Seeing is believing: testing an explicit linking assumption for visual world eye-tracking in psycholinguistics

Anonymous CogSci submission

Abstract

[jd: write]

Keywords: psycholinguistics; experimental pragmatics; scalar implicature; linking functions; visual world; eye-tracking

Introduction

Experimental investigation is integral to scientific inquiry in cognitive science, and experimental data has informed cognitive theory-building for centuries. A fundamental ingredient in using empirical data to put theories to the test is the assumption made by researchers about the mapping between theoretical notions and empirical measurements: the *linking assumption*. Indeed, empirical measurements are only informative if the linking assumption is (sufficiently) clear and justified. We argue that both clarity and justification are often lacking for linking assumptions made in *visual world eye-tracking*, a widely used experimental method in psycholinguistic research. We highlight the role that visual world eye-tracking has played in the burgeoning field of experimental pragmatics, which suffers particularly acutely from a lack of clear and justified linking assumptions. We then test a (usually implicitly) made linking assumption for referential tasks, which we term the *Referential Belief* linking assumption: that the proportion of looks to a referent in a time window reflects participants' degree of belief that the referent is the intended target in that time window. To do so, we compare eye movement data against explicit beliefs collected in an incremental decision task. We make use of a previously collected eye movement dataset on scalar implicature processing (Exp. 3 of ?, ?). In Exp. 1 we collect explicit beliefs to test against ? (?)'s eye movement data. In Exp. 2, we replicate ? (?) in a web-based eye-tracking paradigm using `WebGazer.js`.

Linking assumptions for visual world eye-tracking

Visual world eye-tracking (VWE) is a widely used measure in psycholinguistics, fruitfully driving advances in our understanding of phonetic, lexical, syntactic, prosodic, semantic, and pragmatic processing (?, ?, ?, ?, ?, ?, ?, ?). In standard VWE tasks, participants view displays of objects while listening to spoken sentences while their eye movements are monitored (see Figure **??** for an example). The popularity of VWE stems from one of its very desirable features: eye movements can be interpreted as an indicator of attention that is closely time-locked to the linguistic signal and not subject to voluntary control **[jd: get ref for this]**. Language can guide eye movements to a region of interest in a display within 200 ms (?, ?). By sampling an x/y coordinate every few milliseconds, researchers thus obtain a very temporally fine-grained record of participants' language-directed attention over the course of an unfolding utterance. This property has been particularly useful in resolving questions regarding the time-course of online language processing, which typically cannot be addressed using offline measures like forced choice, truth-value judgments, or even more coarse-grained temporal measures like response times from button presses. Notable VWE findings that could not have been obtained with more coarse-grained measures include the diverse insights that visual context is rapidly integrated into syntactic structure assignment (?, ?, ?), that words are processed incrementally and listeners maintain uncertainty about past input (?, ?, ?), and that listeners anticipate upcoming linguistic material based on selectional restrictions and rapid pragmatic reasoning (?, ?, ?).

These notable successes notwithstanding, there is an elephant in the room: we still have a relatively poor understanding of how to link observed eye movements to the underlying mental processes that generate them (?, ?, ?, ?, ?). The problem of interpretability is compounded by the fact that VWE is used in vastly different tasks. Consider just the difference between active referential tasks (in which participants' goal is to identify and select the speaker's intended referent) and passive predictive tasks (in which participants simply watch a display while listening to language, without an overt task or goal). In the former case, arguments are made that eye movements reflect listeners' active search for or belief in the referent, depending on whether visual information about the display has been integrated. In the latter case, the evidence indicates that eye movements reflect an automatic predictive process (?, ?). But what, exactly, does this mean? What is the generative process by which, e.g., a notion like "prediction" ultimately results in an eye movement to a region at a particular point in time? Few make clear assumptions about the generative process underlying eye movements (but for a principled early example, see ?, ?).

Here, we test an explicit linking assumption for referential tasks, which we term the Referential Belief linking assumption: that the proportion of looks to a referent in a time window reflects participants' degree of belief that the referent is the intended target in that time window. This linking assumption was tested and found not supported in previous work (?, ?). In a re-analysis of an experimental semantics dataset on the processing of relative and absolute gradable adjectives (?, ?), ? (?) found that explicit beliefs collected in an incremental decision task (similar to gating tasks, ?, ?) did not correlate with eye movements, with the exception of one condition. They argued that the reason for the lack of support may have been participants' close to zero expectation for observing the linguistic stimuli used in the original experiment, where expectation was linked to production probabilities elicited in a separate free production task.

Of course, it is also possible that the incremental decision task simply does not capture the beliefs that inform eye movements. We believe this is unlikely, given recent successes in using such tasks to elicit contrastive inferences (?, ?, ?), but acknowledge the possibility. The previous failure to find support for the Referential Belief link, compounded by the concern regarding the validity of the incremental decision task, motivates the current study: to test the Referential Belief link on a different dataset. For this purpose, we replicate Exp. 3 of ? (?) (henceforth, "SB2020") in an incremental decision task rather than an eye-tracking task (Exp. 1) and ask how well the explicit beliefs predict the eye movement data. **[jd: webgazer?]**

Test bed: Sun & Breheny (2020)

We replicate Exp. 3 of SB2020, which addressed a now classic question in experimental pragmatics: is the processing of scalar inferences delayed relative to the processing of literal information (?, ?, ?, ?, ?, ?, ?, ?)? In particular, SB2020 were interested in assessing the possible effect of two factors on the speed with which determiners are processed: first, preexisting low-level associations between determiners and set sizes (i.e., a preference for *all* to be associated with bigger set sizes and for *some* to not show a clear preference, as established in a norming study); and second, whether the application of the determiner to a set of objects can be verified without checking a separate set of objects. For instance, the partial utterance *Click on the boy that has three*, heard in the left display of Figure **??**, requires only verifying that there is a boy with three objects. In contrast, replacing *three* with either *all* or *some* requires additionally verifying that there are no other apples in the display or – if *some* is pragmatically enriched to *not all* – that there is at least one other orange in the display, respectively. That is, *all* and *some* should require additional processing time before settling on the target, but if *some* is immediately enriched to *some, but not all*, verification looks to what SB2020 call the 'residue set' (the remaining objects in the center of the screen) should increase immediately after observing the determiner.

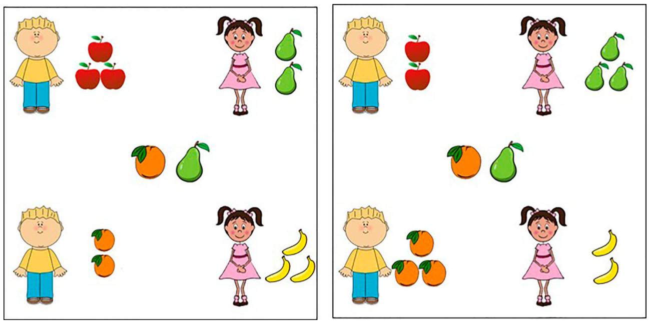

Figure 1: Example displays from Exp. 3 of SB2020. The left image (big *all*/ small *some*) was paired with *Click on the boy that has all/three of Susan's apples* or *Click on the girl that has some/two of Susan's pears*. The right image (small *all*/ big *some*) was paired with *Click on the boy that has all/two of Susan's apples* or *Click on the girl that has some/three of Susan's pears*.

Their predictions were borne out (see target advantage scores in Figure **??**). In the determiner window (200ms after determiner onset to 200ms after name onset) and the name window (200ms after name onset to 200ms after noun onset), number terms led to more (and a faster increase in) target looks than did *all* and *some*, suggesting that the need for verification of the residue set is a source of relatively fewer target looks for *all* and *some*. Moreover, while there was no effect of set size in the number or *some* condition, big *all* led to more target looks than small *all*. Finally, and crucially for their purposes, they found that looks to the residue set in the determiner window increased for both *all* and *some* (but not numbers) equally and at the same time.

Jointly, their results support what they call the fast-pragmatic account: the view that the computation of scalar inferences itself is not delayed compared to literal processing, and that previously reported apparent slowdowns in processing of *some* are instead likely due to joint effects of verification time and low-level set size associations for *all* which facilitate the processing of big *all* compared to small *some*.

For the purpose of testing the Referential Belief linking assumption, this study has both appealing features as well as one glaring problem. The appealing features include the simple 2x3 design, a limited and clearly defined set of referents in each display, and the clarity of the referential task. The glaring problem, which disqualifies the Referential Belief link as a full linking theory from the outset, are the systematic looks to the residue set: the Referential Belief link is only defined for looks to possible referents. There is no plausible argument to be made that participants look to the residue set because they believe it may be the intended target. Thus, we have already identified one way in which, if otherwise supported by the data, the Referential Belief link will have to be extended. We return to this point in the General Discussion.

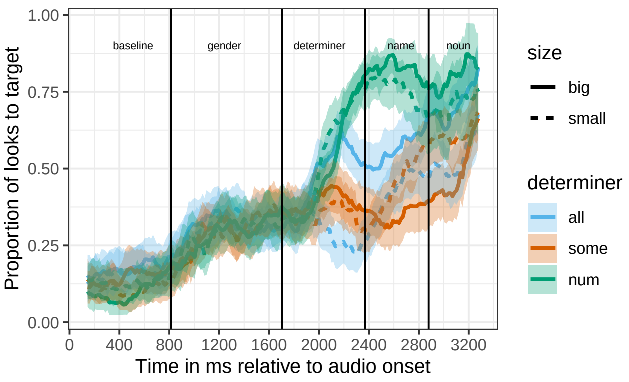

Figure 2: Proportion of target looks (out of target, competitor, and residue looks) from instruction onset. Transparent ribbons indicate 95% bootstrapped confidence intervals. Black vertical lines indicate onsets of analysis windows of interest (window labels at top of graphs). **Top:** Exp. 3 of SB2020. **Bottom:** Our Exp. 2.

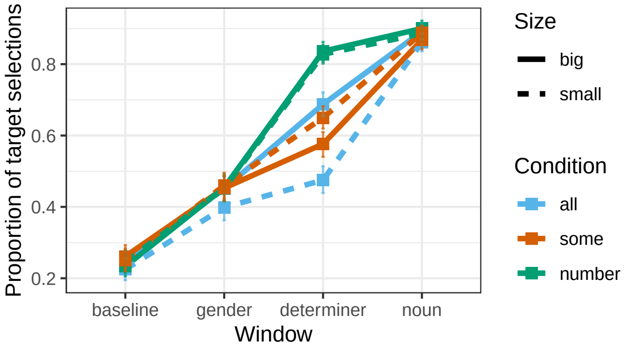

Figure 3: Proportion of target selections in Exp. 1 by quantifier and set size. Error bars indicate 95% bootstrapped confidence intervals.

Exp. 1: replicating Sun & Breheny (2020) using an incremental decision task

In order to measure participants' beliefs about the intended referent at points in the utterance that would allow us to compare explicit beliefs to proportions of looks in SB2020, participants engaged in an incremental decision task (?, ?, ?, ?, ?).

Methods

Participants. We recruited 120 participants on Mechanical Turk, excluding participants with $<$ 95% accuracy (n=29) and trials on which participants selected the wrong referent in the last window (665 trials). All participants were self-reported native English speakers.[1]

Materials and procedure. We measured participants' beliefs about the intended referent for each display shown to participants by SB2020 (see Figure **??** for examples). Participants were told that they were playing a guessing game, and whenever they made a guess, more words would appear. The critical sentences of the form "Click on the GENDER who has DETERMINER of NAME's NOUN" were revealed incrementally. GENDER was one of *boy/girl*, DETERMINER was one of *some/all/two/three*, NAME was one of *Susan/Amy/Michael*, and NOUN was one of *apples/bananas/erasers/scissors/knives/rulers/forks/plates/ spoons/pencils/pears/oranges*. Participants clicked on the presumed target after (a) "Click on the" (baseline window), (b) "GENDER that has" (gender window), (c) "DETERMINER of NAME's" (determiner window), and (d) "NOUN" (noun window). After each click, the next word(s) or display was shown. After 6 practice trials, each participant saw 48 experimental trials, of which 12 were filler trials with the number terms *one* and *four*. The 36 critical trials implemented SB2020's 2 (big vs. small target set) by 3 (*all, some*, number) design.

Results and discussion

Figure **??** shows proportions of target selections out of all selections in each time window and condition.

Data analysis. SB2020 fit separate linear regression models to target advantage scores in time windows of interest. We instead fit logistic mixed effects models predicting target selections. This choice was motivated by logistic regression being the more principled approach to modeling categorical data. It also avoids problems like pre-aggregating data and adding smoothing terms to avoid division by zero or discarding mathematically problematic data points.

SB2020 fit separate models to the baseline, gender, determiner, name, and noun windows (though only the determiner and name windows were of interest). We collapsed the name window into the determiner window because the name does not add disambiguating information. The models predicted target over competitor choices from fixed effects of quantifier (reference level: "number"), centered size (higher value: "big"), by-item and by-subject random intercepts, and random by-subject slopes for condition and size. No effects reached significance in the baseline, gender, and noun window, as expected.[2] In the determiner window, the window of interest, there were main effects of condition, such that target selections were less likely in both the *some* (β=-2.90, *SE*=0.36, *p* <.0001) and *all* (β=-2.92, *SE*=0.36, *p* <.0001) conditions, compared to the number condition. There was no main effect of size, consistent with the visual result that target selections in the number condition are not modulated

[1] Procedure, materials, analyses and exclusions were pre-registered at `https://rb.gy/ecoggz` (Exp. 1) and `https://rb.gy/ednwye` (Exp. 2). The collected sample size for Exp. 2 (183 participants) was larger than the originally pre-registered sample size (102 participants) because 40% of the initially tested 102 participants had a technical issue and weren't able to see the whole display.

[2] In fact, fitting models to the noun window was impossible because participants, with very few exceptions, always chose the target. That is, there was no variance to speak of that a model could be estimated to explain.

by target set size (β=-0.09, *SE*=0.26, *p* <.73). However, we did observe interactions between quantifier and size, such that small sets resulted in more target selections for *some* (β=0.59, *SE*=0.28, *p* <.05) but fewer target selections for *all* (β=-1.27, *SE*=0.29, *p* <.0001), compared to number terms.

Comparison with SB2020: replication analysis. These results constitute a near-perfect replication of SB2020. Most of their effects reported in the determiner window replicated, with two exceptions (see overview in Table **??**): we did not observe a main effect of set size, while they did; and we observed an interaction of size and determiner such that small *some* led to greater target selections than big *some*. These differences are not as big as they may seem: in our dataset, the lack of set size main effect can be explained by the interactions with determiner: while size makes no difference at all for number (the determiner predictor reference level), it has the opposite effect for *all* compared to *some*. A similar tendency can be observed in SB2020's results when taking into account the joint determiner and name windows. In fact, SB2020 report the absence of a main effect for size in the name window, and instead an interaction between size and determiner. While they do not report the same post hoc analyses, visual inspection of Figure **??** suggests that the interaction in the name window is indeed the result of set size having the opposite effect for *all* compared to *some*. Thus, when taking into account their results from both the determiner and name window, which we collapsed into one, the results are qualitatively identical. The different results reported by SB2020 in the two time windows are presumably the result of certain information taking longer to be integrated, something which the incremental decision task by its offline nature cannot capture.

Comparison with SB2020: linking function analysis. Figure **??** shows the correlation between proportion of selections in Exp. 1 and proportions of looks in SB2020. The overall correlation was very high ($r(862) = .87$, $p < .0001$), suggesting preliminary support for the Referential Belief link. To investigate whether the predictive power of explicit beliefs was modulated by additional factors, we conducted a linear regression analysis predicting proportion of looks from fixed effects of proportion of selections (mean-centered) and its 2-way interactions with time window (dummy-coded, reference level: 'determiner') and region of interest (dummy-coded, reference level: 'target'). There was a large and significant effect of selection proportion (β=0.79, *SE*=0.02, *t*=32.31, *p* <.0001), such that stronger beliefs resulted in more looks. This effect was modulated by time window: selection proportion was a less good predictor of looks in the baseline window (β=-0.11, *SE*=0.05, *t*=-2.25, *p* <.05), a better predictor in the noun window (β=0.13, *SE*=0.03, *t*=5.18, *p* <.0001), and there was no evidence for a difference in the gender window (β=-0.02, *SE*=0.04, *t*=-0.61, *p* <.55), compared to the determiner window. The effect was also modulated by the region of interest: selection proportion was a less good predictor of competitor looks (β=-0.46, *SE*=0.03, *t*=-13.16, *p* <.0001) and distractor looks (β=-0.33, *SE*=0.03, *t*=-9.89, *p* <.0001) than of target looks (see Figure **??**).

These results suggest strong support for the Referential Belief link: across the board, subjective referential beliefs quantified as proportion of selections were a good predictor of proportions of looks. This support is tempered by some of the auxiliary findings in interesting ways. For instance, the fact that selections were a better predictor of target than of competitor and distractor looks suggests that looks to these regions may be driven more strongly by other cognitive processes, e.g., verification or noise processes. Similarly, the apparently gradient increase in predictive power of selections from baseline through noun window may reflect the important role that uncertainty plays in guiding eye movements.

Finally, a model that also included interactions with experimental conditions of interest—determiner and set size—did not reveal modulation of the selection proportion effect by experimental condition, in contrast to ? (?)'s results on the re-analyzed adjective processing dataset from ? (?). Those authors hypothesized that the difference in predictive power by condition may be driven by participants' varying expectation for the utterances observed, such that the less surprising the utterance was, the better explicit beliefs predicted proportions of looks in the experimental window of interest (analogous to our determiner+name window). If this reasoning is correct, it suggests that there were no differences in production expectations in SB2020 across conditions, and that the expectation for the actually observed utterances was high across the board, explaining the very strong correlations we observed between proportion of selections and proportion of looks. This hypothesis requires further empirical investigation via a production study eliciting descriptions of SB2020's referential displays.

Finally, a crucial issue the Referential Belief link cannot account for is the fact that participants in SB2020's experiment looked towards the central residue set very systematically as a verification strategy. In Figs. **??** and **??**, proportions of looks to a region are computed out of target, competitor, distractor, and residue looks, which explains why selection proportions consistently under-predict proportions of looks—there are additional looks not captured in the figures.

Exp. 2: replicating Sun & Breheny (2020) using web-based eye-tracking

Having replicated SB2020's main findings in an incremental decision task, we investigate whether the results are also replicable in a web-based paradigm.

Methods

Participants. We recruited 183 participants on Prolific and excluded 21 because accuracy was $< 95\%$. We also excluded trials on which participants selected the wrong referent (303 trials). All participants were native English speakers.

Materials and procedure. Exp. 2 was identical to SB2020's Exp. 3 with one difference: we collected eye move-

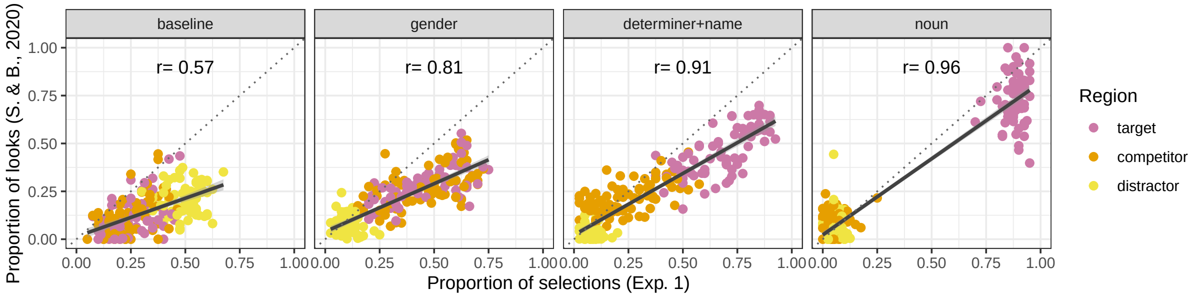

Figure 4: Proportions of looks in SB2020 against proportions of selections in Exp. 1. Facets indicate time windows.

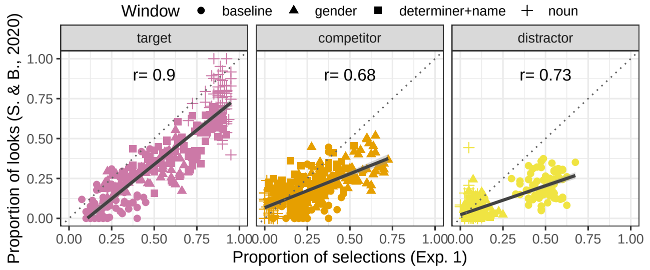

Figure 5: Proportions of looks in SB2020 against proportions of selections in Exp. 1. Facets indicate region of interest.

ments with webcam eyetracking using the WebGazer.js (?, ?) library. Participants were presented with the same experimental displays as the original experiment (Figure **??**). One second after the display onset, participants heard the auditory instruction of the form ”Click on the GENDER who has DETERMINER of NAME’s NOUN”. Their task was to select the correct image according to the instruction. After each click, participants moved to the next trial. At the beginning of the experiment, there were 6 practice trials which familiarized participants with the characters and the task. After the practice trials, participants saw 36 critical trials and 12 fillers in randomized order. The critical trials had a two (target size: big, small) by three (determiner: all, some, number) design. On each trial, eye movements were recorded from the onset of the display until the participant made a selection.

Results

We conducted separate mixed-effects logistic regressions in the 5 time windows of interest (baseline, gender, determiner, name, noun), predicting target over competitor looks from fixed effects of time (scaled and mean-centered), determiner (reference level: ‘number’), centered size (higher value: ‘big’), and their interactions. The random effects structure included by-item and by-participant random intercepts and slopes for all the fixed effects.

The crucial time windows of interest are the determiner and name window. In the determiner window, there was a significant intercept effect, suggesting an overall determiner-driven preference for target over competitor looks (β=0.70, *SE*=0.13, *p* <.0001). Of the other effects reported by SB2020, only the relatively fewer target looks for *some* relative to number replicated (β=-0.53, *SE*=0.17, *p* <.01).

Table 1: Overview of critical effects in determiner (det.) and name windows in SB2020, our Exp. 1 and our Exp. 2. Rows correspond to model predictors. “+”: significantly positive effect; “–”: significantly negative effect; “·”: no evidence of an effect.

	SB2020		Exp. 1	Exp. 2	
Predictor	det.	name	det.	det.	name
all.v.num	–	–	–	·	–
some.v.num	–	–	–	–	–
size	+	·	·	·	·
all.v.num:size	+	+	+	+	·
some.v.num:size	·	·	–	·	–
time	+	?	NA	·	+

In the name window, there was again a significant (and larger) intercept effect (β=1.46, *SE*=0.14, *p* <.0001), suggesting the target bias increased further in this window. In addition, most of the effects reported by SB2020 replicated: there was a main effect of time (β=0.25, *SE*=0.05, *p* <.0001), such that target looks increased over time. There were also main effects of determiner, such that there were fewer target looks in both the *all* condition (β=-0.81, *SE*=0.22, *p* <.001) and the *some* condition (β=-1.08, *SE*=0.24, *p* <.0001). While there was no main effect of size, there was an interaction of size with the *some* contrast, such that there were fewer target looks when the target set was big (β=-0.60, *SE*=0.18, *p* <.001). There was also a trending interaction of size with the *all* contrast in the expected direction, which did not reach significance (β=-0.30, *SE*=0.18, *p* <.11).

Comparison with SB2020: replication analysis. **[jd: fill in]**

General discussion

- main point: in contrast to previous re-analysis of leffel et al, clear support for the Referential Belief link. reasons?
  - most likely: participants in SB expected the observed language (see preliminary production results from Qing et al, but also Kreiss and Degen 2020 results, plus tons of other literature supporting tight link between production

  and comprehension in general, and importance of clearly nameable items (ie, clear production expectations) for VW eye-tracking in particular (go back to mike's papers from the oughts)
  - possibly more power in SB than Leffel

- pick the residue set issue back up – possible extensions of Referential Belief link?

- linking function issue is a problem not just with eye-tracking but with experimental measures in xprag in general (see eg the tvjt literature, though there have been several recent attempts to be explicit abt link, especially with the advent of probabilisitc models that make specifying a clear link easier): (?, ?, ?, ?, ?, ?, ?, ?)

- methodological points:
  - further validation of incermental decision task as useful measure of comprehension
  - The additional replication suggests that web-based eye-tracking using the WebGazer.js library is a) feasible; and b) a faster, cheaper way to collect eye movement data despite the added noise.

  The latency of these effects indicate we are some way from being able to conduct the gamut of eye-tracking experiments that require fine-grained, time-sensitive data remotely. Designs requiring multiple regions of interest (e.g., a four-image display as in Exp. 3) or subtle discrimination between stimuli (as in a reading task) may fare particularly badly given the added noise, especially if effects are subtle. However, our results do show signal amongst this noise, and web-based eye-tracking may be particularly fruitful for studies that do not require precise timing information about the onset of an effect.

  Much more investigation is needed into whether different methodological and implementation decisions might yield more reliable results. Two-image displays that allow for greater distance between the regions of interest may reduce noise, for example, but validation through replication of existing work is needed. More rigorous and frequent accuracy checks throughout the task may also improve the clarity of the resulting data: in Exp. 3, participants completed an initial calibration and accuracy check, and were able to proceed if they scored above 50%. Given the novelty of the medium, we do not know whether increasing this threshold might increase clarity, or whether additional accuracy checks might better ensure that participants' eye movements were still being tracked as reliably as possible.

- theoretical connection: more fine-grained RSA story that can capture the pref for small over big "some" despite "some" being bad for both; goes beyond (against?) the simple (SB and degen tanenhaus 2015/2016) production expectation story that only takes into account the probability of "some" being used for the bigger vs the smaller set. if we apply RSA, where what matters to interpretation is not just the relative probability of "some" being used for one vs the other set size, but it also matters how likely "all" and other alternatives are for the set sizes under consideration, then I think this result is expected as long as we assume that "all" is generally highly dispreferred for the small set and highly preferred for the big set. see example numbers below (assuming a uniform prior), where S is the pragmatic speaker and L is the pragmatic listener:

[jd: daisy's para:]

The latency of these effects indicate we are some way from being able to conduct the gamut of eye-tracking experiments that require fine-grained, time-sensitive data remotely. Designs requiring multiple regions of interest (e.g., a four-image display as in Exp. 3) or subtle discrimination between stimuli (as in a reading task) may fare particularly badly given the added noise, especially if effects are subtle. However, our results do show signal amongst this noise, and web-based eye-tracking may be particularly fruitful for studies that do not require precise timing information about the onset of an effect.

Much more investigation is needed into whether different methodological and implementation decisions might yield more reliable results. Two-image displays that allow for greater distance between the regions of interest may reduce noise, for example, but validation through replication of existing work is needed. More rigorous and frequent accuracy checks throughout the task may also improve the clarity of the resulting data: in Exp. 3, participants completed an initial calibration and accuracy check, and were able to proceed if they scored above 50%. Given the novelty of the medium, we do not know whether increasing this threshold might increase clarity, or whether additional accuracy checks might better ensure that participants' eye movements were still being tracked as reliably as possible.

. . . of course this all just captures the categorical belief data and can't handle verification (eg looks to residue set) at all; must bridge to algorithmic level