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

Judith Degen (jdegen@stanford.edu)
~~Department of Linguistics, 450 Jane Stanford Way Stanford, CA 94305 USA~~

Leyla Kursat (lkursat@stanford.edu)
~~Department of Linguistics, 450 Jane Stanford Way Stanford, CA 94305 USA~~

Daisy Leigh (ddleigh@stanford.edu)
Department of Linguistics, 450 Jane Stanford Way
Stanford, CA 94305 USA

Abstract

Experimental investigation is fundamental to theory-building in cognitive science, but its value depends on the *linking* ~~*assumption*~~*assumptions* ~~: the assumption~~ made by researchers about the mapping between empirical measurements and theoretical constructs. We argue that sufficient clarity and justification are often lacking for linking assumptions made in *visual world eye-tracking*, a widely used experimental method in psycholinguistic research. We test what 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. We do so by comparing eye-tracking data against explicit beliefs collected in an incremental decision task (Exp. 1), which replicates a scalar implicature processing study (Exp. 3 of Sun & Breheny, 2020). In Exp. 2, we replicate Sun and Breheny (2020) in a web-based eye-tracking paradigm using `WebGazer.js`. The results provide ~~strong~~ support for the Referential Belief link and cautious optimism for the prospect of conducting web-based eye-tracking. We discuss limitations on both fronts.

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

Introduction

~~Visual world eye-tracking (VWE~~Eye-tracking in the visual world paradigm (VWP) is a widely used measure in psycholinguistics, fruitfully driving advances in our understanding of phonetic, lexical, syntactic, prosodic, semantic, and pragmatic processing (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Allopenna, Magnuson, & Tanenhaus, 1998; Altmann & Kamide, 1999; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Huang & Snedeker, 2009; Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014). In standard ~~VWE tasks ,~~ VWP tasks participants view displays of objects and listen to ~~spoken sentences~~ speech while their eye movements are monitored (see Fig. 1 for an example). ~~VWE~~ The VWP is popular for good reason: eye movements can be interpreted as 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 (Allopenna et al., 1998). By sampling an x/y coordinate every few milliseconds, researchers thus obtain a 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~~ VWP 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 (Tanenhaus et al., 1995), that words are processed incrementally and listeners maintain uncertainty about past input (Allopenna et al., 1998; Clayards et al., 2008), and that listeners anticipate upcoming linguistic material based on selectional restrictions and rapid pragmatic reasoning (Altmann & Kamide, 1999; Sedivy et al., 1999).

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 (Salverda & Tanenhaus, 2017; Tanenhaus, Magnuson, Dahan, & Chambers, 2000; Allopenna et al., 1998; Magnuson, 2019). The problem of interpretability is compounded by the fact that ~~VWE is used in~~ the VWP is used for vastly different tasks ~~. Consider just~~ (for an overview, see Huettig, Rommers, & Meyer, 2011). Consider the difference between active referential tasks~~(~~, in which participants' goal is to identify and select the speaker's intended referent~~) and passive predictive tasks(~~, and passive listening tasks, in which participants simply watch a display while listening to language, without an overt task or goal~~)~~. In the former case, ~~it is argued that eye movements~~ eye movements are assumed to reflect listeners' active search for or belief in the referent, ~~depending on whether visual information about the display has been integrated. The~~ while the latter case indicates that eye movements ~~reflect an automatic predictive process (Altmann & Kamide, 1999). But what, exactly, does this mean? What is the generative process by which, e.g., a notion like "belief," "search,", or "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.~~ may reflect predictive processes (Altmann & Kamide, 1999).

Here, we test an explicit linking assumption for referential tasks, ~~most clearly~~ first put forward by Allopenna et al. (1998), which we term the ~~Referential Belief linking assumption~~*Referential Belief* link: that the empirical proportion of looks $p_{\text{empirical}}$ to a referent r in a time window in response to a (possibly partial) utterance u reflects participants' degree of belief p_{belief} that the referent is the intended target~~in that time window.~~:

$$p_{\text{empirical}}(r|u) \propto p_{\text{belief}}(r = \text{target}|u) \tag{1}$$

This linking assumption, which implicitly underlies much work in the VWP using referential tasks, was tested and found not supported in previous experimental pragmatics research (Qing, Lassiter, & Degen, 2018). In a re-analysis of an adjective processing dataset (Leffel, Xiang, & Kennedy, 2016), Qing et al. (2018) found that explicit beliefs collected in an incremental decision task (similar to gating tasks, Allopenna et al., 1998) did not correlate with eye movements, with the exception of one condition. They argued that the lack of support may have been the result of participants' negligible expectation[1] for the linguistic stimuli used in the original experiment.

An alternative possibility is that the incremental decision task simply does not capture the beliefs that inform eye movements. We believe this is unlikely, given recent successes using such tasks to elicit contrastive inferences (Kreiss & Degen, 2020; Alsop, Stranahan, & Davidson, 2018). 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 work, which tests the Referential Belief link on a different dataset. For this purpose, we replicate Exp. 3 of Sun and Breheny (2020) (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. ~~We also~~

Besides testing the Referential Belief link, this investigation also serves the purpose of assessing the utility of web-based incremental decision tasks as an alternative to lab-based eye-tracking. To this end, we assess a second alternative to lab-based eye-tracking: in (Exp. 2), we replicate SB2020 in a web-based eye-tracking paradigm ~~(Exp. 2).~~using the WebGazer.js library (Papoutsaki et al., 2016). The importance of evaluating the appropriateness and limitations of web-based alternatives to lab-based VWP studies has been especially highlighted by the pandemic.

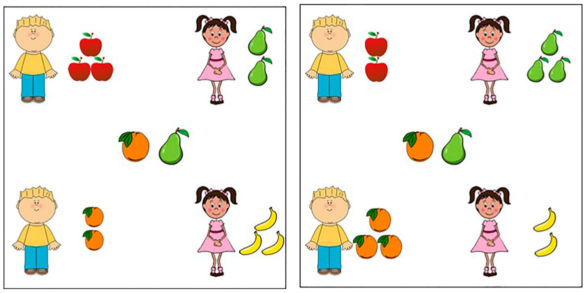

Figure 1: Example displays from Exp. 3 of SB2020. The left image (big *all*/ small *some*) was paired with *Click on the boy that has all/three of Susan's apples* or *Click on the girl that has some/two of Susan's pears*. Right image (small *all*/ big *some*): *Click on the boy that has all/two of Susan's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

Test bed: Sun & Breheny (2020)

SB2020 addressed a now classic question in experimental pragmatics: is the processing of scalar inferences delayed relative to the processing of literal information (Bott & Noveck, 2004; Breheny, Katsos, & Williams, 2006; Huang & Snedeker, 2009; Grodner, Klein, Carbary, & Tanenhaus, 2010; Degen & Tanenhaus, 2016; Tomlinson, Bailey, & Bott, 2013)? In particular, they 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, the determiner used, in particular whether its application 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 Fig. 1, 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, target looks upon hearing *all* and *some* should be delayed, 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.

Indeed, number terms led to more (and a faster increase in) target looks than did *all* and *some* (see proportions of looks in Fig. 2, top) 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). Looks to the residue set (not pictured) increased in the determiner window for *all* and *some* but not numbers, suggesting that the need for verification of the residue set is a source of relatively fewer early 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*.

[1] Expectations were independently estimated in free production.

Jointly, SB2020's 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 link, this study has both appealing features ~~as well as one~~ and a 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~~ No plausible argument can be made that participants look to the residue set because they believe it may be the intended target. ~~Thus, we have~~ We have thus 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.

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

~~Participants~~We measured participants' beliefs about the intended referent ~~were measured~~ in an incremental decision task~~(Allopenna et al., 1998; Qing et al., 2018; Kreiss & Degen, 2(~~ i.e., at ~~different~~ various points in the utterance ~~, allowing us~~ (Allopenna et al., 1998; Qing et al., 2018; Kreiss & Degen, 2020) in order to compare explicit beliefs to proportions of looks in SB2020 .

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.[2]

Materials and procedure. We measured participants' beliefs about the intended referent for each display shown to participants by SB2020 (see Fig. 1 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 (target set size: big vs. small) by 3 (determiner: *all, some*, number) design.

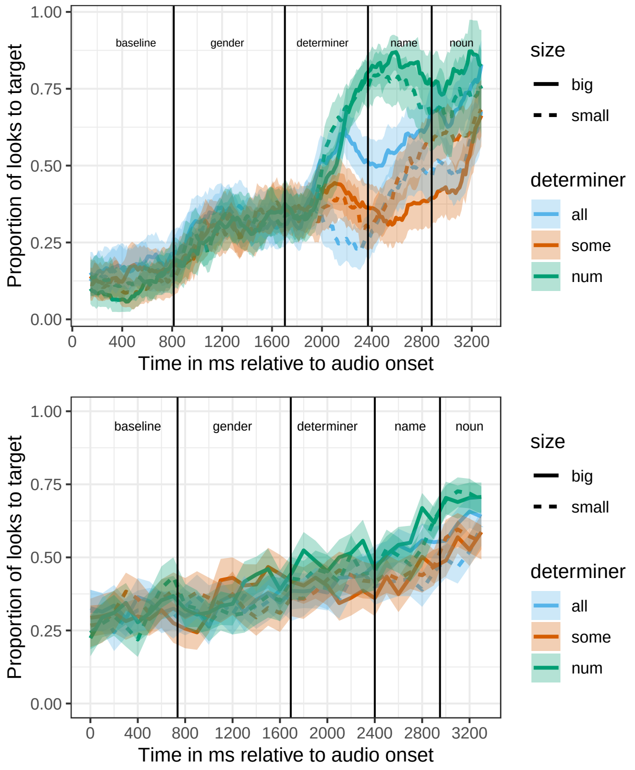

Figure 2: Proportion of target looks (out of target, competitor, and residue looks) from instruction onset. Transparent ribbons indicate 95% bootstrapped confidence intervals. Black vertical lines indicate onsets of analysis windows of interest (window labels at top of graphs). **Top:** Exp. 3 of SB2020.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

Results and discussion

Fig. 3 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

[2] Procedure, materials, analyses and exclusions were pre-registered (see Exp. 1: `https://osf.io/vfgc8`; Exp. 2: `https://osf.io/y2cgb`. Sample size for Exp. 2 (183 participants) was larger than pre-registered (102) because 40% of the initially tested 102 participants had a technical issue and couldn't see the whole display. All experimental materials, anonymized data, and analysis scripts are available at `https://github.com/thegricean/eyetracking_replications`.

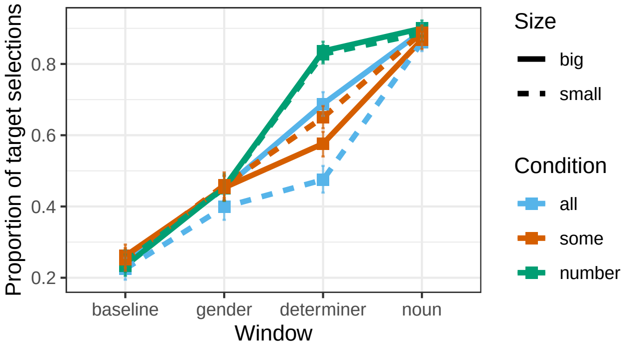

Figure 3: Proportion of target selections in Exp. 1 by determiner and set size. Error bars indicate 95% bootstrapped confidence intervals.

Table 1: Overview of critical effects in determiner (det.) and name windows in SB2020, our Exp. 1 and our Exp. 2. Rows list model predictors. "+": significant positive effect; "–": significant 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	·	+

adding smoothing terms to avoid division by zero or discarding mathematically problematic data points.

We fit separate models to the baseline, gender, determiner, and noun windows (collapsing SB2020's 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 determiner (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.[3] 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 (β=-0.09, *SE*=0.26, ~~consistent with the visual result~~ p <.73), i.e., there was no evidence that target selections in the number condition ~~are not~~ were modulated by target set size (~~β=-0.09, *SE*=0.26, p <.73~~see Fig. 3). However, we did observe interactions between determiner 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 determiner window effects replicated, with two exceptions (see overview in Table 1): we did not observe a main effect of set size, and we observed an interaction of size and determiner such that small *some* led to greater target selections than big *some*, and vice versa for *all*. These differences are ~~negligible~~related. 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 Fig. 2 (top) 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 (jointly corresponding to our determiner window), 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~~ does not capture.

Comparison with SB2020: linking function analysis. Fig. 4 shows the correlation between proportion of selections in Exp. 1 and proportions of looks in SB2020~~. The overall~~ , computed at the level of unique combinations of item, determiner, and size. The 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 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: compared to the determiner window, selection proportion was a worse 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 no different in the gender window (β=-0.02, *SE*=0.04, t=-0.61, p <.55). The effect was also modulated by the region of interest: selection proportion was a worse 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

[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.

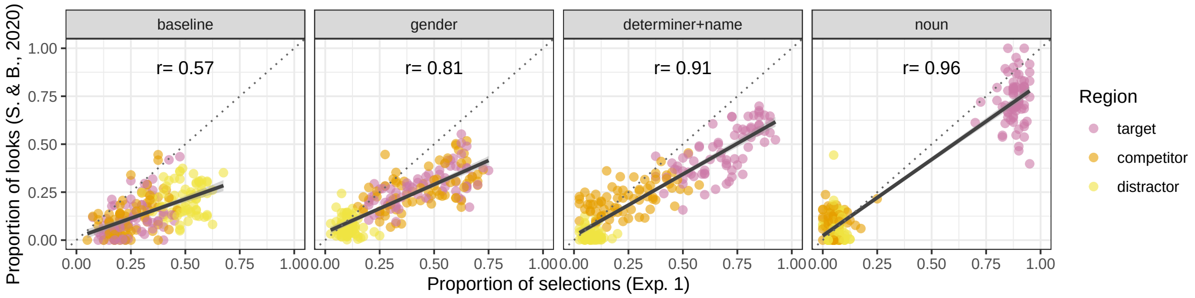

Figure 4: Proportions of looks in SB2020 against proportions of selections in Exp. 1. Facets indicate time windows. In each time window, proportions were computed for each of the 216 unique combinations of item (12), determiner (3), size (2), and region (3). Distractor looks indicate sum of looks to both distractors. Pearson's correlation coefficient for each time window is shown at the top of each corresponding facet.

looks.

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 Qing et al. (2018)'s results on the re-analyzed adjective processing dataset from Leffel et al. (2016). 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, the strong correlations we observed between proportion of selections and proportion of looks suggest that there were no differences in production expectations in SB2020 across conditions, and that the expectation for the observed utterances was high across the board. This hypothesis requires further empirical investigation via a production study eliciting descriptions of SB2020's referential displays.

Finally, a fact the Referential Belief link cannot account for is that participants in SB2020's experiment looked towards the residue set as a verification strategy. In Fig. 4, proportions of looks 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

We next investigate whether SB2020's and the Exp. 1 results also replicate in web-based eye-tracking.

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 movements with webcam eye-tracking using the WebGazer.js library (Papoutsaki et al., 2016). Participants were presented with the same experimental displays as the original experiment (Fig. 1). One second after 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. Upon clicking, the next trial started. The experiment began with 6 practice trials. Participants then completed the same 36 critical trials and 12 filler trials as in Exp. 1, implementing the same 2 (set size) by 3 (determiner) design. On each trial, eye movements were recorded from display onset until a selection was made.

Results

Fig. 2 (bottom) shows proportions of target looks in each condition (computed out of target, competitor, and residue looks in 100ms time bins). The ~~results display striking dissimilarities to SB2020: in particular, all curves are flatter, i.e.,~~ first thing of note is that looks to the target ~~increase~~ increased more slowly than in the original study ~~.~~

~~We~~ across the board. We attribute this to noise in the dependent measure, which we discuss in detail in the General Discussion.

To assess the effect of determiner and set size on target looks, we conducted separate mixed-effects logistic regressions in the 2 time windows of interest (determiner, name), predicting target over competitor looks from fixed effects of time (scaled and 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 fixed effects.

In the determiner window but not in earlier time windows, there was a significant intercept effect, i.e., an overall preference for target over competitor looks (β=0.70, *SE*=0.13, *p* <.0001), suggesting the target preference was driven by hearing the determiner. 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).

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, 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~~ such that there were more target looks when the target set was big (β=-0.30, *SE*=0.18, *p* <.11~~)~~.

~~Overview of critical effects in determiner (det.) and name windows in SB2020, our Exp. 1 and our Exp. 2. Rows list model predictors. "+": significantly positive effect; "–": significantly negative effect; "·": no evidence of an effect. Exp. 1 Predictor det. name det. det. name all.v.num – – – · – some.v.num – – – – – size + · · · · all.v.num:size + + + + · some.v.num:size · · – · – time + ? NA · +~~

Comparison with SB2020: replication analysis. An overview of the effects reported by SB2020 and the effects observed in our Exps. 1 and 2 is shown in Table 1. Most of the effects reported by SB2020 as first emerging in the determiner window did not emerge until the name window (about 700ms later). Of these, ~~however, most replicated~~ the determiner main effects replicated clearly. The replication patterns of the smaller interaction effects between determiner and size interactions were more subtle. The interaction between *some* and size replicated the Exp. 1 result. The trending interaction between *all* and size replicated both SB2020 and Exp. 1 numerically. Given the noise in the web-based eye-tracking measure, a greater sample size may be necessary to detect the more subtle effects reported by SB2020.

General discussion

The contributions of this work are three-fold: first, we twice—once in an incremental decision task and once in a novel web-based eye-tracking paradigm—replicated Sun and Breheny (2020)'s result that there is more uncertainty regarding the intended target for the determiners *all* and *some* than for numbers; and that this uncertainty is modulated by set size such that *all* is associated with larger sets (3 objects) and *some*, if anything, with smaller ones (2 objects).

Second, in contrast to a previous re-analysis (Qing et al., 2018) of an experimental pragmatics eye-tracking dataset ~~(Qing et al., 2018)~~(Leffel et al., 2016), the current re-analysis ~~offers clear~~ of Sun and Breheny (2020)'s data offers support for the Referential Belief link. ~~We believe this difference~~ This is encouraging, given that the Referential Belief link at least implicitly underlies much work with referential tasks in the VWP. This includes work in very different subfields of psycholinguistics, e.g., rhyme effects (Allopenna et al., 1998) or semantic competitor effects (Dahan & Tanenhaus, 2005; Yee & Sedivy, 2006) in word recognition. Such work is typically concerned with questions regarding the features of competitor items that interfere with looks to the target. These questions can be re-cast at the computational level as questions about which features affect degree of belief in various displayed referents being the intended target, while remaining agnostic as to whether such beliefs are the result of automatic activation or priming processes, or more strategic or goal-driven processing.

We believe the difference in support for the Referential Link in SB2020's data compared to Leffel et al. (2016)'s data is most likely the result of participants' ~~s~~ greater expectations for the observed ~~instructions~~ linguistic materials in SB2020's study, though this requires further investigation (see also Huettig & McQueen, 2007; Pontillo, 2017, for debate regarding t ~~The~~

An important limitation of the Referential Belief link is that it does not capture that participants' looks to regions in a display are not just guided by their belief that a region contains the target, but can be subject to other attentional processes (Allopenna et al., 1998). In SB2020's dataset, this problem is exemplified by looks to the residue set, which cannot reflect a possible target belief, since the residue set is never the target. Instead, these looks serve verification purposes – to make sure that objects of a category are left over or not (in the case of *some* and *all*, respectively). A fuller linking theory must integrate ~~such processes~~ referential beliefs with processes related to the deployment of attentional resources for prediction, integration, and verification in a task- and goal-dependent manner (see Pontillo, 2017, for in-depth discussion).

Finally, we ~~show~~ showed that web-based eye-tracking may provide a useful way to collect eye-tracking data for psycholinguistic research. ~~The~~ However, the latency of the observed effects indicates ~~we are some way from being able to conduct the gamut of eye-tracking experiments that require fine-grained,~~ that there are serious methodological and implementational issues that will need to be addressed before we are able to reliably collect time-sensitive data remotely. ~~Designs requiring multiple regions of interest or subtle discrimination between stimuli may fare particularly badly given the added noise, especially if effects are subtle.~~

~~However, our results do show signal amongst this noise. Web-based eye-tracking~~ While far more extensive testing is needed, we suggest several factors may have contributed to this latency.

First, participants' system performance may have played a significant role. The facial detection method used to detect eye position in the current implementation of WebGazer.js can be computationally demanding, as can the regression model WebGazer.js uses to make predictions about gaze location. This could have resulted in a bottleneck in the speed with which WebGazer.js was able to make predictions, leading to slower sampling frequency, or may ~~be particularly fruitful for studies that do not require precise timing information about effect onsets.~~ have led to lags or asynchronicity in the loading of audio and visual stimuli – especially for older, or less powerful machines. Indeed, Semmelmann and Weigelt (2018) found that eye-tracking data collected via web-cam was more susceptible to temporal error when participants completed the task remotely, using their own machines, compared to a more controlled, lab environment in which all participants used MacBook Pros. They suggest this discrepancy was likely due to differences in system performance, and recommend testing machine performance prior to participation, or excluding participants from analysis on this basis. In addition, using an alternative facial detection algorithm may reduce processing demands, though this may result in a trade-off between computing efficiency and accuracy.

~~Much more investigation is needed into how methodological and implementation decisions affect result reliability.~~ However, it is worth noting that sampling frequency in web-cam based eye-tracking cannot currently rival that of most modern eye-trackers, and it may be some time before it does – even if the above measures are implemented. We hope that future work using remote eye-tracking methods will allow for the establishment of latency 'baselines', as a means of approximating the effects we might expect in a lab environment.

Second, our replication followed the original experimental design in including four images in the display. The images were placed fairly close to each other in order to accommodate a variety of screen sizes. This, compounded by variability in the accuracy of WebGazer.js's predictions, may have contributed a considerable source of noise. Two-image displays that allow for greater distance between ~~the~~ regions of interest may ~~reduce noise, for example,~~ fare better, but validation through replication of existing work is needed. ~~More rigorous and frequent accuracy checks~~

Third, drift correction or re-calibration throughout the task~~may also improve the clarity~~ , standard in in-lab experiments, may improve the quality of the resulting data~~:~~ ~~in~~ . In Exp. 2, 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 decrease noise, or whether additional, continuous accuracy checks analogous to standard drift correction practices might better ensure that participants' eye movements are ~~being~~ reliably recorded.

In conclusion, this work demonstrates the utility of carefully investigating linking assumptions in psycholinguistics in general, and in experimental pragmatics in particular (see also Franke, 2016; Waldon & Degen, 2020). Much is still left to do in the quest towards developing linking functions from theory to data. While not glamorous work, it is the bedrock that our scientific inferences depend upon.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *JML*, *38*(4), 419–439.

Alsop, A., Stranahan, E., & Davidson, K. (2018). Testing contrastive inferences from suprasegmental features using offline measures. *Proceedings of the LSA*, *3*(1), 71–1.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Bott, L., & Noveck, I. (2004, oct). Some utterances are underinformative: The onset and time course of scalar inferences. *JML*, *51*(3), 437–457.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*(3), 434–63.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic bulletin & review*, *12*(3), 453–459.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, *40*(1), 172–201.

Franke, M. (2016). Task types, link functions & probabilistic modeling in experimental pragmatics. In F. Salfner & U. Sauerland (Eds.), *Preproceedings of trends in experimental pragmatics* (pp. 6 – 63).

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010, jul). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42–55.

Huang, Y. T., & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*, 376–415.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*(4), 460-482.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151-171.

Kreiss, E., & Degen, J. (2020). Production expectations modulate contrastive inference. In *Proceedings of CogSci 42*.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014). Is it or isnt it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335–342.

Leffel, T., Xiang, M., & Kennedy, C. (2016). Imprecision is pragmatic: Evidence from referential processing. In *Semantics and Linguistic Theory* (Vol. 26, pp. 836–854).

Magnuson, J. S. (2019). Fixations in the visual world paradigm: Where, when, why? *Journal of Cultural Cognitive Science*, *3*(2), 113–139.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). Webgazer: Scalable webcam eye tracking using user interactions. In *Proceedings of IJCAI 25* (pp. 3839–3845).

Pontillo, D. (2017). *Object naming in visual search tasks*. University of Rochester.

Qing, C., Lassiter, D., & Degen, J. (2018). What do eye movements in the visual world reflect? A case study from adjectives. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Salverda, A. P., & Tanenhaus, M. K. (2017). The visual world paradigm. In A. M. B. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics and the neurobiology of language: A practical guide* (pp. 89–110). Wiley.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465.

Sun, C., & Breheny, R. (2020). Another look at the online processing of scalar inferences: An investigation of conflicting findings from visual-world eye-tracking studies. *Language, Cognition and Neuroscience*, *35*(8), 949–979.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, *29*(6), 557–580.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632 – 1634.

Tomlinson, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *JML*, *69*(1), 18–35.

Waldon, B., & Degen, J. (2020). Modeling Behavior in Truth Value Judgment Task Experiments. In *Proceedings of the Society for Computation in Linguistics* (Vol. 3, pp. 10–19).

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 1.