught (explicitly or implicitly) in introductory linguistics classes, but it may be challenging for those participants who are not familiar with thinking about language in formal logical terms. An essential component of mastering this skill is the idea that the acceptability ratings of different readings are independent of each other. For instance, the rating of the non-coreferential reading in (1b) does not improve (nor does it impair) the rating of the coreferential reading in (1a) and vice versa.[3] A related issue is that in a typical 2×2 experiment testing, for example, Condition C, the metalinguistic step is repeated at least 35–40 times. During each repetition, a non-coreferential interpretation similar to (1b) is readily available, which creates the risk that a participant, if tired or distracted at any point during the experiment, may stop following the task focusing on the coreferential reading and instead report the acceptability of a sentence under any interpretation, leading to a false positive.

In turn, the coreference judgment task is built around a logical relationship between different readings. Therefore, mixing binding phenomena that use different sets of such relationships in the same experiment (or in the same item) is problematic. Consider a Condition A violation in (2). The coreferential reading in (2a) is available, but the non-coreferential reading in (2b) is ungrammatical making the modal statement "must refer to different people" undefined irrespective of the status of the coreferential reading, which may confuse some participants. Apart from the non-mixing of binding phenomena,

[2] The experimental literature also offers several variations of this task. Stockwell et al. (2021) give participants two distinct scales to evaluate the "naturalness" of each reading independently, while Keller & Asudeh (2001) offer a single scale for the non-coreferential reading only. It should be clear that both introduce the metalinguistic aspect into the task.

[3] Kaiser & Runner (2023) suggest two remedies for the metalinguistic problem: creating a setting (primarily, in an offline experiment) where a participant and an experimenter can discuss practice items and any related clarification questions and introducing "catch trials" designed to detect that a participant is not following the instructions.

there are a couple of other options. Gordon & Hendrick (1997) suggest adding an extra checkbox for participants to indicate that a sentence is ungrammatical, although this would make the task slightly more complex. Another possible workaround is to only use Condition A configurations that include a local antecedent that can license a reflexive in a non-coreferential reading as in (3). However, these sentences introduce a confound since "uncle" counts as a possible intervener for "Adam" under the coreferential reading.

(2) Adam sometimes surprised himself.

a. Adam$_i$ sometimes surprised himself$_i$. coreferential reading

b. *Adam$_j$ sometimes surprised himself$_i$. non-coreferential reading

(3) Adam's uncle sometimes surprised himself.

a. Adam$_i$'s uncle sometimes surprised himself$_i$. coreferential reading

b. Adam$_j$'s uncle sometimes surprised himself$_i$. non-coreferential reading

The acceptability and coreference judgment tasks are essential for investigating binding phenomena. Each task has its own challenges and, as a result, may interact differently with various binding phenomena. An informed decision on which task to use requires a deeper understanding of the potential implications and ramifications of either option. However, we currently lack a systematic comparison of the two tasks across binding phenomena. The main goal of this paper is to provide guidance in selecting the task that is best suited to the research question, thus promoting robust and reliable results.

The rest of the paper is organized as follows: Section 2 provides the necessary statistical background. Section 3 outlines the experiments. Section 4 presents the results. Section 5 summarizes and contextualizes the findings. Section 6 concludes.

2 Background

In the following sections, we analyze four distinct experiments, each a version of Experiment 5 in Koval & Sprouse (2023). Every experiment uses one of the two tasks (the acceptability judgment task and the coreference judgment task) and one of the two sets of BT Conditions (ABC and BC). Our comparison of the tasks is centered around filler items and pairwise sanity checks testing violations of Condition A, Condition B, Condition C, and Condition C under reconstruction.

The comparison of the two tasks starts with an informal visual comparison of the response distributions for different filler items, which cover a broad range of acceptability levels and binding configurations. This should give us an intuitive understanding of how the two tasks manage the binding phenomena at different points on their respective scales. Next, we are going to estimate the effect sizes of the four binding phenomena at the center of this study: Condition A, Condition B, Condition C, and Condition C + Reconstruction. We examine how the effect size of each binding phenomenon varies with the chosen task in two ways: through z-unit mean differences, where the data is normalized to a standard normal distribution ($\mu = 0, \sigma = 1$), and via Cohen's d, which measures the effect size independent of the scale, standardized by the population standard deviation. During the next step, we conduct a Receiver Operating Characteristic (ROC) performance analysis on the same 4 phenomena. The ROC curve, a graphical representation of the performance of a binary classification system, is used to evaluate task performance by examining the relative trade-offs between the true positive rate and the false positive rate across all possible classification thresholds. Lastly, we carry out a series of resampling power simulations using the same 4 binding phenomena. The

simulation-based power analysis yields estimates of the sample sizes required for each of the phenomena and for each task, given a chosen risk level of Type II errors (β), errors in which a false null hypothesis is not rejected. By further comparing the statistical detection rates of both tasks across binding phenomena of different magnitudes, we obtain a measure of each task's statistical power, thus allowing us to assess the sensitivity of the two tasks. Together, these methods offer a comprehensive and actionable comparison of the acceptability and coreference judgment tasks across several sentence types, which will prove useful to any experimental syntactician interested in binding experiments. The following four subsections provide an in-depth look at each step of the comparison.

2.1 Visual comparison

We begin with a visual comparison of the filler items. This helps us to understand how different binding phenomena at different levels of acceptability are represented on the respective scales of the two tasks. We will pay particular attention to the shape and median of response distributions. A unimodal distribution, which has a single prominent peak (mode), suggests a generally consistent response pattern. A bimodal or multimodal distribution, featuring multiple peaks, may suggest that different groups of participants interpret or respond to the task differently. Finally, a flat distribution, with a particularly wide spread of data points, may suggest uncertainty or confusion among participants when following the task. We will also calculate the median of responses, which provides a robust measure of the central tendency that is less affected by outliers.[4] Regardless of the shape or skewness of the distribution, the median indicates the point below and above which half of the observations fall, allowing us to compare distributions in a manner relatively unaffected by extreme responses.

We will use density functions and histograms to visualize the response distributions. Density functions show the concentration of responses across the standardized scale, and histograms show the frequency of particular responses. Alongside the visual tools, we use Hartigans' dip test (Hartigan & Hartigan 1985), a non-parametric test of unimodality that calculates the maximum discrepancy between the observed distribution and the best-fitted unimodal distribution. This test provides a measure of how much the data deviates from a unimodal distribution, which in our case translates into the likelihood of different response modes among participants. The dip statistic D is defined as follows:

$$D = \max_x \left\{ \max \left[\hat{F}_n(x) - GCM(x), LCM(x) - \hat{F}_n(x) \right] \right\}, \tag{4}$$

where $\hat{F}_n(x)$ is the empirical cumulative distribution function and $GCM(x)$ and $LCM(x)$ are the greatest convex minorant and the least concave majorant of $\hat{F}_n(x)$, respectively.[5] Once calculated, the D statistic is then compared to the critical value obtained through r samples from a simulated unimodal distribution to calculate the p-value of the empirical distribution being unimodal.

Visual comparison is a simple yet powerful tool that provides an initial intuitive grasp of how the tasks perform under different binding conditions and across their respective re-

[4] As a reminder, the median is the exact middle point in an ordered data set that is calculated using the following formula:

$$\text{med}(X) = \frac{1}{2}(x_{\lceil \frac{n}{2} \rceil} + x_{\lfloor \frac{n}{2} \rfloor + 1}), \tag{i}$$

where x_i refers to the i-th value in the vector X with n data points in ascending order, $\lceil n/2 \rceil$ is the smallest integer that is greater than or equal to $n/2$, and $\lfloor n/2 \rfloor$ is the largest integer that is less than or equal to $n/2$.

[5] See Wasserman (2006) and Shorack & Wellner (2013) for the definitions and various applications.

sponse scales. It is an important exploratory tool that highlights unusual patterns, trends, and potential outliers, which may suggest areas for further investigation. However, it provides an overview rather than precise quantification and should be complemented by more detailed analyses.

2.2 Comparison of effect sizes

The discussion in this and the following two sections centers on the comparison of the two binding tasks against four sets of minimal pairs, each representing a violation of Condition A, Condition B, Condition C, or Condition C under reconstruction. Here, we focus on comparing the effect sizes emerging from these binding phenomena in response to the two tasks.

The relationship between the task and the effect size across various phenomena is very nuanced and extends beyond the scope of this project (see Pashler & Wagenmakers 2012; Button et al. 2013; Maxwell et al. 2017). Nevertheless, the results of such a comparison can still be useful when designing experiments. For instance, if we know that using one task for testing some phenomenon results in a larger (observed) effect size than using the other task, the former can be used for studying subadditive effects, while the latter is more appropriate for studies expecting additive or superadditive effects, since it leaves adequate room on the scale for those manipulations. In this way, this comparison can be instrumental in choosing a suitable task.

We use two measures of effect size: mean difference (in z-scores) and Cohen's d. Both statistics quantify the magnitude of differences between groups, but do so in slightly different ways and serve complementary purposes. Z-scores are a measure of how many standard deviations a data point deviates from the mean of the distribution. In our context, the formula for z-scores is as follows:

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}, \tag{5}$$

where x_{ij} is the i-th data point from participant j, and $\overline{x}_j$ and s_j are the mean and standard deviation of j's responses, respectively. Z-scoring standardizes the values of a distribution to have a mean of 0 and a standard deviation of 1. Applied within participants, it removes individual biases while using the scale. We then calculate the difference of mean z-scores for different conditions between participants. Juxtaposing those values from different tasks allows us to compare effect sizes measured on different scales since it eliminates any impacts from the scales themselves.

We also calculate Cohen's d, a measure of the effect size that quantifies the standardized difference between two means, regardless of the scale of measurement. It can be computed using the following formula:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}}, \tag{6}$$

where $\overline{x}_1$ and $\overline{x}_2$ are the means, s_1 and s_2 are the standard deviations, and n_1 and n_2 are the sample sizes of the two conditions. Cohen's d measures the differences between the means of two groups, normalized by the pooled standard deviation, which averages the standard deviations of both groups. This allows comparisons that are also unaffected by the scale of the data.

Using both z-score mean differences and Cohen's d together gives us a more robust and informative comparison of effect sizes across tasks and binding phenomena than relying on just one. Typically, the two measures align very closely since they both quantify scale-free differences between the means of two groups of responses. However, sometimes z-scores and Cohen's d may disagree on the effect size due to differences in how they handle variance in the data.

Z-scoring relies on the standard deviation to standardize individual data points, and so the presence of individual outliers can directly affect the resulting z-scores. For example, if a data set has an overall small standard deviation and also a couple of extreme outlier responses, this can lead to larger z-scores for any given deviation from the mean, which, when averaged, would show up as a smaller mean difference of z-scores and thus underestimate the 'true' effect size. In the same situation, the value for Cohen's d would be larger since it is only sensitive to the relative variability within each group (by using pooled standard deviation). However, an important aspect to keep in mind when interpreting Cohen's d is that when variability within groups is low, even a modest difference in group means can result in a large value of Cohen's d. This does not necessarily mean that the 'true' effect size is large.

In the inverse situation, when the within-group variance is large, z-scores that do not explicitly account for the group-level variance can overestimate the effect size, while Cohen's d, which incorporates this type of variance, would be noticeably smaller. Therefore, if we find that z-scores and Cohen's d are notably different for a certain binding phenomenon and a certain task, the directionality of this discrepancy may tell us something about the group-level variance in the data and, therefore, indirectly, about the task that was used to produce it.

In summary, by using both z-scores and Cohen's d in a comparison of effect sizes across the two tasks, we can better understand the impact of the task on the data, which can help navigate the selection of the task when planning new experiments.

2.3 ROC curve performance analysis

When considering a task for a new experiment, a researcher may want to know how well participants can differentiate between conditions when following the instructions of that task. To answer this, we need to shift our focus from a regression problem to its logical inverse, a classification problem. A regression problem is looking to predict a continuous output (e.g. z-scores) from input examples based on their features. The output space of a regression problem is often infinite. A typical experiment constitutes a regression problem since it tests whether some feature (i.e. an experimental factor) is a good enough regressor for predicting the output. An example of this is shown in (7a). In a classification problem, the goal is to predict a discrete label or category for an input example using a continuous output. The output space for a classification problem is typically finite (and often quite small). This is shown in (7b).

(7) a. `zscore ~ structure + (1|participant) + (1|item)` regression
b. `structure ~ zscore + (1|participant) + (1|item)` classification

It should be clear that solving both problems for the same phenomenon at the same time is impossible.[6] If instead we pick a few well-established phenomena (e.g. Conditions A, B, and C and Condition C under reconstruction), we can directly compare participants' performance as they classified experimental items according to the instruc-

[6] In some cases, it may be necessary to optimize both solutions concurrently; see Ruder (2017) for an overview.

tions of one or another task. In essence, when a group of participants follows a specific task, they generate a classification of the stimuli. This classification is then compared to the ideal classification encoded in the experimental design, and the match/mismatch between them provides a measure of the task performance.

The Receiver Operating Characteristic (ROC) analysis methodology is widely used in medicine and machine learning to evaluate classifier performance (Hanley & McNeil 1982; Swets 1988; Bradley 1997; Fawcett 2006). ROC analysis of a classification starts by creating a confusion matrix (also known as the *contingency table*) as shown in Table 1. In our case, the "predicted" classes correspond to the results of the classification under consideration, while "actual" classes contain the ideal classification, i.e. sentences that do not include a binding violation are considered "actual positives" and ones that do include one become "actual negatives".

	Actual positive	Actual negative
Predicted positive	True positive	False positive
Predicted negative	False negative	True negative

Table 1: A confusion matrix for a binary classification problem

In the next step, the true positive and false positive rates are calculated as follows:

(8) a. $$\text{True positive rate} = \frac{\text{True positive}}{\text{Actual positive}} \quad \text{(TPR)}$$

b. $$\text{False positive rate} = \frac{\text{False positive}}{\text{Actual negative}} \quad \text{(FPR)}$$

Following the computation of TPR and FPR, an important step in the ROC analysis is to define and vary a threshold for the classifier. This threshold represents the cut-off point at which a classifier distinguishes between two classes. TPR and FPR are then calculated for each threshold value. The ROC curve, which plots TPR against FPR for various threshold values, provides a visual representation of the performance of the classifier as its discrimination threshold is varied.

Finally, the area under the ROC curve (AUC-ROC), which measures the classifier performance across all possible thresholds, can be calculated using the following formula:

(9)
$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x))\,\mathrm{d}x$$

Figure 1 shows an ROC curve for a perfect classifier with the AUC-ROC value of 1. This classifier is able to identify all true positives and avoid all false positives at the same time.

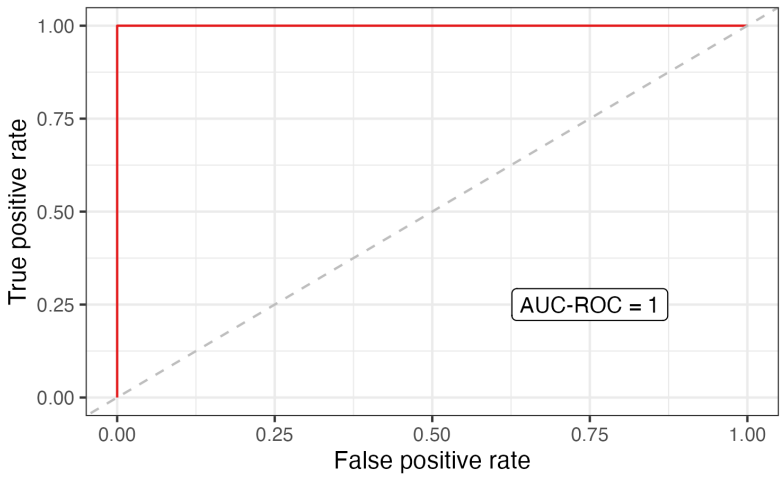

Figure 1: A sample ROC curve of a perfect classifier

ROC analysis provides a clear and systematic approach to evaluating task performance. By comparing the ROC curves and the AUC-ROC values of different tasks, we can identify the one that offers better differentiation between conditions, leading to more accurate experimental results.

2.4 Power analysis

When planning an experiment, it is important to think carefully about the number of participants to recruit. The size of the sample can affect the overall cost of the study. More importantly, it is directly related to the likelihood of missing a true effect if one exists. Thus, oversampling can lead to unnecessary spending, while undersampling runs the risk of missing a valuable theoretical result. Ideally, when getting a null result, we want to feel confident that it is a genuine null effect rather than a consequence of an inadequate sample size. The choice of task for an experiment can influence the required sample size since different tasks can vary in their sensitivity, leading to different effect sizes for the same phenomenon.

The statistical framework of Neyman-Pearson hypothesis testing (NPHT) (Neyman & Pearson 1928a; b; 1933) provides a formal structure to address the undersampling problem (see Sprouse & Almeida (2012b; 2017) for a detailed discussion of this problem in different statistical frameworks). NPHT sees hypothesis testing as the process of making a decision between the null hypothesis (H0) and the alternative hypothesis (H1). Depending on the true state of the world, NPHT distinguishes two types of errors: Type I and Type II. Type I error, or a false positive, is the incorrect rejection of a true null hypothesis.[7] Type II error, or a false negative, occurs when we fail to reject a false null hypothesis. This is summarized in Table 2.

[7] Note that the ROC curve performance analysis discussed in the previous subsection does not provide a separate estimate of Type I error, but instead it analyzes a measure related to it. An ROC curve plots TPR (true positive rate) against FPR (false positive rate) over a range of decision thresholds. FPR can be interpreted as the probability of a false positive, or a Type I error. An ROC curve illustrates the balance between TPR and FPR, which includes considering Type I errors in a broader evaluation of participants' performance when following the task.

The state of the world	Test result	Outcome
H0 is true	Keep H0	True negative
H0 is true	Reject H0	Type I error (false positive)
H1 is true	Keep H0	Type II error (false negative)
H1 is true	Reject H0	True positive

Table 2: Type I and Type II errors in Neyman-Pearson hypothesis testing

NPHT sheds light on the relationship between undersampling and Type II error. Type II error arises when H1 is true of the world, but we (incorrectly) choose to keep H0. This can occur when the sample size of the experiment is insufficient to detect the effect. Therefore, increasing the sample size will reduce the chance of a Type II error. This naturally leads to the notion of statistical power, the probability of a test rejecting H0 when H1 is true of the world (Cohen 1988). Statistical power and Type II error are inversely related: a test with high statistical power has a lower probability of committing a Type II error.

Statistical power depends on the sample size, but also on the effect size and the significance level. The effect size measures the strength of a phenomenon, which can vary depending on the task (see Section 2.2 above). The significance level (α) is the probability of rejecting H0 when it is true, i.e. a Type I error. As a rule of thumb, a larger effect size, a larger sample size, and a higher significance level increase the power. However, one must be cautious applying this rule to an experimental design, as increasing the significance level can also raise the risk of Type I error. The relationship between all three is summarized in Figure 2.

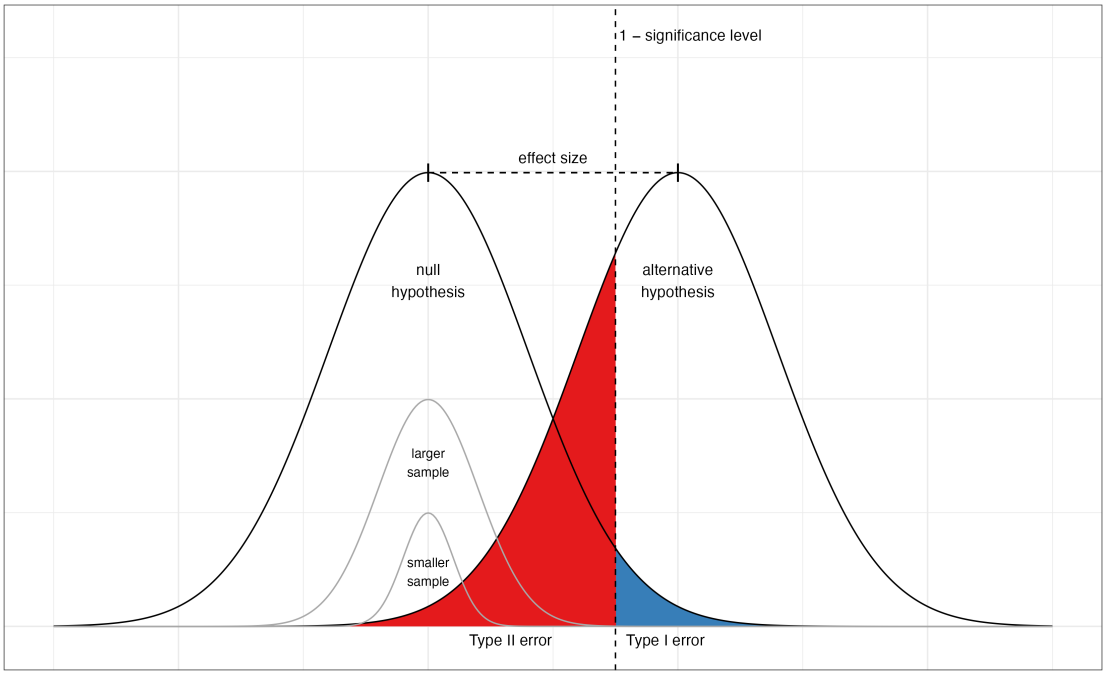

Figure 2: An illustration of a relationship between possible outcomes of the test statistic and Type I and Type II errors, the effect size, the sample size, and the significance level

Power analysis provides a robust way to estimate the statistical power of an experiment given the effect size, sample size, and significance level. One approach to estimating statistical power is through power simulations. This involves running the same statistical test (e.g. fitting a linear mixed-effects model) multiple times for different random subsamples drawn from the same experimental sample. A typical resampling power simulation includes the following steps:

i. Specify a model with the significance level and the effect size of interest.
ii. From the full experimental sample, draw a random subsample of size n using resampling with replacement.
iii. Fit the model to the new subsample and test the null hypothesis at the specified significance level.
iv. Repeat steps 2 and 3 r times, each time with a different subsample of the same size.
v. Estimate the power for the sample size n as the proportion of times the null hypothesis is rejected for the samples of that size.
vi. Repeat steps 2–5 for all sample sizes of interest.

For the purpose of task comparison, power simulations help us navigate the likelihood of correctly rejecting H0 (i.e. statistical power) across a range of sample sizes and identify the minimum sample size necessary to achieve a desired level of power, given a certain effect size and significance level. Both types of information are valuable when reviewing existing experiments and designing new ones. Thus, comparing power simulations from several binding phenomena gives us a more nuanced understanding of the two tasks.

In the power simulations discussed below, we focus on four binding phenomena, three of which have very large effect sizes, namely, Conditions A, B, and C). These phenomena, although fundamental for binding, do not represent the full spectrum of the potential effects of interest. In fact, it is likely that most binding phenomena have smaller effect sizes than those three. Thus, using only the power estimates for large effect sizes when choosing the sample size for a new experiment may lead to undersampling (again) when studying phenomena with medium or small effect sizes. To mitigate this, we adopt a practical significance approach that uses the concept of Smallest Effect Size of Interest (SESOI); see Kumle et al. (2021) for a discussion and related references. SESOI is defined as the smallest effect that would be considered theoretically interesting in the context of a research question. For the purposes of this preliminary report, we set SESOI at 80% for all estimates (β for the linear mixed effects models) during power simulations for Conditions A, B, and C. By setting the SESOI limit this low, we can be sure that our power simulations, if used for planning new experiments, are going to give a more conservative effect size estimate that will be adequate for the majority of medium and large effects.

3 Methods

The four experiments reported here test two different binding tasks: the acceptability judgment task and the coreference judgment task. Each task is evaluated against a range of binding configurations. From the four experiments, two (one for each task) test structures that either satisfy or violate Conditions A, B, and C. The other two experiments only use Conditions B and C.

Exp. #	Task type	BT Conditions
1	acceptability	ABC
2	acceptability	BC
3	coreference	ABC
4	coreference	BC

Table 3: Classification of experiments based on the task and BT Conditions tested

3.1 The tasks

The exact task instructions used in the corresponding pairs of experiments are as follows:

(10) **The acceptability judgment task**
Your task is to imagine that the speaker intended the two underlined words to refer to the same person, and then judge whether this is a grammatical sentence of English. You will rate the sentence on a scale from 1 (Ungrammatical) to 7 (Grammatical).

(11) **The coreference judgment task**
Your task is to determine whether the two underlined words could refer to the same person or whether they must refer to different people. You will rate this from -3 (they must refer to different people) to 3 (they could refer to the same person).

3.2 Materials

The four experiments in this study contain the same experimental items described in Experiment 5 of Koval & Sprouse (2023). For the purpose of the task comparison, the experimental items, as well as the practice items and anchor items, are excluded from further analysis. Our primary focus here is on the filler items and the sanity check items, which, formally, constitute 2×1 experiments. Complete lists of all items are found in the supplemental material.

3.2.1 Fillers

The sets of examples shown in (12) through (14) contain the filler items used in all four experiments, organized by BT Condition. The labels of the fillers reflect their target BT conditions and their expected ratings on a 1–7 scale, e.g. "A2" is a Condition A sentence with an expected rating of 2. To the right, it shows the source of the corresponding sentence type.

We used 8 fillers in the ABC and 10 fillers in the BC experiments. To replace 3 Condition A fillers from the ABC experiments, 1 filler for Condition B (B5) and 2 fillers for Condition C (C2 and C5) were used in the BC experiments. 2 additional fillers (B7 and C3) were included in the BC experiments in place of 2 Condition A sanity check items.

(12) Condition A

A2. Chloe invited Claire to challenge herself. (Reinhart & Reuland 1993)
A5. Margaret expects stories about herself to be flattering. (Chomsky 1981)
A7. Natalie found herself in an awkward situation. (Chomsky 1981)

(13) Condition B

B1. Brian continued letting him down. (Heim 1983)
B3. Who did he say likes kayaking? (May 1985)
B5. In Mason's kitchen he keeps fresh herbs. (Gordon & Hendrick 1997)
B6. Monica introduced Sean to his new trainer. (Gordon & Hendrick 1997)
B7. Margaret still decided to invite her mom and her new partner for Christmas. (Safir 1999)

(14) Condition C

C1. She took a nice picture of Courtney. (Fiengo & May 1994)
C2. James asked her about Claire's parents. (Gordon & Hendrick 1997)
C3. Her father was impressed by Erin. (Gordon & Hendrick 1997)
C5. Which friend that Chloe invited to her birthday party did she like best? (Van Riemsdijk & Williams 1981; Sportiche 1997)
C7. While Luke was working in the backyard, he spotted two hedgehogs. (Reinhart 1976)

3.2.2 Sanity check items

In Koval & Sprouse (2023), three of the four binding phenomena discussed here were introduced as sanity checks for Condition C experiments. For both clarity and continuity, we keep the term 'sanity check items' for them (and extend it to include the fourth). However, it is important to note that for the purpose of this task comparison, these phenomena transition from serving a supplementary role to being the primary focus of our interest. The ABC experiments include all four sanity checks, while the BC experiments use only three sanity checks excluding Condition A.

All four sanity checks test minimal pairs. Therefore, they all use a 2×1 experimental design: the control satisfies the corresponding BT Condition, while the experimental condition violates it. In this design, all structural changes contributing to the violation feed into the same fixed effect. This includes increasing the length of the binding dependency, changing the structural position of the head of the binding dependency, swapping two NPs between the positions of the head and tail of a binding dependency, and keeping a potential intervener in the structure. Because of that, in all four sanity checks, we use the same two levels of the factor dubbed STRUCTURE: 'no violation' and 'violation'.

An example of a minimal pair from the Condition A sanity check is shown in (15). These sentences contrast a binding dependency that consists of a reflexive in the object position and a local subject NP (the no-violation condition) and a binding dependency including a reflexive and a possessor of a local subject (the violation condition). In the latter case, the lack of c-command between the coreferential possessor and reflexive constitutes a Condition A violation, which may also be combined with the effect of having an intervener in the structure. If participants ignore the coreferential interpretation, the sentence in the violation condition should be fully acceptable, since there is a potential local antecedent. Across all items, both the subject and the possessor are matched in gender to make sure that both NPs can serve as a potential antecedent for the reflexive. This guarantees that, in the acceptability judgment task, participants who neglect the metalinguistic step are going to report the violation condition as fully acceptable, while, in the coreference judgment task, it ensures that the sentence with the non-coreferential reading is grammatical and thus does not invalidate the modal statement "can be the same person".

(15) Sanity check: Condition A

a. Bella's sister eventually forgave herself. no violation
b. Bella's sister eventually forgave herself. violation

Shown in (16) is a sample minimal pair from the Condition B sanity check. This pair contrasts two binding dependencies. In the no-violation condition, the tail of the dependency is a pronominal in the object position and the head is a referential NP in the position of the possessor of a local subject. In the violation condition, the referential NP heading the binding dependency is in the subject position and, as a result, the pronominal is c-commanded by a coreferential NP within its local domain, which causes a Condition B violation. Similar to the Condition A sanity check, the possessor and the subject NPs are matched in gender. This allows the sanity check to focus on the structural component of the Condition B violation.

(16) Sanity check: Condition B

a. Hannah's aunt sometimes surprised her. no violation
b. Hannah's aunt sometimes surprised her. violation

(17) shows a sample minimal pair from the Condition C sanity check. Within this pair, we compare two binding dependencies that are mirror images of each other. In the no-violation condition, an R-expression is the head of the dependency and a pronominal is the tail, while both occur in the subject position of the matrix and embedded clause, respectively. In the violation condition, the two NPs are swapped. Since an R-expression is c-commanded by a coreferential pronoun, we expect to find a Condition C violation. In this design, both the dependency reversal and the Condition C violation costs feed into the fixed effect of STRUCTURE.

(17) Sanity check: Condition C

a. Allison added that she liked reggae. no violation
b. She added that Allison liked reggae. violation

Finally, (18) contains the minimal pair of conditions from the Condition C + Reconstruction sanity check. Similar to the Condition C sanity check, two binding dependencies in (18) are mirror images of each other. In the no-violation condition, the head of the binding dependency is an R-expression, while a pronoun inside the fronted PP is the tail. In the violation condition, the two are swapped. After Reinhart (1976); Bruening & Al Khalaf (2019), we expect a fronted PP to obligatorily reconstruct to its base position, causing a Condition C violation for the R-expression-first dependency.

(18) Sanity check: Condition C + Reconstruction

a. Rachel said that [$_{PP}$ ahead of him], the paperboy heard a dog ~~PP~~. no violation
b. Rachel said that [$_{PP}$ ahead of the paperboy], he heard a dog ~~PP~~. violation

In this design, both conditions in (18) share the costs associated with fronting a PP to the edge of the IP and then reconstructing it back to the base position. These costs are expected to push both conditions closer to the lower end of the scale, reducing the room left for the Condition C violation. Furthermore, the linear order effect that we observed in Experiment 5 in Koval & Sprouse (2023) also reduces the space on the scale for identifying

Condition C, since it pushes the pronoun-first, no-violation condition down and closer to the violation condition. As a result, in this configuration, we expect the effect size of a Condition C violation during reconstruction to be relatively small.

3.3 Anchor items and practice items

Both ABC and BC experiments use 2 anchor items and 9 practice items. Sets of items for different types of experiments are shown in (19) and (20) along with the label indicating the associated BT Condition and an expected rating on the 1–7 scale, similar to the fillers.

(19) ABC experiments

a. Anchor items
   C1. She said that Julie enjoys reading.
   A7. Paige promised herself to walk to work.

b. Practice items
   C1. He misunderstood Richard.
   B2. Kristen bought her a new set of chairs.
   C3. I saw him in Jacob's office.
   A4. She likes her family, but herself, Claire simply adores.
   C5. Her brother visited Lisa at college.
   B6. If he does well on the exam, Josh will pass.
   A7. Francesca introduced herself.
   A1. Abigail's cousin respects herself.
   B7. John's roommates met him at the restaurant.

(20) BC experiments

a. Anchor items
   C1. She said that Julie enjoys reading.
   B7. Steven knows that Paige loves him.

b. Practice items
   B1. Richard cheered him up.
   B2. Kristen bought her a new set of chairs.
   C3. I saw him in Jacob's office.
   B4. Kaya promised Noah and Natalie that she would be invited.
   C5. Her brother visited Lisa at college.
   B6. If he does well on the exam, Josh will pass.
   C7. Francesca showed Bill to his new desk.
   C1. She misunderstood Abigail.
   B7. John's roommates met him at the restaurant.

3.4 Survey composition

The surveys in both the ABC and BC experiments comprise a total of 33 items. In ABC experiments, they are organized as follows: 9 practice items in a fixed order are followed by a pseudorandomized sequence of 8 experimental items, 8 fillers, and 8 sanity check items, with 2 items per sanity check. For BC experiments, the item distribution slightly differs: after 9 practice items in a fixed order, the pseudorandomized sequence includes 8 experimental items, 10 fillers, and 6 sanity check items (2 items per sanity check). A Latin square procedure is used to distribute both experimental items and sanity check items

among the experimental lists. To further control for order effects, 4 counterbalanced orders are imposed on the 8 lists of each experiment.

3.5 Participants and presentation

We recruited a total of 280 participants for the 4 experiments, with each experiment assigned a subset of 70 participants. All participants were compensated for their time at an hourly rate of $15 per hour with an estimated completion time of 6 minutes. Each participant saw only one list of one experiment and all the experimental conditions in that experiment. Each sentence was presented on a separate screen and had a separate scale next to it. Participants were also asked to complete a two-question language proficiency questionnaire. Based on the results of the questionnaire, a total of 3 participants per experiment were excluded either because US English was not their first language or because they grew up in a non-monolingual household. Importantly, their responses did not affect their compensation, thus eliminating the incentive for lying. The remaining 67 participants per experiment were self-reported native speakers of US English.

All experiments were conducted online using the Qualtrics survey platform. The participant recruitment was carried out via Amazon Mechanical Turk with the help of a recruitment facilitation service CloudResearch.

3.6 Analysis

All statistical analyses in this study were performed using R version 4.2.3 (R Core Team 2023). Several specialized R packages were used for different parts of the analysis. All plots were created using the `ggplot2` package (Wickham 2016). To identify non-unimodal distributions we applied Hartigans' dip test from the `diptest` package (Maechler 2021) with 100,000 iterations. While determining the effect sizes, we also incorporated information about the p-values. For each task and each sanity check, we constructed a linear mixed-effects model with STRUCTURE as a fixed effect and PARTICIPANT and ITEM as random effects (slope and intercepts) using the `lme4` package (Bates et al. 2015). Associated p-values were derived with the `lmerTest` package (Kuznetsova et al. 2017), leveraging the Satterthwaite approximation for degrees of freedom. ROC curve analysis was carried out using the `pROC` package (Robin et al. 2011). To ensure an accurate comparison across model classes, the classifier scores were standardized to z-scores. Lastly, power simulations for linear mixed-effects models were conducted using the `lme4` (Bates et al. 2015) and `mixedpower` (Kumle et al. 2021) packages. This includes 3000 simulations per sample size, spanning a range between 5 and 100 participants. For each simulation, a linear mixed-effects model was fitted to a newly resampled dataset with the critical value $z = 2$ for the test statistic. The smallest effect size of interest (SESOI) was set at 80% of β for each component.

4 Results

4.1 Visual comparison of fillers across the two tasks

Figure 3 shows the histograms, density functions, and medians of the z-scored responses for all fillers in the four experiments. To complement this, Table 4 provides the p-values and D statistics (shown in parentheses) from applying Hartigans' dip test to each filler in

each experiment. The fillers for Conditions A, B, and C are listed in (12), (13), and (14), respectively.

		Acceptability judgment task		Coreference judgment task	
BT Condition	Label	ABC	BC	ABC	BC
Condition A	A2	.013(0.07)	NA	.686(0.038)	NA
	A5	.976(0.029)	NA	.885(0.033)	NA
	A7	.972(0.029)	NA	.681(0.039)	NA
Condition B	B1	.005(0.074)	<.001(0.09)	.382(0.046)	.661(0.039)
	B3	.212(0.051)	.094(0.057)	.82(0.035)	.99(0.027)
	B5	NA	.867(0.034)	NA	.996(0.023)
	B6	.849(0.034)	.993(0.025)	.161(0.053)	.93(0.032)
	B7	NA	.412(0.045)	NA	.006(0.074)
Conditon C	C1	.002(0.079)	.001(0.085)	.76(0.037)	.84(0.035)
	C2	NA	.062(0.06)	NA	.383(0.045)
	C3	NA	.004(0.076)	NA	.81(0.036)
	C5	NA	.903(0.033)	NA	.904(0.033)
	C7	.984(0.028)	.792(0.036)	.885(0.033)	.988(0.028)

Table 4: The results of Hartigans' dip test for all fillers across all experiments

In Condition A, both acceptability and coreference judgment tasks yield similar median values across all items, except for A2, where acceptability produces a slightly higher value. Visual inspection suggests that A2 in the acceptability judgment task has a bimodal distribution, which is corroborated by the results of the dip test (p = .013). In contrast, all Condition A items in the coreference judgment task, along with A5 and A7 in the acceptability judgment task, are unimodal, with p-values ranging from .681 to .976.

In Condition B, the fillers B3, B5, and B6 have similar median values in both tasks, while B1 gets a higher rating in the acceptability judgment task and B7 scores lower in the coreference judgment task, compared to their expected ratings. Visual inspection reveals that B5 and B6 have unimodal distributions in both tasks (p = .867 and .996 for B5 and .849/.993 and .161/.93 for B6 for both tasks). In contrast, B1 and B3 in the acceptability judgment task deviate from the unimodality (p = .005/<.001 for B1 and .212/.094 for B3), unlike in the coreference judgment task (p = .382/.661 for B1 and .82/.92 for B3). In the coreference judgment task, B7 is clearly not unimodal (p = .006), but it is difficult to determine whether it is bimodal or flat (or both). The source of this effect is also unclear and further investigation may be necessary.

In Condition C, we observe the most significant difference between the two tasks. In the coreference judgment task, the median values align closely with the expected ratings across the scale. However, in the acceptability judgment task, the medians of all items except C7 are grouped around 0. The distributions of C1–C3 appear to be bimodal, which is supported by the results of the dip test, suggesting that these distributions are not unimodal (p = .002/.001 for C1, .062 for C2, .004 for C3). The C5 distribution appears to be flat, and the dip test confirms that this distribution is indeed closer to being unimodal (p = .903). This flat distribution may be due to the participants' disagreement about the acceptability rating of weak crossover during *wh*-movement. In contrast, the same items in the coreference judgment task consistently show unimodal distributions, with the lowest p-value being .383 for C2.

Comparing ABC and BC experiments, the acceptability judgment task shows similar results with the exception of C1, which has a non-unimodal distribution (p = .001 and .002) but a median closer to 0 in the BC experiment. Coreference tasks remain consistent throughout the two experiments, except for B6, which appears much less unimodal in the ABC than in the BC experiment (p = .161 and .93 respectively).

In summary, the acceptability judgment task consistently shows greater variability in filler distributions compared to the coreference judgment task, which produces unimodal distributions across the entire scale and across different binding Conditions. Notably, in the acceptability judgment task, non-unimodal distributions are frequently found in the items with expected ratings in the lower portion of the scale, suggesting that some participants may be ignoring the metalinguistic part of the task and, instead of assessing the acceptability of a sentence under a coreferential interpretation, are simply reporting the general acceptability of a sentence.

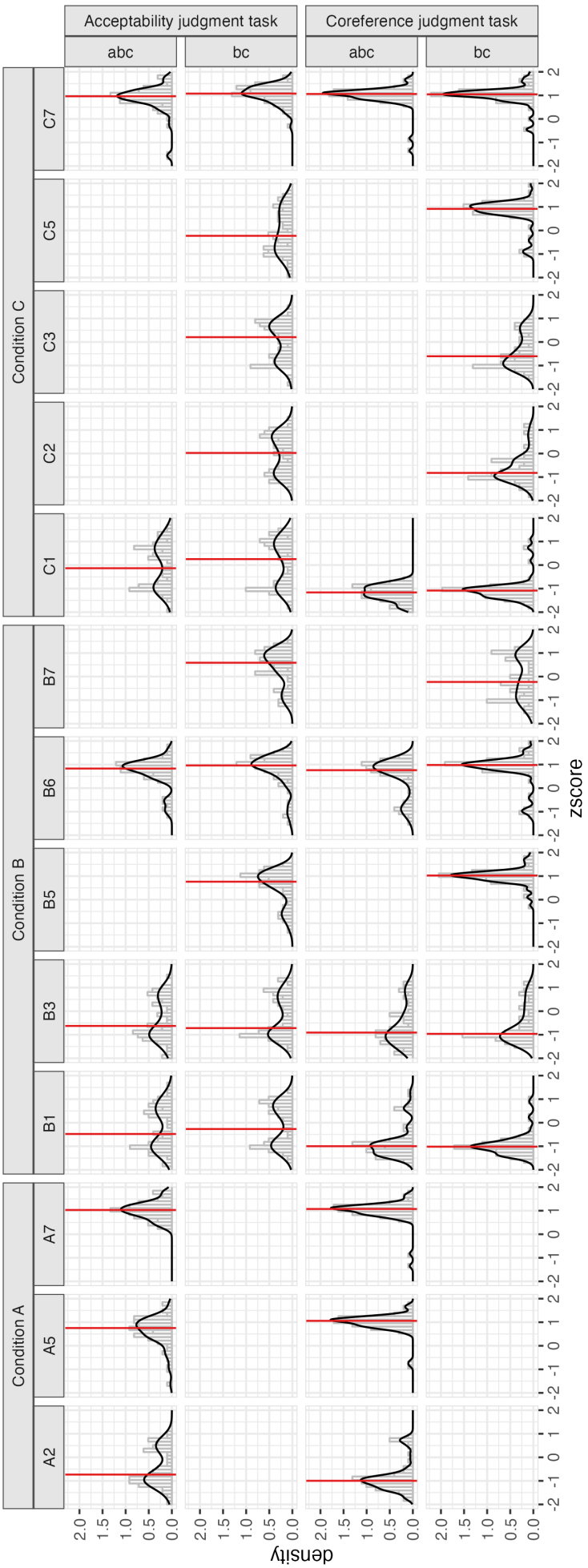

Figure 3: Fillers across different experiments

4.2 The effect sizes of different binding phenomena across the two tasks

Figure 4 shows the mean differences in z-scores between conditions for each of the four sanity checks across the two tasks. The plots also include the numerical value of the mean difference, as well as Cohen's d, the number of participants, and the p-value for the fixed effect of STRUCTURE derived from the linear mixed-effects model.

We consolidated the data from the pairs of ABC and BC experiments that use the same task for three sanity checks: Condition B, Condition C, and Condition C + Reconstruction. Condition A is tested only in ABC experiments. The consolidation is justified by the similarity between the ABC and BC experiments, which have surveys of the same length with the same lists and orders and use mostly the same items. The only difference is that ABC experiments include Condition A items (3 fillers and 2 sanity check items), while BC experiments do not include Condition A items, but include 2 Condition B and 3 Condition C fillers instead.

No methods were used to identify uncooperative participants, so small fluctuations in variance are to be expected, including within-group variance. Because of that, when comparing effect size estimates, we ignore any discrepancies smaller than 0.2 (i.e. one small effect size).

In Condition A, the acceptability judgment task uncovers a statistically significant effect ($p < .001$, n = 67; at the significance level of $p < .05$) with a mean difference of 0.62 and an identical Cohen's d of 0.62, both indicating a medium effect size. In the coreference judgment task, the effect is also statistically significant ($p < .001$, n = 67) with a mean difference of 0.95 and Cohen's d of 0.89, suggesting a large effect. The matching values of effect size estimates for each task suggest that the participants' responses were consistent irrespective of the task.

In Condition B, we find a statistically significant effect in the acceptability judgment task ($p < .001$, n = 134) with a mean difference of 0.57 and Cohen's d of 0.56, which both suggest a medium effect size. The coreference judgment task produces a very large effect size (mean diff. = 1.29, Cohen's d = 1.22), which is also significant ($p < .001$, n = 134). Both effect size estimates align closely in both tasks, suggesting similar within-group variance in both tasks.

In Condition C, the acceptability judgment task revealed a large and significant effect ($p < .001$, n = 134, mean diff. = 1.18, Cohen's d = 1.3). The coreference judgment task also identifies a significant effect ($p < .001$, n = 134) with a massive effect size (mean diff. = 1.93, Cohen's d = 3.46). In the coreference judgment task, we observe a much higher value of Cohen's d compared to the mean difference, which indicates a very small within-group variance, suggesting a high level of agreement among participants using this task.

In Condition C + Reconstruction, both effect sizes are much smaller than in the other sanity checks. The acceptability judgment task produces a small, non-significant effect (p = .096, n = 134, mean diff. = 0.17, Cohen's d = 0.17), while the coreference judgment task identifies a small, but significant effect (p = .004; n = 134, mean diff. = 0.29, Cohen's d = 0.26). Comparison of effect size estimates for each task indicates a similar amount of within-group variance.

Our results show that the coreference judgment task produces larger effect sizes across all sanity checks than the acceptability judgment task. Both tasks found significant effects in Conditions A, B, and C. In Condition C + Reconstruction, only the coreference judgment task uncovered a small, yet significant effect. The coreference judgment task also yields a much smaller within-group variance for Condition C, suggesting that it fits this task particularly well.

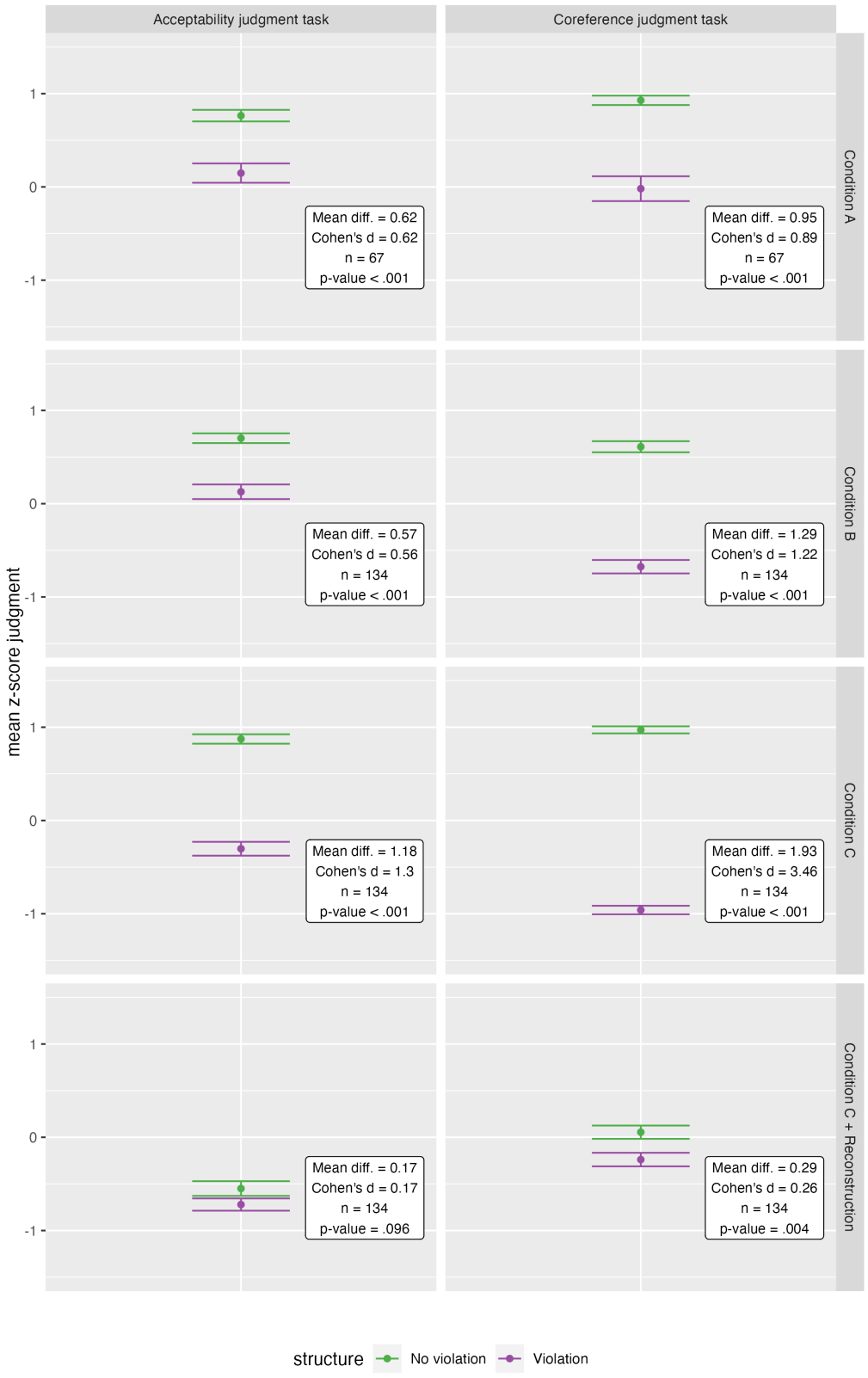

Figure 4: Effect sizes for the two tasks across several binding phenomena

4.3 The results of the ROC curve performance analysis for the two tasks

Figure 5 shows the ROC curves for each sanity check and the corresponding AUC-ROC values for the two binding tasks. As a reminder, the closer the AUC-ROC value is to 1, the better the participants following the task instructions distinguish between the two classes of items, while an AUC-ROC value of 0.5 indicates that they have the same discriminative ability as a coin toss.

Across all conditions, the coreference judgment task consistently shows superior discriminative ability compared to the acceptability judgment task. It shows slightly better performance in Condition A (AUC-ROC = 0.741 vs. 0.719), a substantial improvement in Condition B (AUC-ROC = 0.869 vs. 0.683), and near-perfect discrimination in Condition C (AUC-ROC = 0.982 vs. 0.86). Although both tasks receive lower AUC-ROC values in Condition C + Reconstruction, the coreference judgment task still outperforms the acceptability judgment task (AUC-ROC = 0.603 vs. 0.553).

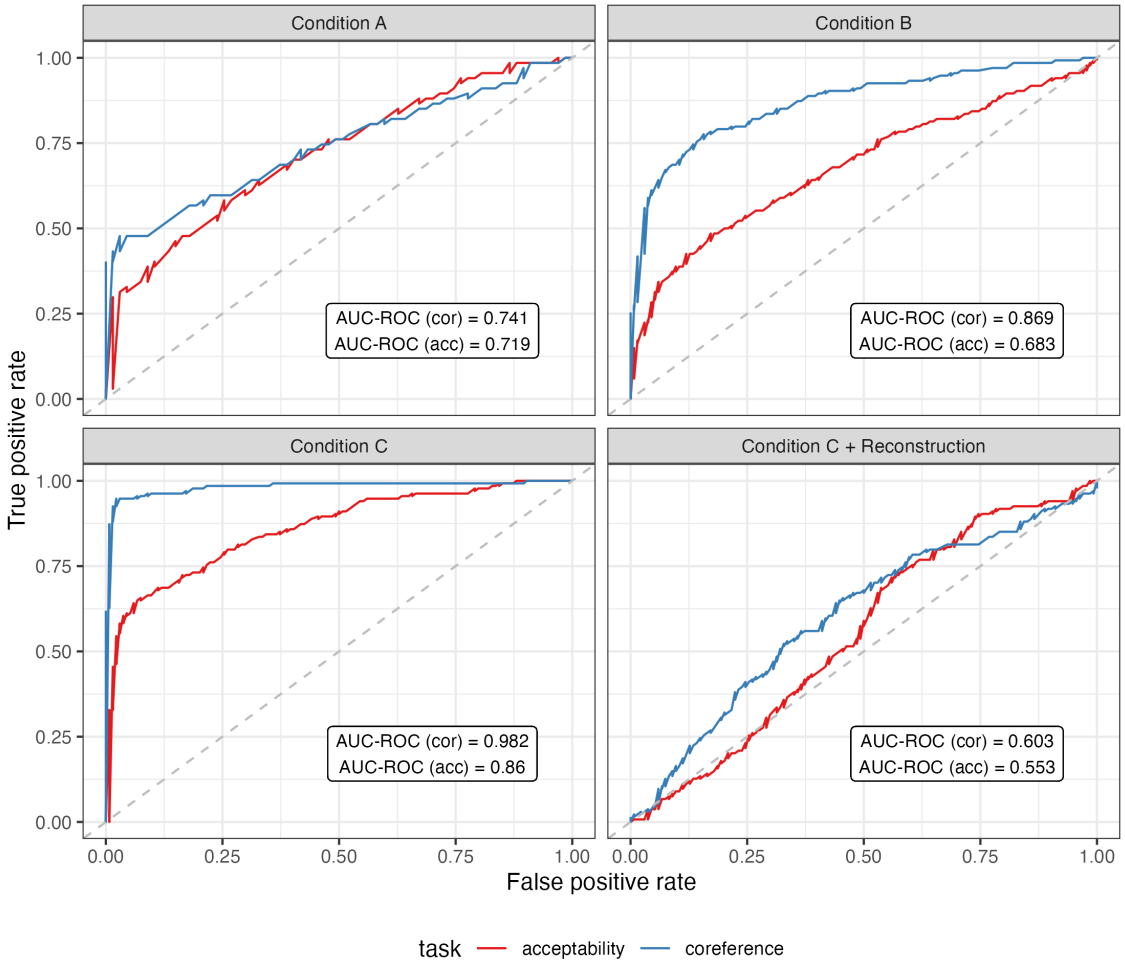

Figure 5: ROC curves for the two tasks across several binding phenomena

4.4 The results of the power simulations for the two tasks

Figure 6 shows the results of the power simulations for the two tasks. We conducted 3000 simulations for each sample size across a wide range of sample sizes, from 5 to 100

participants, using a linear mixed-effects model with a critical value of $z = 2$. For Conditions A, B, and C, the smallest effect size of interest (SESOI) was set to 80% of the beta coefficients (β) for the fixed effect and the intercept. For Condition C + Reconstruction, unadjusted values of β were used. The simulations for each task under each condition were conducted separately.

In terms of the sample required to reach a recommended minimal threshold of 80% statistical power, all the sanity checks demonstrate the same trend for the two tasks. For Condition A, the acceptability judgment task requires 31 participants, while the coreference judgment task requires only 19. In Condition B, the acceptability judgment task needs 45 participants, compared to only 8 for the coreference judgment task. Condition C shows the lowest participant requirements, with 10 needed for the acceptability judgment task and just 5 for the coreference judgment task. For Condition C + Reconstruction, neither task achieved 80% power within the tested range of 5 to 100 participants, though the power of the acceptability judgment task increases more slowly than that of the coreference judgment task. Overall, the coreference judgment task consistently requires fewer participants to reach 80% power in the binding configurations tested.

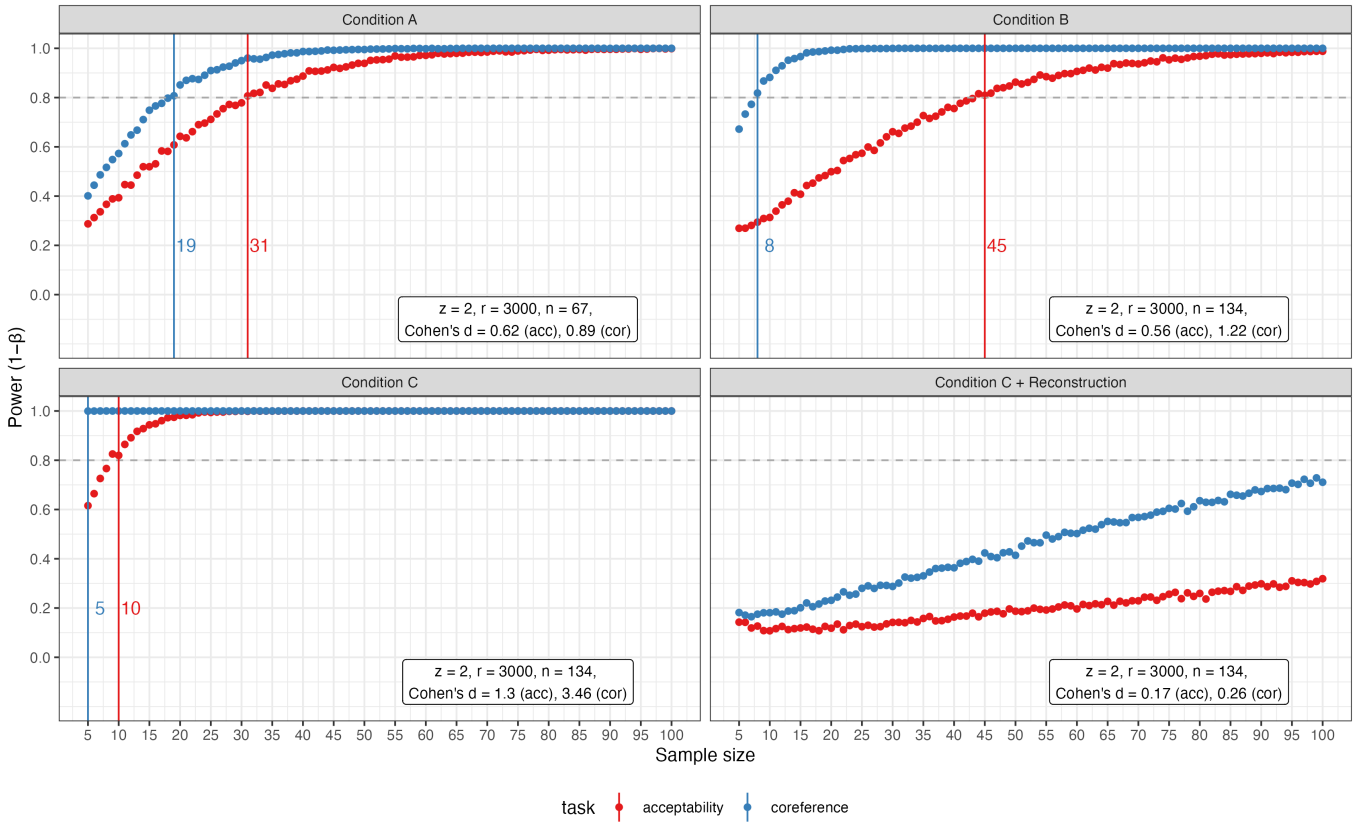

Figure 6: Power simulations for the two tasks across several binding phenomena

5 Discussion

In this study, we compared two binding tasks, the coreference judgment task and the acceptability judgment task, using several statistical techniques. Our results indicate that the coreference judgment task is superior in all respects. We saw that the coreference judgment task produced far fewer non-unimodal distributions for fillers, indicating that

participants were able to follow the task instructions easily and get similar results each time. It also yielded larger effect sizes across all four binding phenomena that were tested using minimal pairs. Additionally, we discovered that the within-group variance for Condition C was particularly low when using this task, suggesting that the task is especially well-suited for studying Condition C and its subtypes. Furthermore, the task showed better performance across the same four binding phenomena and required a smaller sample size in each case.

In contrast, using the acceptability judgment task presents a number of serious challenges. We observed multiple non-unimodal distributions, especially among fillers with expected ratings in the lower part of the scale. One possible interpretation for this is that participants did not fully engage with the metalinguistic part of the task and instead reported the general acceptability of a sentence. Potential remedies for this situation include: incorporating a protocol to identify uncooperative participants, extending the training period to ensure that all participants understand all components of the task, and introducing the functionality for participants to contact the experimenter with clarification questions in real-time. Finally, we found that the acceptability judgment task task produced smaller effect sizes and also needed larger sample sizes, but both of these trends might change if more participants followed the task instructions.

The overarching conclusion of our study is that the coreference judgment task offers more advantages for experiments that test binding phenomena. This is surprising, since the expected ratings for the fillers and the four phenomena tested with minimal pairs were all taken from the linguistic literature and were presumably generated by professional linguists using the same procedure as the acceptability judgment task. However, when naive participants are presented with the same task, it appears to be more difficult and therefore produces poorer results compared to the coreference judgment task. Nevertheless, the acceptability judgment task may still be necessary for research questions that explore possible interactions between binding phenomena and other grammatical phenomena that can only be assessed using acceptability, but not coreference. In all other cases, the coreference judgment task is the better option.

6 Conclusion

In summary, our study found that the coreference judgment task is a better choice than the acceptability judgment task for studying binding phenomena. It provides much more consistent results, with fewer confusing patterns in participant responses and stronger effects in our tests, and requires fewer participants to get reliable results. However, the acceptability task may still be useful for some specific research questions. One common scenario is exploring the interaction of binding with other phenomena that do not require assessing different readings. In these situations, acceptability is the only measure of the interaction of the phenomena. When using the acceptability judgment task for binding, it is important to consider increasing the sample size and implementing various outlier detection techniques to exclude participants who have difficulty following the task.

References

Bates, Douglas & Mächler, Martin & Bolker, Ben & Walker, Steve. 2015. Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01

Bradley, Andrew P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30(7). 1145–1159.

Bruening, Benjamin & Al Khalaf, Eman. 2019. No argument–adjunct asymmetry in reconstruction for Binding Condition C. *Journal of Linguistics* 55(2). 247–276.

Button, Katherine S. & Ioannidis, John P. & Mokrysz, Claire & Nosek, Brian A. & Flint, Jonathan & Robinson, Emma S. & Munafò, Marcus R. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14(5). 365–376.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Foris.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. Routledge.

Fawcett, Tom. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27(8). 861–874.

Fiengo, Robert & May, Robert. 1994. *Indices and identity*, vol. 24. MIT Press.

Gordon, Peter C. & Hendrick, Randall. 1997. Intuitive knowledge of linguistic coreference. *Cognition* 62(3). 325–370.

Hanley, James A. & McNeil, Barbara J. 1982. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology* 143(1). 29–36.

Hartigan, John A. & Hartigan, Pamela M. 1985. The dip test of unimodality. *The annals of Statistics* 70–84.

Heim, Irene. 1983. File change semantics and the familiarity theory of definiteness. *Semantics Critical Concepts in Linguistics* 108–135.

Kaiser, Elsi & Runner, Jeffrey. 2023. Acceptability judgments of binding and coreference: Methodological considerations. In Sprouse, Jon (ed.), *The oxford handbook of experimental syntax*, 29–52. Oxford University Press.

Kazanina, Nina & Lau, Ellen F. & Lieberman, Moti & Yoshida, Masaya & Phillips, Colin. 2007. The effect of syntactic constraints on the processing of backwards anaphora. *Journal of Memory and Language* 56(3). 384–409.

Keller, Frank & Asudeh, Ash. 2001. Constraints on linguistic coreference: Structural vs. pragmatic factors. In *Proceedings of the annual meeting of the cognitive science society*, vol. 23.

Koval, Pasha & Sprouse, Jon. 2023. The target of relative clause extraposition: An experimental investigation of c-command effects. Manuscript.

Kumle, Levi & Võ, Melissa L-H & Draschkow, Dejan. 2021. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior research methods* 53(6). 2528–2543.

Kuznetsova, Alexandra & Brockhoff, Per B. & Christensen, Rune H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82(13). 1–26.

Lasnik, Howard. 1989. *Essays on anaphora*, vol. 16 (Studies in Natural Language and Linguistic Theory). Kluwer.

Linzen, Tal & Oseki, Yohei. 2018. The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics* 3(1).

Maechler, Martin. 2021. *diptest: Hartigan's Dip Test Statistic for Unimodality - Corrected*. https://CRAN.R-project.org/package=diptest. R package version 0.76-0.

Marty, Paul & Chemla, Emmanuel & Sprouse, Jon. 2020. The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: a journal of general linguistics* 5(1). 72.

Maxwell, Scott E. & Delaney, Harold D. & Kelley, Ken. 2017. *Designing experiments and analyzing data: A model comparison perspective*. Routledge.

May, Robert. 1985. *Logical form: Its structure and derivation*, vol. 12. MIT Press.
Neyman, Jerzy & Pearson, Egon S. 1928a. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* 20(1/2). 175–240.
Neyman, Jerzy & Pearson, Egon S. 1928b. On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika* 20(3/4). 263–294.
Neyman, Jerzy & Pearson, Egon Sharpe. 1933. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231(694-706). 289–337.
Pashler, Harold & Wagenmakers, Eric-Jan. 2012. Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science* 7(6). 528–530.
R Core Team. 2023. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.
Reinhart, Tanya. 1976. *The syntactic domain of anaphora*. Cambridge, MA: Massachusetts Institute of Technology Doctoral dissertation.
Reinhart, Tanya & Reuland, Eric. 1993. Reflexivity. *Linguistic inquiry* 24(4). 657–720.
Robin, Xavier & Turck, Natacha & Hainard, Alexandre & Tiberti, Natalia & Lisacek, Frédérique & Sanchez, Jean-Charles & Müller, Markus. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12. 77.
Ruder, Sebastian. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* .
Safir, Ken. 1999. Vehicle change and reconstruction in ā-chains. *Linguistic inquiry* 30(4). 587–620.
Shorack, G. R. & Wellner, J. A. 2013. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.
Sportiche, Dominique. 1997. Reconstruction, movement, and scope. *Ms. UCLA* .
Sprouse, Jon & Almeida, Diogo. 2012a. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48(3). 609–652.
Sprouse, Jon & Almeida, Diogo. 2012b. Power in acceptability judgment experiments and the reliability of data in syntax. *Ms., University of California, Irvine & New York University Abu Dhabi* .
Sprouse, Jon & Almeida, Diogo. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa* 2(1). 1.
Sprouse, Jon & Schütze, Carson T. & Almeida, Diogo. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134. 219–248.
Stockwell, Richard & Meltzer-Asscher, Aya & Sportiche, Dominique. 2021. There is reconstruction for Condition C in English questions. *North East Linguistic Society (NELS 51)* .
Swets, John A. 1988. Measuring the accuracy of diagnostic systems. *Science* 240(4857). 1285–1293.
Temme, Anne & Verhoeven, Elisabeth. 2017. Backward binding as a psych effect: A binding illusion? *Zeitschrift für Sprachwissenschaft* 36(2). 279–308.
Van Riemsdijk, Henk & Williams, Edwin. 1981. NP-structure. *The Linguistic Review* 1.
Wasserman, Larry. 2006. *All of nonparametric statistics*. Springer Science & Business Media.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.