# 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 a key method of scientific inquiry in cognitive science, and experimental data has informed cognitive theory-building for centuries. A key ingredient in using empirical data to put theories to the test is the assumption made by researchers about how the mapping between theoretical notions and empirical measurements: the *linking assumption*. Indeed, empirical measurement, and the data resulting from it, are only useful and informative if the assumed linking assumption is (sufficiently) clear and justified. We argue that both clarity and justification are currently typically lacking for linking assumptions made for *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) assumed 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 (Experiment 3 ?, ?), **[jd: which we also replicate in a web-based eye-tracking paradigm using `webgazer.js` (Experiment 1)]**. We collect explicit beliefs to test against the original and replicated eye movement data in Experiment 2.

## 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 **[jd: at least X]** features: eye movements are taken to be an indicator of attention that is closely time-locked to the linguistic signal. 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 (?, ?, ?, ?, ?), a problem that is exacerbated when little attention is paid to the further variability in interpretation of eye movements introduced by the vastly different tasks employed in the literature. Consider just the difference between active referential tasks (in which participants' goal is to identify the speaker's intended referent, and frequently to select it) and passive predictive tasks (in which participants simply watch a display while listening to language, without an over task or goal). In the former case, an argument can be 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 un-

derlying 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 free production task.

**[jd: CONTINUE HERE, make explicit that lack of support may be because incremental decision task doesn't work (then cite work that does)(?, ?, ?, ?, ?). here we test again..., and we'll also try to replicate the original results in a web-based paradigm]**

## Test bed: Sun & Breheny (2020)

We replicate Experiment 3 of ? (?), 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, ? (?) were interested in assessing the possible effect of two factors on the speed with which determiners are processed: first, pre-existing 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 ? (?) call the 'residue set' (the remaining objects in the center of the screen) should increase immediately after observing the determiner.

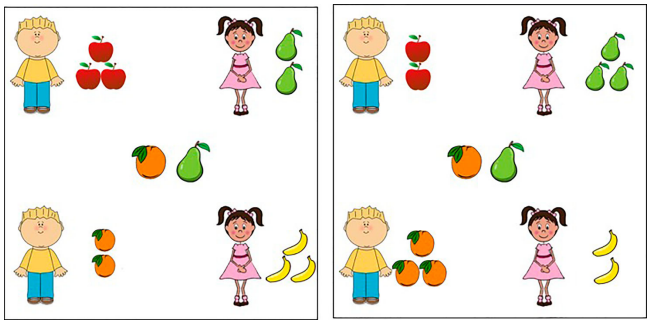

Figure 1: Example displays from Experiment 3 of ? (?). The left image (big *all*/ small *some*) was paired with *Click on the boy that has all/three of Susan's apples* or *Click on the girl that has some/two of Susan's pears*. The right image (small *all*/ big *some*) was paired with *Click on the boy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

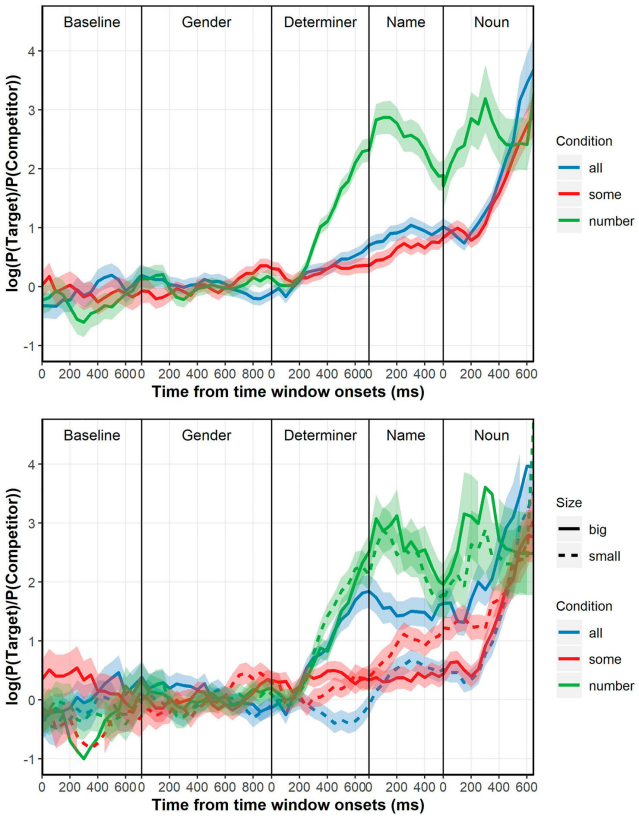


Figure 2: Eye movement results for Experiment 3 of ? (?). Shown are target preference scores from instruction onset to instruction offset. Top: target preference scores by determiner type. Bottom: target preference scores by determiner type and target set size. Transparent ribbons indicate standard error. **[jd: do we really need both these plots? the bottom one has the crucial results.]**

## Exp. 1: replicating Sun & Breheny (2020) using web-based eye-tracking

**[jd: . . . will it work? tbd.]**

### Methods

**Participants.** We recruited 183 participants on Prolific, of which 21 were excluded because their accuracy was lower than 95%. We also excluded trials on which the participants clicked the wrong target (33 trials). All participants were self-reported native English speakers.[1]

**Materials and procedure.** Exp. 1 was identical to ? (?)'s Exp. 3 except for one difference: we collected eye movements with webcam eyetracking using the webgazer library. Participants were presented with the same experimental displays as the original experiment (Figure **??**). One second after the display onset, paricipants 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 instuction. 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 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

## Exp. 2: 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 ? (?), participants engaged in an incremental decision task (?, ?, ?, ?).

### Methods

**Participants.** We recruited 120 participants on Mechanical Turk, of which **[jd: XXX]** were excluded because **[jd: X Y, Z]**.

**Materials and procedure.** We measured participants' beliefs about the intended referent for each display shown to participants by ? (?) (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

[1] Procedure, materials, analyses and exclusions were preregistered at https://osf.io/y2cgb. The collected sample size for Exp. 1 (183 participants) was larger than the originally preregistered sample size (102 participants) because 40% of the initally tested 102 participants had a technical issue and weren't able to see the whole display. **[lk: added prereg footnote here]**

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 ? (?)'s 2 (big vs. small target set) by 3 (*all, some*, number) design.

## Results

Following ? (?), we computed the target preference score $\ln(\frac{p(\text{target})}{p(\text{competitor})})$ for each time window, where $p(\text{target})$ and $p(\text{competitor})$ refer to the proportion of target and competitor selections, respectively (see Figure **??**). Proportions were aggregated by subject, determiner, and set size.

**Data analysis**. ? (?) fit separate regression models to the baseline, gender, determiner, name, and noun windows. We collapsed the name window into the determiner window because the name does not add information. We fit mixed-effects logistic regression models[2] to the baseline, gender, determiner, and noun window,[3] predicting 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. 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 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 led to more target selections for *some* (β=0.59, *SE*=0.28, $p < .05$) but to fewer target selections for *all* (β=-1.27, *SE*=0.29, $p < .0001$), compared to number terms.

[2] ? (?) ran linear models on the computed target advantage scores. We instead ran logistic models because logistic regression is the more principled approach to modeling categorical data and moreover avoids the problems of having to pre-aggregate data and add smoothing terms to avoid division by zero. The results are qualitatively identical if we instead run linear models on target preference scores. **[jd: make sure this is true or delete]**

[3] 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.

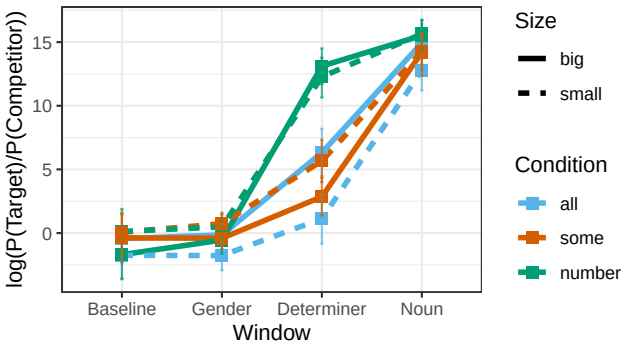


Figure 3: Target advantage scores in Experiment 2 by quantifier and set size. Error bars indicate 95% bootstrapped confidence intervals.

**Comparison with ? (?): replication analysis**. These results constitute a near-perfect replication of ? (?). Most of their effects reported in the determiner window replicated, with two exceptions: 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 ? (?)'s results when taking into account the joint determiner and name windows. In fact, ? (?) 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 ? (?) 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. **[jd: if time, re-analyze their results with logistic models?]**

**Comparison with ? (?): linking function analysis**. The overall correlation between proportion of selections and proportions of looks, where proportions were calculated separately by region of interest (target, competitor, distractors), time window (baseline, gender, determiner+name, noun), determiner (*all, some*, number), and set size (big, small), was high ($r(70) = .69$). At the individual

**[jd: CONTINUE HERE DESCRIBING LINKING FUNCTION RESULTS]**

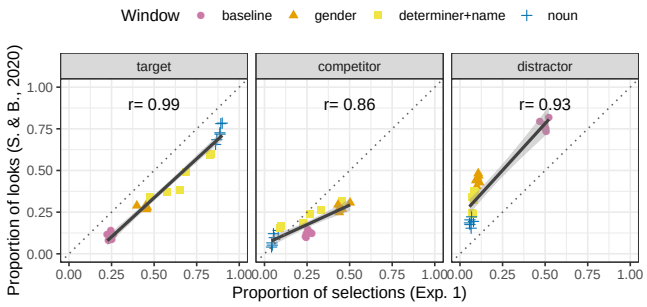


Figure 4: Proportions of looks in ? (?) against proportions of selections in Experiment 1. Facets indicate regions, colors indicate time window.

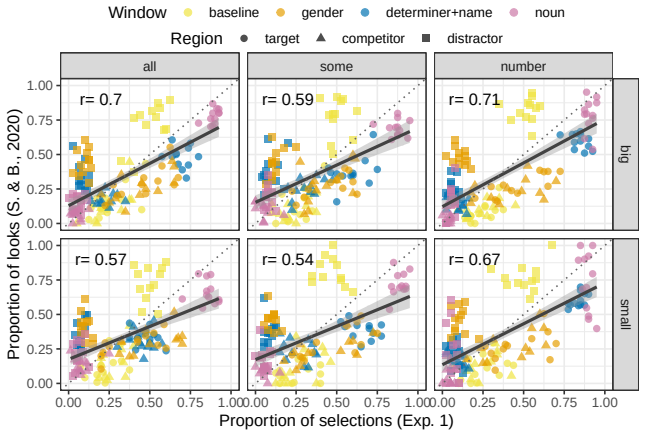


Figure 5: Proportions of looks in ? (?) against proportions of selections in Experiment 1. Facets indicate regions, colors indicate time window.

## 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 library is a) feasible; and b) a faster, cheaper way to collect eye movement data despite the added noise.
- 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:

Hypothetical speaker distribution for small set:

$$S(\text{"two"}|small) = .8$$
$$S(\text{"three"}|small) = 0$$
$$S(\text{"all"}|small) = .05$$
$$S(\text{"some"}|small) = .15$$

Hypothetical speaker distribution for big set:

$$S(\text{"two"}|big) = 0$$
$$S(\text{"three"}|big) = .6$$
$$S(\text{"all"}|big) = .3$$
$$S(\text{"some"}|big) = .1$$

The above speaker assumptions result in (observed) small target advantage for small-some over big-some:

$$L(small|\text{"some"}) = \frac{S(\text{"some"}|small)}{(S(\text{"some"}|small) + S(\text{"some"}|big))} = .15/(.15+.1) = .6$$

$$L(big|\text{"some"}) = \frac{S(\text{"some"}|big)}{(S(\text{"some"}|small) + S(\text{"some"}|big))} = .1/(.15+.1) = .4$$

The above speaker assumptions result in (observed) large target advantage for big-all over small-all:

$$L(small|\text{"}all\text{"}) = \frac{S(\text{"}all\text{"}|small)}{(S(\text{"}all\text{"}|small) + S(\text{"}all\text{"}|big))}$$
$$= .05/(.05+.3) = .14$$
$$L(big|\text{"}all\text{"}) = .3/(.05+.3) = .86$$

Numbers behave categorically (assuming exact semantics):

$$L(small|\text{"}two\text{"}) = 1$$
$$L(big|\text{"}two\text{"}) = 0$$
$$L(small|\text{"}three\text{"}) = 0$$
$$L(big|\text{"}three\text{"}) = 1$$

. . . 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