COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e12926

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12926

# The Division of Labor in Communication: Speakers Help Listeners Account for Asymmetries in Visual Perspective

Robert D. Hawkins,[a] Hyowon Gweon,[a] Noah D. Goodman[a,b]

[a]*Department of Psychology, Stanford University*
[b]*Department of Computer Science, Stanford University*

Received 28 August 2019; received in revised form 17 September 2020; accepted 4 November 2020

**Abstract**

Recent debates over adults' theory of mind use have been fueled by surprising failures of perspective-taking in communication, suggesting that perspective-taking may be relatively effortful. Yet adults routinely engage in effortful processes when needed. How, then, should speakers and listeners allocate their resources to achieve successful communication? We begin with the observation that the shared goal of communication induces a natural division of labor: The resources one agent chooses to allocate toward perspective-taking should depend on their expectations about the other's allocation. We formalize this idea in a *resource-rational* model augmenting recent probabilistic weighting accounts with a mechanism for (costly) control over the degree of perspective-taking. In a series of simulations, we first derive an intermediate degree of perspective weighting as an optimal trade-off between expected costs and benefits of perspective-taking. We then present two behavioral experiments testing novel predictions of our model. In Experiment 1, we manipulated the presence or absence of occlusions in a director–matcher task. We found that speakers spontaneously modulated the informativeness of their descriptions to account for "known unknowns" in their partner's private view, reflecting a higher degree of speaker perspective-taking than previously acknowledged. In Experiment 2, we then compared the scripted utterances used by confederates in prior work with those produced in interactions with unscripted directors. We found that confederates were systematically less informative than listeners would initially expect given the presence of occlusions, but listeners used violations to adaptively make fewer errors over time. Taken together, our work suggests that people are not simply "mind-blind"; they use contextually appropriate expectations to navigate the division of labor with their partner. We discuss how a resource-rational framework may provide a more deeply explanatory foundation for understanding flexible perspective-taking under processing constraints.

*Keywords:* Theory of mind; Pragmatics; Resource rationality; Communication

Correspondence should be sent to Robert D Hawkins, Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544. E-mail: rxdhawkins@gmail.com

## 1. Introduction

Our success as a social species depends on our ability to understand, and be understood by, different social partners across different contexts. *Theory of mind*—the ability to represent and reason about others' mental states (Premack & Woodruff, 1978)—is considered to be the key cognitive mechanism that supports such context sensitivity in our everyday social interactions. Being able to infer 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. These inferences do not necessarily come for free, however. Behavioral, developmental, and neural evidence increasingly suggests that at least some aspects of theory of mind use are computationally costly, requiring effortful processing under cognitive control (Bradford, Jentzsch, & Gomez, 2015; Brown-Schmidt, 2009b; Ferguson, Apperly, Ahmad, Bindemann, & Cane, 2015; Jouravlev et al., 2019; Long, Horton, Rohde, & Sorace, 2018; Low & Perner, 2012; Nilsen & Graham, 2009; Ryskin, Benjamin, Tullis, & Brown-Schmidt, 2015; Saxe, Schulz, & Jiang, 2006; Symeonidou, Dumontheil, Chow, & Breheny, 2016; but see Rubio-Fernández, Mollica, Ali, & Gibson, 2019).

How, then, should agents allocate their cognitive resources to successfully communicate with one another? One prominent proposal is that agents cope with these constraints by using egocentric heuristics (Barr, 2014; Keysar, 2007; Keysar, Barr, Balin, & Brauner, 2000; Keysar, Barr, & Horton, 1998). An "anchor-and-adjust" heuristic, in particular, allows agents to anchor on their own easily available perspective and effortfully adjust in the direction of another perspective to the extent that sufficient cognitive resources are available (Epley, Keysar, Van Boven, & Gilovich, 2004). Because the adjustment process satisfices at some threshold, heuristic accounts predict that optimal perspective-taking is rarely observed and communicative behavior is marked by some degree of egocentric bias (for a related two-stage account, see Barr, 2008). These accounts have provided algorithmic explanations for a variety of key phenomena, such as the increase of egocentric biases under cognitive load and the effect of individual differences in working memory (Lin, Keysar, & Epley, 2010; Roxβnagel, 2000). However, they have also been challenged by apparently contradictory evidence. A number of subsequent eye-tracking studies suggested that people are sensitive to other perspectives from the earliest moments of processing, precisely when the egocentric bias is predicted to be the strongest (Brown-Schmidt & Tanenhaus, 2008; Hanna, Tanenhaus, & Trueswell, 2003; Heller, Grodner, & Tanenhaus, 2008; Nadig & Sedivy, 2002).

Alternative accounts have been proposed to reconcile these contradictions. Under the prominent *simultaneous integration* account, listeners (Heller, Parisien, & Stevenson, 2016) and speakers (Mozuraitis, Stevenson, & Heller, 2018) consider both their own private perspective and their partner's perspective at the same time (for reviews of the broader class of constraint-based theories, which allow multiple competing sources of information to combine during online processing, see Brown-Schmidt & Heller, 2018; Degen & Tanenhaus, 2019). The simultaneous integration account is formalized as a Bayesian probabilistic weighting model, where the degree to which each perspective

contributes to the combination is given by a weighting parameter. An intermediate value of this parameter, weighting each perspective about equally, has been found to account for prior results better than a purely egocentric or purely perspective-taking strategy. This proposal offers a computational-level explanation (Marr, 1982) for why prior eye-tracking studies have found early traces of the agent's own perspective *and* their partner's.

Yet probabilistic weighting models also leave open an important question: *Why* do people use the weighting they do in a given context? What determines the degree to which people weight their egocentric perspective in different communicative scenarios? Without considering algorithmic-level processes, for example, it is difficult to explain what leads to apparently different weightings under cognitive load (Lin et al., 2010) or time constraints (Horton & Keysar, 1996), or as a function of individual differences in working memory. Heller et al. (2016) and Mozuraitis et al. (2018) discuss a potential role for the cognitive demands of inhibiting one's own perspective, but no explicit model has yet emerged that explains the flexible weighting of different perspectives in terms of more general principles of human cognition.[1]

### 1.1. The division of labor in communication

We argue in this paper for a *resource-rational* account of perspective-taking in communication that formally fills this explanatory gap. The recent development of resource-rational analysis (Griffiths, Lieder, & Goodman, 2015; Lieder & Griffiths, 2019; Shenhav et al., 2017) has provided a framework for understanding a range of costly but important cognitive functions, including attention (Padmala & Pessoa, 2011), working memory maintenance (Howes, Duggan, Kalidindi, Tseng, & Lewis, 2016), planning (Callaway et al., 2018), and decision-making under uncertainty (Lieder, Griffiths, & Hsu, 2018), through the application of rational principles under cognitive constraints. Computational-level accounts are often under-constrained: There are many solutions to the computational problem that could be considered equally "optimal" a priori regardless of how costly or intractable the required computations are. Resource-rational analyses attempt to place stronger constraints on these accounts by incorporating processing considerations. The key insight, motivated by recent work on the mechanisms of cognitive control, is that agents consider both the functional value of a computation as well as its *costs* (Kool & Botvinick, 2018; Shenhav, Botvinick, & Cohen, 2013), and behave in a way that is consistent with an approximately optimal trade-off between them. In other words, "the question of interest has begun to shift from whether an individual is *capable* of exerting cognitive effort to whether the individual will choose to do so" (Kool & Botvinick, 2013). This broader shift is consistent with recent mechanistic frameworks for language processing that argue for a central role of executive control and recurrent processing (Ferreira, 2019).

Communication presents a novel and interesting test case for resource-rational analysis because it is a fundamentally cooperative, multi-agent activity. Participants in a typical interaction share the same joint goal, and their ability to achieve this goal depends on the *joint effort* they each contribute. Collaboratively minimizing joint effort therefore sets up

a natural division of labor in communication (Clark, 1996; Ferreira, 2008; Tomasello, 2009): The effort one participant ought to exert depends on how much effort they expect others to exert. This mutual dependency poses a nontrivial representational and inferential challenge for participants. We propose a resource-rational formulation of this problem, which shares with simultaneous integration accounts the basic assumption that agents may be attending to and weighting their partner's perspective even at the outset of an interaction. Indeed, as we show in Section 2, our proposal can be seen as a straightforwardly extending the family of probabilistic weighting models. Unlike previous models, however, we provide an explicit computational explanation for how perspective weightings are set, in terms of a principled resource-rational trade-off between the expected costs and benefits of perspective-taking. Our consideration of cost also addresses the algorithmic-level concerns that motivated egocentric heuristic models. Rather than assuming agents are "reflexively mindblind" with no control over their default egocentric biases, however, resource rationality predicts that agents can anticipate the perspective-taking needs of the interaction based on various contextual factors and make flexible decisions about the resources they dedicate toward perspective-taking.

We further suggest that the appropriate consideration of contextual factors can be derived from principles of Gricean reasoning (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016). Assigning a higher weight to a partner's perspective may be expected to lead to gains in expected communicative accuracy even as it incurs a proportionally higher processing cost (i.e., in terms of cognitive resources allocated). Critically, the expected gain in accuracy depends on pragmatic inferences about the other agent's underlying effort in the current context, and the overall cost may depend on environmental modulations such as cognitive load. Hence, this model is capable of systematic context-and partner-sensitivity in the effort an agent chooses to exert. In the following section, we analyze the specific Gricean considerations at play in the director–matcher task (Keysar et al., 2000), beginning with the unique challenges facing the director. This Gricean analysis forms the basis for the initial expectations a listener ought to hold about a speaker's behavior, which in turn informs the expected value of perspective-taking for the listener.

### *1.2. Referring under uncertainty about the visual context*

The Gricean notion of cooperativity (Grice, 1975) refers to the idea that speakers try to avoid saying things that are confusing or unnecessarily complicated given the current context, and that listeners expect them to do so. 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* expression, like "the little terrier with the blue collar," even though it is more costly to produce (Brennan & Clark, 1996; van Deemter, 2016). Importantly, you might also prefer more specific labels even when

you yourself see only one dog at the moment. As long as there may be other dogs from the *listener's* point of view, or uncertainty about what the listener can see, a cooperative speaker might want to be more specific to ensure that the listener identifies the correct dog.

While sensitivity to uncertainty about a partner's visual context is natural in everyday conversations, it has often been overlooked in the design of lab experiments. We argue that the influential director–matcher paradigm (Keysar et al., 2000; Keysar, Lin, & Barr, 2003) places the speaker in an analogous situation to the speaker at the dog park. In this task, a speaker instructs a listener to move objects around a grid. Certain cells of the grid are covered to prevent the speaker from seeing some of the objects. It is therefore highly salient to the speaker that there exist hidden objects she herself cannot see but her partner can. The speaker must generate a description such that a listener can identify the correct object among distractors, even though the speaker cannot be sure what all of the distractors are.

More generally, it is helpful to differentiate between three states that may in principle be considered by each agent in a director–matcher task: (A) the contents of one's own private view, which are known to oneself but not necessarily one's partner; (B) the contents of the shared view, which are known to both oneself and one's partner; and (C) the contents of the partner's private view, which are known to one's partner but not oneself (see Fig. 1 for an illustration). For example, the version of the task introduced by Keysar et al. (2000) only placed occluders on the speaker's side of the display and focused on the extent to which listeners distinguish between A and B. C was not of interest: Because nothing was occluded from the listener's point of view (i.e., the display only used the red occluder from Fig. 1), the listener knew that C, the speaker's private view, was exactly the same as B, the shared view. Extensive work has also examined how *speakers* adjust their utterances (or fail to adjust their utterances) depending on their *own* private

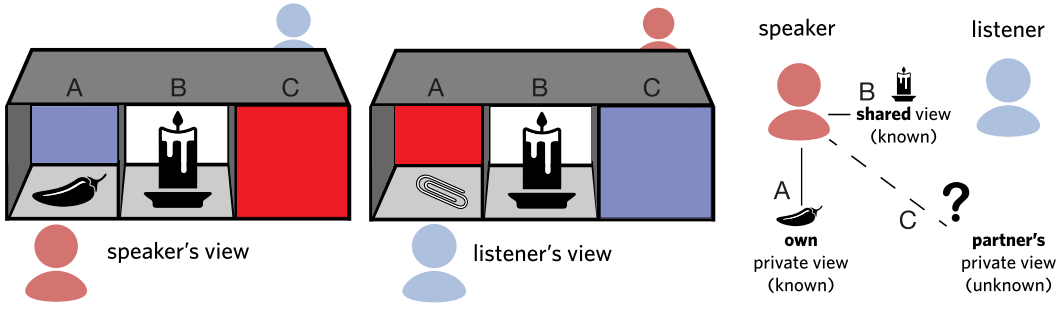


Fig. 1. Schematic illustrating the three possible states that may be considered in a director–matcher task, where both parties may have objects in their own private view that are inaccessible to the other. In the presence of occlusions, agents must not only represent the known contents of their own private view (A) versus the content shared with their partner (B), but *also* the unknown contents of their partner's private view (C). In practice, most studies place occlusions only on the speaker's side (red only) or only on the listener's side (blue only).

information (e.g., Nadig & Sedivy, 2002; Wardlow Lane, Groisman, & Ferreira, 2006), thus evaluating the extent to which the speaker accounts for differences between A and B in their production. Again, C was not of interest: Analogously to the listener studies, nothing was occluded from the speaker's point of view (i.e., the display only used the blue occluder in Fig. 1), so the speaker knew that C, the listener's private view, was exactly the same as B, the shared view.

Yet we still understand relatively little about the extent to which speakers naturally consider their own uncertainty about C, their *partner's* private information, in scenarios like the one used by Keysar et al. (2000), where C is not identical to B. The possible objects behind the occluder are salient "known unknowns" that may influence a Gricean speaker's choice of referring expression, even if they have no private information of their own, that is, even if A and B are identical. It also remains unclear how Gricean listeners ought to account for such behavior in their own initial expectations. Because prior work investigating listener perspective-taking has commonly used confederates in the speaker role, it is possible that confederate behavior may have interacted with these expectations.

### *1.3. The current work*

Our first goal is to derive and test Gricean predictions about how speakers should produce referring expressions under conditions of uncertainty about the listener's visual context. As shown below, our model predicts that a speaker will compensate for her uncertainty about the listener's visual context by increasing the informativity of her utterance to some extent beyond what she would produce in a completely shared context. In Experiment 1, we directly test this prediction by manipulating the presence and absence of occlusions in a simplified variant of the director–matcher task.

Our second goal is to examine the consequences of this observation for the listener's allocation of effort. The behavior observed in Experiment 1 establishes reasonable baseline expectations that listeners should use when deciding how much perspective-taking effort to allocate in the director–matcher task. In Experiment 2, we conduct a replication of the landmark study reported by Keysar et al. (2003). We compare the replicated findings in this *scripted* condition with a new *unscripted* condition to evaluate the gap between the scripted referring expressions used by confederate speakers in prior work and what a naive speaker without a script would naturally say in the same interactive context (Bavelas & Healing, 2013; Brown-Schmidt, 2009a; Kuhlen & Brennan, 2013; Tanenhaus & Brown-Schmidt, 2008). Our model predicts that listeners will initially make more errors with confederate speakers (who are less informative than expected under a natural division of labor) compared with naive speakers. Critically, it also predicts that the gap will decrease over time; listeners in the confederate condition will gradually devote more effort to perspective-taking as they learn that the confederate is devoting less effort.

Taken together, this work aims to establish the plausibility of a resource-rational basis for some degree of perspective neglect on the part of both speaker and listener, and to emphasize the role of pragmatic expectations in determining this division of

labor. It is important to note that our aim is to extend the explanatory power of recent probabilistic weighting models, not to falsify them. In fact, if we are successful in deriving from more basic principles the perspective-weighting proportions that were previously fit to empirical data, our model will necessarily make similar behavioral predictions for those experiments. Consequently, our experiments were designed to expose and test the novel predictions of our extension, placing probabilistic weighting models on a firmer foundation, not necessarily to construct scenarios challenging the broader simultaneous integration view. We clarify this theoretical relationship in the following section and return to the broader implications and predictions of the resource-rational view in the discussion.

## 2. A resource-rational analysis of perspective-taking

In this section, we formally derive the core predictions of our resource-rational analysis. We begin with a brief review of the Rational Speech Act (RSA) framework, which formalizes pragmatic reasoning as recursive probabilistic inference, and define a new "ideal observer" model of perspective-taking under uncertainty about a partner's visual context. This model can then be mixed with an egocentric model, using the same probabilistic weighting mechanism proposed by Heller et al. (2016). Finally, we conduct an analysis of the optimal parameter value for this mixture model given the additional assumption that there is higher cognitive cost to higher perspective-weighting.

### 2.1. Preliminaries

The RSA framework derives language behavior from basic Gricean mechanisms of recursive social reasoning (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016; Goodman & Stuhlmüller, 2013; Kao, Wu, Bergen, & Goodman, 2014). In this framework, a pragmatic speaker $S$ is 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 $U$, capturing the trade-off between the cost, or effort, of producing an utterance and the usefulness of each utterance for an imagined listener agent. In the context of the director–matcher task, the listener is a matcher who hears a referring expression $u$ in a context $C$ containing different objects and must select the target object $o$. They do so by inverting their generative model of the speaker. This formulation introduces a mutually recursive dependency between the speaker and the listener. A key idea of the RSA framework is to introduce a "base case" where this recursion bottoms out. Specifically, we define a "literal listener" $L_0$ who updates their beliefs about which object is the target of reference using the literal meaning of the utterance, $\mathscr{L}(u,o)$. In our referential context, $\mathscr{L}$ simply represents a simple lexical semantics for $u$: If $u$ is true of $o$ (i.e., if $u$ is "square" and $o$ is actually a square), then $\mathscr{L}(o,u) = 1$; otherwise, $\mathscr{L}(o,u) = 0.01$. The literal listener then serves as the foundation for a chain of additional layers of recursive reasoning:

$$\begin{aligned} P_{L_0}(o|u,C) &\propto \mathscr{L}(o,u)P(o) \\ P_{S_1}(u|o,C) &\propto \exp\{\alpha U(u;o,C)\} \\ U(u;o,C) &= \log P_{L_0}(o|u,C) - \text{cost}(u) \\ P_{L_1}(o|u,C) &\propto P_{S_1}(u|o,C)P(o) \end{aligned} \tag{1}$$

where normalization takes place over objects $o \in C$ or utterances $u \in \mathscr{U}$.

### 2.2. Reasoning about asymmetries in visual access

This basic set-up assumes that the speaker reasons about a listener sharing the full context $C$ in common ground, that is, that the entire display is in state (B) of Fig. 1. But how does a speaker refer to a target object when they know their partner has additional, unknown distractor objects in their private view, as in the scenario from Keysar et al. (2000)? Models which contrast the egocentric domain of reference against what is shared in common ground would predict no difference in speaker production between this scenario and one with no occlusions at all. After all, because the speaker is not shown any private information, the information in the speaker's egocentric perspective, state (A), is equivalent to the information they know to be in common ground, state (B): All visible objects in the speaker's view are also clearly visible to the listener. The relevant perspective at issue for evaluating the speaker's perspective-taking in this scenario is not the content of the shared view, but instead the (unknown) private contents of the *listener's* visual field, state (C).

In the RSA framework, speaker uncertainty about the listener's visual field is represented straightforwardly by a probability distribution. For example, Goodman and Stuhlmüller (2013) examined a case where the speaker has limited perceptual access to the objects they are describing, and derived how a pragmatic listener who is taking the speaker's perspective should interpret the speaker's utterances in light of such limited access. In the case of the director–matcher task studied in this paper, the latent state of the world is the space of objects $\mathscr{O}$ seen by one's partner. Because the speaker knows that objects may be behind occluders, we introduce uncertainty $P(o_h)$ over which object $o_h \in \mathscr{O}$, if any, is hidden behind each occlusion. The speaker ought to then marginalize over these alternatives when reasoning about which object a literal listener will select from the set of objects in their view. This gives us a speaker utility under conditions of *asymmetries in visual access*:

$$U_{S_1}^{\text{asym}}(u;o,C) = \sum_{o_h \in \mathscr{O}} P(o_h) \log P_{L_0}(o|u, C \cup o_h) - \text{cost}(u), \tag{2}$$

where $C$ still denotes the set of objects that the agent knows to be in common ground. Conversely, we can define an *egocentric* speaker who ignores the possible existence of hidden objects that only the listener can see and only seeks to be informative relative to

the objects in their own view (which, again, happens to be identical to the common ground):

$$U_{S_1}^{\text{ego}}(u;o,C) = \log P_{L_0}(o|u,C) - \text{cost}(u) \tag{3}$$

The analogous asymmetry-aware and egocentric models for the listener are more straightforward. Because nothing is occluded from their own view, they have full information about exactly which objects are known to each person. The egocentric listener model chooses from the full set of objects in their view while the asymmetry-aware listener excludes any objects that are private, only considering the set of objects in the speaker's view.

$$\begin{aligned} P_{L_i}^{\text{asym}}(o|u,C) &= P_{L_i}(o|u,C-\{o_h\}) \\ P_{L_i}^{\text{ego}}(o|u,C) &= P_{L_i}(o|u,C) \end{aligned} \tag{4}$$

### 2.3. A probabilistic weighting model

The utility in Eq. 2 represents an "ideal" perspective-taking speaker, while the utility in Eq. 3 represents a completely egocentric speaker. Next, we follow Heller et al. (2016) in allowing for a probabilistic mixture between these two perspectives using an interpolation weight $w_S \in [0,1]$:

$$U_{S_1}^{\text{mix}}(u;o,C,w_S) = w_S \cdot U_{S_1}^{\text{asym}}(u;o,C) + (1-w_S) \cdot U_{S_1}^{\text{ego}}(u;o,C) \tag{5}$$

When $w_S = 0$, the speaker using this utility is purely "occlusion-blind" or egocentric: She assumes her partner sees exactly the same objects she herself does.[2] When $w_S = 1$, this speaker is purely "occlusion-sensitive": She assumes there may be additional objects in her partner's view that she cannot see behind the occlusions. Similarly, we define a mixture model for the listener, with $w_L = 0$ corresponding to the purely egocentric domain and $w_L = 1$ corresponding to the objects in common ground (i.e., the speaker's perspective):

$$P_{L_i}^{\text{mix}}(o|u,C,w_L) \propto w_L \cdot P_{L_i}^{\text{asym}}(o|u,C) + (1-w_L) \cdot P_{L_i}^{\text{ego}}(o|u,C) \tag{6}$$

A critical point of difference between Heller et al. (2016) and our recursive RSA model formulation, however, is that we assume that occlusion-aware speakers and listeners account for the fact that their partner is also a mixture model with some (unknown) mixture weight. Introducing this dependency between agents, in terms of maintaining explicit beliefs about a partner's mixture weight, is a key step toward formalizing the division of labor. We therefore revise Eqs. 2 and 4 as follows:

$$\begin{aligned} U^{\text{asym}}_{S_1}(u;o,C,w_L) &= \sum_{o_h \in \mathcal{O}} P(o_h) \log P^{\text{mix}}_{L_0}(o|u, C \cup o_h, w_L) - \text{cost}(u) \\ P^{\text{asym}}_{L_1}(o|u,C,w_S) &\propto P^{\text{mix}}_{S_1}(o|u, C - \{o_h\}, w_S) P(o) \end{aligned} \tag{7}$$

and update Eqs. 5 and 6 to pass this parameter through:

$$\begin{aligned} U^{\text{mix}}_{S_1}(u;o,C,w_S,w_L) &= w_S \cdot U^{\text{asym}}_{S_1}(u;o,C,w_L) + (1-w_S) \cdot U^{\text{ego}}_{S_1}(u;o,C) \\ P^{\text{mix}}_{L_i}(o|u,C,w_L,w_S) &\propto w_L \cdot P^{\text{asym}}_{L_i}(o|u,C,w_S) + (1-w_L) \cdot P^{\text{ego}}_{L_i}(o|u,C) \end{aligned} \tag{8}$$

These equations derive speaker and listener behavior for a *particular* expected value of their partner's mixture weight. Next, we assume agents have uncertainty about the exact weight their partner is using, and marginalize over it when choosing an action. In this way, we obtain the theoretical dependency between mixture weights that is characteristic of a division of labor: One agent's behavior at a particular mixture weight will differ depending on the mixture weight they think their partner is using. The final models are given as follows:

$$\begin{aligned} P_{S_1}(u|o,C,w_S) &\propto \exp\left\{ \alpha \int_{w_L} P(w_L) U^{\text{mix}}_{S_1}(u;o,C,w_S,w_L) dw_L \right\} \\ P_{L_1}(o|u,C,w_L) &\propto \int_{w_S} P(w_S) P^{\text{mix}}_{L_1}(o|u,C,w_L,w_S) dw_S \end{aligned} \tag{9}$$

To build intuition about the behavior of these models, it is useful to consider a few example cases. First, consider the behavior of the literal listener at the extreme values of $w_L$: When $w_L$ is close to 1, the listener fully considers the speaker's perspective and will never select an occluded object, even if it perfectly matches the description. When $w_L$ is close to 0, it will select an occluded object that matches the description exactly half of the time. Intermediate values of $w_L$ interpolate between these cases, leading to lower but non-zero probability of selecting the occluded object.

Now, consider the behavior of a pragmatic speaker model that decides which utterance to produce by reasoning about this literal listener. If the speaker's mixture weight $w_S$ is close to 0, then it does not consider the possible existence of occluded objects and produces a description that is only sufficient to disambiguate the target from alternatives in its *own* view. If $w_S$ is close to 1 then the speaker's decision depends purely on the *mixture weight the literal listener is expected to be using*. When $w_L = 1$, the listener will always correctly pick the sole object that matches the description in the speaker's view, irrespective of how minimal a description is given, so there is no benefit to producing a more detailed utterance. Conversely, when $w_L = 0$, then shorter utterances are risky: There are more possible hidden objects $o_h$ that would match a shorter description. Every additional feature the speaker mentions helps guard against a broader class of potential hidden objects, so it may be worth incurring the additional production cost to add information (for a more extensive proof of this

behavior, see Appendix A). When the speaker marginalizes over their prior expectations about the value of $w_L$, these behaviors are combined: The speaker model errs on the side of more informative utterances, to hedge against the risks of lower values of $w_L$.

### 2.4. Resource-rational analysis

We now conduct a resource-rational analysis of these mixture models. We find the optimal weight, accounting for both costs and benefits of allocating cognitive resources to perspective-taking. We consider the trade-off between one specific benefit (the expected value of communicative accuracy) and one specific cost (the cognitive cost of perspective-taking). We define the former value as the expected probability of the listener choosing the true target. At each level of speaker perspective-taking $w_S$, the speaker agent will prefer some utterance $u^*$; they can then compute the probability of $L_0$ selecting the target after hearing this utterance. Similarly, at each level of listener perspective-taking $w_L$, the listener agent will have some likelihood of selecting the target upon hearing the different speaker utterances. In both cases, the agents have uncertainty about their partner's level of perspective-taking and must therefore compute expected accuracy by marginalizing over the weight prior.

If communicative accuracy were the only consideration, it would always be preferable to use maximal perspective-taking (i.e., $w_S = w_L = 1$), since higher perspective-taking leads to higher accuracy. In a resource-rational model, however, these benefits are traded off against the costs of perspective-taking. For simplicity, we assume that cost is linear in the degree of perspective-taking and use $\beta$ to denote the slope of this linear term. It is unclear whether there exists a process-level algorithm for perspective-taking where this linearity holds exactly, but our analysis holds under the weaker condition that cost is strictly increasing (for now, we maintain an abstract notion of "cost" encompassing multiple processing considerations; see General Discussion for a more detailed interpretation of these costs).

Our analysis proceeds by running the $S_1$ and $L_1$ models in Eq. 9 with different choices of $w_S$ and $w_L$, respectively. In cognitive terms, this corresponds to an introspective speaker and listener meta-cognitively simulating the costs and benefits of exerting each amount of perspective-taking effort.

$$\begin{aligned} U_{S_{RR}}(w_S) &= \mathbb{E}_{P(w_L)}[P_{L_0}(o|u^*, C, w_L)] - \beta \times w_S \\ U_{L_{RR}}(w_L) &= \mathbb{E}_{P(w_S)}[P_{L_1}(o|u^*, C, w_L)] - \beta \times w_L \end{aligned} \tag{10}$$

where in both cases $u^*$ is the utterance produced by the speaker model using weight $w_S$:

$$u^* = \arg\max_u P_{S_1}(u|o, C, w_S).$$

We define the optimal weights as the arguments for which this utility is maximized:

$$\begin{aligned} w_S^* &= \arg\max_{w_S}[U_{S_{RR}}(w_S)] \\ w_L^* &= \arg\max_{w_L}[U_{L_{RR}}(w_L)] \end{aligned} \tag{11}$$

To derive concrete simulation results, we set $\alpha = 2$ and $\text{cost}(u) = 0.03$ for all $u$, and sweep over different values of $\beta$. The utterance space, object space, and context $C$ are based on the ones we use below in Experiment 1: Objects varied in shape, color, and texture, and the speaker model was able to produce any combination of shape, color, and texture descriptors. To simplify analytic enumeration over these spaces, we set the target to be a particular setting of features (i.e., "color 1, texture 1, shape 1") and represented other objects and utterances in terms of whether they match the target on each dimension (e.g., "same color, different texture, different shape"). We used uniform priors over the identity of a single hidden object, $P(o_h)$, as well as when taking internal expectations over $w_S$ and $w_L$.

Results of this analysis are shown in Fig. 2. As suggested above, when there is no cost associated with perspective-taking (i.e., $\beta = 0$), the expected likelihood of communicative success increases monotonically as a function of perspective-taking weight and there is no reason to weight one's own perspective (i.e., $w_S = w_L = 1$ are optimal). Once we factor in a non-zero cost for perspective-taking, however, the increased likelihood of communicative success at higher weights begins to be offset by the corresponding increase in

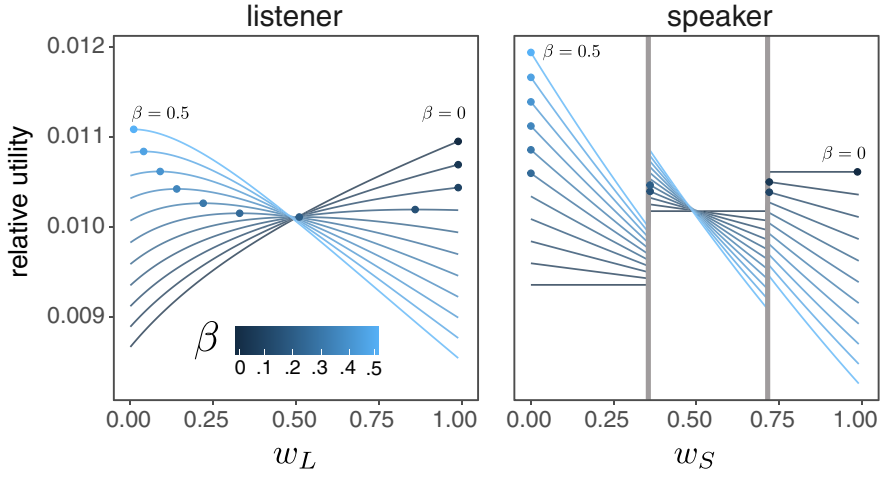


Fig. 2. Resource-rational analysis of speaker and listener models. Each curve represents the relative utility of adopting different weights at a particular cost regime, and the point on each curve represents the weight where utility is maximized. Above a certain value of β (i.e., if perspective-taking is sufficiently effortful), an intermediate weighting of perspective-taking is boundedly optimal. The discontinuities in the speaker plot occur when a higher level of perspective-taking motivates the speaker to switch to a longer utterance (e.g., "the blue square" instead of "the square" at $w_S = 0.36$, followed by "the blue checked square" at $w_S = 0.72$).

effort required to achieve it. Above a certain β, we find that an intermediate perspective weighting is optimal for both speaker and listener. That is, once perspective-taking has a certain cost, a resource-rational agent will choose to weight their partner's perspective to a lesser extent. For instance, at $\beta = 0.2$, we find that the optimal speaker weight is $w_S^* = 0.36$ and the optimal listener weight is $w_L^* = 0.51$. At even higher values of β, the optimal weighting eventually drops to zero, theoretically reaching a regime where any degree of perspective-taking at all is too costly to be justified. This simulation demonstrates the explanatory logic of the resource-rational framework, deriving the conditions under which the intermediate probabilistic weightings empirically measured by Heller et al. (2016) and Mozuraitis et al. (2018) emerge from deeper underlying computational principles: specifically, the trade-off between the costs and benefits of different degrees of perspective-taking.

#### *2.4.1. Two qualitative predictions*

We highlight two key predictions of this formulation which motivate our experiments. First, our proposal for a basic asymmetric speaker utility in Eq. 2 already leads to a novel prediction about speaker behavior in the presence of "known unknowns" hidden by occlusions. This formulation goes beyond the speaker model of Mozuraitis et al. (2018), which only considers the case where the speaker has perfect knowledge of the mismatch between their own private information and the listener's private information. Specifically, as we show analytically in Appendix A, our model qualitatively predicts that speakers will anticipate possible confusion from the listener's perspective, and produce additional information beyond what would be necessary from their own viewpoint. Note that such additional information would be unnecessary if the listener were expected to use perfect perspective-taking (i.e., if the speaker believed $w_L = 1$); the functional need to increase informativity arises only when speakers assign nonzero probability to the possibility that listeners would act egocentrically. This prediction is not strictly a consequence of the speaker's own resource-rational trade-off (it is expected to emerge to some degree at any $w_S > 0$); however, it is a foundational assumption on which the rest of our resource-rational modeling rests and is therefore the first target of our empirical investigation in Experiment 1.

Second, a key prediction distinguishing the resource-rational framework from a "fixed capacity" egocentric heuristic model is that agents may flexibly adjust the effort dedicated to perspective-taking depending on contextual factors. The optimal level of perspective-taking for one agent depends on reasoning about expected communicative success. Expected success, in turn, depends on the perspective-taking weight being used by the other agent. Both agents bring into the interaction some prior expectations about this weight, but by comparing their partner's behavior to what would be expected at different levels of perspective-taking, they can update these beliefs. These updated beliefs lead to different expectations of future communicative success and may therefore shift the optimal level of their own perspective-taking. In other words, our model predicts that agents will adapt their own perspective-taking effort to their partner's to maintain a resource-rational trade-off.

We suggest that these mechanisms may help shed further light on the errors made by listeners (matchers) in Keysar et al. (2003). Specifically, the scripted referring expressions produced by confederate speakers in the director role may have been *less informative* than what listeners in the matcher role would naturally expect from a cooperative speaker, leading to an initially mis-calibrated level of listener perspective-taking. In Appendix B, we simulate a resource-rational listener agent playing a director–matcher task with a speaker who systematically produces less informative utterances than expected under the prior. As expected, we find that the listener model gradually increases their own perspective-taking weight as they make stronger inferences about their partner's effort. In Experiment 2, we test this prediction in two ways. First, we evaluate the actual gap between natural speaker behavior and confederate speaker behavior. Second, we evaluate the extent to which listeners adapt over subsequent rounds.

## 3. Experiment 1: Speaker production under uncertainty about the listener's visual context

Occlusions blocking the speaker's view were originally used to evaluate the effort required of the listener, who must think about which cells in their own private view are visible from the speaker's view. However, our model highlights that the same occlusions also demand perspective-taking, vis-à-vis pragmatic audience design, on the part of the speaker. The speaker must anticipate the level of informativity that would be most appropriate given the possibility of hidden distractors behind the occlusions, which are visible only to the listener. To test this novel prediction of our asymmetric speaker model, we designed a simplified version of the director–matcher task that allows us to causally isolate the effect of occlusions on production.

Our task used a space of stimuli that varied along a fixed set of feature attributes, similar to previous work using shape and color contrasts (Hanna & Brennan, 2007; Hanna et al., 2003) or size contrasts (Heller et al., 2016; Nadig & Sedivy, 2002). To allow participants to interact in real time, we developed a multiplayer web experiment allowing participants to be automatically paired and to communicate with one another through an instant-messaging interface (Hawkins, 2015). Instant-messaging via text differs in many ways from face-to-face verbal communication (for a detailed discussion of these differences, see Section 4.3). Critically, however, our online task environment preserves real-time interactivity between naive participants, which is known to produce significantly different perspective-taking behavior than designs using prerecorded utterances (e.g., Brown-Schmidt, 2009a).

Note that this task is not designed to ask whether speakers produce perfectly "optimal" referring expressions by some absolute standard—it is implausible that they would know the true underlying distribution of hidden objects within the context of this task, and as our model formalizes, they would face their own resource constraints even if they did. Instead, our prediction is qualitative: Do speakers spontaneously produce more

informative referring expressions in the presence of occlusions than they do in the absence of occlusions?

### 3.1. Methods

#### 3.1.1. Participants

We recruited 102 pairs of participants from Amazon Mechanical Turk. 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 addition to a fixed base payment of $1.00, each participant could receive a performance bonus of up to 24 cents to incentivize engagement.

#### 3.1.2. Materials and procedure

Participants were automatically paired into dyads and placed in a virtual environment containing a $3 \times 3$ grid of objects on the right side of the screen and a standard instant-messaging interface on the left (for a screenshot of the full graphical interface, see Fig. S1). The objects shown in the grid were drawn from a stimulus space of objects varying along three discrete features (*shape*, *texture*, and *color*), each of which took four possible values for a total of $4 \times 4 \times 4 = 64$ possible objects. One participant in the dyad was randomly assigned to the *director* role, and the other was assigned to the *matcher* role. They proceeded through a series of trials of a director–matcher game. On each trial, one object in the grid was privately highlighted for the director as the *target*. They were instructed to use the free chat interface to communicate the identity of the target to the matcher. Matchers were also able to freely and interactively respond through the chat box. After the matcher clicked one of the objects in their private display, both participants received feedback before advancing to the next trial. The identity of the true target was revealed to the matcher, and the matcher's selection was revealed to the director. The complete text of the instructions given to participants is available in our open materials. Participants were explicitly told that they would play a game with another human, asked not to use degenerate spatial locations (e.g., "2nd row from the top, 3rd column from the left"), and informed on some trials, curtains will appear that block the view of the speaker. To make the graphical depiction of occlusions clear, we showed an example of the speaker's and listener's views side by side, where there is a green circle on the listener's side that the speaker cannot see. They were told that they will see a 2D view, but they can imagine that they are seated across a table from their partner with cubby holes in between. Each participant was required to pass a comprehension quiz covering these instructions before proceeding to the task.

There were four conditions, forming a within-pair $2 \times 2$ factorial design. The key manipulation was the presence or absence of occlusions (see Fig. 3, rows). On "occlusion-absent" trials, all objects were seen by both participants, but on "occlusion-present" trials, two randomly selected cells of the grid were covered with occluders (curtains) from

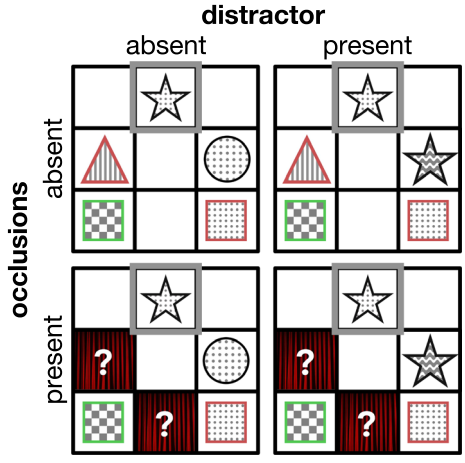


Fig. 3. Stimuli in 2 × 2 design used in Experiment 1 as seen by the speaker. Gray square indicates target.

the speaker's viewpoint such that only the listener could see the contents of the cell. As a baseline, we also included an explicit informativity manipulation (e.g., Brennan & Clark, 1996; Dale & Reiter, 1995; Monroe, Hawkins, Goodman, & Potts, 2017; Pechmann, 1989). On "distractor-absent" trials, the target was the only object with a particular shape; on "distractor-present" trials, there was a distractor with the target's shape in view for both participants, differing only in color or texture (see Fig. 3, columns). When this distractor was present, a shape-only referring expression (e.g., "star") would no longer be sufficient to discriminate the target even among *visible* objects.

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 8 trials. Displays were procedurally generated on the fly to satisfy the constraints of the given trial type, using the following algorithm. First, a target was randomly selected from the full space of 64 objects. Second, a set of distractors were selected from the remaining 63 objects. On trials in the "distractor-present" condition, one of these distractors was constrained to have the same shape as the target. Otherwise, distractors were chosen to be fillers with a different shape and randomly selected colors and textures. We randomized the total number of distractors in the display (between 2 and 4) as well as the number of those distractors covered by curtains (1 or 2) on "occlusion-present" trials. This randomization procedure prevented participants from picking up on statistical patterns of the identity or quantity of hidden objects on any particular trial. If there were only two distractors, we did not allow both of them to be covered: There was always at least one mutually visible distractor. Because the distractor-present condition required the distractor with the same shape to be mutually visible, one consequence of the design was that there was never a hidden distractor with the same shape as the target.

Finally, we collected mouse-tracking data as a window into the matcher's real-time decision-making process. Mouse-tracking is commonly used as a continuous measure of

competition in psycholinguistics (Freeman et al., 2011; Spivey, Grosjean, & Knoblich, 2005), including in prior studies of perspective-taking (van der Wel, Sebanz, & Knoblich, 2014). While mouse-tracking measures differ from eye-tracking measures in several ways —for one, mouse movements are represented by continuous trajectories while eye movements are represented by discrete saccades—the two measures are still tightly related. For example, cursor movements have been found to be correlated with gaze in web browsing (Chen, Anderson, & Sohn, 2001; Rodden, Fu, Aula, & Spiro, 2008). To collect mouse movements, we asked the matcher to wait on an empty grid at the beginning of each trial while the director typed their 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 100 Hz from the matcher's mouse in the decision window after this click, until the point when they started to move one of the objects. While we did not intend to analyze these data for Experiment 1, we anticipated using it in our second experiment below and decided to use the same procedure across experiments for consistency.

### 3.2. Results

Our primary measure of speaker behavior is the length (in words) of naturally produced referring expressions sent through the chat box. We tested differences in speaker behavior across conditions using a mixed-effect regression predicting the number of words produced on each trial. We included dummy-coded fixed effects of distractor presence and occlusion presence, as well as their interaction. Following Barr, Levy, Scheepers, and Tily (2013), we included the maximal random effect structure that converged, including random intercepts as well as random slopes for both distractor presence and occlusion presence at the dyad level. Because we procedurally generated unique displays on each trial, there was no finite set of "items" with clustered data to include in the model. A model minimally adding random intercepts for each of the 64 target objects failed to converge. We approximate degrees of freedom using Satterthwaite's method.

First, we examine the key simple effect of occlusion in "distractor-absent" contexts (Fig. 3, left column), which are most similar to the displays used in prior work using the director–matcher task. We found that speakers used significantly more words on average ($d = 1.3$ words) when they knew that additional objects could potentially be visible to their partner ($t(120.3) = 8.8$, $p < .001$). Second, we examined the simple effect of whether a distractor of the same shape as the target was present in an unoccluded display (Fig. 3, top row). We found that speakers used significantly more words on average ($d = 0.6$ words) when a distractor was present ($t(206) = 5.7$, $p < .001$; see Fig. 4A). This finding is consistent with extensive previous work evaluating speaker informativity in the experimental pragmatics literature. In (unoccluded) scenes where multiple objects share attributes, speakers naturally modulate their utterances to disambiguate target objects along contrastive dimensions (Brennan & Clark, 1996; Davies & Arnold, 2019; van Deemter, 2016), even in larger displays (Brown-Schmidt & Konopka, 2011). Lastly, we found a significant interaction ($b = -0.49$, $t(1742) = 4.1$, $p < .001$) where the effect of occlusion was larger in distractor-absent trials, likely reflecting a ceiling on the level of

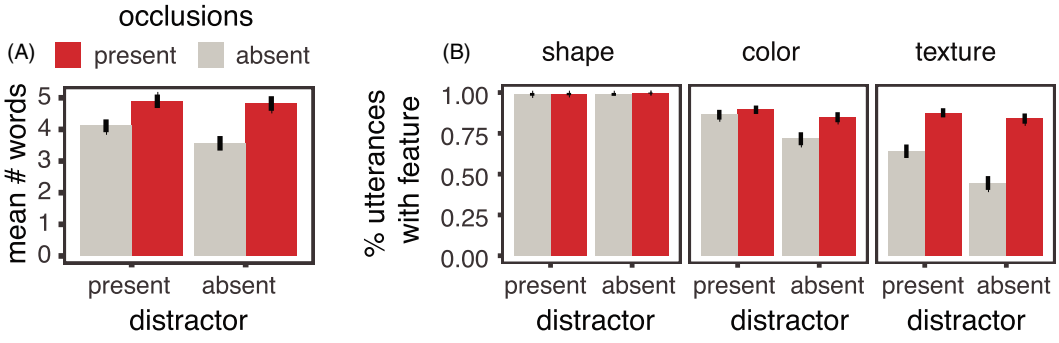


Fig. 4. Results for Experiment 1. (A) Speakers used significantly more words when occlusions were present. (B) Utterances broken out by feature mentioned. Error bars on empirical data are bootstrapped 95% confidence intervals; model error bars are 95% credible intervals.

informativity required to individuate objects in our simple three-dimensional 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, and color). Because speakers nearly always mentioned shape (e.g., "star," "triangle") as the head noun of their referring expression regardless of context (~99% of trials), differences in utterance length across conditions must be due to differentially mentioning the other two features (color and texture). To test this observation, we ran separate mixed-effect logistic regressions to predict color and texture mentions. We included fixed effects of occlusion, distractor, and their interaction. We again included random intercepts and slopes for each speaker, but no random interaction. We found simple effects of occlusion in distractor-absent contexts for both features ($b = 1.6$, $z = 3.2$, $p = .001$ for color; $b = 5.6$, $z = 6.8$, $p < .001$ for texture; see Fig. 4B). In other words, in displays like the left column of Fig. 3 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 was sufficient to disambiguate the target from visible distractors in both cases. The baseline asymmetry between production of color and texture modifiers in unoccluded contexts is consistent with prior work on over-specification (e.g., Degen, Hawkins, Graf, Kreiss, & Goodman, 2020; Tarenskeen, Broersma, & Geurts, 2015). Listener errors were rare: The target failed to be selected on only 2.5% of trials, and we found no significant difference in error rates across the four conditions ($\chi^2(3) = 1.23$, $p = .74$).

Finally, we inferred the speaker's probabilistic perspective weighting parameter using a quantitative Bayesian model comparison (for details, see Appendix C). We found that the inferred mixture was near the maximal endpoint allowed by our model ($w_S \approx 1$), suggesting that people's behaviors were better described by an occlusion-sensitive speaker model that considers possible hidden objects (i.e., Eq. 2), relative to an egocentric speaker model that considers only the objects in its own view (i.e., Eq. 3), or a mixture of the two.

### 3.3. Discussion

Our results provide strong evidence supporting our model's foundational prediction that speakers increase their level of specificity in the face of occlusions. Speakers spontaneously spent additional time and keystrokes to give further information beyond what they produced in unoccluded contexts, even though that information would be redundant given the visible objects in their own display. The effect of occlusions on referring expressions was even larger than the classic pragmatic effect of having similar distractors in the display. Critically, rather than planning their utterance purely in light of objects shared in *common ground*, which was held constant across occlusion conditions, this finding shows that speakers plan their utterance relative to their uncertainty about what the *listener* privately knows.

Sensitivity to the listener's private information may also manifest in speaker behavior in other ways. For example, Brown-Schmidt, Gunlogson, and Tanenhaus (2008) found that participants naturally asked their partner questions about occluded objects in a task that required mutual knowledge about these objects for success. Indeed, actively highlighting and resolving known sources of uncertainty about a partner's private information may be one of the primary functions of questions in discourse (Hawkins, Stuhlmüller, Degen, & Goodman, 2015; Rothe, Lake, & Gureckis, 2018). Further work is needed to determine the resource-rational trade-offs of asking explicit questions about hidden distractors as opposed to implicitly increasing the specificity of one's referring expressions to account for them, as we found here. Especially in the setting we consider, where the exact identity of the distractors was not task-relevant, simply increasing informativity may be a less costly strategy.

At the same time, the evidence for an intermediate mixture of perspectives was less clear in our task; rather, the inferred weight that speakers placed on their partner's perspective was at ceiling, $w_S = 1$, and the mixture model was rejected in a direct comparison. This contrasts with the findings from Mozuraitis et al. (2018), where the inferred speaker weight was close to the midpoint, and the endpoint value of $w_S = 1$ was explicitly rejected as inconsistent with the data. Our inferred weight is closer to the findings of Heller & Stevenson (2018), which found a much higher estimated weight of $w_S = 0.92$ in a different task. In that case, however, the endpoint was still statistically rejected in favor of the mixture. One explanation for the ceiling levels of perspective-taking we observed is that our simplified variant of the director–matcher task was too "easy": It did not place participants under sufficiently high cognitive load for resource considerations to play a meaningful role in their decisions about perspective-taking. Indeed, our resource-rational analysis in Section 2.4 predicted high levels of perspective-taking by both speakers and listeners in regimes where the cost of perspective-taking (i.e., the $\beta$ parameter) is sufficiently low.

This suspicion was further supported by pilot work in which we attempted to examine matcher errors using the same simplified design (for further details about this pilot experiment, see Appendix D). When paired with an (artificial) confederate who deliberately produced referring expressions that were ambiguous between a mutually visible object

and an occluded one, participants in the matcher role were able to avoid selecting the occluded objects with near-perfect accuracy. These pilot results indicated that listener perspective-taking weights $w_L$ were *also* near ceiling, consistent with prior work finding extremely low error rates in similarly simple displays (e.g., Hanna et al., 2003; Nadig & Sedivy, 2002). Taken together, these data suggest that the simplified director–matcher task we used in Experiment 1 is not ideal for testing the further resource-rational model predictions outlined in Section 2.5, as resource management considerations may only be expected to become relevant in higher cost regimes (Lin et al., 2010). Thus, in Experiment 2, we returned to the original details of the paradigm reported by Keysar et al. (2003) where confederate speakers were able to successfully elicit high rates of listener errors.

## 4. Experiment 2: Manipulating speaker informativity

Keysar et al. (2000, 2003) argued that if listeners were reliably using theory of mind in the director–matcher task, they would rule out referents that were not visible to the speaker and only consider *mutually* visible objects as possible targets of reference. This required a design where, on some trials, the speaker's referring expression was ambiguous between an object in the listener's private view and another object that was mutually visible. For instance, on one trial, a roll of Scotch tape was mutually visible and a cassette tape was hidden from the speaker's view. When the confederate speaker produced the ambiguous utterance, "tape," participants should still interpret it as a reference to the mutually visible roll of Scotch tape even if it would fit the hidden cassette the same or better. Surprisingly, Keysar et al. (2003) found that participants attempted to move the hidden item in 30% of cases: 71% of participants attempted to move this hidden item at least once out of four "critical" trials where an ambiguous distractor was present. Additionally, eye-tracking data showed that participants fixated on the competing hidden item more often and for longer on critical trials than would be expected from their baseline eye movements on other trials. These results were taken as evidence of an egocentric bias, establishing limits on spontaneous theory of mind use in conversation.

Subsequent work has criticized this interpretation from several angles. Hanna et al. (2003) argued that the viewpoint asymmetry paradigm itself is somewhat unnatural: Common ground is typically built incrementally over the course of an interaction rather than presented all at once, and it is rare for a shared display to differ in perceptual accessibility. Heller et al. (2008) observed that in many cases, the hidden object was designed to be a better fit for the referring expression than the one in common ground (e.g., the hidden bottom-most block vs. the shared block on the second-to-bottom row for "the bottom block"), making the hidden object a priori more likely to be the referent. It would be fairer to compare two objects that fit the referring expression equally well.[3] Brown-Schmidt and Hanna (2011) summarized these concerns, adding that Keysar et al. failed to include an important comparison condition where the critical distractor (e.g., the hidden cassette tape) was *also* in common ground. That is, the paradigm was set up to reject the

null hypothesis that processing is guided only by perspective information (vs. lexical competition in the private view), but it did not allow a test of the converse null hypothesis that participants *fail* to consider perspective: for example, by observing whether participants moved the critical distractor less frequently when it was hidden.

In addition to these considerations, many aspects of the design used by Keysar et al. differed from the simplified designs used in subsequent work. Some of these choices may have increased the overall cognitive load on participants, creating a regime where resource considerations become more relevant. In Experiment 2, we thus adopted the exact stimuli and design used by Keysar et al. to examine the downstream consequences of the pragmatic speaker behavior we observed in Experiment 1. In the resource-rational framework, the deployment of effort is guided by expectations about the value of that effort: Additional cost must be justified by commensurate benefits. Although a participant in the matcher role may begin the task with certain expectations about the director's share of the division of labor in the face of occlusions, the expected benefits of additional perspective-taking effort may shift as they obtain further evidence of the director's behavior. We suggest that these dynamics may provide a further explanation for listener errors. If the confederate directors in prior work were less informative than listeners (rationally) expected at the outset, then the listener's initial allocation of perspective-taking effort may have been mis-calibrated, with detrimental consequences for their performance. However, our model also predicts that listeners should gradually readjust their effort, resulting in fewer critical errors over the course of the experiment.

We tested both of these predictions in a close replication of Keysar et al. (2003) using the same interactive instant-messaging web interface we used in Experiment 1. In addition to this *scripted* condition, where speakers used the same scripted referring expressions used by confederates in the original study, we introduced a new *unscripted* condition where speakers were free to generate their own referring expressions. Our first goal was to use the scripted condition to ensure that prior findings successfully replicate in our online instant-messaging setting. Our second goal was to compare the specificity of utterances naturally produced in the unscripted condition with the scripted utterances previously used by confederates. We predicted that naive speakers would spontaneously provide more informative referring expressions than confederate directors used in prior work. A difference in listener error rate between these conditions would indicate that confederates deviated from the naturally expected division of labor. A decrease in listener errors over the course of the experiment would suggest that participants are indeed able to adapt their own allocation of effort to maintain successful communication.

### *4.1. Methods*

#### *4.1.1. Participants*

We recruited 200 pairs of participants from Amazon Mechanical Turk. Due to a server outage, 58 pairs were unable to complete the game and were thus excluded. Following our preregistered exclusion criteria, we removed 24 pairs who reported confusion,

violated our instructions, or made multiple errors on filler items, as well as 2 additional pairs containing non-native English speakers. This left 116 pairs in our final sample.

#### 4.1.2. Materials and procedure

The materials 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 (we discuss the potential impact of these differences below). Directors used a chat box to communicate where to move a privately cued target object in a $4 \times 4$ grid with five occluded cells (see Fig. 5). We used exactly the same graphical representation of occlusions as in Experiment 1. After receiving a message, the listener attempted to click and drag the intended object to the intended cell. In each of eight object 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 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.

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; see Table 1) automatically appeared

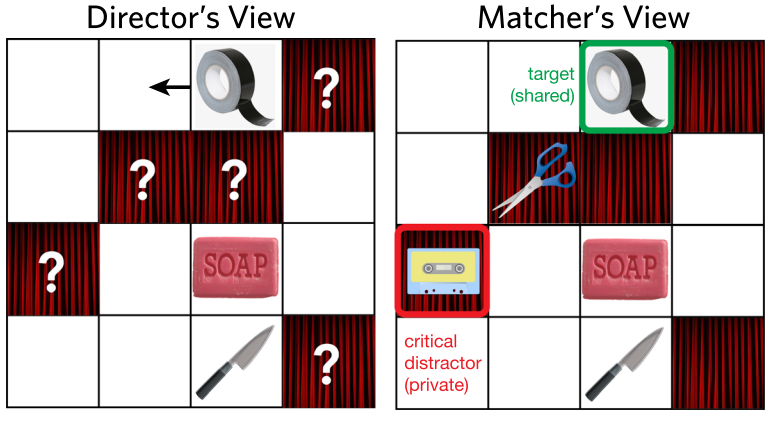


Fig. 5. Critical trial of director–matcher task using the ambiguous utterance "the tape": A roll of tape is in view of both players, but a *cassette tape* is occluded from the speaker's view.

Table 1
Critical stimuli and instructions used in Experiment 2, reproduced from Keysar et al. (2003)

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 | Item 8 |
|---|---|---|---|---|---|---|---|---|
| Instruction | "Glasses" | "Bottom block" | "Tape" | "Large measuring cup" | "Brush" | "Eraser" | "Small candle" | "Mouse" |
| Target | Sunglasses | Block (3rd row) | Cassette | Medium cup | Round hairbrush | Board eraser | Medium candle | Computer mouse |
| Hidden distractor | Glasses case | Block (4th row) | Scotch-tape | Large cup | Flat hairbrush | Pencil eraser | Small candle | Toy mouse |

in their chat box on exactly half of trials (the eight critical trials and about half of the fillers). To maintain an interactive environment, we allowed the director to freely produce referring expressions on the remainder of filler trials. Hence, the *scripted* condition served as a close replication of Keysar et al. (2003), ported to our online text-messaging interface. In the "unscripted" condition, directors were unrestricted and free to send whatever messages they deemed appropriate on all trials, although as in Experiment 1 we explicitly asked participants not to use purely spatial descriptions (e.g., "row 3, column 2 to row 4, column 2"). In both conditions, listeners were free to respond through the bidirectional chat interface. In addition to analyzing the director's messages and the matcher's errors, we again collected matcher mouse-tracking data.

### 4.2. Results

#### 4.2.1. 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 significantly by more than half, $p_1 = .51$, $p_2 = .20$, $\chi^2(1) = 43$, $p < .001$ in a binomial test (Fig. 6A). While we found substantial heterogeneity in error rates across object sets (just three of the eight object sets accounted for the vast majority of remaining unscripted errors; see Fig. S3), listeners in the unscripted condition made fewer errors for nearly every critical item. To more rigorously account for these sources of variance, we conducted a logistic mixed-effects model including a fixed effect of condition, random intercepts for each dyad, and random slopes and intercepts for each object set. We found a significant difference in error rates across conditions ($z = 2.6$, $p = .008$).

It is possible that participants in the unscripted condition still *considered* the hidden objects just as often as those in the scripted condition, even though they made fewer actual errors. To address this possibility, we conducted an analysis of our mouse-tracking data. We computed the mean (log-) amount of time spent hovering over the hidden

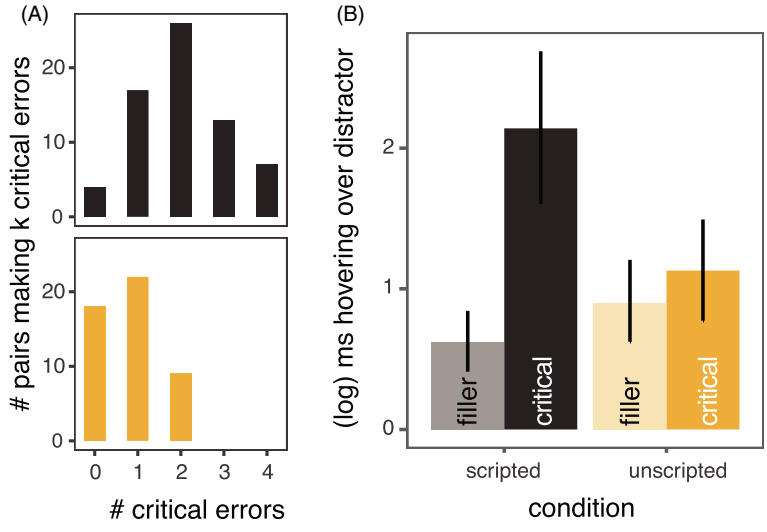


Fig. 6. Listener results for Experiment 2. (A) Distribution of errors with scripted and unscripted instructions. Participants in the unscripted condition made significantly fewer errors. (B) Even when they were correct, listeners in the scripted condition were more likely to hover their mouse cursor over the distractor relative to baseline while the unscripted condition shows no difference.

distractor and found a significant interaction between condition and the contents of the hidden cell ($t(10.2) = 3.6$, $p < .001$; Fig. 6B) in a mixed-effects regression including intercepts for each dyad and maximal random effect structure (intercepts, main effects, and interaction) for each object. That is, while listeners in the *scripted* condition spent more time hovering over the hidden cell when it contained a confusable distractor, relative to baseline (suggesting they considered the hidden object), listeners in the unscripted condition showed no difference from baseline.[4]

### *4.2.2. Adaptation over time*

Next, we examined how error rates change over the course of the interaction. If the effort a listener chooses to exert depends on their expectations about the speaker's informativity, our resource-rational account predicts that they may gradually recalibrate their expectations through repeated observations of the speaker's behavior (see Appendix B, Fig. B1). That is, listeners (and speakers in unscripted interactions) may learn that the allocation of perspective-taking they initially adopted is not sufficient and flexibly adjust the extent to which they weight their partner's perspective, leading to fewer errors on later trials.

To test this hypothesis, we ran a mixed-effects logistic regression predicting whether or not each participant made an error on each critical trial with fixed effects of the trial's position in the sequence (coded one through four) and condition (scripted vs. unscripted), including random intercepts for each pair of participants. We found a significant main effect of trial number, $z = 2.6$, $p < 0.01$, indicating that listener errors decrease over the

course of the experiment. We found no support for an interaction between trial number and condition *in a nested likelihood test,* $\chi^2$ *(1) = 0.07, p=0.79.*

Because the maximal random effect structure was too complex to converge using standard maximum likelihood methods, we further tested this effect using a fully Bayesian fitting procedure (Bürkner, 2017). In this model, we also included random intercepts and random effects of trial number at both the dyad-level and the item-level. We again found a reliable decrease in the probability of critical errors (i.e., attempting to move hidden objects) across both unscripted and scripted conditions ($b = 0.35$, 95% CI: [0.05, 0.69]) from an average of 43% on the first critical trial to only 30% on the fourth and final trial.

#### *4.2.3. Speaker informativity*

Finally, we test whether higher listener accuracy in the unscripted condition is accompanied by more informative speaker behavior than allowed in the scripted condition. 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 ($b = 0.54$, $t(13.8) = 2.6$, $p = .019$ in a mixed-effects regression on difference scores using a fixed intercept and random intercepts for object and dyads). 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 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) so that they could feasibly judge spatial or relative references like "bottom block." We excluded four judges for guessing with response times <1 s. Inter-rater reliability was moderately high, with intra-class correlation coefficient of 0.54 (95% CI = [047, 0.61]). 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 Supplemental Appendix S1 for details on our multi-level bootstrap procedure). 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, even though the speaker was not aware of the hidden distractor (see Fig. 7 A). In other words, the scripted labels used in Keysar et al. (2003) were systematically less informative than the expressions speakers would normally produce to refer to the same object in this context.

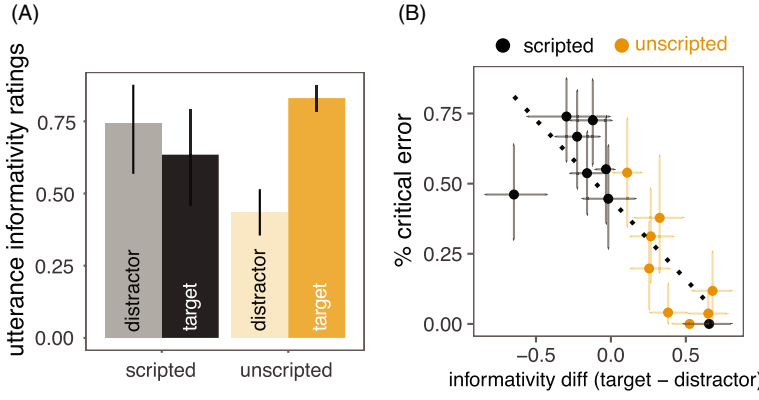


Fig. 7. Speaker results for Experiment 2. (A) While speakers in the scripted condition were forced to use utterances that were judged to fit target and distractor roughly equally (by design), speakers in the unscripted condition naturally produced utterances that fit the target much better than the distractor. (B) The extent to which an utterance fits the target relative to the distractor predicts error rates (dotted line is linear regression fit, each point is an item). All error bars are bootstrapped 95% confidence intervals.

Taken together, these results 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. 7B). In other words, a large proportion of the variance in listener error rates across different items can be explained by how well utterances fit each object in their own egocentric view, consistent with a division of labor relying on higher speaker informativity.

### *4.3. Discussion*

Building on Experiment 1, which aimed to identify pragmatic speaker behavior in the presence of occlusions, Experiment 2 aimed to test the downstream consequences of such behavior for listener perspective-taking. More specifically, given that speakers differentially allocate effort to produce more informative utterances in the presence of occlusions, we predicted that resource-rational listeners should expect this and exert differential effort toward visual perspective-taking. To test this hypothesis, we used a design that has been shown to elicit high levels of listener perspective-taking failure (Keysar et al., 2003). By comparing the utterances produced by a naive speaker to the scripted utterances produced by confederates in prior work, we found further evidence that naive speakers spontaneously produced costlier and more informative utterances, establishing the natural level of informativity that naive listeners may have expected. Listeners, in turn, make fewer errors when playing with naive, unscripted speakers than they do when playing with scripted speakers.

Note that while the scripted utterances developed by Keysar et al. (2003) were explicitly designed to be ambiguous between the target and hidden distractor, they were *not*

necessarily designed to violate Gricean maxims of quantity governing how a speaker ought to refer to the target in the presence of occlusions. Thus, while it may be unsurprising that listeners make more errors given under-specified utterances (e.g., Arts, Maes, Noordman, & Jansen, 2011), it was not obvious a priori that scripted referring expressions would in fact be less informative than expected from natural speakers in context. For example, if the hidden distractor were more similar to the target (e.g., a second roll of Scotch tape rather than a cassette), then the confederate could have used an appropriately specific utterance, closer to those produced by naive speakers (e.g., "clear roll of tape") to avoid pragmatic violations, while still maintaining ambiguity. Most importantly, error rates decreased over the course of interaction, suggesting that even if listeners' initial expectations about the speaker's level of effort were violated, they could still adaptively increase their perspective-taking effort to compensate. These findings raise several specific issues regarding choices of stimuli and procedure.

#### *4.3.1. Implications of stimulus choices*

First, our use of the stimuli and procedure from Keysar et al. (2003) successfully elicited listener errors, while our attempted conceptual replication in the simplified Experiment 1 task did not (see Appendix D). While there are several reasons why the simpler task may have reduced cognitive load (e.g., a smaller grid with fewer objects, fewer occlusions, a finite set of feature dimensions, and so on), it is particularly important to emphasize the differences between the stimuli used in our two experiments, which correspond to two prominent methodological threads in the literature. Experiment 1 used clean property contrasts between features like color, texture, and shape, similar to the geometric stimuli used by Hanna et al. (2003) and the pure size contrasts used by Heller et al. (2008). Experiment 2 used the much more heterogeneous items from Keysar et al. (2003), which included homonyms ("mouse" for a visible stuffed animal and hidden computer device), basic-level terms for different subordinate instances (e.g., "brush" for a visible round brush and a hidden flat brush), size contrasts (e.g., "large candle" for a visible large candle and an even larger hidden candle), and position contrasts (e.g., "top block" for a visible block on the second-to-top row and a hidden block on the top row).

Each of these stimulus choices has its advantages and disadvantages. On the one hand, there are concerns about the generalizability of simpler variants. Findings in narrower stimulus spaces may not straightforwardly extend to more crowded, high-variability contexts where there are not such salient and consistent dimensions along which items in each display vary. It is also possible that these design features of simpler variants have the effect of easing the overall cognitive load on participants. On the other hand, the heterogeneity of the eight items from Keysar et al. (2003) also creates serious difficulties for evaluating perspective-taking. We found that listener errors varied systematically across the items (see Fig. S3), as did the informativity of the scripted utterances, and it is challenging to place behavior across the items on the same scale, as each may be associated with distinct pragmatic considerations (e.g., relative contrast using modifiers, homonym processing, typicality of basic-level membership). This heterogeneity may also explain many of the remaining critical errors in the unscripted condition. Naive speakers

often made the effort to mention multiple redundant properties given the presence of occlusions (e.g., "the clear audio cassette tape" when there was only one thing that could be described as "tape" from their view), but because they could not know the relevant dimensions for distinguishing the target from the hidden distractors, their additional effort did not always pay off. For example, the highest proportion of errors made in the unscripted condition occurred on the "brush" item, where the target and hidden distractor were so similar that almost any increase in specificity would fail to distinguish them.

This limitation also emphasizes an important consequence of the referential context. While the relatively small number of features along which the finite stimulus space varied in Experiment 1 made it straightforward for speakers to anticipate the identity of hidden objects and provide maximally distinguishing expressions, it is computationally implausible that speakers could enumerate all possible hidden distractors in the open-ended space of objects used in Experiment 2. What algorithm speakers use to nevertheless produce more informative descriptions in this open-ended space remains an open question. One possibility is that speakers use the distribution of *visible* objects as a cue to the distribution of hidden objects, or that visible objects serve as anchors in a truncated search of semantic space. Another possibility is that speakers do not consider specific distractors at all and instead respond to the worst-case scenario or use the uncertainty introduced by occlusions as a generic cue to increase their production effort along the most salient properties.

#### *4.3.2. Implications of procedural choices*

While our results closely matched those of Keysar et al. (2003), and our dyadic instant-messaging interface preserved key aspects of interactive communication, including real-time feedback, several key differences between the procedure of our web experiment and earlier in-person work must be considered. Most prominently, the textual and verbal modalities differ in many ways, with implications for the listener's processing mechanisms and the speaker's cost of production.

First, listeners in an in-lab verbal version are able to make eye movements toward possible targets before the utterance has been completed, reflecting the incrementality of comprehension, while participants in our version had to fixate on and read the message in its entirety after it had been sent. Speakers may also have access to additional backchannel feedback in face-to-face verbal communication for the same reason: Listener utterances like "mm-hmm" or "uhh" may be initiated during speaker production, rather than needing to be sent after receiving one of the speaker's message in its entirety. Anecdotally, we found that some participants spontaneously broke up a longer message into multiple shorter "chunks" sent in rapid succession, which may mimic the incrementality of natural speech in some ways. Second, we have observed in other replications of in-person verbal communication tasks using a similar instant-messaging interface on the web (e.g., Hawkins, Frank, & Goodman, 2020) that typing tends to yield shorter descriptions overall than found in the lab, suggesting that production cost in terms of effort per word may be higher for typing, all else being equal. Third, face-to-face communication supports a variety of additional cues that are not available when participants cannot see one another. For

example, listeners may use the speaker's eye gaze (Hanna & Brennan, 2007) and head orientation (Hanna, Brennan, & Savietta, 2020) to disambiguate intended meanings, which may reduce the overall cognitive load of perspective-taking. Still, despite these clear differences, our ability to reproduce the core results of Keysar et al. (2003) in a (real-time, interactive) written modality suggests that the basic mechanisms at issue may be broadly preserved across modalities. As instant-messaging via text becomes increasingly common as the site of everyday interactive communication, it is important to distinguish it from more traditional settings of written communication which are non-interactive and intended to be processed offline (Arts et al., 2011; Herring, Stein, & Virtanen, 2013; Krauss & Fussell, 1991).

Two additional differences arise in comparison to the specific design of Keysar et al. (2003). First, a physical grid with real curtains may make occlusions more salient than virtual depictions of curtains on a computer screen. It is possible that the slightly greater overall number of errors we observed relative to Keysar et al. (2003) were due to a subset of participants not understanding how the occlusions worked. However, because the same depiction and instructions about occlusions were used across every condition, in both Experiments 1 and 2, these misunderstandings are unlikely to affect the comparisons of interest. Similar virtual occlusions have been used in previous work studying face-to-face verbal communication in the lab (Brown-Schmidt, 2009b, 2012; Brown-Schmidt et al., 2008; Rubio-Fernández, 2017), so this concern is not specific to a web interface. Second, because we were not able to precisely match the scripted instructions for *filler* items (or even the identity of filler objects), it is possible that listeners in our scripted condition obtained different input between critical items than participants in the original study. In particular, we observed that speakers in our scripted conditions used highly specific descriptions for the portion of trials on which they were allowed to freely send messages (e.g., "the red over ear headphones" when there was only one pair of headphones). These filler trials perhaps set even stronger expectations of hyper-informativity leading to larger prediction error when scripted labels were substituted in.

Finally, it is important to note that our findings are not intended to be a criticism of the use of a confederate or the choice of scripted utterances in prior work; using scripted directions manipulates the input received by the listener and allows measurements of how the listener engages in perspective-taking under different conditions. Experiments are useful precisely because they allow behavior to be observed under conditions that may not naturally occur. Rather, our results help identify an unintended consequence of an uncooperative director manipulation and clarify how it affects downstream listener behavior. More specifically, the informativity gap between unscripted and scripted utterances highlights the role of the listener's initial expectations about speaker informativity in their allocation of effort, and how an apparent violation of these expectations may have unintended pragmatic consequences.

These expectations become especially important under higher cognitive load where the appropriate division of labor is constrained by resource-rational considerations on both sides; in such contexts, it is particularly important for both parties to consider the other's allocation of effort. While we found near-ceiling levels of speaker and listener

perspective-taking in Experiment 1, this experiment, with its relatively higher cognitive load, enforced a higher pressure to establish a division of labor, as speakers could not reasonably be expected to produce perfectly unambiguous utterances. Even with an unscripted partner, some adaptation may be required to recalibrate to the challenges of the context. Under such conditions, we were able to identify clear effects of speaker informativity.

## 5. General discussion

The long-standing debate over the role of theory of mind in communication has largely centered on the extent to which listeners (or speakers) deviate from "optimal" perspective-taking toward egocentric influences (Barr & Keysar, 2016; Hanna et al., 2003). Our work aims to present a more nuanced analysis of how resource-constrained speakers and listeners nonetheless make reasonable decisions about how to allocate their resources based on contextual expectations. In particular, 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). One important case is when the speaker has uncertainty over what the listener can *see*, as in the director–-matcher task. Our resource-rational 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; Experiment 1 presents direct evidence in support of this hypothesis.

Importantly, when the director is expected to contribute effort to be additionally informative, communication can be successful even when the matcher contributes less than maximal perspective-taking effort. Indeed, the matcher will actually strike the optimal trade-off between minimizing joint effort and maximizing communicative success by *not* weighting the director's visual perspective. This suggests a resource-rational explanation of *when* and *why* resource-constrained listeners down-weight 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, although the listener's "failures" may indeed be failures to identify the correct items, they are not necessarily *failures of theory of mind*; rather, these inaccuracies are consistent with listeners using their theory of mind to decide when (and how much) they should expect the speaker to be cooperative and informative, and allocating their resources accordingly (Griffiths et al., 2015). Experiment 2 is consistent with this hypothesis; when speakers (directors) used under-informative scripted instructions taken from prior work, listeners made significantly more errors than when speakers were allowed to provide referring expressions at their natural level of informativity. Furthermore, listeners were able to adapt to the speaker's level of informativity to make fewer errors over time.

To be clear about our theoretical stance, these results do not imply that speakers are generally expected to shoulder more of the work, or that Gricean considerations free

listeners of all effort. Indeed, speakers often use vague or ambiguous language that reduces their own production cost, especially when they can rely on listeners to infer the intended meaning from context (Peloquin, Goodman, & Frank, 2020; Wasow, 2015). For example, Piantadosi, Tily, and Gibson (2012) found in a large corpus study that more efficient words (i.e., shorter and easier for speakers to produce) tended to be more overloaded with meanings (i.e., homophony and polysemy), suggesting that languages shift some of the division of labor onto the listener rather than requiring speakers to do all the work of disambiguating. Similarly, everyday adjectives like "tall" or "expensive" may be less costly for the speaker to produce than precise estimates of height or price but shift the division of labor to the listener's ability to use world knowledge about the relevant comparison class (Lassiter & Goodman, 2017). In the resource-rational elaboration of the simultaneous integration view we are advancing, the perspective-taking effort each person chooses to exert is rarely all or none: It is a matter of *degree* (Heller et al., 2016). There is in principle a continuum of many acceptable divisions of labor, and no single division should be considered the "rational" yardstick. Instead, the resource-rational weighting for one agent should in principle depend on a number of contextual factors, including the relationship between the agents; the other agent's capacity, perspective, belief, and knowledge; the ability to avoid further clarification exchanges or repair; and the current cognitive load imposed by the environment. It may be asymmetric when one partner is able to take on more costly processing than the other, and it should be continually adjusted throughout the course of an interaction.

This flexibility is a key feature of the resource-rational framework. An important direction for future work is to more directly explore how perspective-taking effort adjusts dynamically given aspects of the scenario (Grodner & Sedivy, 2011; Pogue, Kurumada, & Tanenhaus, 2016; Ryskin, Kurumada, & Brown-Schmidt, 2019). While this hypothesized form of "effort adaptation" is similar to context-specific adaptation previously studied at the phonetic (Kleinschmidt & Jaeger, 2015), syntactic (Fine, Jaeger, Farmer, & Qian, 2013), semantic (Schuster & Degen, 2020), or pragmatic (Grodner & Sedivy, 2011; Pogue et al., 2016) levels, it is a subtly more specific mechanistic proposal about exactly what is being adapted. Our account raises the possibility that observations of a partner's behavior not only allow agents to update their expectations directly about that particular behavior, but also provide information about an additional latent variable: the degree of effort their partner is exerting. After updating one's beliefs about this underlying quantity, it may be appropriate to change one's own allocation of effort, leading to downstream changes in one's surface-level behavior via the mechanism of effort. While further work is needed to test this hypothesis, we provided preliminary evidence that, given sufficient evidence of an unusually under-informative partner, listeners may realize that devoting additional attention to which objects are occluded from their partner's view is necessary to maintain communicative success. Conversely, given evidence of an over-informative partner, listeners may be able to get away with exerting less effort in a high-cost context. Dynamic adaptation of perspective-taking effort could be particularly functionally important in light of pervasive individual differences in working memory or executive control (Brown-Schmidt, 2009b; Wardlow, 2013): Variability in the capabilities of different

partners should lead to variability in the appropriate division of labor, and it may not be possible to anticipate at the outset of an interaction. Indeed, recent work by Ryskin, Stevenson, and Heller (2020) has found substantial variability in the best-fitting probabilistic weighting parameter $w$ used by each speaker in a large population. Our resource-rational account predicts that these different weights—corresponding to different division of labor—may arise systematically from such individual differences. While we have focused on adaptation, it is also possible that background knowledge about a partner leads to differing resource allocations even at the outset of the interaction. For instance, an adult may expect to shoulder more of the division of perspective-taking labor when interacting with a child (Leung, Hawkins, & Yurovsky, 2020) and an expert may shoulder more of the labor when interacting with a novice in a technical field (Bromme, Jucks, & Wagner, 2005). Further work may test this hypothesis by manipulating initial expectations about effort allocation.

Our theoretical framework relies on an abstract computational notion of "effort" or "cost." We remain agnostic about the precise source of these costs at the algorithmic level; the director–matcher task, like many other standard tasks used to evaluate theory of mind abilities (Quesque & Rossetti, 2020), involves the coordination of many cognitive systems, and the available data do not allow us to isolate a specific cause for poor performance (Rubio-Fernández, 2017). We expect that the abstract cost associated with using a higher mixture weight in our model represents a range of different costs associated with general executive control, working memory, selective attention, and other processes, as well as whatever cost may be specifically associated with forming and maintaining representation of a partner's likely behavior given their perspective. For instance, it is possible that the listener can take a small number of samples from their posterior about the speaker's likely behavior and use the resulting estimate of communicative success to decide to devote less persistent attention to which cells are occluded. If this is the case, the primary effort at stake is attentional, with the deployment of attentional resources guided by theory of mind use. In any case, it is clear that solving the full constrained optimization problem at the core of the resource-rational account (Eq. 11) from scratch in every situation would be intractable: The additional effort required to compute the appropriate level of effort across these processes would exceed the resulting savings. This has been a general challenge for resource-rational accounts, which argue that this optimization problem is solved by learning over longer (e.g., developmental) timescales (Lieder & Griffiths, 2019); an intriguing possibility is that speakers amortize the optimization across many different partners, with relatively inexpensive adjustments based on local evidence (Bustamante, Lieder, Musslick, Shenhav, & Cohen, in press; Lieder, Shenhav, Musslick, & Griffiths, 2018). Further work in the resource-rational framework is needed to formulate explicit algorithmic theories of the "mental labor" associated with different processes, and how these processes are integrated to support success in communication. Among these processes, it is particularly important to identify the respective costs associated with different aspects of theory of mind use. For example, two-system theories distinguish cheap and fast forms of perspective-taking from more costly but flexible forms (Apperly, 2010). As in other domains where dual-process theories have been proposed, resource

rationality may provide a useful way of explaining why such processes may arise from the computational problems facing social agents under resource constraints (e.g., Milli, Lieder, & Griffiths, 2018).

Our work also adds to the growing literature on the debate over the role of pragmatics in the director–matcher task. Recently, Rubio-Fernández (2017) has suggested that listeners monitor the speaker's level of informativity and become suspicious of the director's visual access when the director shows unexpectedly high levels of specificity in their referring expressions. Our results further bolster the argument that pervasive pragmatic reasoning about expected levels of informativity is an integral aspect of perspective-taking in the director–matcher task (and communication more generally). We note, however, that in this work participants became suspicious about the experimenter, while in our study participants simply adapted their expectations about informativity; a more detailed look at differences between experimental paradigms is necessary to better understand why participants drew different inferences (see Rubio-Fernández & Jara-Ettinger, 2018). Prior work also suggests that although speakers tend to be over-informative in their referring expressions (Degen et al., 2020; 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 "known unknowns" in the listener's private view and may increase informativity to disambiguate the referent relative to these possible contexts.

While our experiments have focused directly on the demands of asymmetries in *visual* perspective, closely following the design of Keysar et al. (2003), variations on this basic paradigm have also manipulated other dimensions of nonvisual knowledge asymmetry, including those based on spoken information (Hanna et al., 2003; Keysar, Barr, Balin, & Paek, 1998), spatial cues (Galati & Avraamides, 2013; Schober, 1993), private pre-training on object labels (Wu & Keysar, 2007), cultural background (Isaacs & Clark, 1987), and other task-relevant information (Hanna & Tanenhaus, 2004; Yoon, Koh, & Brown-Schmidt, 2012). We expect that each of these variants introduces subtly different processing demands and pragmatic expectations, and resource-rational analysis may be a useful framework for understanding how variance in these demands leads to variance in perspective-taking behavior. Individual differences in basic cognitive function (e.g., Ryskin et al., 2015) and the cognitive demands imposed by different tasks or environments (Lin et al., 2010) can be viewed as real differences in the underlying $\beta$ parameter, shifting the agent's decisions about perspective-taking, which may provide new traction on the problem of explaining and predicting the precise relationship between the two. Similarly, studies of how speakers *inhibit* private knowledge during production may involve specific processing mechanisms involving costly executive control (e.g., Ferreira, 2019) and resource-rational considerations may yield predictions about the extent to which private information leaks into speaker utterances (see also Brown-Schmidt & Tanenhaus, 2008; Heller, Gorman, & Tanenhaus, 2012; Nadig & Sedivy, 2002; Savitsky, Keysar, Epley, Carter, & Swanson, 2011; Wardlow Lane et al., 2006; Yoon & Brown-Schmidt, 2014).

More broadly, we suggest that a resource-rational approach may provide a more constructive and principled standard for what should constitute "rational" perspective-taking

behavior in conversation. As discussed by Brown-Schmidt and Heller (2018), previous work arguing for egocentric heuristics has tended to use a strong classical standard of rationality. Any deviation from error-free perspective-taking is then taken as evidence of "irrational" biases that motivate a rejection of the entire rational analysis framework. By contrast, a more bounded standard of rationality preserves the advantages of these unifying frameworks, namely the ability to formalize the functional problem facing communicative agents at the computational level of analysis, but moves beyond the question of *if* people are classically rational to ask *when* and *how* they make approximately optimal decisions about allocating their resources. In other words, the resource-rational framework allows the comparison of formal proposals about which factors the agent considers when making decisions about how much perspective-taking effort to allocate, and may help to illuminate how people are so flexible across contexts. In this way, we seek to push computational-level probabilistic weighting models toward process-level consideration of cognitive resources, forming a bridge to the initial concerns of egocentric heuristic accounts.

## Acknowledgments

This manuscript is based in part on work presented at the 38th Annual Conference of the Cognitive Science Society. An early pilot of Experiment 2 was originally conducted with input from Michael Frank, Desmond Ong, and Long Ouyang. We are grateful to Victor Ferreira, Herb Clark, Barbara Tversky, Fred Callaway, Roger Levy, and Judith Fan for thoughtful comments and conversations and to Boaz Keysar for providing selected materials for our replication in the scripted condition of Experiment 2. 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/division_of_labor.

## Open Research badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/qwkmp/.

## Notes

1. In technical terms, the weighting parameter has previously been treated as an "exogeneous" variable determined by factors outside the scope of the model. The problem we consider of determining it as a function of other factors originating *within* the model is known as "endogenization" (Mankiw, 2003).

2. Note that this could correspond to either an "egocentric" domain of reference or a "common ground" domain, which are equivalent for the speaker in the classic variant of the director–matcher task we are considering.
3. We validate this argument in Experiment 2 by empirically measuring relative fit of the expressions to the target and distractor items.
4. Hover time was exactly zero for many trials in both conditions, which skewed the overall distribution of hover times; to address potential issues comparing the means of such zero-inflated distributions, we conducted a follow-up analysis examining the binarized *proportion* of trials that listeners hovered over the hidden distractor at all, and found the same pattern of results, $z = -2.1$, $p = 0.035$. We also pre-registered an analysis of an additional measure—the response latency before *first* hovering over the target—but due to unexpectedly poor precision in aligning response times to the beginning of the trial, we did not pursue this analysis further.
5. Note that this use of Bayesian statistics in analyzing and evaluating our cognitive model is dissociable from the assumption of Bayesian recursive reasoning within the model.
6. We are grateful to an anonymous reviewer for suggesting this experiment.

## References


Apperly, I. (2010). *Mindreaders: The cognitive basis of "theory of mind"*. Hove, UK: Psychology Press.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*, 361–374.

Barr, D. J. (2008). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*, 18–40.

Barr, D. J., & (2014). Perspective taking and its impostors in language use: Four patterns. T. M. Holtgraves In *The Oxford handbook of language and social psychology* (pp. 98–110). Oxford: Oxford University Press.

Barr, D. J., & Keysar, B. (2006). Perspective taking and the coordination of meaning in language use. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 901–938). Amsterdam: Elsevier.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

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

Bradford, E. E., Jentzsch, I., & Gomez, J.-C. (2015). From self to social cognition: Theory of mind mechanisms and their relation to executive functioning. *Cognition*, *138*, 21–34.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1482.

Bromme, R., Jucks, R., & Wagner, T. (2005). How to refer to 'diabetes'? Language in online health advice. *Applied Cognitive Psychology*, *19*, 569–586.

Brown-Schmidt, S. (2009a). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*, 171–190.

Brown-Schmidt, S. (2009b). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*, 893–900.

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, *27*, 62–89.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*, 1122–1134.

Brown-Schmidt, S., & Hanna, J. E. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialogue and Discourse*, *2*, 11–33.

Brown-Schmidt, S., & Heller, D. (2018). Perspective taking during conversation. In S. Rueschemeyer & M. G. Gaskell (Eds.), *Oxford handbook of psycholinguistics* (pp. 551–574). Oxford: Oxford University Press.

Brown-Schmidt, S., & Konopka, A. E. (2011). Experimental approaches to referential domains and the on-line processing of referring expressions in unscripted conversation. *Information*, *2*, 302–326.

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

Bürkner, P.-C. (2017). Advanced Bayesian multilevel modeling with the r package brms. *arXiv Preprint* arXiv:1705.11123.

Bustamante, L. A., Lieder, F., Musslick, S., Shenhav, A., & Cohen, J. D. (in press). Learning to overexert cognitive control in a stroop task. *Cognitive Affective & Behavioral Neuroscience*. https://doi.org/10.31234/osf.io/3rynj

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. (2018). A resource-rational analysis of human planning. In C. Kailash M. Rau J. Zhu & T. Rogers (Eds.), *Proceedings of the 40th annual conference of the Cognitive Science Society* (pp. 178–183). Austin, TX: Cognitive Science Society.

Chen, M. C., Anderson, J. R., & Sohn, M. H. (2001). What can a mouse cursor tell us more?: Correlation of eye/mouse movements on web browsing. M. Tremaine In *CHI'01 extended abstracts on human factors in computing systems* (pp. 281–282). New York: ACM.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*, 233–263.

Davies, C., & Arnold, J. (2019). Reference and informativeness: How context shapes referential choice. In C. Cummins & N. Katsos (Eds.), *Handbook of experimental semantics and pragmatics* (pp. 474–493). Oxford: Oxford University Press.

Degen Judith, Hawkins Robert D., Graf Caroline, Kreiss Elisa, Goodman Noah D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions.. *Psychological Review*, *127*, 591–621.

Degen, J., & Tanenhaus, M. (2019) Constraint-based pragmatic processing. In C. Cummins & N. Katsos (Eds.), *Handbook of experimental semantics and pragmatics* (pp. 21–38). Oxford: Oxford University Press.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*, 327.

Ferguson, H. J., Apperly, I., Ahmad, J., Bindemann, M., & Cane, J. (2015). Task constraints distinguish perspective inferences from perspective use during discourse interpretation in a false belief task. *Cognition*, *139*, 50–70.

Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation*, *49*, 209–246.

Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annual Review of Psychology*, *70*, 29–51. https://doi.org/10.1146/annurev-psych-122216-011653

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS One*, *8*, e77661.

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

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift Für Sprachwissenschaft*, *35*, 3–44.

Freeman, J., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*.

Galati, A., & Avraamides, M. N. (2013). Flexible spatial perspective-taking: Conversational partners weigh multiple cues in collaborative tasks. *Frontiers in Human Neuroscience*, *7*, 618.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Science*, *20*, 818–829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

Goodman, N. D., & Stuhlmller, A. (2014). *The design and implementation of probabilistic programming languages*. http://dippl.org

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. *Topics in Cognitive Science*, *7*, 217–229.

Grodner, D., & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. In N. Pearlmutter & E. Gibson (Eds.), *The processing and acquisition of reference* (pp. 239–272). Cambridge, MA.: MIT Press.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*, 596–615.

Hanna, J. E., Brennan, S. E., & Savietta, K. J. (2020). Eye gaze and head orientation cues in face-to-face referential communication. *Discourse Processes*, *57*, 201–223.

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

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

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

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the Dynamics of Learning in Repeated Reference Games. *Cognitive Science*, *44*, e12845.

Hawkins, R. X. D., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? Good questions provoke informative answers. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th annual conference of the Cognitive Science Society* (pp. 878–883). Austin, TX: Cognitive Science Society.

Heller, D., Gorman, K. S., & Tanenhaus, M. K. (2012). To name or to describe: Shared knowledge affects referential form. *Topics in Cognitive Science*, *4*, 290–305.

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

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

Heller, D., & Stevenson, S. (2018). Modelling reference production using the simultaneity approach: A new look at referential success. In C. Kalish M. Rau J. Zhu & T. Rogers (Eds.), *Proceedings of the 40th annual conference of the Cognitive Science Society* (pp. 481–486). Austin, TX: Cognitive Science Society.

Herring, S., Stein, D., & Virtanen, T. (2013). *Pragmatics of computer-mediated communication*. Berlin: Walter de Gruyter.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*, 91–117.

Howes, A., Duggan, G. B., Kalidindi, K., Tseng, Y.-C., & Lewis, R. L. (2016). Predicting short-term remembering as boundedly optimal strategy choice. *Cognitive Science*, *40*, 1192–1223.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, *116*, 26.

Jouravlev, O., Schwartz, R., Ayyash, D., Mineroff, Z., Gibson, E., & Fedorenko, E. (2019). Tracking colisteners' knowledge states during language comprehension. *Psychological Science*, *30*, 3–19.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 12002–12007.

Keysar, B. (2007). Communication and miscommunication: The role of egocentric processes. *Intercultural Pragmatics*, *4*, 71–84.

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

Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). ) Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory and Language*, *39*, 1–20.

Keysar, B., Barr, D. J., & Horton, W. S. (1998). ) The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science*, *7*, 46–49.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25–41.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*, 148.

Kool, W., & Botvinick, M. (2013). The intrinsic cost of cognitive control. *Behavioral and Brain Sciences*, *36*, 697–698.

Kool W., Botvinick M. (2018). Mental labour. *Nature Human Behaviour*, *2*, 899–908.

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

Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, *9*, 2–24.

Kuhlen, A. K., & Brennan, S. E. (2013). Language in dialogue: When confederates might be hazardous to your data. *Psychonomic Bulletin & Review*, *20*, 54–72.

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, *194*, 3801–3836.

Leung, A., Hawkins, R., & Yurovsky, D. (2020). Parents scaffold the formation of conversational pacts with their children. In S. Denison M. Mack Y. Xu & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the Cognitive Science Society* (pp. 1022–1028). Austin, TX: Cognitive Science Society.

Lieder F., Griffiths T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, 1–60.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*, 1.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology*, *14*, e1006043.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*, 551–556.

Long, M. R., Horton, W. S., Rohde, H., & Sorace, A. (2018). Individual differences in switching and inhibition predict perspective-taking across the lifespan. *Cognition*, *170*, 25–30.

Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: State of the art. *British Journal of Developmental Psychology*, *30*, 1–13.

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*, 3–42.

Mankiw, N. G. (2003). *Macroeconomics*. Macmillan: New York.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Milli, S., Lieder, F., & Griffiths, T. (2018). A rational reinterpretation of dual-process theories. https://doi.org/10.13140/RG.2.2.14956.46722/1

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. *Cognitive Science*, *42*, 974–1008.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*, 329–336.

Nilsen, E. S., & Graham, S. A. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cognitive Psychology*, *58*, 220–249.

Padmala, S., & Pessoa, L. (2011). Reward reduces conflict by enhancing attentional control and biasing visual cortical processing. *Journal of Cognitive Neuroscience*, *23*, 3419–3432.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*, 89–110.

Peloquin, B. N., Goodman, N. D., & Frank, M. C. (2020). The interactions of rational, pragmatic agents lead to efficient language structure and use. *Topics in Cognitive Science*, *12*, 433–445.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

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

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

Quesque F., Rossetti Y. (2020). What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice. *Perspectives on Psychological Science*, *15*, 384–396.

Rodden, K., Fu, X., Aula, A., & Spiro, I. (2008). Eye-mouse coordination patterns on web search results pages. In *CHI'08 extended abstracts on human factors in computing systems* (pp. 2997–3002). New York: ACM.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*, 69–89.

Roxβnagel, C. (2000). Cognitive load and perspective-taking: Applying the automatic-controlled distinction to verbal communication. *European Journal of Social Psychology*, *30*, 429–445.

Rubio-Fernández, P. (2017). The director task: A test of theory-of-mind use or selective attention? *Psychonomic Bulletin & Review*, *24*, 1121–1128.

Rubio-Fernández, P., & Jara-Ettinger, J. (2018). Joint inferences of speakers' beliefs and referents based on how they speak. In C. Kalish M. Rau R. Zhu & T. Rogers (Eds.), *Proceedings of the 40th annual conference of the Cognitive Science Society* (pp. 991–996). Austin, TX: Cognitive Science Society.

Rubio-Fernández, P., Mollica, F., Ali, M. O., & Gibson, E. (2019). How do you know that? Automatic belief inferences in passing conversation. *Cognition*, *193*, 104011.

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, *144*, 898.

Ryskin, R., Kurumada, C., & Brown-Schmidt, S. (2019). Information integration in modulation of pragmatic inferences during online language comprehension. *Cognitive Science*, *43*, e12769. https://doi.org/10.1111/cogs.12769

Ryskin, R., Stevenson, S., & Heller, D. (2020). Probabilistic weighting of perspectives in dyadic communication. In S. Denison M. Mack Y. Xu & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the Cognitive Science Society* (pp. 252–258). Austin, TX: Cognitive Science Society.

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology*, *47*, 269–273.

Saxe, R., Schulz, L. E., & Jiang, Y. V. (2006). Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. *Social Neuroscience*, *1*, 284–298.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, *47*, 1–24.

Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, *203*, 104285.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*, 99–124.

Spivey, M., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 10393–10398.

Symeonidou, I., Dumontheil, I., Chow, W.-Y., & Breheny, R. (2016). Development of online use of theory of mind during adolescence: An eye-tracking study. *Journal of Experimental Child Psychology*, *149*, 81–97.

Tanenhaus, M. K., & Brown-Schmidt, S. (2008). Language processing in the natural world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*, 1105–1122.

Tarenskeen, S., Broersma, M., & Geurts, B. (2015). Overspecification of color, pattern, and size: Salience, absoluteness, and consistency. *Frontiers in Psychology*, *6*, 1703.

Tomasello, M. (2009). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

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

van der Wel, R. P., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*, 128–133.

Wardlow, L. (2013). Individual differences in speakers' perspective taking: The roles of executive control and working memory. *Psychonomic Bulletin & Review*, *20*, 766–772.

Wardlow Lane, L., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants! Speakers' control over leaking private information during language production. *Psychological Science*, *17*, 273–277.

Wasow, T. (2015). Ambiguity avoidance is overrated. In S. Winkler (Ed.), *Ambiguity: Language and Communication* (pp. 29–48). Berlin: De Gruyter.

Wu, S., & Keysar, B. (2007). The effect of information overlap on communication effectiveness. *Cognitive Science*, *31*, 169–181.

Yoon, S. O., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 919.

Yoon, S. O., Koh, S., & Brown-Schmidt, S. (2012). Influence of perspective and goals on reference production in conversation. *Psychonomic Bulletin & Review*, *19*, 699–707.

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

**Fig. S1:** Screenshot of experiment interface.

**Fig. S2:** Parameter posteriors for best-fitting occlusion-sensitive model in Appendix C. All parameters shown on log scale. MAP estimates with 95% highest posterior density intervals are as follows: $\alpha = 55.7$, HDI = [51, 58.2]; $c_{color} = 6.9 \times 10^{-5}$, HDI = [$4.5 \times 10^{-5}$, $3.6 \times 10^{-4}$]; $c_{shape} = 5.2 \times 10^{-5}$, HDI = [$4.5 \times 10^{-5}$, $1.2 \times 10^{-4}$]; $c_{texture} = 9.9 \times 10^{-3}$, HDI = [$8.1 \times 10^{-3}$, $1.2 \times 10^{-2}$.

**Fig. S3:** Heterogeneity in errors across the 8 object sets used in Experiment 2 (reproduced 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 $\chi^2$ test.

**Appendix S1:** Details of multi-stage bootstrap procedure used in Experiment 2 analyses.

## Appendix A: Mathematical derivation of qualitative speaker predictions

The key novel prediction motivating Experiment 1 is that speakers should attempt to be more informative when there is an asymmetry in visual access. Here, we prove analytically that the predicted increase in informativity holds under fairly unrestrictive conditions. We define "specificity" extensionally, in the sense that if an utterance $u_0$ 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. Recall that $\mathscr{L}$ is a (soft) truth-conditional semantics returning 0.01 or 1.

**Definition 1:** The *extension* of an utterance $u$ is the set $E_u = \{o \in \mathcal{O} | \mathscr{L}(u,o) = 1\}$.

**Definition 2:** Utterance $u_0$ is said to be *more specific* than $u_1$ iff $E_{u_0} E_{u_1}$, where we define $\mathcal{O}^* = E_{u_1} \setminus E_{u_0}$ denoting the set difference: the objects in the wider extension of $u_1$ that are not in the narrower extension of $u_0$.

We now show that our "ideal" 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 *egocentric* utility.

**Theorem 1:** If $u_0$ is more specific than $u_1$ then the following holds for any target $o^t$ and shared context $C$:

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

**Proof 1:** 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

$$U_{\text{asym}}(u_0;o^t,C) - U_{\text{asym}}(u_1;o^t,C) > U_{\text{ego}}(u_0;o^t,C) - U_{\text{ego}}(u_1;o^t,C)$$

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

$$U_{\text{asym}}(u_0;o^t,C) - U_{\text{asym}}(u_1;o^t,C) = \sum_{o_h \in \mathcal{O}} p(o_h)[\log L(o|u_0, C \cup o_h) - \log L(o|u_1, C \cup o_h)]$$

$$= \sum_{o^* \in \mathcal{O}^*} p(o^*) \log \frac{L(o^t|u_0, C \cup o^*)}{L(o^t|u_1, C \cup o^*)} \tag{A1}$$

$$+ \sum_{o_h \in \mathcal{O} \setminus \mathcal{O}^*} p(o_h) \log \frac{L(o^t|u_0, C \cup o_h)}{L(o^t|u_1, C \cup o_h)} \tag{A2}$$

By the definition of $\mathcal{O}^*$ we have $\mathcal{L}(u_0, o_h) = \mathcal{L}(u_1, o_h)$ for objects $o_h$ in the complement $\mathcal{O} \setminus \mathcal{O}^*$. Therefore, for Eq. A2, $L(o^t|u_i, C \cup o_h) = L(o^t|u_i, C)$, 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 Eq. A1, we can substitute the definition of the listener $L$ and simplify:

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

Thus,

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

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

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

since the same constant appears on both sides of inequality. It also follows that a speaker using any mixture of the asymmetric and egocentric utilities (i.e., $w_S U_{\text{ego}} + (1 - w_S) U_{\text{asym}}$ where $w_S > 0$) will monotonically prefer more informative utterances than a purely egocentric speaker.

## Appendix B: Model prediction for flexibility over extended interaction

Another key prediction that distinguishes a resource-rational framework from a "fixed capacity" egocentric heuristic account is that agents may flexibly adjust the effort dedicated to perspective-taking depending on contextual factors. In this section, we derive the prediction that listeners adapt their own perspective-weighting effort over the course of several rounds where the speaker is less informative than initially expected. The basic mechanism for this adaptation in our model is an inference about the underlying perspective-taking weighting being used by the speaker, based on observations of the speaker's behavior. The speaker is expected to behave differently under different settings of the parameter $w_S$, so data, $D = \{(u, o)\}$, from repeated observations of the speaker's choice of utterance $u$ for targets $o$ provides a statistical signal about which $w_S$ they are likely to be using. Using Bayes rule, the posterior over $w_S$ is given by $D$:

$$\begin{aligned} P(w_S|D) &\propto P(D|w_S)P(w_S) \\ &= P(w_S) \cdot \prod_i P_{S_1}(u_i|o_i, C, w_S) \end{aligned} \tag{B1}$$

We now conduct a resource-rational analysis of a listener using this posterior instead of the uniform prior $P(w_S)$. Specifically, we examine the listener's posterior after they

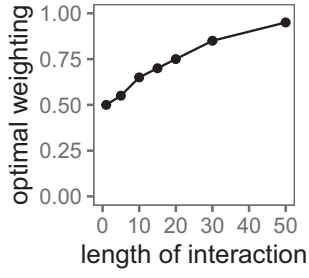


Fig. B1. Our model predicts that the listener should flexibly increase the effort they dedicate to perspective-taking as they infer from the speaker's short utterances that the speaker is dedicating less effort. For these simulations, we set $\beta = 0.1$.

observe the speaker provide a single-word utterance to refer to the target over a fixed number of rounds. Note that, as in the director-matcher task we used in Experiment 2, this single-word utterance is completely sufficient to distinguish the target given the objects in common ground (i.e., in the speaker's view), so it is only "under-informative" relative to what we previously established a Gricean speaker would do to account for the fact the listener may see hidden objects they do not. Results are shown in Fig. B1. As the listener observes more and more evidence that the speaker is exerting a low level of perspective-taking effort, the boundedly optimal setting of their own perspective-taking effort grows higher. In other words, the division of communicative labor gradually shifts onto the listener to preserve communicative success.

## Appendix C: Quantitative model comparison for Experiment 1

In this section, we conduct a quantitative model comparison using our empirical data from Experiment 1 to further bolster the qualitative speaker predictions derived in the previous section. Specifically, we describe the details of a Bayesian data analysis evaluating our mixture model on the empirical data, and comparing it to the purely egocentric (or "occlusion-blind") baseline model (Eq. 3), which does not reason about the possible existence of hidden objects behind occlusions.

The implementation of the director–matcher task for the model was the same as we used for the resource-rational simulations presented in Section 2. 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 again collapsed across these details and only provided the model which features of each distractor differed from the target on each trial. After this simplification, there were four possible kinds of contexts: *distractor-absent* contexts, where the other objects differed in every dimension, and three varieties of *distractor-present* contexts, where the critical distractor differed in *only shape*, *shape and color*, or *shape and texture*. In addition, we provided the model information about whether each trial had cells occluded or not. The space of predicted utterances for the speaker model was the same as our feature annotations: for each trial, the speaker model selected among

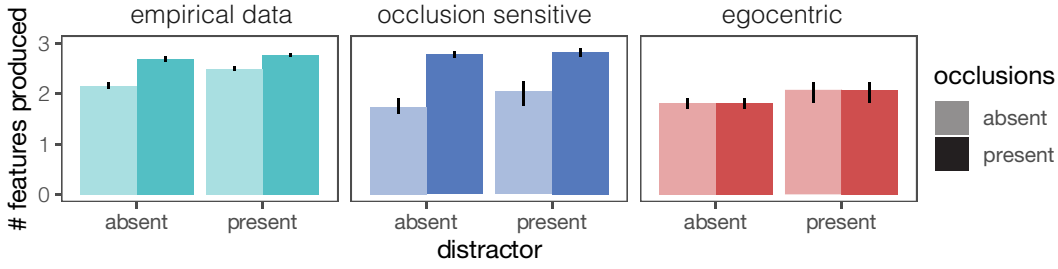


Fig. C1. Quantitative modeling results for Experiment 1. Posterior predictives of each model are projected to the mean number of features produced in each condition. Error bars on empirical data are bootstrapped 95% confidence intervals; model error bars are 95% credible intervals.

seven 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 $\mathcal{O}$ that the speaker marginalizes over, we used a uniform prior over all combinations of sharing the same or different properties as the target (i.e., the same as the possible distractors).

Table C1
Model comparison conditioned on Experiment 1 data. Marginal likelihoods estimated using annealed importance sampling (AIS).

| Model | Marginal Likelihood |
|---|---|
| Egocentric ($w_S = 0$) | −4037 |
| Occlusion-sensitive ($w_S = 1$) | **−2997** |
| Mixture | −3153 |

Our full mixture model has five free parameters which we infer from the data using Bayesian inference.[5] 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 shifts to maximizing. In addition, to allow for the differential production of the three features (i.e., Fig. 4B), we assume separate production *costs* for each feature: a texture cost $c_t$, a color cost $c_c$, and a shape cost $c_s$. Finally, we also fit the speaker's mixture weight $w_S$. We use (uninformative) uniform priors for all parameters:

$$\begin{aligned} \alpha &\sim \quad \mathrm{Unif}(0, 1000) \\ w_S &\sim \quad \mathrm{Unif}(0, 1) \\ c_t, c_c, c_s &\sim \quad e^{\mathrm{Unif}(-10,1)} \end{aligned}$$

We obtained predictions from our speaker model (i.e., a distribution over the possible utterances) for a particular setting of parameters using analytic enumeration. These predictions were mixed with a 5% chance that participants randomly guess among the utterances to obtain a likelihood function for scoring the empirical data. Finally, we obtained a posterior over parameters using MCMC. We discarded 1,000 burn-in samples and then drew 1,000 samples from the posterior with a lag of 5. Posterior predictives were computed by sampling parameters from these posteriors and taking the expected number of features produced by the speaker, marginalizing over possible noncritical 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 obtain marginal likelihoods for a model comparison, we averaged 20 runs of annealed importance sampling (AIS) for each model, taking 10,000 steps per run. We implemented our models and conducted inference in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014). All code necessary to reproduce our model results is available at the project Github repository.

Our primary model comparison is to compare the full mixture model to the endpoints, with $w_S = 0$ corresponding to a purely egocentric or "occlusion-blind" speaker, and $w_S = 1$ corresponding to our occlusion-sensitive speaker. First, we found extremely strong support for the pure occlusion-sensitive model relative to the pure occlusion-blind model, providing quantitative backing to the qualitative failure of an egocentric model to predict differences between occlusion-present and occlusion-absent trials. Somewhat surprisingly, however, we also found support for the pure occlusion-sensitive speaker over the mixture model: The Bayesian Occam's razor determined that the additional model complexity contributed by the mixture parameter was not justified by sufficient increases in predictive accuracy and prefers the simpler model. This result, along with the corresponding listener results reported in Appendix D, suggests that the simplified variant of the director–matcher task used in Experiment 1 may not be sufficiently cognitively demanding to elicit (resource-rational) failures of perspective-taking in either speakers or listeners, and may correspond to the optimal levels of perspective-taking predicted at lower levels of perspective-taking cost $\beta$ (see Fig. 2).

Next, 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 distractor-absent versus distractor-present 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. C1). Finally, we examined parameter posteriors for the best-fitting occlusion-sensitive model with $w_S = 1$ (see Fig. S2): The inferred production cost for *texture* was significantly higher than that for the other features, accounting for why participants were overall less likely to include texture in their descriptions relative to color.

## Appendix D: Supplemental experiment

To further motivate our rationale for using the original materials and design from Keysar et al. (2003) in Experiment 2, we conducted a version of the same listener manipulation using the stimuli from Experiment 1.[6] We recruited $N = 72$ participants on Amazon Mechanical Turk and placed them into the same environment used in Experiment 1 with several key changes to the trial sequence. First, we removed the occlusion-absent condition, so every trial contained occlusions, generated randomly on each trial to cover two cells. Second, in every block of eight trials, we included two "critical trials" where we placed an occluded distractor in the listener's private view with the same shape as the target. Third, we added a "practice" block of four noncritical trials at the front of the trial sequence, leading to a total of 28 trials. Otherwise, the experiment design and stimuli were held constant.

Instead of recruiting real speakers for real-time, multiplayer interaction, as in Experiments 1 and 2, we used a simple bot as our scripted confederate. On critical trials, it produced an ambiguous utterance mentioning only the shape (e.g., "the square"). When an object with the same shape as the target appeared in common ground, it would produce

an utterance mentioning a perfectly distinguishing attribute (e.g., "the blue square" if there were no other blue objects) or produce an exhaustive three-word utterance if distractors existed on each dimension. Otherwise, to prevent short utterances from being suspicious, it produced shape-only utterances on two-thirds of filler trials, and added an additional modifier on the other one-third.

As in Experiment 2, our primary measure is the proportion of errors on critical trials. Unlike in Experiment 2, we found no evidence that errors on critical trials, requiring the use of theory of mind, were higher than on filler trials. Excluding practice trials, we found an error rate of 4.9% on critical trials and an error rate of 8.4% on filler trials. If anything, we find that the error rate on critical trials was significantly *lower* than on filler trials, $\chi^2(1) = 5.9$, $p = .015$. When we implement the strict exclusion criterion used in Experiment 2, excluding $N = 25$ participants who made more than one error on filler trials (under the rationale that these participants may be generally unattentive), we find that only 9 of the remaining 49 participants made any critical errors at all, at any point in the experiment, and the error rate was still not significantly higher than the error rate on filler items (4.6% for critical trials, 3.3% for filler items, $\chi^2(1) = 1.02$, $p = .312$). Under both analyses, the prevalence of errors was dramatically lower than that reported by Keysar et al. (2003) or in our Experiment 2, using the Keysar stimuli. The presence of this ceiling effect suggests that this simple stimulus space may not be sufficiently cognitively demanding for listeners (due to a variety of possible design factors) to allow us to ask more detailed questions about failures of perspective-taking, so we did not proceed to run the corresponding "unscripted" condition.