Quantifying weak and strong crossover for wh-crossover and proper names[1]

Hayley ROSS — *Harvard University*
Gennaro CHIERCHIA — *Harvard University*
Kathryn DAVIDSON — *Harvard University*

Abstract. Despite the status of crossover phenomena as key data in the theoretical study of pronominal syntax/semantics, crossover structures have been subjected to relatively little experimental testing. We investigate *wh*-crossover, quantificational crossover, and analogous structures involving proper name cataphora in behavioral rating studies, deploying a novel experimental paradigm which contrasts the acceptability of multiple readings of a target sentence. Using this more sensitive experimental task, our results favour an acceptability distinction between weak and strong *wh*-crossover, and provide a baseline against which to compare more controversial cases of crossover. We further find that proper name cataphora display a remarkably similar crossover effect, including a distinction between "strong" and "weak" cases.

Keywords: binding, crossover, anaphora, cataphora, semantics, experimental semantics

1. Introduction

This paper investigates crossover phenomena from an experimental perspective. We develop a general method to measure the deviance of crossover sentences with respect to each other and to grammatical controls. This is useful and necessary in order to address the many empirical controversies that still surround this classical topic, and hinder progress in understanding why crossover constraints exist. Crossover refers to a specific syntactic configuration in which pronouns cannot be interpreted as coconstrued with a particular antecedent, even though that antecedent ostensibly has scope over the pronoun. Classical cases of crossover involve either *wh*-words (Postal, 1971) or quantifiers Chomsky (1976):

(1) *Wh*-crossover
   a. *The teacher couldn't remember **which**$_i$ of the students they$_i$ said __ didn't need to hand in the essay.
   b. *The teacher couldn't decide **which**$_i$ student's poem topic they$_i$ liked __ the most.
   c. *?The teacher wondered **which**$_i$ student **their**$_i$ project topic frustrated __ the most.
   d. The teacher wondered **which**$_i$ of the students __ enjoyed the essay topic **they**$_i$ had chosen.
   e. The teacher couldn't decide **which**$_i$ student's poem topic __ frustrated **them**$_i$ the most.
   f. The teacher wondered **which**$_i$ of the students __ enjoyed **their**$_j$ project topic.

(2) Quantificational crossover
   a. *The lawyer noticed that **they**$_i$ forgot something about **every witness**$_i$' statement.
   b. *?The lawyer noticed that **their**$_i$ statement at the trial was upsetting for **every witness**$_i$.
   c. The lawyer noticed that **every witness**$_i$' statement at the trial was upsetting for **them**$_i$.
   d. The lawyer noticed that **every witness**$_i$ forgot something about **their**$_i$ statement.

[1]We would like to thank four anonymous reviewers for Sinn und Bedeutung as well as the members of the Meaning and Modality Lab at Harvard University for their feedback. Special thanks go to research assistant Sara Manning.

Traditionally, (1a), (1b) and (2a) are marked ungrammatical and referred to as *strong crossover* while the examples (1c) and (2b) are often deemed less severely deviant and referred to as *weak crossover*, a term coined by Wasow (1972). In strong crossover, the pronoun c-commands the quantifier or the gap corresponding to the *wh*-word, and thus cannot bind the pronoun, in contrast to when the gap c-commands the pronoun (as in (1d), (1e), (1f), (2c) and (2d)). In weak crossover, the pronoun does not c-command the gap or quantifier. Some authors further divide strong crossover into "true" strong crossover as in (1a) versus secondary strong crossover (Postal, 1971), as in (1b) and (2a), where the *wh*-word or quantifier is nested inside a larger constituent. The paradigm in (3) constitutes a minimal set that vividly illustrates the problem specifically with quantifier scope. The inverse scope construal is straightforward in (3a), yet pronoun binding in the very same configuration in (3b) and (3c) appears to be unavailable.

(3) Scope vs. binding
   a. The teacher thought that **someone** liked **every student**'s essay topic.
      '∀ over ∃' construal possible
   b. *The teacher thought that **they**$_i$ liked **every student**$_i$'s essay topic.
   c. *?The teacher thought that **their**$_i$ friends liked **every student**$_i$'s essay topic.

One of our goals is to find a way to measure whether *wh*-crossover is as deviant as quantificational crossover with respect to grammatical controls and whether strong crossover is as deviant as weak crossover. We use two grammatical controls. The first type involves exactly the same sentences (i.e. (1a), (1b), (1c), (2a), and (2b)) except that the pronoun is not understood as bound by the *wh*-phrase or the quantifier. Instead, we target the reading where it corefers with the matrix subject NP (*teacher* / *lawyer*) which we refer to as the "distractor NP". By providing an alternate referent for the pronoun explicitly, we not only create a grammatical control but also avoid participants accommodating such a referent, which can be a confound for acceptability judgement studies on binding/crossover (Kush, 2013; Kush et al., 2017). The second of type of control involves sentences in which the structural relation between the pronoun and the gap is inverted to form a standard binding configuration (as in (1d), (1e), (1f), (2c), and (2d)).

Quantitatively measured judgements will help us decide whether strong vs. weak crossover are on par or, as tradition has it, that weak crossover is less deviant. It will also provide us a with a methodological basis to address controversial cases. For example, there is disagreement on the extent to which strong vs. weak relative clause variants of *wh*-crossover are deviant: weak crossover in relative clauses is sometimes found to be less deviant than its *wh*-question counterpart (Lasnik and Stowell, 1991). Postal (1993) claims that relative clauses in French display no weak crossover effects at all (unlike English relative clauses, which he claims do). Further, a plethora of languages including German, Kiswahili, Hindi, Malayalam and Mandarin are argued not to show weak crossover effects under certain conditions, such as scrambling (Bresnan, 1998; Fanselow et al., 2005; Lyu, 2017; Bhatt and Keine, 2019). This area is clearly ripe for experimental work. To showcase how our design can be applied to such controversies, we study cases structurally parallel to those in (1)-(2) involving coreference with proper names.

(4) Constraints on coreference (cataphora with proper names)
   a. ***He**$_i$ claimed that **Daniel**$_i$ was an amazing chef.
   b. *The chef knew that **he**$_i$ was disappointed by the soup **Daniel**$_i$ made.
   c. ?The chef$_j$ knew that **his**$_i$ soup had disappointed **Daniel**$_i$.

In (4a) and (4b), which are parallel to strong crossover configurations, the pronoun cannot be understood as coconstrued with the proper name *Daniel* that it c-commands. This is particularly interesting since no constraint on binding can per se automatically prevent two referential expressions from coreferring. By contrast, coreference/coconstrual of the pronoun with the proper name appear to be much more accessible in weak crossover configurations such as (4c). However, these basic intuitions can be challenged, with the readings available in these constructions affected by contextual factors, and by phenomena like focus (Heim, 1998, 2007; Bianchi, 2009; Moulton et al., 2018; Gor and Syrett, 2019, Gor and Syrett), and so we apply our method to probe crossover severity quantitatively to these cases as well.

It turns out that the most widely used, perhaps at this point even "classic", experimental method for investigating semantics, namely acceptability or truth-value judgements given a context, is difficult to implement for binding or coreference phenomena[2]. This is because providing a context is not sufficient for participants to read the target sentence with the desired bound reading: if they find the sentence acceptable, it's often still possible they have instead understood the sentence with the pronoun referring to a distractor NP, as in (5a) and (5b). Moreover, they may actually judge the sentence as *unacceptable* relative to a context that describes the bound reading (since they understand the sentence as having a different reading).

(5) a. The lawyer$_j$ noticed that **they**$_j$ forgot something about **every witness**$_i$' statement.

b. The lawyer$_j$ noticed that **their**$_j$ statement at the trial was upsetting for **every witness**$_i$.

Instead we deploy a "meaning availability" design inspired by the unambiguous paraphrases commonly used in syntax/semantics teaching, described in Section 2, which directly compares the two available readings side-by-side. A direct comparison of this approach to the classic sentence acceptability design, using the same experimental items, can be found in Appendix A.

The structure of this paper is as follows: Experiment 1 and Experiment 2 investigate *wh*-crossover and quantificational crossover respectively. These showcase our experiment design on relatively well-understood cases of crossover, create a baseline for more controversial cases of crossover to be compared against, and allow us to probe whether *wh*-words and quantifiers really behave identically. In brief, we find a significant impact of crossover on pronoun binding for both *wh*-binding (Experiment 1) and quantifier binding (Experiment 2). We also find a significant difference between weak and strong cases, but only for *wh*-crossover. This will be relevant in assessing the empirical coverage of theories which do vs. don't distinguish strong and weak crossover. We then move on to a more controversial case, investigating coreference between pronouns and proper names in Experiment 3. We also find a significant effect of cataphora, analogous to crossover, and a significant difference between weak and strong configurations ((4b) is more deviant than (4c)). These results have implications for theories relating coreference and binding, such as Rule I (Grodzinsky and Reinhart, 1993) and its derivatives.

[2]An acceptability design is used by Kush (2013); we suspect that this, along with issues with plausibility, may explain why he found no difference between weak and strong crossover. While he did find an effect of crossover overall, this is arguably confounded by the implausibility of his crossover sentences under their bound reading.

2. Experiment 1: *wh*-crossover

Our first experiment investigates *wh*-crossover by comparing possible crossover interpretations to two controls: the same sentence where the pronoun is interpreted as the distractor NP, and structurally similar sentences involving straightforward (possible) binding.

2.1. Method

We use a "meaning availability" design which sidesteps the idea of sentence acceptability and jumps immediately to the question at the heart of crossover, namely what the pronoun is co-construed with. Participants see the target sentence along with two unambiguous paraphrases corresponding to the two readings, and are asked to rate "To what degree can this mean...?" each paraphrase on a Likert scale of 1-5. A screenshot from this study showing this meaning availability design (used in all subsequent experiments) is given in Figure 1. Presenting the two paraphrases side by side helps participants be aware that the sentence may be ambiguous, and consider a second reading they may not initially have spotted[3]. We use a 2x3x2 cross, crossing the two orders *wh*...[gap]...pronoun (corresponding to binding) and *wh*...pronoun...[gap] (corresponding to crossover) with three structural configurations corresponding to strong, secondary strong and weak crossover, and each of those with two potential readings[4].

(1) *Wh*-crossover and *wh*-binding (repeated from the introduction)

a. The teacher$_j$ couldn't remember **which**$_i$ of the students they$_{i/j}$ said __ didn't need to hand in the essay.

b. The teacher$_j$ couldn't decide **which**$_i$ student's poem topic they$_{i/j}$ liked __ the most.

c. The teacher$_j$ wondered **which**$_i$ student **their**$_{i/j}$ project topic frustrated __ the most.

d. The teacher$_j$ wondered **which**$_i$ of the students __ enjoyed the essay topic **they**$_{i/j}$ had chosen.

e. The teacher$_j$ couldn't decide **which**$_i$ student's poem topic __ frustrated **them**$_{i/j}$ the most.

f. The teacher$_j$ wondered **which**$_i$ of the students __ enjoyed **their**$_{i/j}$ project topic.

To better control the domain of quantification and noun gender, we use *which NP* phrases instead of plain *who*. We note that *which NP* constructions (D-linked *wh*-phrases) have been claimed to result in weaker (weak) crossover effects than *who* (Wasow, 1972; Pesetsky, 1987; Falco, 2007), but do not anticipate this as an issue since it will apply to all our items equally.

A second notable feature in our experiments is the use of singular *they* rather than gendered pronouns. Most traditional examples of crossover in the literature involve masculine pronouns; however, many anecdotal reports among younger English speakers suggest a dispreference for coconstrual between masculine pronouns and *wh*-words like *who* in English. Singular *they* has been in use for centuries precisely in these epicene cases, where the gender is unknown or irrelevant (Balhorn, 2004; Bjorkman, 2017; Conrod, 2019). We include masculine, feminine and singular *they* pronouns for a final 2x3x2x3 design. Finally, we conducted a separate norming study on the plausibility of the paraphrases used, which we omit for space, but which ensured that readings were similarly plausible across sentences.

[3]Unlike the forced-choice design of Felser and Drummer (2017) (studying crossover in German), this allows participants to rate both readings high if they feel both to be available, meaning that the possibility for the pronoun to refer to something other than the *wh*-word does not occlude the ratings for it being bound by the *wh*-word.

[4]The full set of experimental items (108 sentences) is available on OSF at `https://osf.io/dh2qb`.

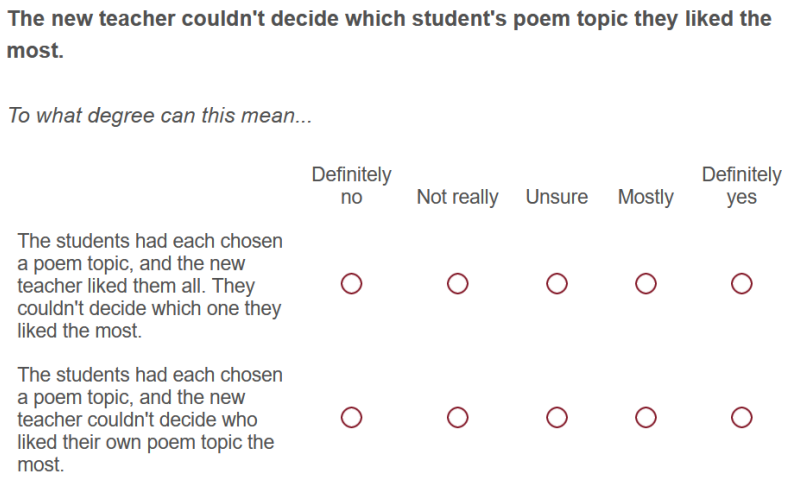


Figure 1: Screenshot of a secondary strong crossover sentence in the meaning availability design (Experiment 1). The first paraphrase describes the distractor NP reading (*they* coconstrued with *the new teacher*), while the second describes the bound reading (*they* coconstrued with *which student*). We expect high ratings for the first paraphrase and low ratings for the second.

We preregistered this study on OSF[5]. We recruited 144 self-reported native English speakers on Prolific, of which we excluded 7 due to failed attention checks and 1 due to having learned English later than age 5. Participants saw 6 target items and 6 fillers in random order. These participants and participants in all studies reported in this paper were paid for their time *pro rata* at a rate of $12/hr.

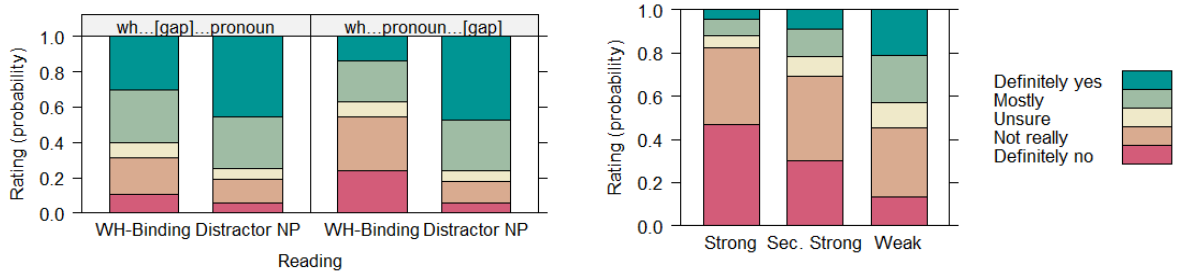
2.2. Results

We fit an ordinal mixed effects model in R (R Core Team, 2021) using the `ordinal` package (Christensen, 2019) with an interaction between gap/pronoun order and reading, as well as random effects for participant ID, average rating of each paraphrase from the norming study and the tendency of participants to identify ambiguous filler items (some participants always rated both readings of ambiguous fillers high, while others only ever rated one of the two readings high)[6]. Figure 2a shows the model's proportions of ratings for each condition.

We see no significant effect of reading alone ($p = 0.14$), but a clear, significant effect of pronoun-before-gap on the bound reading (i.e. crossover vs. binding) between columns 1 and 3 of Figure 2a. This decreases the odds of a high rating by a factor of 0.38 ($SE = 1.31$, $p < 0.05$). We also see a significant positive effect of pronoun-before-gap with the distractor reading (compare columns 3 and 4), which increases the odds of a high rating by a factor of 2.81 ($SE = 1.49$, $p < 0.05$). In other words, we find that across strong, secondary strong and weak crossover combined, we do indeed see that the crossover (pronoun before gap) configuration significantly reduces the availability of coconstrual. These results also show that it is specifically the coconstrued reading causing the low ratings, not the syntax of the sentence per se, since the distractor reading receives high ratings and the sentences are identical apart from the indexation/interpretation of the pronoun.

[5] `https://osf.io/9p3ws`

[6] Later AIC tests found that the ambiguity availability on fillers did not actually improve the model's overall performance; however, we leave this random effect in since it was used throughout the original analysis.



(a) Effect of gap/pronoun order on reading

(b) Strong vs. weak crossover

Figure 2: Results for Experiment 1. We see a significant effect of crossover (columns 1 and 3 of (a)) and a significant effect of strong vs. weak crossover.

To quantify the effect of strong vs. weak crossover, we fit a second ordinal mixed effects model on just the bound reading of pronoun-before-gap (crossover) items, with the same random effects as before, shown in Figure 2b. Notably, this effect is significant: weak crossover roughly doubles the likelihood of a high rating, increasing the odds of a high rating by a factor of 2.19 ($SE = 1.26$, $p < 0.05$). The difference between strong and secondary strong shows a trend for secondary strong to be weaker than true strong crossover, but this is not significant ($p = 0.30$).

The above results are shown combined across all pronoun genders (masculine, feminine and singular *they*). We also investigated the effect of pronoun gender on binding, and found that in fact, singular *they* provides the highest ratings for the bound reading (and the most balanced between the bound and distractor reading). Participant age has no significant effect on the availability of bound singular *they*. More details can be found in Appendix B.

2.3. Discussion

Experiment 1 demonstrates the utility of the meaning availability design: we are able to reliably detect crossover and even a significant difference between strong and weak crossover, supporting long-held intuitions (Wasow, 1972) that weak crossover violations are less severe than strong crossover. This also matches the work of Felser and Drummer (2017) on German *wh*-crossover. That said, the overall ratings for weak crossover are still relatively low, with 66% of ratings "Unsure" or lower (compared to 77% for strong crossover and 76% for secondary strong crossover) so this does not mean that weak crossover violations are not still violations. Finally, we find that singular *they* yields the crispest results in this experiment design on standard binding, motivating the use of just singular *they* in subsequent experiments.

3. Experiment 2: Quantificational crossover

3.1. Method

Experiment 2 uses exactly the same method as Experiment 1[7], namely the meaning availability design with a 2x2x2 design crossing the two orders *wh*...[gap]...pronoun (corresponding to binding) and *wh*...pronoun...[gap] (corresponding to crossover) with two sentence types corresponding to secondary strong and weak crossover, and each of those with two potential

[7]We did not preregister Experiment 2, but used an identical method and plan of analysis to Experiment 1.

readings[8]. We used singular *they* throughout. As before, we also conducted a separate norming study on the plausibility of the paraphrases, which ensured that readings were similarly plausible across sentences.[9]

(2) Quantificational crossover and binding (repeated from the introduction)
   a. The lawyer$_j$ noticed that **they**$_{i/j}$ forgot something about **every witness**$_i$' statement.
   b. The lawyer$_j$ noticed that **every witness**$_i$' statement at the trial was upsetting for **them**$_{i/j}$.
   c. The lawyer$_j$ noticed that **their**$_{i/j}$ statement at the trial was upsetting for **every witness**$_i$.
   d. The lawyer$_j$ noticed that **every witness**$_i$ forgot something about **their**$_{i/j}$ statement.

We recruited 60 self-reported native English speakers on Prolific, of which we excluded 1 due to failed attention checks and 1 due to having learned English later than age 5. Participants saw 6 target items and 6 fillers in random order.

3.2. Results

We fit an ordinal mixed effects model with an interaction between gap/pronoun order and reading, as well as random effects for participant ID, average rating of each paraphrase from the norming study and the ability of participants to identify ambiguous filler items. Figure 3a shows the model's proportions of ratings for each condition.

This time, we do see an effect of reading, ($p < 0.05$), with the distractor NP reading being significantly dispreferred by a factor of 0.27 over the bound reading for binding (non-crossover) sentences. As for *wh*-crossover, we see a clear, significant effect of pronoun-before-quantifier on the bound reading. This decreases the odds of a high rating by a factor of 0.12 ($SE = 1.55$, $p < 0.05$). We also see a significant positive effect of pronoun-before-gap with the distractor reading (compare columns 3 and 4), which increases the odds of a high rating by a factor of 52.34 ($SE = 1.88$, $p < 0.05$). In other words, we find that across secondary strong and weak crossover combined, we do indeed see that the crossover (pronoun before quantifier) configuration significantly reduces the availability of coconstrual.

We fit a second ordinal mixed effects model on just the bound reading of pronoun-before-quantifier (crossover) items to quantify the effect of secondary strong vs. weak crossover, with the same random effects as before. Unlike for *wh*-crossover, we find no significant difference between strong and weak quantificational crossover ($p = 0.69$). In fact, there is even a (non-significant) tendency for weak crossover to be rated slightly *worse* than strong crossover[10].

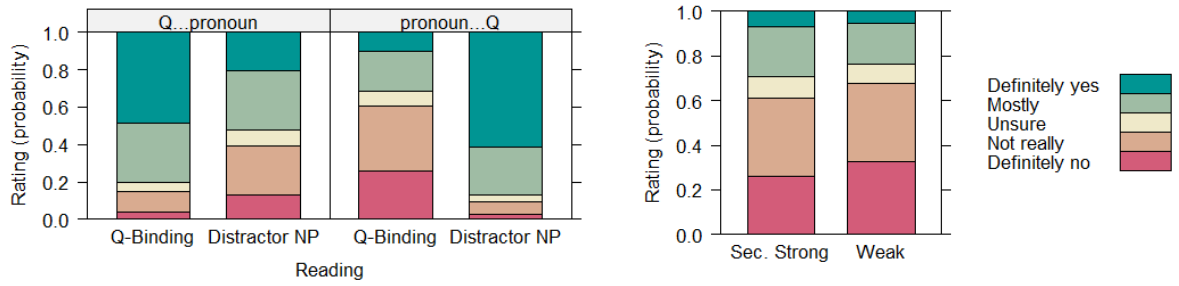
[8]We also tested "true strong" crossover involving nested sentences and no possessive pronouns, in parallel with Experiment 1. However, quantifier scope restrictions rule out coconstrual in these cases independently of crossover. We were largely unable to design "true strong" quantificational crossover sentences that did not violate scope restrictions within the confines of the experiment paradigm (which must not use reflexives to preserve the ambiguity between readings), but did include one set of "snake" sentences (Rooryck and Vanden Wyngaerd, 2011 i.a.), which use locative prepositions following particular verbs (we use perception verbs).

(i) The wildlife expert$_j$ said that **they**$_{i/j}$ spotted a stick insect in front of **every visitor**$_i$ eventually.

Snake sentences are said to contain a small clause, not a full clause (Bryant, 2022), though the ability of the quantifier to scope out of it has not been tested experimentally.

[9]The full set of experimental items (30 sentences) is available on OSF at `https://osf.io/dh2qb.`.

[10]For snake sentences, where we can construct "true" strong crossover as well as secondary strong crossover, we do find a significant effect of strong vs. weak crossover, however it is in the opposite direction to what the theory



(a) Effect of quantifier/pronoun order on reading

(b) Secondary strong vs. weak crossover

Figure 3: Results for Experiment 2. We see a significant effect of crossover (columns 1 and 3 of (a)), but no significant effect of strong vs. weak crossover.

3.3. Discussion

Experiment 3 creates a controlled comparison between *wh*-crossover and quantifier crossover in English. The ratings for the coconstrued reading of the quantificational crossover sentences are strikingly similar to the *wh*-setting, which is especially surprising given the variation in stimuli used and the seeming superficial difference between *wh*-words and quantifiers. We find, however, that the two differ in two respects: in sentences where the *wh*-word c-commands (could bind) the pronoun, the bound reading is typically dispreferred compared to the distractor NP reading, while it is significantly preferred for quantifiers. This has no bearing on our theoretical concerns but remains a point of curiosity.

Secondly, we find a statistically significant difference between strong and weak crossover only for *wh*-crossover. This supports the early work on crossover and coining of the term by Wasow (1972), who worked on *wh*-crossover, but raises questions about why quantificational crossover behave differently – current theoretical accounts typically treat them identically. We believe our experiment used sufficient participants and a sufficiently granular scale to detect a difference, but a replication may help settle this.

4. Impact on theories of crossover

First and foremost, our results lay a baseline quantifying the strength of crossover violations in the two standard cases of *wh*-crossover and quantificational crossover in English. While the fact that crossover has a significant effect on pronoun coconstrual obviously does not come as a surprise to the theoretical community, our results and experimental methodology allow for future comparison of these two standard cases against the more contested cases discussed in the introduction, including other languages and other constructions such as relative clauses and coreference with proper names.

Our strikingly similar results for *wh*-crossover and quantificational crossover in English (when viewed as weak and strong crossover combined) support the long-held theoretical view (Chomsky, 1976) that the two should be treated in parallel. Our results on strong vs. weak crossover

predicts. Judging by the norming study, we believe this is because the coconstrued reading for the weak crossover sentence was simply rather implausible (unfortunately the only item in the experiment where we could not come up with a truly equiplausible scenario), so this is open for more work. Nevertheless, we do find a significant effect of crossover overall for snake sentences.

are harder to interpret. In so far as we know, no theory makes a distinction between quantifier vs. *wh*-binding. The lack of difference between strong and weak crossover for quantifiers may be due to insufficient experimental power or other experimental flaws, or perhaps due to processing effects related to linear order or embedding depth (see Kush et al., 2017 and references therein for effects of crossover on processing). Further experiments will have to probe this in more detail. For *wh*-crossover, where we find a clear difference, theories of crossover can be divided into three broad camps along these lines. One camp subsumes all crossover under a single principle (Safir, 2004). An example of this is Safir's *Independence Principle* .

(6) *Independence Principle* (Safir, 2004)
If x depends on y, then x cannot c-command y.

Under Safir's definition of *depends*, both *they* in (1a) (strong crossover) and *their project topic* in (1c) (weak crossover) c-command the trace of the *wh*-word while also depending on it. Safir argues extensively against the strong/weak crossover distinction based on various cases of resumptive pronouns, weakest crossover and other related constructions.

(1) a. *The teacher couldn't remember **which**$_i$ of the students **they**$_i$ said __ didn't need to hand in the essay.
c. *?The teacher wondered **which**$_i$ student **their**$_i$ project topic frustrated __ the most.

A second camp holds that weak crossover is governed by an entirely different principle (Koopman and Sportiche, 1982; Safir, 1984). For example, Koopman and Sportiche's *Bijection Principle* is applied only to weak crossover, while they derive strong crossover from their definition of variable alongside Principles A and B.

(7) *Bijection Principle* (Koopman and Sportiche, 1982)
There is a bijective correspondence between variables and $\bar{\text{A}}$ positions.

The third camp uses an overarching crossover principle but holds that strong crossover violates some additional principle such as Principle C (Reinhart, 1983; Grodzinsky and Reinhart, 1993; Ruys, 2000; Shan and Barker, 2006; Chierchia, 2020). Reinhart's classic c-command-based binding theory defines co-indexation such that only bound variables can be interpreted when coindexed. This rules out strong and weak crossover alike as uninterpretable since they fail to bind the pronoun. Principle C or the Chain Condition (Reinhart and Reuland, 1993) is then invoked in addition to incur a "double" violation for strong crossover.

(8) a. *Binding* (Reinhart, 1983)
A node α is bound by a node β iff α and β are coindexed and β c-commands α.
b. *Translation definition* (Reinhart, 1983)
An NP is a variable iff either (i) it is empty and $\bar{\text{A}}$-bound, or (ii) it is A-bound and lacks lexical content. Other cases of NP coindexation are uninterpretable.

Only the latter two camps support an explanation for our difference between strong and weak *wh*-crossover, and only the third camp explicitly captures that strong crossover is "stronger" (the second camp would need to stipulate that violating their weak crossover principle is less severe). That said, the issue of the absence of a strong/weak contrast for quantificational crossover remains open, and in light of that, it also remains an open question whether capturing differences in severity should be a responsibility of semantic/syntactic theories, or rather whether

this is the responsibility of other domains such as pragmatics or processing.

5. The relationship between binding and coreference

We turn now from standard crossover to a case study showcasing how our method can be applied to other phenomena related to crossover which have been difficult to resolve with appeal to intuition. Specifically, we will focus on anaphora and cataphora involving proper names. The following sentences parallel crossover syntactically, though they involve coreference between referential expressions instead of binding:

(4) Constraints on coreference (cataphora with proper names)
   a. *He$_i$ claimed that **Daniel**$_i$ was an amazing chef.
   b. *The chef knew that **he**$_i$ was disappointed by the soup **Daniel**$_i$ made.
   c. ?The chef$_j$ knew that **his**$_i$ soup had disappointed **Daniel**$_i$.

Judgements on sentences of this form vary: while sentences likes (4c) are often reported as acceptable in the literature (Chomsky, 1976; Lasnik and Stowell, 1991; Ruys, 2004), especially with appropriate supporting context (consider: *Sue is not a popular girl, but **her**$_i$ mother loves **Sue**$_i$ unconditionally*[11]), they seem to be somewhat degraded compared to the reversed order *Sue's mother loves her*. Further, Chomsky (1976) argues that such sentences are unacceptable when the proper name bears focus, though this condition has been shown to be too simplistic (Bianchi, 2009; Moulton et al., 2018). Experiment 4 uses our meaning availability design to probe the relative severity of violation (or absence thereof) in these two cases of proper name cataphora. We will call the sentence configuration of (4c) "weak" in parallel with weak crossover, and refer to the c-command configuration in (4a) and (4b) as "strong" (collapsing the distinction between strong and secondary strong).

These sentences are examples of coreference between the proper name and the pronoun, not binding (at least under most accounts), and yet there is a striking parallel between these and cases of binding. This connection is capitalised on by Reinhart's Rule I (Grodzinsky and Reinhart, 1993), which essentially states that you can't have coreference between two NPs if you would have got the same meaning by one binding the other:

(9) *Rule I* (Grodzinsky and Reinhart, 1993)
NP A cannot corefer with NP B if replacing A with C, C a variable A-bound by B, yields an indistinguishable interpretation.

Rule I hinges on Reinhart's definition of binding, which involves c-command. Thus, it does not apply to cases like (4c). While we can get bound interpretations in configurations like (4c) and many other cases of "indirect binding" (see Barker (2012) for an overview), these must involve some other mechanism for achieving that reading, such as e-type pronouns, dynamic approaches, or a revision of the definition of c-command (the strategy chosen by Reinhart in her 1983 book). The parallel for (4c) are simple possessor binding sentences:

(10) Every boy's mother thinks he's a genius. *(adapted slightly from Higginbotham, 1980)*

Grodzinsky and Reinhart's paper on Rule I makes no mention of the extended version of c-

[11] There may be an important difference to be made between supporting contexts that mention the proper name in advance of the pronoun, allowing coreference between the pronoun and the previously mentioned name, and ones which do not (genuine cataphora). We will not investigate this further in this paper, though we are careful not to mention the proper name before the pronoun in all our experimental items.

command adopted in Reinhart (1983), and so we will treat Rule I as using the classical definition of c-command, and thus not covering these "weak" cases. If it does extend to possessor binding, then it makes the prediction that sentences such as (4c) should be ungrammatical; otherwise, they should be grammatical. Since intuitive judgements in this area vary, this makes it ripe for experimental study.

6. Experiment 3: Backwards cataphora with proper names

6.1. Method

We apply the same meaning availability design as used for the crossover experiments, taking advantage of the fact that these pronouns too may refer to some distractor NP other than the target proper name. We use a 2x2 design which crosses proper name and pronoun order with "strong" and "weak" (possessive) configurations, balanced for pronoun gender. We investigate these sentences in their most simple form, without special focus on the proper name and without previous mention of the proper name or any other description of its referent in the context[12]. We use six sets of examples, including the following (repeated from the introduction)[13]:

(11) a. The chef$_j$ knew that **Daniel**$_i$ was disappointed by the soup **he**$_{i/j}$ made.
b. The chef$_j$ knew that **Daniel**$_i$'s soup had disappointed **him**$_{i/j}$.
c. The chef$_j$ knew that **he**$_{i/j}$ was disappointed by the soup **Daniel**$_i$ made.
d. The chef$_j$ knew that **his**$_{i/j}$ soup had disappointed **Daniel**$_i$.

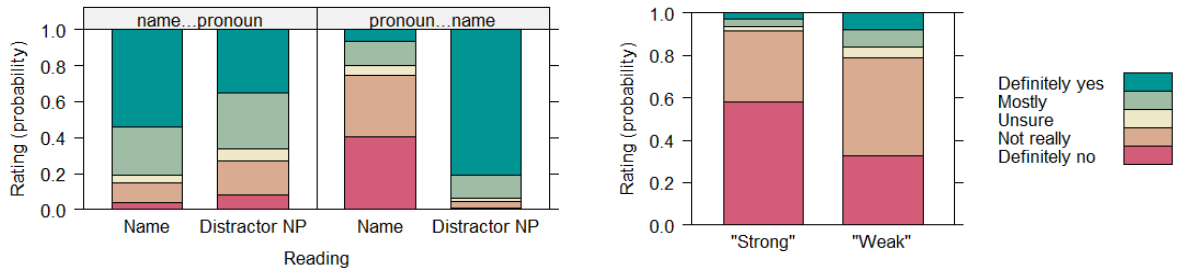
We preregistered this study on OSF[14]. We recruited 48 self-reported native English speakers on Prolific, of which we excluded 1 due to failed attention checks. Participants saw 6 target items and 6 fillers in random order.

6.2. Results

We fit an ordinal mixed effects model with an interaction between name/pronoun order and reading, as well as random effects for participant ID, participants' tendency to notice ambiguity in fillers, and scenario (in lieu of plausibility ratings). The results are shown in Figure 4a. As for *wh*-crossover, we see no significant effect of the reading alone ($p = 0.08$) but a significant effect of pronoun-before-name on the name reading, which decreases the odds of a high rating by a factor of 0.06 ($p < 0.05$). We again see a significant effect of pronoun-before-name with the distractor reading, which increases the odds of a high rating by a factor of 133.76 ($p < 0.05$). This shows that just like for crossover, cataphora sentences like (11c) and (11d) are only unacceptable on the cataphoric (crossover) reading.

These results are aggregated across the strong and weak configurations, following our analysis for Experiments 1 and 2. To split them apart, we fit an ordinal mixed effects model with just a fixed effect of strength and random effects as before. We see a significant effect of "strong" vs. "weak" in Figure 4b, which increases the odds of a high rating by 2.90 ($p < 0.05$). This is roughly 1.5x the size of the weak/strong crossover effect. That said, simply looking at Figure 4b shows that weak cataphora readings are still rated quite low in this setting: 71.4% of items are rated "Not really" or "Definitely no" (for strong cataphora, the figure is 88.7%).

[12]Moulton et al. (2018), by contrast, sets up their contexts such that the proper name, or its referent, is always assumed to be already known in the discourse by the time the cataphor is produced.
[13]The full set of items (24 sentences) is available on OSF at `https://osf.io/dh2qb`.
[14]`https://osf.io/w28vt`



(a) Effect of name/pronoun order on reading (b) "Strong" vs. "weak" proper name cataphora

Figure 4: Results for Experiment 3. We see a significant effect of crossover (columns 1 and 3 of (a)) and a significant effect of "strong" vs. "weak".

6.3. Discussion

Given the many reports in the literature that "weak" cataphora like (11d) are not just better but genuinely acceptable, the results from Experiment 3 are surprising. A natural response is that this must be the fault of the experiment design. One objection is that the distractor NP, in matrix subject position, may simply be too prominent and cause participants to not perceive the theoretically possible cataphor reading. While this is possible, we note that when the name precedes the pronoun (anaphora), this prominence does not prevent coconstrual of the pronoun with the target proper name (in fact, this coconstrual is preferred over coconstrual with the distractor). This was also not an issue for binding in Experiments 1 and 2. Nevertheless, to address this, we are planning a future experiment with no distractor NP, where the second paraphrase simply involves "someone else". Another objection is that we did not support these cataphor sentences with proper context. This is an interesting objection, especially as it is unclear whether a "proper context" includes the proper name in advance (as in e.g. Moulton et al., 2018). Do we expect a difference between (12a) and (12b), and if so, should that be accounted for by syntax/semantics or rather by discourse structure (pragmatics)? If cataphora are disfavoured without proper support, does that mean we should see no effect at all of our theory (e.g. Rule I) without a proper context, or just a smaller effect?

(12) a. Once upon a time there was a teacher called **Maria**$_i$. **Her**$_i$ students always emailed **Maria**$_i$ about the homework that she assigned. So she ...

b. Let me tell you a story: **Her**$_i$ students always emailed **Maria**$_i$ about the homework that she assigned. So she ...

With this in mind, there are two possible interpretations of these results, depending where we draw the line between grammatical and ungrammatical. If we take the results as showing that both "strong" and "weak" cataphor sentences are in general ungrammatical (on that coconstrual), then we need to revise Rule I or any derivative theory such as Marty (2017) to apply to cases of indirect binding (or at minimum, to possessives) as well as direct binding. This would predict that cataphora are ruled out if and only if some kind of binding between the two positions is possible – an interesting and testable prediction. We might also stipulate that Rule I with indirect binding results in a weaker violation than Rule I applied to direct (c-command) binding, to explain the statistically significant difference found in Experiment 4.

Alternatively, we might argue that we are primarily interested in is the difference between "strong" and "weak" cataphora, rather than their absolute ratings, which are difficult to interpret at the best of times. Perhaps there is a pragmatic principle militating against cataphora of any kind (supporting arguments about needing the right context) which decreases the ratings of both strong and weak crossover in an experiment like this. Pragmatic principles such as focus, topic and salience have indeed been shown to increase or decrease coconstrual availability for weak proper name cataphora (Gordon and Hendrick, 1997; Moulton et al., 2018; Gor and Syrett, 2019; Gor and Syrett to appear). In this case, the fact that "strong" c-command cataphora are significantly more deviant supports leaving Rule I as is, and suggests that direct (c-command) and indirect binding are very different creatures with different effects on coreference. Moreover, this experiment gives us a baseline allowing us to compare future experiments on other configurations of cataphora (such as *when*-clauses or conjunction) to these two standards, and in turn sharpen our definition of what must count as direct vs. indirect binding. For example, if sentences such as (13), which appear to be sharply ungrammatical, can be shown to pattern with c-command (strong) cataphora, this may suggest a revision of what counts as binding to include conjunction, as in dynamic semantic frameworks such as Chierchia (2020).

(13) *$\mathbf{He}_i$ walked in and $\mathbf{John}_i$ sat down.

7. Conclusion

We present a novel experimental paradigm to measure strong and weak crossover, and more broadly any phenomenon involving multiple possible coindexations or readings. We show that for binding phenomena, the side-by-side presentation of multiple unambiguous paraphrases is more effective than trying to fix these readings using a context paragraph. We find a significant difference in meaning availability between strong and weak *wh*-crossover in English, contra Kush (2013) who did not find a difference using acceptability judgements. Taken at face value, this supports theories which distinguish strong and weak crossover such as Koopman and Sportiche (1982), Safir (1984), Grodzinsky and Reinhart (1993) or Ruys (2000). Unified accounts such as Safir (2004) must rely more heavily on their argument that such differences do not need to be accounted for by the theory. However, we do not find such a difference for quantificational crossover, raising questions as to how these two kinds of crossover may differ. We further find that proper names display a significant crossover effect similar to *wh*-crossover, supporting Rule I (Grodzinsky and Reinhart, 1993) and its derivatives. This is significantly less severe in weak configurations, but these are still quite degraded under our experimental setup, leaving an open question as to how they should be handled theoretically. More broadly, we propose a robust, adaptable methodology to test disputed cases of crossover and cataphora, including relative clauses (Postal, 1993), variation in weak crossover across languages (Bresnan, 1998 and many others) and cataphora over conjunction.

References

Balhorn, M. (2004). The Rise of Epicene They. *Journal of English linguistics 32*(2), 79–104.

Barker, C. (2012). Quantificational binding does not require c-command. *Linguistic inquiry 43*(4), 614–633. Publisher: MIT press.

Bhatt, R. and S. Keine (2019). Secondary strong crossover in Hindi and the typology of movement. In *49th Annual Meeting of the North East Linguistic Society*.

Bianchi, V. (2009). A note on backward anaphora. *Rivista di Grammatica Generativa 34*, 3–34.

Bjorkman, B. M. (2017). Singular they and the syntactic representation of gender in English. *Glossa: a journal of general linguistics 2*(1).

Bresnan, J. (1998). Morphology competes with syntax: Explaining typological variation in weak crossover effects. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis, and D. Pesetsky (Eds.), *Is the best good enough?: optimality and competition in syntax*, pp. 59–92. Cambridge, Mass.: MIT Press : MITWPL.

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal 9*(2), 378–400.

Bryant, S. (2022). *Lost in Space: Pronoun Choice in English Locative Prepositional Phrases*. Ph.D., Harvard University, United States – Massachusetts.

Chierchia, G. (2020). Origins of weak crossover: when dynamic semantics meets event semantics. *Natural Language Semantics 28*(1), 23–76.

Chomsky, N. (1976). Conditions on Rules of Grammar. *Linguistic Analysis 2*(4), 303–351.

Christensen, R. H. B. (2019). ordinal—regression models for ordinal data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Conrod, K. (2019). *Pronouns raising and emerging*. PhD Thesis, University of Washington.

Falco, M. (2007). Weak crossover, specificity and LF chains. In *Coreference, Modality, and Focus: Studies on the Syntax-semantics Interface*, pp. 19–44. John Benjamins Publishing.

Fanselow, G., R. Kliegl, and M. Schlesewsky (2005). Syntactic variation in German wh-questions: Empirical investigations of weak crossover violations and long wh-movement. *Linguistic Variation Yearbook 5*(1), 37–63.

Felser, C. and J.-D. Drummer (2017). Sensitivity to Crossover Constraints During Native and Non-native Pronoun Resolution. *Journal of Psycholinguistic Research 46*(3), 771–789.

Gor, V. and K. Syrett. Principle C is not the final word: Experimental evidence for the influence of conceptual plausibility and pronominal position on coconstrual relations. To appear.

Gor, V. and K. Syrett (2019). Beyond Principle C: (Not)-at-issueness and plausibility influence acceptability of coconstrual. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, Volume 54, pp. 109–123. Chicago Linguistic Society.

Gordon, P. C. and R. Hendrick (1997). Intuitive knowledge of linguistic co-reference. *Cognition 62*(3), 325–370.

Grodzinsky, Y. and T. Reinhart (1993). The Innateness of Binding and Coreference. *Linguistic Inquiry 24*(1), 69–101.

Heim, I. (1998). Anaphora and semantic interpretation: A reinterpretation of reinhart's approach. *The interpretive tract 25*, 205–246.

Heim, I. R. (2007). Forks in the Road to Rule I. In M. Abdurrahman, A. Schardl, and M. Walkow (Eds.), *NELS 38 : Proceedings of the Thirty-Eigth Annual meeting of the North East Linguistic Society*.

Higginbotham, J. (1980). Pronouns and Bound Variables. *Linguistic Inquiry 11*(4), 679–708.

Koopman, H. and D. Sportiche (1982). Variables and the Bijection Principle. *Linguistic review 2*(2), 139–160.

Kush, D., J. Lidz, and C. Phillips (2017). Looking forwards and backwards: The real-time processing of Strong and Weak Crossover. *Glossa: a journal of general linguistics 2*(1), 70.

Kush, D. W. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing*. PhD Thesis, University of Maryland, College Park.

Lasnik, H. and T. Stowell (1991). Weakest crossover. *Linguistic inquiry 22*(4), 687–720.

Lyu, J. (2017). Weak Crossover in Chinese—now you see it, now you don't. In *Proceedings of the 17th Texas Linguistic Society*, pp. 54–64.

Marty, P. P. (2017). *Implicatures in the DP domain*. Thesis, Massachusetts Institute of Technology.

Moulton, K., Q. Chan, T. Cheng, C.-h. Han, K.-m. Kim, and S. Nickel-Thompson (2018). Focus on Cataphora: Experiments in Context. *Linguistic Inquiry 49*(1), 151–168.

Pesetsky, D. (1987). Wh-in-situ: Movement and unselective binding. In E. Reuland and A. t. Meulen (Eds.), *The representation of (in) definiteness*, Volume 98, pp. 98–129. Cambridge, MA: MIT Press.

Postal, P. (1971). *Crossover phenomena*. Transatlantic series in linguistics. Holt, Rinehart and Winston.

Postal, P. M. (1993). Remarks on weak crossover effects. *Linguistic Inquiry 24*(3), 539–556.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reinhart, T. (1983). *Anaphora and semantic interpretation*. Croom Helm linguistics series. London: Croom Helm.

Reinhart, T. and E. Reuland (1993). Reflexivity. *Linguistic Inquiry 24*(4), 657–720.

Rooryck, J. and G. Vanden Wyngaerd (2011). *Dissolving Binding Theory*. Oxford University Press.

Ruys, E. G. (2000). Weak Crossover as a Scope Phenomenon. *Linguistic inquiry 31*(3), 513–539.

Ruys, E. G. (2004). A Note on Weakest Crossover. *Linguistic Inquiry 35*(1), 124–140.

Safir, K. (1984). Multiple variable binding. *Linguistic Inquiry 15*, 603–638.

Safir, K. J. (2004). *The Syntax of (In)dependence*. The MIT Press.

Shan, C.-C. and C. Barker (2006). Explaining Crossover and Superiority as Left-to-right Evaluation. *Linguistics and Philosophy 29*(1), 91–134.

Wasow, T. (1972). *Anaphoric relations in English*. PhD Thesis, Massachusetts Institute of Technology Cambridge.

A. Experiments 4a and 4b: Comparing response types

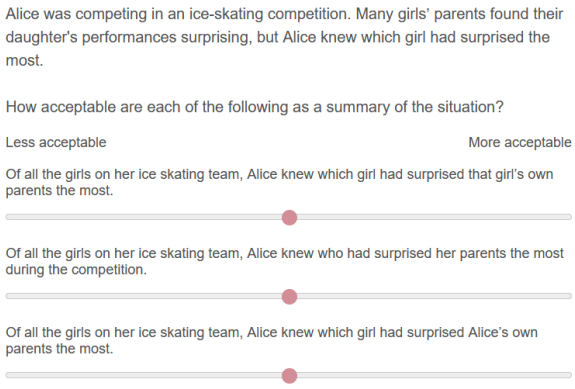
We directly compare a classic acceptability task (Experiment 4a) to our "meaning availability" design, which is inspired by the unambiguous paraphrases commonly used in syntax/semantics teaching (Experiment 4b). We study the effect of each design on the ratings for the bound reading of *wh*...[gap]...pronoun sentences (the standard binding configuration). Since crossover is a case where binding fails, to study crossover it is crucial that an experiment be able to capture binding success and thus (perhaps) differentiate crossover. For each design, we use the same experimental items and the same 2x2 cross, varying only the response type. We cross two structural configurations of binding which parallel secondary strong and weak crossover with two potential readings[15]. An even split of masculine and feminine pronouns is used.

We recruited 100 self-reported native English speakers on Prolific for Experiment 4a and another 100 for Experiment 4b. Participants saw 8 target items and 16 fillers in random order (Experiment 4a) and 4 target items and 8 fillers in random order (Experiment 4b). We preregistered this study on OSF[16] and carried out the analysis exactly as planned, with the exception

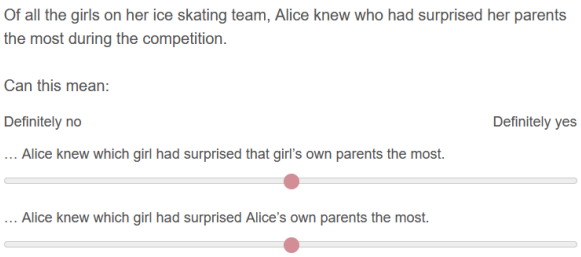
[15]The original experiments also included items using the *wh*...pronoun...[gap] order (corresponding to crossover), but we will not analyse those here.

[16]`https://osf.io/g4z52`

of excluding participants due to attention checks, which turned out to be impractically strict.



(a) Screenshot of Experiment 4a.



(b) Screenshot of Experiment 4b.

Figure 5: "Weak" binding sentence in the sentence acceptability design (Experiment 4a) and in the meaning availability design (Experiment 4b). In (a), the context describes the bound reading. The middle sentence is the target sentence, the top sentence is an unambiguous paraphrase of the bound reading, and the bottom sentence is an unambiguous paraphrase of the "distractor NP" reading. We expect the top two sentences to receive high ratings, with the target sentence receiving the bound reading. In (b), the first paraphrase represents the bound reading; the second represents the "distractor NP" reading. We expect both to get high ratings.

A.1. Experiment 4a: Sentence acceptability design

This experiment follows a traditional experimental semantics design by presenting participants with a short context paragraph describing a situation where exactly one of the readings (*wh*-binding or distractor NP) holds. Participants are then given the target sentence, as well as two control sentences, and asked to rate on a sliding scale from 0 (less acceptable) to 100 (more acceptable) how acceptable each sentence is as a description of the situation. The control sentences are identical to the unambiguous paraphrases used in Experiment 4b, and serve as checks that the participant is interpreting the context paragraph in the way we expect. A screenshot of an experiment item is shown in Figure 5a[17].

A.2. Experiment 4b: Meaning availability design

We compare the sentence acceptability design with the "meaning availability" design described in Experiment 1. The only difference from Experiment 1 is the use of a slider bar from 0 (definitely no) to 100 (definitely yes) instead of Likert scale (discussed in the Results section). A screenshot of an experiment item is shown in Figure 5b.

A.3. Results

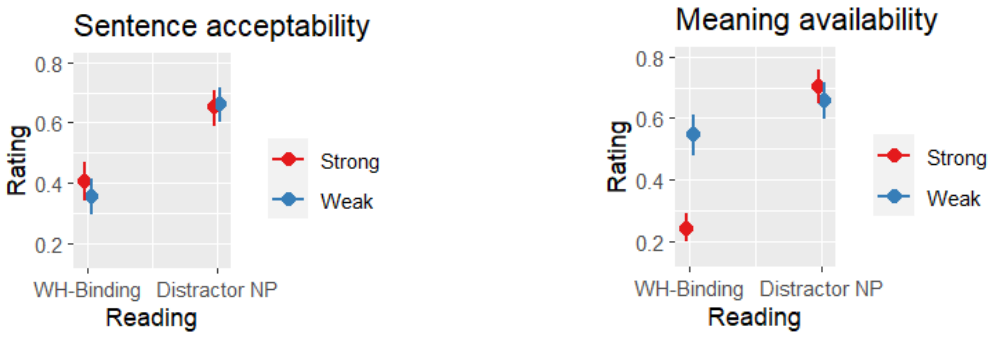
For each experiment, we fit a mixed effects beta regression with an interaction between strong vs. weak and *wh*-binding vs. distractor NP reading, as well a random effect for participant ID using the `glmmTMB` (Brooks et al., 2017) package. We use a beta regression since we have

[17]The paraphrases used in this set of items are somewhat stilted; we improved on this and the general target item quality substantially in subsequent experiments (Experiments 1, 2 and 3 above).

many ratings at the ends of the scale (0 or 100) and so a linear regression results in a poor fit. For the sentence acceptability design, we focus here only on the ratings of the target sentences, setting aside the data on the two unambiguous paraphrases, which showed that the contexts were generally interpreted as expected.

In the traditional sentence acceptability design of Experiment 4a, we found a significant effect of reading/context ($p < 0.05$), with the target sentences in the Distractor NP context being rated substantially higher than the same sentences in the *wh*-binding context for both "strong" and "weak" cases, shown in Figure 6a. This is unexpected given that *wh*-binding reading of these sentences should be perfectly possible according to the theory, and so should be available in the case where the context describes the bound reading. We found no significant effect of "strong" vs. "weak" ($p = 0.25$, as expected) and no interaction ($p = 0.32$).

In the meaning availability design of Experiment 4b, we also found a significant effect of reading ($p < 0.05$), as well as a significant effect of "strong" vs. "weak" and an interaction between reading and "strong" vs. "weak" (both $p < 0.05$), shown in Figure 6b. Importantly, however, unlike the sentence acceptability design, bound readings for the "weak" binding sentences are rated high, similar to the ratings for the Distractor NP reading, showing that participants do allow the bound reading for at least these binding sentences.



(a) Experiment 4a: Sentence acceptability

(b) Experiment 4b: Meaning availability

Figure 6: Ratings for binding sentences across response types (Experiments 4a and 4b).

A.4. Discussion

Experiments 4a-b show that the sentence acceptability design, at least in the form used here, does not yield results consistent with prior observations of binding sentences. Since the availability of binding of this sort is not a controversial phenomenon, we take this to be an issue with the experimental design. Moreover, since crossover is a case where binding fails, to study crossover it is crucial that an experiment be able to capture binding success and thus (perhaps) differentiate crossover. While the meaning availability design also shows some signatures of low ratings for "strong" binding, this seems to be an artifact of specific sentences used: in one of the three scenarios used here, the strong bound reading is simply much less plausible than the distractor NP reading. This same issue is not present for the "weak" sentences (or in Experiments 1 and 2 above), and we see correspondingly high ratings in the meaning availability design. In subsequent experiments (Experiments 1, 2 and 3), we significantly improved the quality of the target sentences and ensured that both readings were similarly plausible.

We believe that the main flaw behind the sentence acceptability design is that reading a context paragraph describing one coconstrual does not appear to (sufficiently) dispose participants towards giving the target sentence that particular reading. Participants appear to have interpreted the pronoun as coconstrued with the distractor NP regardless of the context, and thus rated the sentence as a poor description of the context describing coconstrual with the *wh*-word. Since no obvious way of rephrasing or restructuring this design appears to avoid this issue, we settled on the meaning availability design for the remainder of our experiments.

In addition, we observed that the sliding scale from 0-100 caused participants to overwhelmingly use the endpoints 0 and 100, resulting in a lack of granularity of responses. Subsequent experiments thus used a 5-point Likert scale and drew participant's attention to the option of using the intermediate points during the training period.

B. Effect of pronoun gender on binding

As part of Experiment 1, we investigate whether pronoun gender has any effect on binding. We fit a mixed effects ordinal regression on just the binding sentences of Experiment 1, with an interaction between reading and pronoun gender and the same random effects as in Experiment 1, shown in Figure 7. We find significant effects of masculine and feminine pronouns which decrease the odds of a high rating for the bound reading by a factor of 0.54 and 0.61 respectively ($p < 0.05$, $SE = 1.25$ for both), compared to the ratings for singular *they*. Moreover, each has an interaction with the reading, with each substantially increasing the odds of a high rating for the distractor reading compared to the bound reading for that pronoun (a factor of 4.96 for *he* and 2.07 for *she*; $SE = 1.39$ and 1.37 respectively, $p < 0.05$). By contrast, singular *they* (treated as the base case in the model) has no significant effect of reading, meaning that its readings are rated similarly to each other. Quantitatively, singular *they* has the highest ratings for bound readings given otherwise identical sentences, as well as the best balance between bound and distractor NP ratings. This supports the use of singular *they* for experiments detecting binding.

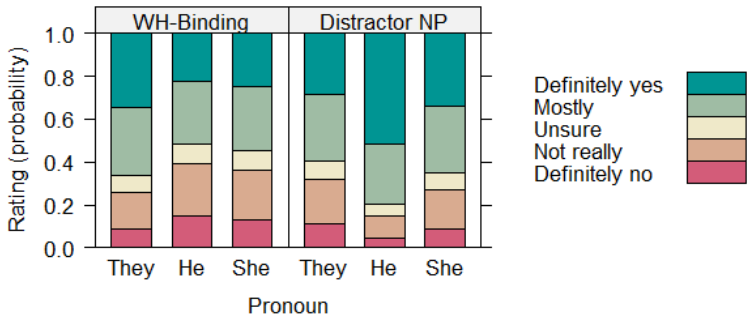


Figure 7: Effect of pronoun choice on reading

We further fit a mixed effect ordinal regression on just the singular *they* binding sentences with participant age group as a fixed effect (interacting with reading and pronoun gender) to investigate whether perhaps only younger participants view singular *they* as an appropriate pronoun. We find that there are no significant interactions of the age group with any of the other factors, supporting Balhorn (2004) and others in suggesting that this epicene use of singular *they* is not a new phenomenon.