Speakers account for asymmetries in visual perspective so listeners don't have to

Robert Hawkins, Hyowon Gweon, Noah Goodman
Stanford University

Abstract

Debates over adults' theory of mind use have been fueled by surprising failures of visual perspective-taking in simple communicative tasks. Motivated by recent computational models of context-sensitive language use, we reconsider the evidence in light of the nuanced pragmatics of these tasks: the differential informativity expected of a speaker depending on the context. Our model predicts that cooperative speakers faced with asymmetries in visual access ought to adjust their utterances to be more informative. In Exp. 1, we explicitly manipulated the presence or absence of occlusions and found that speakers systematically produced longer, more specific referring expressions than required given their own view when they have uncertainty about what their partner is seeing. In Exp. 2, we compare the utterances used by confederates in prior work with those produced by unscripted speakers in the same task. We find that confederates are systematically less informative than expected, leading to more listener errors. In addition to demonstrating a sophisticated form of speaker perspective-taking, these results suggest a deeper pragmatic explanation for why listeners may sometimes neglect to consider visual perspective: Failures of visual perspective-taking may in fact be explained by sophisticated pragmatic expectations about communicative behavior—that is, to successful use of theory of mind.

Keywords: theory of mind, pragmatics, interaction, communication, social cognition, replication

Introduction

Our success as a social species depends on our ability to understand, and be understood by, different communicative partners across different contexts. *Theory of*

mind—the ability to represent and reason about others' mental states—is considered to be the key mechanism that supports such context-sensitivity in our everyday social interactions. Being able to reason about what others see, want, and think allows us to make more accurate predictions about their future behavior in different contexts and adjust our own behaviors accordingly (Premack & Woodruff, 1978). Over the past two decades, however, there has been sustained controversy over the extent to which adults actually make of use theory of mind in communication: in some cases, the evidence appears to be more consistent with an egocentric or "mind-blind" account (Keysar, Barr, Balin, & Brauner, 2000; Lin, Keysar, & Epley, 2010; Keysar, Lin, & Barr, 2003).

Much of this debate has centered around the influential *director-matcher* paradigm, where a confederate speaker gives participants instructions about how to move objects around a grid. By introducing an asymmetry in visual access—certain cells of the grid are covered such that the listener can see objects that the speaker cannot (E.g. Fig. 4)—the task is designed to expose cases where listeners either succeed or fail to take into account what the speaker sees. While there have been numerous rounds of methodological criticism and reinterpretation of the evidence (e.g. Hanna, Tanenhaus, & Trueswell, 2003; Hanna & Tanenhaus, 2004; Heller, Grodner, & Tanenhaus, 2008; Brown-Schmidt & Tanenhaus, 2008; Mozuraitis, Stevenson, & Heller, 2018; Heller, Parisien, & Stevenson, 2016; Rubio-Fernández, 2017), prior interpretations have been limited in two intersecting ways. First, they have focused on the *listener*'s behavior while neglecting demands on the *speaker*'s theory of mind in the same interaction. Second, they have primarily focused on only one aspect of theory of mind use — considering the speaker's visual perspective.

We argue in this paper that the mental models supporting theory of mind use in communication are richer than previously considered. Just as making sense of an agent's physical behaviors requires a broad, accurate model of how the agent's visual access, beliefs, and intentions translate into motor plans (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017), making sense of an agent's *linguistic* behaviors depends on an accurate mental model of what a speaker would say, or what a listener would understand, in different situations (Bergen & Grodner, 2012; Goodman & Frank, 2016; Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Franke & Jäger, 2016). From this perspective, theory of mind use not only incorporates people's mental models of a partner's knowledge or visual access but also their expectations about how their partner would behave in a communicative context.

The Gricean notion of cooperativity (Grice, 1975; Clark, 1996) refers to the idea that speakers intend to avoid saying things that are confusing or unnecessarily

---

This report is based in part on work presented at the 38th Conference of the Cognitive Science Society. Correspondence concerning this article should be addressed to Robert X.D. Hawkins, e-mail: rxdh@stanford.edu

complicated given the current context, and that listeners expect this. For instance, imagine trying to help someone spot your dog at a busy dog park. It may be literally correct to call it a "dog," but as a cooperative speaker you would understand that the listener would have trouble disambiguating the referent from many other dogs. Likewise, the listener would reasonably expect you to say something more informative than "dog" in this context. You may therefore prefer to use a more specific or *informative* expressions, like "the little terrier with the blue collar." (Brennan & Clark, 1996; van Deemter, 2016). Critically, you might do so even when you happen to *see* only one dog at the moment, but know there are likely to be other dogs from the listener's point of view. In other words, in the presence of uncertainty about their partner's visual context, a cooperative speaker may tend toward specificity.

Now, what level of specificity is pragmatically appropriate in the director-matcher task (Keysar et al., 2003)? This task requires the speaker to generate a description such that a listener can identify the correct object among distractors, but several cells are hidden from the speaker's view (e.g. Fig. 1, bottom). It is thus highly salient to the speaker that there are hidden objects she cannot see but her partner can. Gricean reasoning, as realized by recent formal models (Goodman & Frank, 2016; Frank & Goodman, 2012; Franke & Jäger, 2016), predicts that a speaker in this context will compensate for her uncertainty about the listener's visual context by increasing the informativity of her utterance beyond what she would produce in a completely shared context. (See SI Text A for a formal model of pragmatic reasoning in this situation and a mathematical derivation of the informativity prediction.) From this perspective, the director-matcher task is not only challenging for the listener; it also requires a sophisticated use of theory of mind, *vis a vis* pragmatic reasoning, on the part of the speaker.

In the following experiments, we ask whether people, as speakers, show such sensitivity to others' visual access, and whether the listener's pragmatic expectations about this sensitivity can explain why prior work has found frequent listener mistakes in director-matcher tasks. First, we directly test our model's prediction by manipulating the presence and absence of occlusions in an interactive, natural-language reference game. Second, we conduct a replication of (Keysar et al., 2003) with an additional *unscripted* condition to evaluate whether the scripted referring expressions used by confederate speakers in prior work accord with what a real speaker would say in the same interactive context (Kuhlen & Brennan, 2013; Bavelas & Healing, 2013; Tanenhaus & Brown-Schmidt, 2008). If confederate speakers were using comparatively uncooperative and underinformative scripts, this might explain why participants made mistakes as listeners; they *rationally* ignored visual perspective, relying instead on an expectation of adequately informative utterances.

Experiment 1: Speaker behavior under uncertainty

How does an unscripted speaker change her communicative behavior when there is uncertainty about visual access? To address this question empirically, we randomly

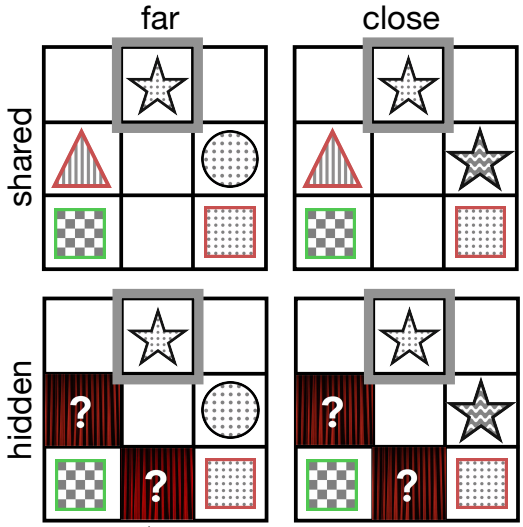

Figure 1. Design used in Exp. 1 (from speaker's view; grey square indicates target).

assigned participants to the roles of speaker and listener and paired them over the web to play an interactive communication task (Hawkins, 2015). On each trial, both players were presented with a 3×3 grid containing objects. One *target* object was privately highlighted for the speaker, who freely typed a message into a chat box in order to get the listener to click the intended referent. The objects varied along three discrete features (*shape*, *texture*, and *color*), each of which took four discrete values (64 possible objects). See Fig. S1 for a screenshot of the interface.

There were four types of trials, forming a within-pair 2×2 factorial design. We manipulated the presence or absence of occlusions and the closeness of *shared* distractors to the target (see Fig. 1). On 'shared' trials, all objects were seen by both participants, but on 'hidden' trials, two cells of the grid were covered with occluders (curtains) such that only the listener could see the contents of the cell. On 'far' trials, the target is the only object with a particular shape; on 'close' trials, there is also a shared distractor with the target's shape, differing only in color or texture (see Materials & Methods).

Behavioral results

Our primary measure of speaker behavior is the length (in words) of naturally produced referring expressions sent through the chat box. First, as a baseline, we examined the *simple* effect of close vs. far contexts in trials with no occlusions. We found that speakers used significantly more words on average when there was a distractor in context that shared the same shape as the target ($b = 0.56, t = 5.1, p < 0.001$; see Fig. 2A). This replicates the findings of prior studies in experimental pragmatics (Brennan & Clark, 1996; Monroe, Hawkins, Goodman, & Potts, 2017, e.g). Next, we turn to the simple effect of occlusion in far contexts (which are most

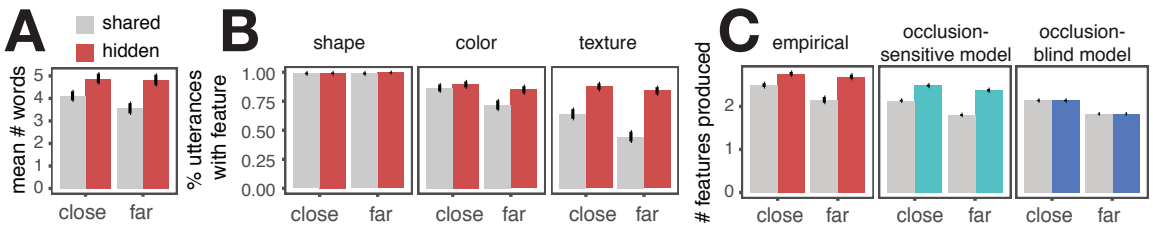

Figure 2. Results for Exp. 1. (A) Speakers used significantly more words when occlusions were present. This effect is larger than the simple pragmatic effect of a close distractor. (B) Utterances broken out by feature mentioned. (C) Posterior predictives of each model, compared to data. Error bars on empirical data are bootstrapped 95% confidence intervals; model error bars are 95% credible intervals.

similar to the displays used in the director-matcher task which we adopt in Exp. 2 (Keysar et al., 2003)). Speakers used 1.25 additional words on average when they knew their partner could potentially see additional objects ($t = 7.5, p < 0.001$). Finally, we found a significant interaction ($b = -0.49, t = 3.8, p < 0.001$) where the effect of occlusion was larger in far contexts, likely indicating a ceiling on the level of informativity required to individuate objects in our simple stimulus space.

What are these additional words used for? As a secondary analysis, we annotated each utterance based on which of the three object features were mentioned (shape, texture, color). Because speakers nearly always mentioned shape (e.g. 'star', 'triangle') as the head noun of their referring expression regardless of context ($\sim 99\%$ of trials), differences in utterance length across conditions must be due to differentially mentioning the other two features (color and texture). Confirming this observation, we found simple effects of occlusion in far contexts for both features ($b = 1.33, z = 2.9, p = 0.004$ for color; $b = 4.8, z = 6.4, p < 0.001$ for texture, see Fig. 2B). In other words, in displays like the left column of Fig. 1 where the target was the only 'star', speakers were somewhat more likely to produce the star's color—and much more likely to produce its texture—when there were occlusions present, even though shape alone is sufficient to disambiguate the target from visible distractors in both cases. Finally, we note that listener errors were rare: 88% of listeners made only one or fewer errors (out of 24 trials), and there was no significant difference in error rates across the four conditions ($\chi^2(3) = 1.23, p = 0.74$). We test the connections between context-sensitive speaker behavior and listener error rates more explicitly in Exp. 2.

Model comparison

While our behavioral results provide qualitative support for a Gricean account over an egocentric account, formalizing these two accounts in computational models allows a stronger test of our hypothesis by generating graded quantitative predic-

tions. We formalized both accounts in the probabilistic Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016; Franke & Jäger, 2016; Kao, Wu, Bergen, & Goodman, 2014; Goodman & Stuhlmüller, 2013), which has successfully captured a variety of other pragmatic phenomena. In this framework, speakers are decision-theoretic agents attempting to (soft-)maximize a utility function balancing parsimony (i.e., a preference for shorter, simpler utterances) with informativeness (i.e., the likelihood of an imagined listener agent having the intended interpretation). The only difference between the two accounts in the RSA framework is how the asymmetry in visual access is handled: the 'occlusion-blind' speaker simply assumes that the listener sees the same objects as she herself sees, while the 'occlusion-sensitive' speaker represents uncertainty over her partner's visual context. In particular, she assumes a probability distribution over the possible objects that might be hidden behind the occlusions and attempts to be informative *on average*. The two models have the same four free parameters: a speaker optimality parameter controlling the soft-max temperature, and three parameters controlling the costs of producing the features of shape, color, and texture (see SI Text B for details).

We conducted a Bayesian data analysis to infer these parameters conditioning on our empirical data, and computed a Bayes Factor to compare the models. We found extremely strong support for the occlusion-sensitive model relative to the occlusion-blind model ($BF = 2.2 \times 10^{209}$; see Fig. S3 for likelihoods). To examine the pattern of behavior of each model, we computed the posterior predictive on the expected number of features mentioned in each trial type of our design. While the occlusion-blind speaker model successfully captured the simple effect of close vs. far contexts, it failed to account for behavior in the presence of occlusions. The occlusion-sensitive model, on the other hand, accurately accounted for the full pattern of results (see Fig 2C). Finally, we examined parameter posteriors for the occlusion-sensitive model (see Fig. S4): the inferred production cost for *texture* was significantly higher than that for the other features, reflecting the asymmetry in production of texture relative to color.

Experiment 2: Comparing confederates to natural speakers

Experiment 1 directly tested the hypothesis that speakers increase their specificity in contexts with asymmetry in visual access. We found that speakers are not only context-sensitive in choosing referring expressions that distinguish target from distractors in a shared context, but are *occlusion-sensitive*, adaptively compensating for uncertainty. Critically, this resulted in systematic differences in behavior across the occlusion conditions that are difficult to explain under an egocentric theory: in the presence of occlusions, speakers were spontaneously willing to spend additional time and keystrokes to give further information beyond what they produce in the corresponding unoccluded contexts, even though that information is equally redundant given the visible objects in their display.

These results validate our prediction that speakers appropriately increase their

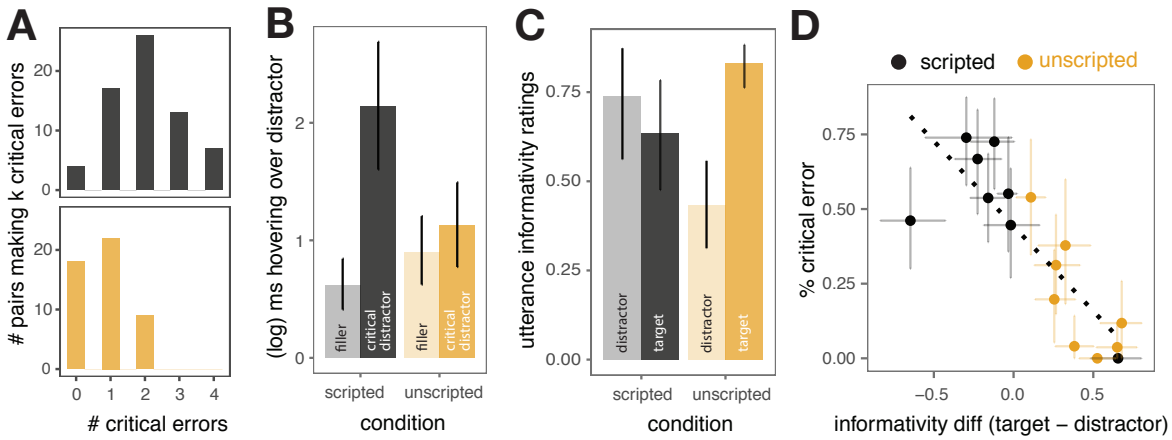

Figure 3. (A-B) Listener results for Exp. 2. (A) Distribution of errors with scripted and unscripted instructions. Participants in the unscripted condition made significantly fewer errors. (B) Even when they were correct, listeners in the scripted condition were more likely to hover their mouse cursor over the distractor relative to baseline while the unscripted condition shows no difference. (C-D) Speaker results for Exp. 2. (C) While speakers in the scripted condition were forced to use utterances that were judged to fit target and distractor roughly equally, speakers in the unscripted condition naturally produced utterances that fit the target much better than the distractor. (D) The extent to which an utterance fits the target more than the distractor is highly predictive of error rates at an item-by-item level. All error bars are bootstrapped 95% confidence intervals.

level of specificity in contexts containing occlusions. In Experiment 2, we recruited pairs of participants for an online, interactive version of the original director-matcher task (Keysar et al., 2003) which used occluded contexts to demonstrate limits on visual perspective-taking for the *listener*. Given the results of Exp. 1, we predicted that participants in the director role (i.e. speakers) would naturally provide more informative referring expressions than the confederate directors used in prior work. This would suggest that the confederate directors in prior work were pragmatically infelicitous, violating listeners' expectations. This violation of listeners' cooperative expectations may have led to detrimental consequences for listener performance.

The stimuli and procedure were chosen to be as faithful as possible to those reported in (Keysar et al., 2003) while allowing for interaction over the web. Directors used a chat box to communicate where to move a privately cued target object in a 4×4 grid (see Fig. 4). The listener then attempted to click and drag the intended object. In each of 8 objects sets, mostly containing filler objects, one target belonged to a 'critical pair' of objects, such as a visible cassette tape and a hidden roll of tape that could both plausibly be called 'the tape.'

We used a between-subject design to compare the scripted labels used by confederate directors in prior work against what participants naturally say in the same role. For participants assigned to the director role in the 'scripted' condition, a pre-

scripted message using the precise wording from (Keysar et al., 2003) automatically appeared in their chat box on half of trials (the 8 critical trials as well as nearly half of the fillers). Hence, the scripted condition served as a direct replication of (Keysar et al., 2003). To maintain an interactive environment, the director could freely produce referring expressions on the remainder of filler trials. In the 'unscripted' condition, directors were unrestricted and free to send whatever messages they deemed appropriate on all trials. In addition to analyzing messages sent through the chat box and errors made by matchers (listeners), we collected mouse-tracking data in analogy to the eye-tracking common in these paradigms (see Materials & Methods).

Listener errors

Our scripted condition successfully replicated the results of (Keysar et al., 2003) with even stronger effects: listeners incorrectly moved the hidden object on approximately 50% of critical trials. However, on *unscripted* trials, the listener error rate dropped by more than half, $p_1 = 0.51, p_2 = 0.20, \chi^2(1) = 43, p < 0.001$ (Fig. 3A). While we found substantial heterogeneity in error rates across object sets (just 3 of the 8 object sets accounted for the vast majority of remaining unscripted errors; see SI Fig. S2), listeners in the unscripted condition made fewer errors for nearly every critical item. We found a significant difference in error rates across conditions in a (logistic) mixed-effects model controlling for both pair- and item-level variability ($z = 2.6, p = 0.008$; see Materials & Methods).

Even if participants in the unscripted condition make fewer actual errors, they may still be *considering* the hidden object just as often on trials where they go on to make correct responses. As a proxy for the eye-tracking analyses reported by (Keysar et al., 2003), we conducted a mouse-tracking analysis. We computed the mean (logged) amount of time spent hovering over the hidden distractor and found a significant interaction between condition and the contents of the hidden cell ($t = 3.59, p < 0.001$; Fig. 3B).Listeners in the *scripted* condition spent more time

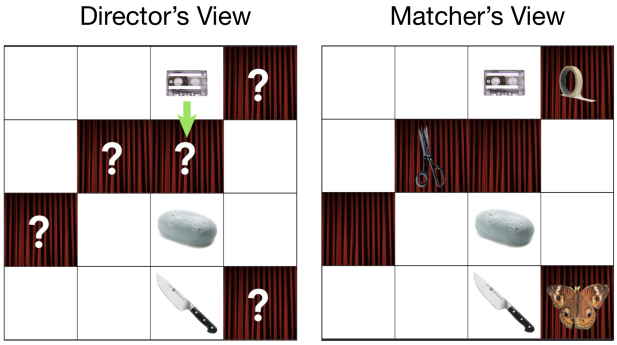

Figure 4. Screenshots of Exp. 2 interface on a critical trial: a cassette tape is in view of both players, but a roll of tape is occluded from the speaker's view.

hovering over the hidden cell when it contained a confusable distractor relative to baseline, again replicating (Keysar et al., 2003). In the unscripted condition there was no difference from baseline.[1]

Speaker informativity

Next, we test whether these improvements in listener performance in the unscripted condition are accompanied by more informative speaker behavior than the scripted utterances allowed. The simplest measure of speaker informativity is the raw number of words used in referring expressions. Compared to the scripted referring expressions, speakers in the unscripted condition used significantly more words to refer to critical objects ($b = 0.54, t = 2.6, p = 0.019$) However, this is a coarse measure: for example, the shorter "Pyrex glass" may be more specific than "large measuring glass" despite using fewer words. For a more direct measure, we extracted the referring expressions generated by speakers in all critical trials and standardized spelling and grammar, yielding 122 unique labels after including scripted utterances. We then recruited an independent sample of judges (see Norming Study in Materials and Methods) to rate how well every label fit both the target and the hidden distractor. We computed the *informativity* of an utterance (the *tape*) as the difference in how well it was judged to apply to the target (the cassette tape) relative to the distractor object (the roll of tape).

Our primary measure of interest is the difference in informativity across scripted and unscripted utterances. We found that speakers in the unscripted condition systematically produced more informative utterances than the scripted utterances ($d = 0.5$, 95% bootstrapped CI $= [0.27, 0.77], p < .001$; see SI Text C for details). Scripted labels fit the hidden distractor just as well or better than the target, but unscripted labels fit the target better and the hidden distractor much worse (see Fig. 3C). In other words, the scripted labels used in (Keysar et al., 2003) were less informative than expressions speakers would normally produce to refer to the same object in this context.

These results strongly suggest that the speaker's informativity influences listener accuracy. In support of this hypothesis, we found a strong negative correlation between informativity and error rates across items and conditions: listeners make fewer errors when utterances are a better fit for the target relative to the distractor ($\rho = -0.81$, bootstrapped 95% CI $= [-0.9, -0.7]$; Fig. 3D). This result suggests that listener behavior is driven by an expectation of speaker informativity: listeners interpret utterances proportionally to how well they fit objects in context.

[1]Mean hover time was exactly zero for the majority of trials; we thus conducted a follow-up analysis examining the binarized *proportion* of trials, and found the same pattern of results. We also pre-registered an analysis of time elapsed before *first* hovering over the target but due to unexpectedly poor precision in timing measurements, we did not pursue this analysis further.

General Discussion

Are human adults expert mind-readers, or fundamentally egocentric? The long-standing debate over the role of theory of mind in communication has largely centered around whether listeners (or speakers) with private information consider their partner's *visual* perspective (Barr & Keysar, 2006; Hanna et al., 2003; Heller et al., 2008). Our work presents a more nuanced picture of how a speaker and a listener use theory of mind to modulate their pragmatic expectations. The Gricean cooperative principle emphasizes a natural division of labor in how the *joint effort* of being cooperative is shared (Clark, 1996; Mainwaring, Tversky, Ohgishi, & Schiano, 2003). It can be asymmetric when one partner is expected to, and able to, take on more complex reasoning than the other, in the form of visual perspective-taking, pragmatic inference, or avoiding further exchanges of clarification and repair. One such case is when the speaker has uncertainty over what the listener can see, as in the director-matcher task. Our Rational Speech Act (RSA) formalization of cooperative reasoning in this context predicts that speakers (directors) naturally increase the informativity of their referring expressions to hedge against the increased risk of misunderstanding; Exp. 1 presents direct evidence in support of this hypothesis.

Importantly, when the director (speaker) is expected to be appropriately informative, communication can be successful even when the matcher (listener) does not reciprocate the effort. If visual perspective-taking is effortful and cognitively demanding (Lin et al., 2010), the matcher will actually minimize joint effort by *not* taking the director's visual perspective. This suggests a less egocentric explanation of *when* and *why* listeners neglect the speaker's visual perspective; they do so when they expect the speaker to disambiguate referents sufficiently. While adaptive in most natural communicative contexts, such neglect might backfire and lead to errors when the speaker (inexplicably) violates this expectation. From this point of view, the "failure" of listeners in these tasks is not really a failure; instead, it suggests that both speakers and listeners know when (and how much) they should expect others to be cooperative and informative, and allocate their resources accordingly (Griffiths, Lieder, & Goodman, 2015). Exp. 2 confirms this hypothesis; when directors used underinformative scripted instructions (taken from prior work), listeners made significantly more errors, and speaker informativeness strongly modulated listener error rates.

Our work adds to the growing literature on the debate over the role of pragmatics in the director-matcher task. A recent study questions the communicative nature of the task itself by showing that selective attention alone is sufficient for successful performance on this task, and that listeners become suspicious of the director's visual access when the director shows unexpectedly high levels of specificity in their referring expressions (Rubio-Fernández, 2017). Our current results are not inconsistent with this idea, and presents another way in which listeners show a sophisticated use of theory of mind; they make errors when the director is unexpectedly under-informative. Prior work also suggests that although speakers tend to be over-informative in their

referring expressions, (Koolen, Gatt, Goudbeek, & Krahmer, 2011), a number of situational factors (e.g., perceptual saliency of referents) can modulate this tendency. Our work hints at an additional principle that guides speaker informativity: speakers maintain uncertainty about the listener's visual context and their ability to disambiguate the referent in that context.

Additionally, recent *constraint-based* models (Heller et al., 2016; Mozuraitis et al., 2018) have proposed that participants probabilistically weight their own egocentric visual perspective alongside their partner's perspective. While this model of production focused on cases where the *speaker* has private information unknown to the listener (Mozuraitis et al., 2018), our model focuses on the original director-matcher case where it is mutually known that the *listener* has additional private information (Keysar et al., 2003). Importantly, these constraint-based models leave open a key question: how is the weighting parameter determined in a given context? Our results suggest one possible answer: under a Gricean account, this weight may reflect the division of labor that can be tuned up or down depending on how informative the listener expects the speaker to be (and vice versa).Yet, whether the allocation of resources, and ensuing perspective neglect, is a fixed strategy or one that adjusts dynamically remains an open question: given sufficient evidence of an unusually underinformative partner, listeners may realize that vigilance about which objects are occluded yields a more effective strategy for the immediate interaction. An important direction for future work is to directly explore listener adaptability in adjusting their use of visual perspective-taking as a function of Gricean expectations for a given partner (Grodner & Sedivy, 2011; Pogue, Kurumada, & Tanenhaus, 2016).

In sum, our findings suggest that language use is well-adapted to contexts of uncertainty and knowledge asymmetry. The pragmatic use of theory of mind to establish division of labor is also critical for other forms of social cooperation, including pedagogy (Shafto, Goodman, & Griffiths, 2014) and team-based problem solving (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010; Krafft, 2018). Enriching our notion of theory of mind use to encompass these pragmatic expectations, not only expectations about what our partner *knows* or *desires*, may shed new light on the flexibility of social interaction more broadly.

Materials & Methods

Data availability

Unless otherwise mentioned, all analyses and materials were preregistered at https://osf.io/qwkmp/. Code and materials for reproducing the experiment as well as all data and analysis scripts are open and available at https://github.com/hawkrobe/pragmatics_of_perspective_taking.

Exp. 1 experimental design

We recruited 102 pairs of participants from Amazon Mechanical Turk and randomly assigned speaker and listener roles. After we removed 7 games that disconnected part-way through and 12 additional games according to our pre-registered exclusion criteria (due to being non-native English speakers, reporting confusion about the instructions, or clearly violating the instructions), we were left with a sample of 83 full games. In order to make it clear to the speaker that there could really be objects behind the occluders without providing a statistical cue to their identity or quantity on any particular trial, we randomized the total number of distractors in the grid on each trial (between 2 and 4) as well as the number of those distractors covered by curtains (1 or 2). If there were only two distractors, we did not allow both of them to be covered: there was always at least one visible distractor. Each trial type appeared 6 times for a total of 24 trials, and the sequence of trials was pseudo-randomized such that no trial type appeared more than twice in each block of eight trials. Participants were instructed to use visual properties of the objects rather than spatial locations in the grid.

Exp. 1 analyses

We tested differences in speaker behavior across conditions using a mixed-effect regression of context and occlusion on the number of words produced, with maximal random effect structure containing intercept, slopes, and interaction. For feature-level analyses, we ran separate mixed-effect logistic regressions for color and texture predicting mention from context; due to convergence issues, the maximum random effect structure supported by our data contains only speaker-level intercepts and slopes for the occlusion effect.

Exp. 2 methods

We recruited 200 pairs of participants from Amazon Mechanical Turk. 58 pairs were unable to complete the game due to a server outage. Following our preregistered exclusion criteria, we removed 24 games who reported confusion, violated our instructions, or made multiple errors on filler items, as well as 2 additional games containing non-native English speakers. This left 116 pairs in our final sample. We displayed instructions to the director as a series of arrows pointing from some object to a neighboring unoccupied cell. Trials were blocked into eight sets of objects, with four instructions each. As in (Keysar et al., 2003), we collected baseline performance by replacing the hidden alternative (e.g. a roll of tape) with a filler object that did not fit the critical instruction (e.g. a battery) in half of the critical pairs. The assignment of items to conditions was randomized across participants, and the order of conditions was randomized under the constraint that the same condition would not be used on more than two consecutive items. All object sets, object placements, and corresponding instruction sets were fixed across participants. In case of a listener

error, the object was placed back in its original position; both participants were given feedback and asked to try again.

Exp.2 analyses

We used maximal mixed-effects regression for all analyses. To test differences in error rates across conditions, we used a logistic model with fixed effect of condition, random intercepts for each dyad, and random slopes and intercepts for each object set. For mouse-tracking analyses, we used dyad-level and object-level random intercepts and slopes for the difference from baseline. To compare speaker message length across conditions, we ran a regression on difference scores with a fixed intercept and random intercepts for object and dyads.

Norming study

We recruited 20 judges on Amazon Mechanical Turk to rate how well each label fit the target and hidden distractor objects on a slider from "strongly disagree" (meaning the label "doesn't match the object at all") to "strongly agree" (meaning the label "matches the object perfectly"). They were shown objects in the context of the full grid (with no occlusions) such that they could feasibly judge spatial or relative references like "bottom block." We excluded 4 judges for guessing with response times $< 1s$. Inter-rater reliability was relatively high, with intra-class correlation coefficient of $0.54\,(95\%CI = [0.47, 0.61])$.

Mouse-tracking

In both experiments, we asked the matcher to wait until the director sent a message; when the message was received, the matcher clicked a small circle in the center of the grid to show the objects and proceed with the trial. We recorded at 100Hz from the matcher's mouse in the decision window after this click, until the point where they clicked and started to drag one of the objects.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 0064.

Barr, D. J., & Keysar, B. (2006). Perspective taking and the coordination of meaning in language use. In *Handbook of psycholinguistics (second edition)* (pp. 901–938). Elsevier.

Bavelas, J., & Healing, S. (2013). Reconciling the effects of mutual visibility on gesturing: A review. *Gesture*, *13*(1), 63–92.

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *J Exp Psychol Learn Mem Cogn*, *38*(5), 1450.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J Exp Psychol Learn Mem Cogn*, *22*(6), 1482.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cogn Sci*, *32*(4), 643–684.

Clark, H. H. (1996). *Using language*. Cambridge university press Cambridge.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, *35*(1), 3–44.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends Cogn Sci*, *20*(11), 818 - 829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Top Cogn Sci*, *5*(1), 173–184.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (pp. 43–58). New York: Academic Press.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Top Cogn Sci*, *7*(2), 217–229.

Grodner, D., & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. *The processing and acquisition of reference*, 239.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cogn Sci*, *28*(1), 105–115.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *J Mem Lang*, *49*(1), 43–61.

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966-976.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.

Heller, D., Parisien, C., & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, *149*, 104.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends Cogn Sci*, *20*(8), 589–604.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proc Natl Acad Sci USA*, *111*(33), 12002–12007.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychol Sci*, *11*(1), 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults.

Cognition, *89*(1), 25 - 41.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J Pragmat*, *43*(13), 3231–3250.

Krafft, P. M. (2018). A simple computational theory of general collective intelligence. *Topics in cognitive science*.

Kuhlen, A. K., & Brennan, S. E. (2013). Language in dialogue: when confederates might be hazardous to your data. *Psychon Bull Rev*, *20*(1), 54–72.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *J Exp Soc Psychol*, *46*(3), 551–556.

Mainwaring, S. D., Tversky, B., Ohgishi, M., & Schiano, D. J. (2003). Descriptions of simple spatial scenes in english and japanese. *Spatial Cognition and Computation*, *3*(1), 3–42.

Monroe, W., Hawkins, R. X. D., Goodman, N. D., & Potts, C. (2017). Colors in context: A pragmatic neural model for grounded language understanding. *arXiv preprint arXiv:1703.10186*.

Mozuraitis, M., Stevenson, S., & Heller, D. (2018). Modeling reference production as the probabilistic combination of multiple perspectives. *Cogn Sci*.

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under-and over-informative prenominal adjective use. *Front Psychol*, *6*, 2035.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(04), 515–526.

Rubio-Fernández, P. (2017). The director task: A test of theory-of-mind use or selective attention? *Psychon Bull Rev*, *24*(4), 1121–1128.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cogn Psychol*, *71*, 55–89.

Tanenhaus, M. K., & Brown-Schmidt, S. (2008). Language processing in the natural world. *Philos Trans R Soc Lond B Biol Sci*, *363*(1493), 1105–1122.

van Deemter, K. (2016). *Computational models of referring: a study in cognitive science*. MIT Press.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

Derivation of qualitative model prediction

Our experiments are motivated by the Gricean observation that speakers should attempt to be more informative when there is an asymmetry in visual access. In this appendix, we formalize this scenario in a computational model of communication as recursive social reasoning and prove that the predicted increase in informativity qualitatively holds under fairly unrestrictive conditions.

Following recent advances in the Rational Speech Act (RSA) framework, we define a speaker as a decision-theoretic agent who must choose a referring expression u to refer to a target object o in a context C by (soft)-maximizing a utility function U:

$$S(u|o, C) \propto \exp\{\alpha U(u; o, C)\}$$

Definition. The *basic* utility used in RSA models captures the informativeness of each utterance to an imagined *literal listener* agent L who is attempting to select the target object from alternatives in context:

$$U_{basic}(u; o, C) = \log L(o|u, C)$$

This information-theoretic expression measures how certain the listener becomes about the intended object after hearing the utterance. The literal listener is assumed to update their beliefs about the target object according to Bayesian inference, conditioning on the literal meaning of the utterance being true of it:

$$L(o|u, C) \propto \mathcal{L}(o, u)P(o)$$

where normalization takes place over objects $o \in C$ and $\mathcal{L}$ represents the lexical semantics of u. If u is true of o then $\mathcal{L}(o, u) = 1$; otherwise, $\mathcal{L}(o, u) = 0$.

This basic setup assumes that the speaker reasons about a listener sharing the same context C in common ground. How should this framework be extended to handle asymmetries in visual access between the speaker and listener, where the speaker has uncertainty over the possible distractors behind the occlusions? The most straightforward way to represent such uncertainty in a Bayesian framework is to posit a space of alternative objects $\mathcal{O}$, place a prior $P(o_h)$ over which object $o_h \in \mathcal{O}$, if any, is hidden behind an occlusion, and marginalize over these alternatives when reasoning about the listener.

Definition. This gives us a utility for conditions of *asymmetries in visual access*:

$$U_{asym}(u; o, C) = \sum_{o_h \in \mathcal{O}} P(o_h) \log L(o|u, C \cup o_h)$$

where C denotes the set of objects in context that the speaker perceives.

We define "specificity" extensionally, in the sense that if u_0 is more specific than u_1, then the objects for which u_0 is true is a subset of the objects for which u_1 is true:

Definition. Utterance u_0 is said to be *more specific* than u_1 iff $\mathcal{L}(u_0, o_h) \leq \mathcal{L}(u_1, o_h)$ $\forall o_h \in \mathcal{O}$ and there exists a subset of objects $\mathcal{O}^* \subset \mathcal{O}$ such that $\sum_{o^* \in \mathcal{O}^*} P(o^*) > 0$ and $\mathcal{L}(u_0, o^*) < \mathcal{L}(u_1, o^*)$ for $o* \in \mathcal{O}^*$.

We now show that the recursive reasoning model predicts that speakers should prefer more informative utterances in contexts with occlusions. In other words, that the *asymmetry* utility leads to a preference for more specific referring expressions than the *basic* utility.

Theorem. *If u_0 is more specific than u_1, then the following holds for any target o^t and shared context C:*

$$\frac{S_{asym}(u_0|o^t, C)}{S_{asym}(u_1|o^t, C)} > \frac{S_{basic}(u_0|o^t, C)}{S_{basic}(u_1|o^t, C)}$$

Proof. Since $S(u_0|o^t, C)/S(u_1|o^t, C) = \exp(\alpha \cdot (U(u_0; o^t, C) - U(u_1; o^t, C)))$ it is sufficient to show

$$U_{asym}(u_0; o, C) - U_{asym}(u_1; o, C) > U_{basic}(u_0; o, C) - U_{basic}(u_1; o, C)$$

We first break apart the sum on the left-hand side:

$$\begin{aligned} U_{asym}(u_0|o^t, C) - U_{asym}(u_1|o^t, C) &= \sum_{o_h \in \mathcal{O}} p(o_h) \left[\log L(o|u_0, C \cup o_h) - \log L(o|u_1, C \cup o_h)\right] \\ &= \sum_{o^* \in \mathcal{O}^*} p(o^*) \log \frac{L(o^t|u_0, C \cup o^*)}{L(o^t|u_1, C \cup o^*)} \quad (1) \\ &\quad + \sum_{o_h \in \mathcal{O} \setminus \mathcal{O}^*} p(o_h) \log \frac{L(o^t|u_0, C \cup o_h)}{L(o^t|u_1, C \cup o_h)} \quad (2) \end{aligned}$$

By the definition of "more specific" and because we defined $o^* \in \mathcal{O}^*$ to be precisely the subset of objects for which $\mathcal{L}(u_0, o^*) < \mathcal{L}(u_1, o^*)$, for objects o_h in the complementary set $\mathcal{O} \setminus \mathcal{O}^*$ we have $\mathcal{L}(u_0, o_h) = \mathcal{L}(u_1, o_h)$. Therefore, for 2, $L(o^t|u_i, C \cup o_h) = L(o^t|u_i, C)$, giving us $\log \frac{L(o^t|u_0,C)}{L(o^t|u_1,C)} \sum_{o_h \in \mathcal{O} \setminus \mathcal{O}^*} p(o_h)$

For the ratio in 1, we can substitute the definition of the listener L and simplify:

$$\begin{aligned} \frac{L(o^t|u_0, C \cup o^*)}{L(o^t|u_1, C \cup o^*)} &= \frac{\mathcal{L}(o^t, u_0)[\sum_{o \in C \cup o^*} \mathcal{L}(o, u_1)]}{\mathcal{L}(o^t, u_1)[\sum_{o \in C \cup o^*} \mathcal{L}(o, u_0)]} \\ &= \frac{\mathcal{L}(o^t, u_0)[\sum_{o \in C} \mathcal{L}(o, u_1) + \mathcal{L}(o^*, u_1)]}{\mathcal{L}(o^t, u_1)[\sum_{o \in C} \mathcal{L}(o, u_0) + \mathcal{L}(o^*, u_0)]} \\ &< \frac{\mathcal{L}(o^t, u_0)[\sum_{o \in C} \mathcal{L}(o, u_1)]}{\mathcal{L}(o^t, u_1)[\sum_{o \in C} \mathcal{L}(o, u_0)]} \\ &= \frac{L(o^t|u_0, C)}{L(o^t|u_1, C)} \end{aligned}$$

Thus,

$$\begin{aligned} U_{asym}(u_0|o^t, C) - U_{asym}(u_1|o^t, C) &< \log \tfrac{L(o^t|u_0,C)}{L(o^t|u_1,C)} \left(\sum_{o^* \in \mathcal{O}^*} p(o^*) + \sum_{o_h \in \mathcal{O} \setminus \mathcal{O}^*} p(o_h) \right) \\ &= \log L(o^t|u_0, C) - \log L(o^t|u_1, C) \\ &= U_{basic}(u_0|o^t, C) - U_{basic}(u_1|o^t, C) \end{aligned}$$

□

Note that this proof also holds when an utterance-level cost term cost(u) penalizing longer or more effortful utterances is incorporated into the utilities

$$\begin{aligned} U_{asym}(u; o, C_s) &= \textstyle\sum_{o_h \in \mathcal{O}} \log L_0(o|u, C_s \cup o_h) P(o_h) - \text{cost}(u) \\ U_{basic}(u; o, C) &= \log L(o|u, C) - \text{cost}(u) \end{aligned}$$

since the same constant appears on both sides of inequality. In principle, it can also be extended to real-valued meanings $\mathcal{L}$, though additional assumptions must be made.

Quantitative model fit for Exp. 1

In addition to the qualitative predictions derived in the previous section, our speaker model makes direct quantitative predictions about Exp. 1 data. Here, we describe the details of a Bayesian Data Analysis evaluating this model on the empirical data, and comparing it to an occlusion-blind model which does not reason about possible hidden objects.

Because there were no differences observed in production based on the particular levels of target features (e.g. whether the target was blue or red), we collapse across these details and only feed the model which features of each distractor differed from the target on each trial. After this simplification, there were only 4 possible contexts: *far* contexts, where the distractors differed in every dimension, and three varieties of *close* contexts, where the critical distractor differed in *only shape*, *shape and color*, or *shape and texture*. In addition, we included in the model information about whether each trial had cells occluded or not.

The space of utterances used in our speaker model is derived from our feature annotations: for each trial, the speaker model selected among 7 utterances referring to each combination of features: only mentioning the target's shape, only mentioning the target's color, mentioning the shape *and* the color, and so on. For the set of alternative objects $\mathcal{O}$, we used the full 64-object stimulus space used in our experiment design, and we placed a uniform prior over these objects such that the occlusion-sensitive speaker assumed they were equally likely to be hidden.

Our model has four free parameters which we infer from the data using Bayesian inference[2]. The speaker optimality parameter, α, is a soft-max temperature such that at $\alpha = 1$, the speaker produces utterances directly proportional to their utility, and as $\alpha \to \infty$ the speaker maximizes. In addition, to account for the differential production of the three features (see Fig. 2B), we assume separate production costs for each feature: a texture cost c_t, a color cost c_c, and a shape cost c_s. We use (uninformative) uniform priors for all parameters:

$$\begin{aligned} \alpha &\sim \text{Unif}(0, 50) \\ c_t, c_c, c_s &\sim \text{Unif}(0, 10) \end{aligned}$$

[2]Note that this use of Bayesian statistics in analyzing and evaluating our cognitive model is completely dissociable from the assumption of Bayesian recursive reasoning within the model.

We compute speaker predictions for a particular parameter setting using (nested) enumeration and infer the posterior over parameters using MCMC. We discard 5000 burn-in samples and then take 5000 samples from the posterior with a lag of 2. Our posterior predictives are computed from these posteriors by taking the expected number of features produced by the speaker marginalizing over parameters and possible non-critical distractors in context (this captures the statistics of our experimental contexts, where there was always a distractor sharing the same color or texture but a different shape as the target). Finally, to precisely compute the Bayes Factor, we enumerated over a discrete grid of parameter values in the prior. We implemented our models and conducted inference in the probabilistic programming language WebPPL (Goodman & Stuhlmuller, 2014). All code necessary to reproduce our model results are available at the project github: https://github.com/hawkrobe/pragmatics_of_perspective_taking.

Multi-stage bootstrap procedure for Expt. 2

The statistical dependency structure of our ratings was more complex than standard mixed-effect model packages are designed to handle and the summary statistic we needed for our test was a simple difference score across conditions, so we instead implemented a simple multi-stage, non-parametric bootstrap scheme to appropriately account for different sources of variance. In particular, we needed to control for effects of *judge*, *item*, and *speaker*.

First, to control for the repeated measurements of each judge rating the informativity of all labels, we resampled our set of sixteen *judge* ids with replacement. For each label, we then computed informativity as the difference between the target and distractor fits within every judge's ratings, and took the mean across our bootstrapped sample of judges. Next, we controlled for item effects by resampling our eight *item* ids with replacement. Finally, we resampled *speakers* from pairs within each condition (scripted vs. unscripted), and looked up the mean informativity of each utterance they produced for each of the resampled set of items. Now, we can take the mean within each condition and compute the difference across conditions, which is our desired test statistic. We repeated this multi-stage resampling procedure 1000 times to get the bootstrapped distribution of our test statistic that we reported in the main text. Individual errors bars in Fig. 4 are derived from the same procedure but without taking difference scores.

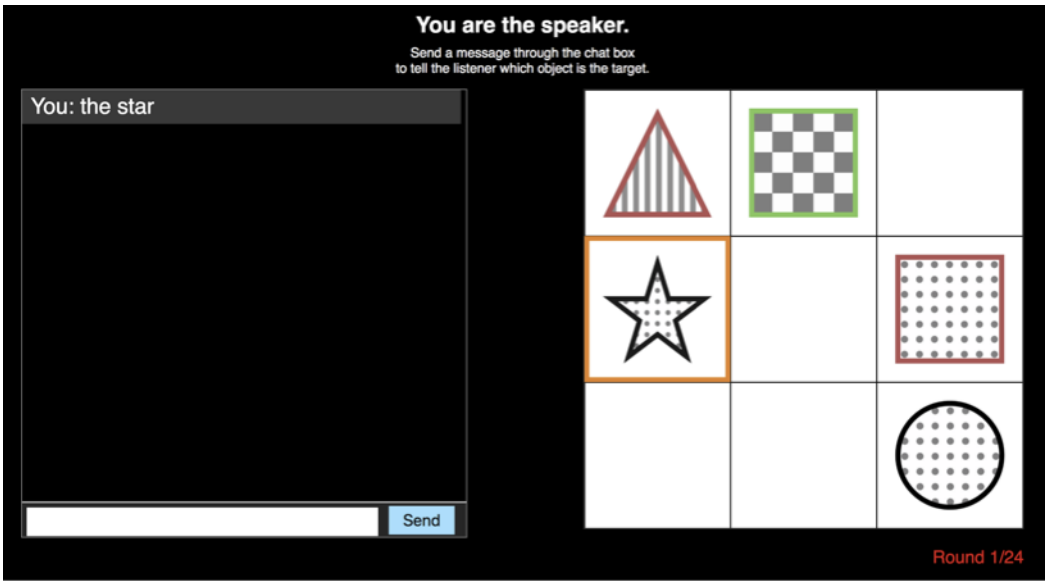

Figure 5. Screenshot of experiment interface.

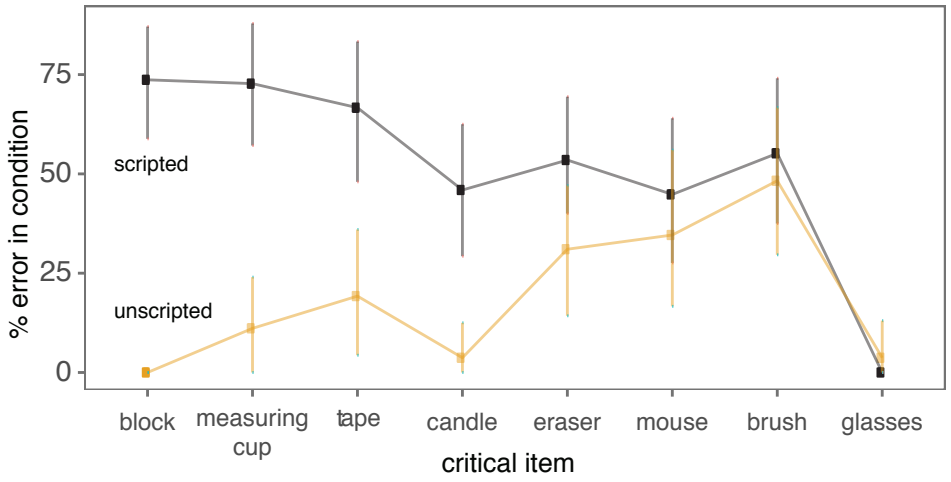

Figure 6. Supplementary figure of heterogeneity in errors across the 8 object sets used in Exp. 2 (from Keysar, 2003). Error rates across object diverge significantly from a uniform distribution in both scripted ($\chi^2 = 55, p < 0.001$) and unscripted ($\chi^2 = 36, p < 0.001$) conditions under a non-parametric χ^2 test.

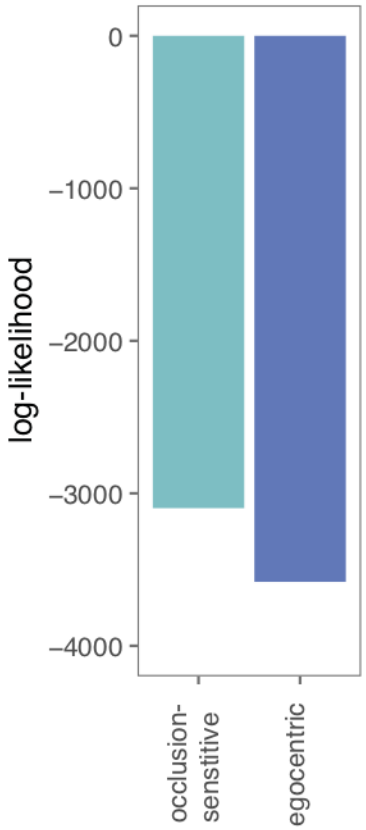

Figure 7. Supplementary figure of model likelihoods.

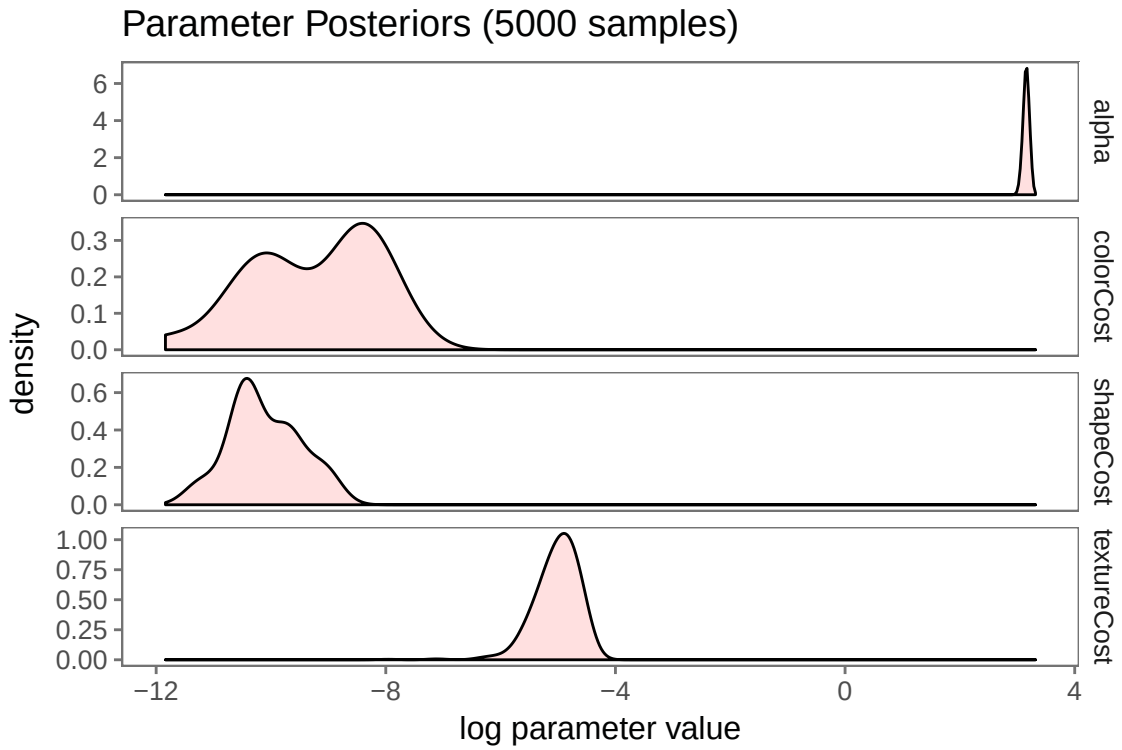

Figure 8. Supplementary figure of parameter posteriors. All parameters shown on log scale. MAP estimates with 95% highest posterior density intervals are as follows: $\alpha = 23.9, HDI = [21.0, 25.9]$; $c_{color} = 3.9 \times 10^{-5}, HDI = [2.5 \times 10^{-5}, 3.9 \times 10^{-4}]$; $c_{shape} = 2.9 \times 10^{-5}, HDI = [1.3 \times 10^{-5}, 1.1 \times 10^{-4}]$; $c_{texture} = 6.6 \times 10^{-3}, HDI = [3.1 \times 10^{-3}, 1.2 \times 10^{-2}]$