

Psychological Review

The Social Basis of Referential Communication: Speakers Construct Physical Reference Based on Listeners' Expected Visual Search

Julian Jara-Ettinger and Paula Rubio-Fernandez

Online First Publication, December 30, 2021. <http://dx.doi.org/10.1037/rev0000345>

CITATION

Jara-Ettinger, J., & Rubio-Fernandez, P. (2021, December 30). The Social Basis of Referential Communication: Speakers Construct Physical Reference Based on Listeners' Expected Visual Search. *Psychological Review*. Advance online publication. <http://dx.doi.org/10.1037/rev0000345>

The Social Basis of Referential Communication: Speakers Construct Physical Reference Based on Listeners' Expected Visual Search

Julian Jara-Ettinger¹ and Paula Rubio-Fernandez^{2, 3}

¹ Department of Psychology, Yale University

² Department of Philosophy, Classics, History of Art and Ideas, University of Oslo

³ Department of Brain and Cognitive Sciences, MIT

A foundational assumption of human communication is that speakers should say as much as necessary, but no more. Yet, people routinely produce redundant adjectives and their propensity to do so varies cross-linguistically. Here, we propose a computational theory, whereby speakers create referential expressions designed to facilitate listeners' reference resolution, as they process words in real time. We present a computational model of our account, the Incremental Collaborative Efficiency (ICE) model, which generates referential expressions by considering listeners' real-time incremental processing and reference identification. We apply the ICE framework to physical reference, showing that listeners construct expressions designed to minimize listeners' expected visual search effort during online language processing. Our model captures a number of known effects in the literature, including cross-linguistic differences in speakers' propensity to over-specify. Moreover, the ICE model predicts graded acceptability judgments with quantitative accuracy, systematically outperforming an alternative, brevity-based model. Our findings suggest that physical reference production is best understood as driven by a collaborative goal to help the listener identify the intended referent, rather than by an egocentric effort to minimize utterance length.

Keywords: computational modeling, social cognition, pragmatics, reference

Supplemental materials: <https://doi.org/10.1037/rev0000345.supp>


According to the Gricean Maxim of Quantity, speakers should provide their listeners with as much information as required for the purpose of the exchange, but not more (Grice, 1975). That is, speakers should produce unambiguous messages while remaining succinct. This maxim is central to successful communication, as it enables listeners to go beyond the literal meaning of what speakers say. Imagine, for example, that your partner asked you to pass them “the plastic mug.” Holding your partner to the Maxim of Quantity would allow you to infer that there are several mugs of different materials to choose from, and that your partner had a specific one in mind. Similarly, if, in the same situation, your partner requested “the mug,” you could conclude that they were unaware that there were multiple mugs available and that they had, therefore, failed to make themselves understood.

Despite its intuitive appeal, evaluating whether everyday communication is structured around the Maxim of Quantity is far from

straightforward; it requires formalizing how much information is necessary for the purpose of an exchange, as well as quantifying how much additional information might be included in different messages. Classical experimental (Engelhardt et al., 2011; Sedivy, 2003, 2005) and computational (Dale & Reiter, 1995; Gatt et al., 2014; Krahmer & van Deemter, 2012; van Deemter et al., 2012) studies initially interpreted the Gricean maxim in terms of a balance between informativity (say as much as necessary ...) and brevity (... but say no more). That is, messages should contain enough information to be unambiguous, but remain brief. For example, if a red plastic mug and a blue porcelain mug were on a table, the expression “the mug” would be *underinformative* (as it fails to uniquely identify the referent) and “the red plastic mug” would be *overinformative* (as it says more than is needed to uniquely identify the mug). However, what about choosing between color modification (“the red mug”) or material modification (“the plastic mug”)? This basic example shows how a brevity-based interpretation of the Gricean maxim can face a first challenge: it fails to distinguish between descriptions matched in word length (e.g., “the plastic mug” vs. “the red mug”), which speakers do not treat as being equivalent (Sedivy, 2005).

A second challenge to this simple interpretation of the Gricean Maxim of Quantity is that speakers routinely violate it: while speakers are typically motivated to be sufficiently informative, they often produce overinformative expressions, particularly when using color words (e.g., referring to “the red mug” in a context with a single mug in sight). In response to this puzzle, researchers have argued that people might produce overinformative expressions to preempt potential ambiguities that they are unaware of (Belke, 2006; Belke & Meyer, 2002; Engelhardt et al., 2006; Fukumura & Carminati, 2021; Hawkins et al., 2021; Koolen et al., 2013; Pechmann, 1989). For example, people might choose to say “the red mug,” even when they see a single mug, to

Julian Jara-Ettinger  <https://orcid.org/0000-0002-6167-1647>

Paula Rubio-Fernandez  <https://orcid.org/0000-0003-1622-0967>

Model code, experiment data, and analyses files are available at <https://osf.io/bezua/>. These studies were not preregistered. Some of the data in the manuscript were presented at the 34th Annual CUNY Conference on Human Sentence Processing. A preprint of this manuscript is available at the PsyArxiv preprint server.

We thank Maddie Long for help collecting data, and Vishakha Shukla and Vrinda Bhatia for help coding the data. We thank Herb Clark, Sammy Floyd and Emiel Krahmer for useful comments on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Julian Jara-Ettinger, Department of Psychology, Yale University, 2 Hillhouse Avenue, New Haven, CT 06511, United States. Email: julian.jara-ettinger@yale.edu

account for the possibility that there might be other mugs that they did not notice. Similarly, inviting someone to sit by “the newly painted table” in the context of one table could be a form of cooperative behavior (Dale & Reiter, 1995).

Here, we advance an alternative interpretation of the Gricean maxim, inspired by the Principle of Least Collaborative Effort: Interlocutors share the same goal of minimizing the total effort spent on both the production and resolution of a message (Brennan & Clark, 1996; Clark, 1996; Clark et al., 1983; Clark & Marshall, 1981; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986; Horton & Brennan, 2016). While the Principle of Least Collaborative Effort was originally proposed to account for extended conversations (and we return to this point in the discussion), it carries an important implication for referential communication: speakers are motivated to help the listener identify the intended referent with a limited expenditure of *time* and *effort*. Therefore, speakers may choose overinformative expressions whenever they believe that their use will benefit collaborative communication. That means that amongst two equally informative descriptions, speakers should not necessarily favor the shorter one, and they should choose instead whichever expression that they believe will help the listener identify the referent more efficiently (e.g., “the red mug” instead of “the plastic mug,” if color is more visually salient than material), allowing speakers to use adjectives in a *nonrestrictive* manner.

The idea that speakers are motivated to help listeners identify an intended referent with limited time and effort highlights our two proposed pillars of collaborative referential communication: First, a sensitivity to how listeners process language in real time and, second, a sensitivity to the processes that listeners undergo, as they seek to resolve reference. To evaluate this proposal, we present a computational framework, where referential expressions are produced based on their communicative efficiency for listeners. To achieve this, our framework presents a complete computational interpretation of the idea that speakers consider how listeners will process and seek to resolve reference, using a mental model of real-time visual search.

Our proposal aims to explain a growing body of recent empirical work showing patterns of reference production that are difficult to explain through a brevity interpretation of the Gricean maxim (and we review these effects in detail below). These studies support the idea that reference production is shaped by a goal to facilitate processing for the listener, rather than minimize production costs for the speaker (Arts et al., 2011; Gwendolyn et al., 2021; Long et al., 2020; Mangold & Pobel, 1988; Paraboni et al., 2007; Paraboni & van Deemter, 2014; Rubio-Fernandez, 2016, 2019; Rubio-Fernandez et al., 2020; Sonnenschein & Whitehurst, 1982; Tourtouri et al., 2019). Our proposal, therefore, adds to a growing body of work that has found a brevity-based interpretation of the Gricean Maxim to be inadequate, advancing more nuanced interpretations (Dale & Reiter, 1995; Degen et al., 2020; Rubio-Fernandez, 2016). In this paper, we present a first test of our model at Marr’s computational level of analysis (Chater & Oaksford, 1999; Marr, 1982): we evaluate whether our model’s computations capture reference production, leaving open questions about the algorithmic nature of these computations (which we discuss in Referential Communication Beyond a Computational Level of Analysis section). Our model and tests focus on the domain of physical reference—the area with most empirical work on the Maxim of Quantity. This enables us to evaluate our proposal with respect to published data sets and standard experimental tasks. However, not

all reference is about physical objects. Therefore, in the discussion, we return to an analysis of other types of reference and discuss the implications of our work in these areas (see section “Collaborative Reference in More Complex Situations”).

In the remainder of this paper, we begin by reviewing a general framework for modeling reference and social reasoning (see section “Computational Models of Reference”), and introduce our proposal in detail, which is centered on the idea that speakers act towards a collaborative goal considering listeners’ real-time processing (see section “The Incremental Efficiency Theory”). Next, we present the Incremental Collaborative Efficiency (ICE) model, which produces referential expressions by considering listeners’ real-time visual search (see section “Computational Framework”); and we discuss its key differences from related approaches to modeling reference (see section “Related Advances in Models of Reference”). Finally, we show that our model explains production and acceptability judgments better than a brevity-based reference model, and that it unifies a range of empirical phenomena around over-specification and cross-linguistic variation (see sections “Validating Speaker–Listener Alignment,” “Can the ICE Model Explain Referential Over-Specification?,” and “Can the ICE Model Explain Graded Acceptability Judgments?”). We end with a discussion of the implications of our results for our understanding of reference as a collaborative process (Discussion section).

Computational Models of Reference

The simplest interpretation of the Gricean Maxim of Quantity can be operationalized as the construction of an utterance that simultaneously maximizes the probability that a listener correctly identifies the intended meaning in a semantic space (i.e., making the message as informative as required) while minimizing its production cost (i.e., while saying no more). Formally, given an intended meaning m , the Maxim of Quantity can be implemented as the process of selecting a utterance u that maximizes the utility function (Equation 1):

$$U(u, m) = \overbrace{f(p_L(m|u))}^{\text{say as much as necessary} \dots} - \underbrace{C(u)}_{\dots \text{but no more}}, \quad (1)$$

where $U(u, m)$ is the utility of producing utterance u to communicate meaning m . $p_L(m|u)$ is the probability that a listener would identify meaning m upon hearing utterance u , and f is a measure of communicative success that depends on this probability. For example, f might be a binary indicator of whether the utterance is unambiguous, or it could encode the information-theoretic measure of surprisal (Degen et al., 2020; Goodman & Frank, 2016; Kao et al., 2014). The final term, $C(u)$, expresses the cost associated with producing utterance u .

Under this framework, the first term, $f(p_L(m|u))$, captures the idea that speakers ought to be sufficiently informative. Returning to the example from the introduction, an ambiguous expression such as “the mug” would have a low utility, because the listener would have a lower chance of identifying the intended mug, relative to more informative utterances, such as “the red mug,” “the plastic mug,” “the red plastic mug,” or even relative clauses, such as “the mug that

is red and made of plastic.” While this first term identifies underinformative utterances, it fails to distinguish between utterances that are sufficiently informative, such as “the red mug” and “the mug made of plastic.” The second term, $C(u)$, implements the pressure to be succinct. Thus, the cost of production would favor “the red mug” and “the plastic mug,” because they ensure communicative success in shorter utterances relative to “the red plastic mug” or “the mug that is red and made of plastic.”

This general approach of modeling social reasoning as utility-maximization has enjoyed wide success in capturing referential communication in both speakers and listeners (Degen et al., 2020; Frank & Goodman, 2012; Goodman & Frank, 2016; Kao et al., 2014) and it shares a similar structure to nonlinguistic social inferences that, from infancy, structure how people reason about each other’s goals and mental states (Jara-Ettinger, 2019; Jara-Ettinger et al., 2016; Jern et al., 2017; Liu et al., 2017; Lucas et al., 2014).

While the role of information gain and its tradeoff with costs have received substantial theoretical and empirical attention, this framework also implicitly posits a conceptual distinction between the two processes: the pressure to be sufficiently informative is allocentric—determined by considering how a listener might interpret an utterance—while the pressure to be brief is egocentric—estimated as the speaker’s effort in producing an utterance, with no regard for a listener’s perspective. In the next section, we introduce our framework, centered on allocentric cost representations that track the time and effort that listeners might need to resolve different referential expressions.

The Incremental Efficiency Theory

Under the framework presented above, interpreting the Gricean Maxim of Quantity as one of brevity would dictate that the cost function $C(u)$ ought to be egocentric and penalize utterances based on length. This formalization, however, faces two empirical challenges. First, people often produce redundant adjectives, especially color (Arts et al., 2011; Koolen et al., 2013; Rubio-Fernandez, 2016, 2019; Sedivy, 2003, 2005). For example, your partner may ask you for “the red mug,” even when there is only one mug in the scene. Second, English speakers tend to produce more redundant color adjectives than Spanish speakers when referring to the same objects in the same visual displays (Rubio-Fernandez, 2016, 2019; Rubio-Fernandez et al., 2020; Wu & Gibson, 2021). The brevity interpretation of the Gricean maxim fails to explain why people sometimes choose to use redundant adjectives, nor why speakers of languages with prenominal modification (such as English, e.g., “the red mug”) use color adjectives more frequently than speakers of languages with postnominal modification (such as Spanish, e.g., “la taza roja”) in situations, where color is equally overinformative.

This view also faces two theoretical challenges. First, the brevity interpretation does not distinguish between referential expressions of equal length that may be equally informative (although this concern can be alleviated through algorithms that preferentially consider one adjective type over others, independent of informational content, e.g., color before material; Dale & Reiter, 1995). Second, it fails to take into account the order in which words are made available to listeners, based on the language’s word order. Such an analysis, therefore, does not distinguish “the red mug” and “la taza roja” (“the mug red” in Spanish), contrary to empirical evidence, showing that the two expressions are not treated as equivalent by speakers and are processed differently by listeners

(as they make information available to listeners in a different order, despite ultimately having the same meaning in English and Spanish; Rubio-Fernandez et al., 2020).

The view of referential communication that we propose here—the *incremental efficiency theory*—posits that speakers aim to produce referential expressions that are incrementally efficient for their listeners. That is, an expression’s efficiency ought to be computed by evaluating how it guides listeners towards the intended referent, as they process words incrementally. This theory predicts that, amongst equally informative descriptions, speakers should prefer the one that ensures communicative success most efficiently. In the case of physical reference, this should lead to a preference for encoding visually salient properties, to the extent that they help the listener find the object. For example, since color is normally more visually salient than material and can be identified from one’s visual periphery (whereas identifying material normally requires fixating on the object), speakers should favor “the red mug” over “the plastic mug” to save the listener time. The complexity of the visual scene (e.g., whether it is cluttered or sparse) and the language’s word order (whether adjectives are prenominal or postnominal) are also important when determining the relative efficiency of a modifier. Thus, this view of referential communication naturally integrates an expression’s informativity (e.g., which properties may uniquely identify the mug you want to request), perceptual factors (e.g., whether the color red is perceptually salient in a given scene) as well as incrementality (e.g., whether you will formulate your request in English or in Spanish; Rubio-Fernandez et al., 2020).

By grounding communication in real-time cognitive processes, the incremental efficiency theory is, therefore, structured around two ideas. First, the relative efficiency of a referential expression is conceptualized incrementally rather than globally. Second, utterance costs are allocentric, capturing the estimated difficulty of listeners’ visual search for the referent (rather than capturing egocentric production costs).

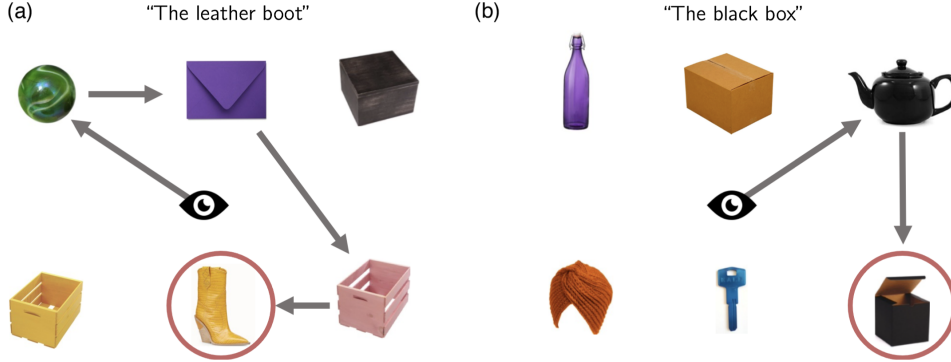
Computational Framework

Model code is available on our project’s OSF repository (<https://osf.io/bezua/>; Jara-Ettinger & Rubio-Fernandez, 2021a), and additional information is available in our Supplemental Materials. To illustrate the logic of the ICE model, Figure 1 shows examples of the types of situations that we consider. Each event consists of a grid of objects, one of which is the target referent. Throughout, we consider four different ways to describe the target: by providing the noun, by providing the noun accompanied by either its material or its color, and by providing the noun accompanied by both its material and color. We do not consider other properties that speakers could rely on (such as texture, shape, or size), although our framework can be extended to include additional adjectives (provided that the visual search strategy associated with the adjective type is known).

We take as a starting point the utility-based frameworks that capture human commonsense psychology (Jara-Ettinger et al., 2016, 2020; Liu et al., 2017) and referential communication (Degen et al., 2020; Frank & Goodman, 2012; Goodman & Frank, 2016; see Computational Models of Reference section for a brief review). Our model embraces these utility-based frameworks, but focuses on implementing the driving forces of our account (see The Incremental Efficiency Theory section). To model the sensitivity to real-time language processing, we compute utilities through a listener model used to simulate real-time listener visual search, as an expression is processed

Figure 1

Sample Displays From Our Experiment With Qualitative Visualizations of Our Model's Visual Search Component (Using a Fixation Rate of One Fixation Per Word)



Note. (a) Example visual trace as the listener hears “the leather boot.” Because leather encodes a material property, the listener randomly fixates on objects in search for leather objects when it has heard “the leather . . .” After hearing the full utterance, the listener model continues searching for the object until identifying one that matches both properties (a boot that is made of leather). (b) Example visual trace as the listener hears “the black box.” Upon hearing “black,” the listener can immediately direct its fixations to black objects (as these can be detected from the periphery). When hearing the full utterance “the black box,” the listener evaluates whether the fixated object is a box, and if it is not, it redirects its attention to other black objects in the scene until identifying the referent. Our model draws samples of this listener model to determine whether an utterance will successfully identify the target referent, and to estimate the expected burden imposed on the listener (quantified in terms of number of fixations). See the online article for the color version of this figure.

incrementally. This listener model enables the speaker to identify utterances designed to minimize the listener’s time and effort (rather than conceptualizing speaking as a fundamentally egocentric costly activity).

Formally, given a target object t in a visual scene, our model estimates the utility of different referential expressions, $U(r, t)$, and then assigns a probability to each expression by softmaxing this utility function, such that the probability associated with referential expression r is given by Equation 2:

$$p(r) \propto \exp(U(r, t)/\tau), \quad (2)$$

where $\tau \in (0, \infty)$ is the softmax parameter, which determines the speaker’s capacity to identify the utterance with the highest utility (see Model Parameters section for a discussion of how parameters affect our model). We define the utility of referential expression r to communicate target t as

$$U(r, t) = R(t)p_L(t|r) - C(T; t, r). \quad (3)$$

$p_L(t|r)$ is the probability that the listener will correctly identify the target t upon hearing referential expression r , and $R(t)$ is the speaker’s subjective reward for successfully communicating target t . Thus, their product captures the referential expression’s expected reward. We define $p_L(t|r)$ as $1/n$, where n is the number of objects in the visual scene that are consistent with expression r . For example, in Figure 1a, the expressions “the boot,” “the yellow boot,” “the leather boot,” and “the yellow leather boot” all assign probability 1 to the boot. By contrast, in Figure 1b, the expressions “the box,” and “the cardboard box” only assign a probability of 0.5 to the target, because the display contains two cardboard boxes. However, the expressions “the black box” and “the black cardboard box” assign a probability of 1 to the target, because the second box is brown (not black).

The second term, $C(T; t, r)$, is a measure of the expected time T that it will take the listener to identify a valid target t , as they process the referential expression r . This term is computed through a mental model of real-time visual search (described below in Listener Visual Search Model section). Thus, this term captures the key idea of an allocentric cost, where the speaker is motivated to convey the meaning to the listener in an efficient manner.

Equation 3 can be used to compute partial utterances and it, therefore, supports tracking how a referential expression’s utility changes over time. However, speaker preferences in our model are guided by the final utility associated with each expression. This final utility captures the usefulness of an expression, accounting for the fact that it will be processed and resolved by the listener incrementally.¹

Listener Visual Search Model

Our visual search model captures the expectation that listeners integrate the meaning of each word, as they hear it, and use it to guide their visual search for the referent. Our visual search model only aims to approximate an intuitive theory of visual search, allowing us to formalize and test the incremental efficiency theory. However, this initial model is likely missing details in people’s intuitive theory of visual search that may make its predictions more nuanced. We return to this point in the discussion (see section “Referential Communication Beyond a Computational Level of Analysis”).

For clarity, we begin with a qualitative description of our visual search model and then turn to a more technical presentation

¹ Another alternative would be to build an expression incrementally, by selecting at each point whichever word adds the most utility to an expression. However, such an approach would capture the idea that speakers are choosing words one by one as they speak, without thinking about the final message, which is not what our theory proposes.

(Technical Description section, which readers not interested in precise implementation details can skip). Given a visual scene, our model searches for the referent as soon as the speaker begins their referential expression (for simplicity, we do not process the word “the,” as it does not discriminate the target in the contexts that we consider here). In Figure 1a, for example, our model first searches for leather-made objects (when it has processed “the leather . . .”) and then updates its search for leather-made boots (when it has processed “the leather boot”).

To model this search, we implemented a distinction between object properties that can be detected from the visual periphery and properties that only be evaluated by fixating on the object. Specifically, we assumed that listeners can detect an object’s color from the periphery, but that an object’s material and category can only be revealed by directly fixating on it (see Validating Speaker–Listener Alignment section for an empirical validation of this assumption).

Equipped with this distinction, our visual search model can direct its fixations to objects with a target color, but must scan the scene when trying to identify objects with a target material or category. Returning to the example in Figure 1a, because “leather” encodes a material property, our model (upon hearing “the leather . . .”) begins by fixating on random objects to check whether they are made of leather or not. When the speaker completes the expression (“the leather boot”), the model updates its criteria and evaluates every object that it fixates on by checking whether it is made of leather and whether it is a boot. Whenever the model encounters a material match before hearing the noun (e.g., finds a leather object when the speaker has only said “the leather . . .”), it pauses momentarily to hear the speaker’s next word and confirm that it found the match (and retriggers its visual search if the selected objects turns out not to be the referent).

The example in Figure 1b shows how this visual search strategy changes when color information is available. Here, because “black” encodes a color property, our model (upon hearing “the black . . .”) can immediately identify black objects in the scene, and directs its fixations to one of them. When the speaker completes the expression (“the black box”), the model updates its criteria, and checks that if the black object, it is considering is indeed a box. If it is not, the model redirects its fixation to another black object and continues to do until finding a match.

The speed with which the search model identifies a referent depends on the linguistic information available to the listener (and the order in which the speaker makes it available, as determined by their language’s word order), the complexity of the visual scene, and the visual salience of the properties encoded in each word. In addition, due to the probabilistic nature of our model, the same expression in the same visual context can produce variable search times, depending on whether the visual search model happens to fixate on the right objects early in the search or not. Thus, in our complete model, we approximated the expected number of fixations required to identify the target via Monte Carlo simulations (using 1,000 samples). We then used this estimate as a proxy for the expected time a listener engages in visual search for any referential expression in any visual context. This component, therefore, aims to capture speakers’ intuitions about the relative ease or difficulty of visual search at a computational level of analysis (Marr, 1982), without claiming that the speakers quantify listeners’ search cost in terms of fixations at an algorithmic level (see Referential Communication Beyond a Computational Level of Analysis section).

Technical Description

Algorithm 1 shows a more technical schematic of our visual search model. We define a visual scene as a set of objects \mathcal{O} , where each object $o \in \mathcal{O}$ is a 3-tuple $o = \{o_c, o_m, o_t\}$ that captures the object’s color (o_c), material (o_m), and type (or category; o_t). In Figure 1, for example, the target objects are represented as {yellow, leather, boot} and {black, cardboard, box}, in panels a and b, respectively.

The listener’s computations are guided by its working memory (WM)² that stores the utterance (integrating new words as they are heard), and an evaluation function Eval which returns True if an object $o \in \mathcal{O}$ is consistent with the information encoded in the heard utterance (WM) and False otherwise.

Fixations are modeled as a function that samples objects from \mathcal{O} without replacement. This function can either sample objects uniformly from the full set ($\text{Fixation} \leftarrow \{o | o \in \mathcal{O}\}$, used to model visual scan) or it can sample objects conditioned on properties that can be detected from the periphery ($\text{Fixation} \leftarrow \{o | o \in \mathcal{O}, o_c = C\}$, where C is a target color). Therefore, when an utterance in WM contains color information, the visual search model uses it to selectively sample color-matched objects, and samples objects from the entire set otherwise. We track the relative speed of fixations to words through a fixations per word (FPW) parameter, which determines how many fixations the listener can perform between each word that the speaker produces.

Put together, our model performs a search loop, where it continuously samples objects in the scene and tests whether they match the information encoded in WM, which continuously expands as the utterance unfolds over time. When our model finds a partial match, it temporarily halts searching and waits to hear the speaker’s next word (retriggering a search if the partial match turns out to be incorrect), and it returns the object that matches all properties in the utterance once the utterance is complete.

Alternative Brevity Account

To better understand how our model behaves relative to a brevity interpretation of costs, we also implemented an alternative model, where the utility of a referential expression is given by

$$U(r, t) = R(t)p_L(t|r) - C(r). \quad (4)$$

As reviewed in the introduction, a brevity-based interpretation of the Gricean maxim of quantity already faces several empirical and theoretical challenges (Dale & Reiter, 1995). The purpose of developing this model was only to have a reference point from which to evaluate the ICE model predictions. Thus, we see a brevity model based on a literal interpretation of the Gricean Maxim of Quantity as a useful control: although such a model is unlikely to generate human-like data given speakers’ tendency to produce overinformative descriptions, it helps shed light on how, why, and when our ICE model deviates from a brevity-based referential system.

To make the models comparable, the brevity-based model has the same structure as our ICE model (Equation 3), differing only in that we replace the allocentric cost function with an egocentric one, $C(r)$, which assigns a cost based on utterance length, estimated in number of

² Although it is easy to add a working memory limitation, we did not include one in our implementation as all referential expressions that we consider are short.

Algorithm 1*Visual Search Algorithm*

```

1: Initialize  $\mathbb{O}$  ▷ Collection of all objects in scene
2: Initialize WM ▷ Working memory
3: Search ← TRUE
4: Fixation count ← 0
5: while reference is unresolved do
6:   if Fixation count mod FPW = 0 then ▷ Listener hears next word
7:     WM ← WM + next word
8:   end if
9:   if Search = TRUE then ▷ Search for objects matching information in WM
10:    if WM contains color information then
11:      Fixation ← { $o \mid o \in \mathbb{O}, o_c = C$ } ▷ Sample color-matching object
12:    else
13:      Fixation ← { $o \mid o \in \mathbb{O}$ } ▷ Sample object from scene
14:    end if
15:  end if
16:  if EVAL(Fixation, WM) is TRUE then
17:    Search ← FALSE ▷ Sample is a good candidate referent
18:  else ▷ Sample is inconsistent with information in WM
19:    Remove  $o$  from  $\mathbb{O}$  ▷ Remove candidate from consideration
20:    Search ← TRUE ▷ Continue searching for referent
21:  end if
22:  if Utterance is complete & EVAL(Fixation, WM) is TRUE then
23:    return Fixation ▷ Utterance complete and referent identified
24:  end if
25:  Fixation count ← Fixation count + 1
26: end while

```

words (Equation 4). Combined, the first term leads the speaker to produce expressions that uniquely identify the target, while the second term incentivizes the speaker to favor shorter expressions over longer ones.

Model Parameters*ICE Parameters*

Our ICE model has three parameters: the reward associated with communicating successfully ($R(t)$), the softmax parameter (τ), and the rate of FPW, which specifies the speed of fixations relative to the speed of speech.

How much power do these parameters have over our model? The first parameter, $R(t)$, determines whether an agent is sufficiently motivated to communicate unambiguously. When $R(t)$ is low, our speaker model might choose to forego clarity whenever unambiguous utterances would be prohibitively costly for the listener to process. As long as $R(t)$ is sufficiently high, however, the speaker will consistently favor clarity over efficiency, with the exact value of $R(t)$ having no further impact. Therefore, although $R(t)$ is a continuous variable, it effectively encodes a categorical distinction that determines whether clarity is negotiable or not.

The second parameter, τ , determines the strength with which the speaker prefers the highest utility utterance relative to alternative expressions. This parameter cannot change our model's qualitative predictions (i.e., if our model prefers one utterance over another,

τ cannot reverse this preference), but it modulates the strength of the speaker's preference. When τ is low, the speaker model shows an overwhelming preference for the utterance with the highest utility, even when alternative expressions have comparably high utilities. As τ increases, the speaker model relaxes its preference, finding utterances more or less acceptable as a function of their utility (but always preferring higher utility expressions over lower utility ones).

Finally, the third parameter (FPW) determines how many fixations (and, consequently, object evaluations) the speaker can perform between hearing each word. When FPW is low, its exact value has little effect on our model (e.g., if the listener can perform one or two fixations between each word, the speaker can still plan for how to minimize the listener's time and effort resolving the referent). However, when FPW is set to an unrealistically high value, our model breaks down: At high enough levels, FPW enables the listener to fixate on every single object between each word, allowing the listener to have near-instant perfect knowledge about every object (including its material and category), creating a situation where visual salience ceased to be relevant, as the listener already knows all properties of all objects.

Because these parameters have little influence on our model predictions (except when set to extreme values that capture unrealistic situations, as described above), we were able to fix all of their values qualitatively, and prior to data collection. This is a standard practice in Bayesian models of Theory of Mind—from which our model builds—as the core predictions are derived by having

nuanced models of other agents that cannot be heavily affected by parameter choice (Baker et al., 2009; Jara-Ettinger et al., 2020; Jern et al., 2017; Ullman et al., 2009). The reward for communicating was set to a constant $R(t) = 30$. This value was chosen, so that it was high enough to outweigh a cost of up to 30 fixations (set as a conservative estimate of nearly twice the amount necessary, as our model evaluations consider scenes with up to 16 objects). For simplicity, we assumed an equal speed of fixations and speech ($FPW = 1$). We next set the softmax parameter to $\tau = 2$, chosen qualitatively to ensure that the model produced probability distributions with variable strengths of beliefs (such that they could be correlated with participant judgments; a τ level too low would produce distributions that place all weight on the highest utility with no variability, while a τ too high would produce uniform distributions with no information about the underlying utilities).

Finally, to ensure that these predetermined parameter choices do not inadvertently bias our results, we additionally fit these models to participant data as a secondary analysis. We report the results of this fitting process in the Results section of our main experiment.

Brevity Model Parameters

Similar to the ICE model, the Brevity alternative model also has three parameters. The first two are identical to the ones from our main model: a reward $R(t)$ associated with communicating successfully and a softmax parameter τ . However, rather than having a fixations per-word parameter in the cost function, the alternative model has a parameter word cost (WC) that determines the cost of producing each word.

All three parameters have a parallel effect to the ones from our ICE model, with the exception of WC, which in this model shapes how strongly the model values brevity. To make the models comparable, we aligned the alternative Brevity model with the ICE model, such that their relative degrees of confidence and uncertainty were matched (e.g., if the ICE model was fully confident about which utterance to select on n of the trials in our experiments, then the Brevity model should also be confident on n of the trials, although the trials, where each model is confident would naturally be different ones). This enables us to compare the models directly while removing the possibility that one model might outperform the other due to differences in how often they express different degrees of confidence. To achieve this alignment, we generated our ICE model's predictions for the stimuli used in our main experiment (see Can the ICE Model Explain Graded Acceptability Judgments? section) and we then found the Brevity model parameters that matched the range of ICE model predictions as closely as possible (using an Euclidean measure between the two probability distributions; see Supplemental Materials for details). This resulted in a communication reward of $R(t) = 1$, a word cost of $WC = 0.09$, and $\tau = 0.12$ (note that because each model's range of utilities is different, the corresponding softmax parameter τ that produce equal variability will also be different).

Related Advances in Models of Reference

Before turning to our model evaluations, it is useful to consider recent advances in models of reference production that aim to explain similar phenomena to the ones we do here. In particular, two models have been recently proposed to account for the observation that people

often produce over-specific referential expressions. The first, the Probabilistic Referential Over-specification (PRO) model, posits a generative model that probabilistically chooses to include or omit adjectives (van Gompel et al., 2019). By adjusting the probability of over-specification for different adjective types, this model successfully captures patterns in referential communication. Nonetheless, PRO is agnostic about the psychological causes behind the predicted probability of including or omitting adjectives. From this standpoint, our model can be thought of as being consistent with the PRO model, with our contribution being on advancing a model of the cognitive processes that give rise to the probabilities of over-specifying, and modeling how they change as a function of the visual context, a language's word order, and their usefulness for visual search.

Second, the Rational Speech Act (RSA) framework (Goodman & Frank, 2016) has also been used to successfully model reference (Frank & Goodman, 2012), handling incremental processing (Waldon & Degen, 2021), and referential over-specification (Degen et al., 2020). In RSA models, speakers generate referential expressions through recursive social reasoning, considering the likelihood that a listener will recover the intended referent, who in turn, considers why the speaker chose the words that they did (Frank & Goodman, 2012; Goodman & Frank, 2016). These models classically use a deterministic association between words and meanings and did not directly account for over-specification (i.e., a speaker should not say "red apple" when "apple" already reveals the intended target). continuous semantics RSA (cs-RSA) extends RSA using a noisy probabilistic association between words and meanings. In doing so, cs-RSA produces a tendency to over-specify, with the goal of reducing noise in the utterance (e.g., cs-RSA might select "red apple" when a single apple is in view, to avoid the chance that a listener might select the wrong item under the bare description "apple" due to the noise in the meaning of the word "apple").

Our work differs from cs-RSA in two critical ways. First, at a theoretical level, cs-RSA explains over-specification through a mechanism, where perceptual salience affects the informativity of words (Degen et al., 2020). By contrast, we propose that perceptual salience affects speaker estimates of listener visual processing, rather than the informativity of words. Second, although Degen et al. (2020) account for visual salience through its "noisy semantics," this effect is instantiated through a parameter fit to data, whereas our model includes a full computational interpretation of how a speaker might estimate the difficulty of the listener's visual search.

Our work, therefore, adds to this area of research by introducing a formal computational theory of referential efficiency that aims to explain over-specification by computing allocentric costs. Our work, therefore, moves beyond past work by introducing a computational model which includes a detailed listener model of visual search during incremental language processing.

Validating Speaker–Listener Alignment

Our computational model relies on two critical assumptions: (a) color is more visually salient than material and (b) speakers are aware of how this affects visual search. Eye-tracking studies have revealed that listeners are quick to fixate on color-matching objects as soon as they hear the corresponding color adjective (Eberhard et al., 1995; Rubio-Fernandez & Jara-Ettinger, 2020; Sedivy et al., 1999), confirming that color is a visually salient cue during language processing. The above assumptions are also consistent with recent work on referential

over-specification. Listeners are often quicker to identify a referent when they are provided with a (technically redundant) color word, despite the utterance being longer (Arts et al., 2011; Gwendolyn et al., 2021; Mangold & Pobel, 1988; Paraboni et al., 2007; Paraboni & van Deemter, 2014; Sonnenschein & Whitehurst, 1982; Rubio-Fernandez, 2021; Tourtouri et al., 2019). Consistent with the findings from language comprehension studies, speakers are more likely to use redundant color words as a function of their visual salience (Long et al., 2021; Rubio-Fernandez, 2021; Rubio-Fernandez et al., 2020).

While vision research has long established the salience of color (e.g., Davidoff, 1991, 2001; Adams & Chambers, 2012; Bramão et al., 2011), the comparison with material properties has not been extensively investigated. Sedivy (2005) showed that speakers tend to use color adjectives redundantly more often than material adjectives. In an eye-tracking task, she also showed that listeners are more likely to interpret material adjectives contrastively (which allowed them to anticipate the right member of a pair) than color adjectives (which did not elicit anticipatory looks to the target). A recent study by Kursat and Degen (2021) revealed faster recognition of color than material in a property verification task, and higher rates of redundant color adjectives than material. However, while generally consistent with the assumption that color is more visually salient than material, these studies did not directly compare speakers' production of color and material adjectives with listeners' visual search by color and material during adjective processing. Thus, the visual salience of color relative to material, and speakers' exact sensitivity to how visual salience affects visual search, has not been directly quantified. We, therefore, began with an experiment that tested for the robustness of assumptions (a) and (b) above by employing the same visual materials with both speakers and listeners, and eye-tracking measures during language processing.

The first model assumption was tested in an eye-tracking task in which listeners had to identify a target in a display of objects following either a color or a material description. If assumption (a) is correct, color descriptions should result in faster target identification than material descriptions. The second model assumption was tested in a description-rating task in which speakers had to rate the same color and material descriptions according to how likely they would be to use them to refer to the target in the display. If assumption (b) is correct, speakers should rate color descriptions higher than material descriptions, giving higher scores to those descriptions that resulted in faster target identification in the eye-tracking task. All studies were approved by MIT's Institutional Review Board under protocol "Development of Visual Perception" (#0403000050R016).

Method

Participants

Thirty-two participants ($M_{\text{age}} = 34.6$) were recruited from the Cambridge area for the eye-tracking study. Additional 25 participants ($M_{\text{age}} = 32.9$) from the U.S. (as determined by their IP address) were recruited via Amazon's Mechanical platform for the description-rating task. No participants were excluded from the tasks.

Stimuli

Thirty displays with pictures of four objects were created, such that all objects were of different colors and materials. Each display

contained a target and a contrast object that was of the same kind as the target, but differed in color and material (e.g., a red leather chair vs. a green plastic chair; see Figure 2a). To ensure that the color and material of the target and the contrast object were discriminable in the pictures, the materials were pretested on 10 pilot participants recruited in the Cambridge area.

For the eye-tracking task, each display was paired with two descriptions of the target object, one mentioning its color (e.g., "The red chair") and another mentioning its material (e.g., "The leather chair"). The recordings were made by a female native speaker of American English who was blind to the experiment and used neutral, noncontrastive intonation in all trials.

The description-rating task employed the same color and material descriptions of the target, but presented both in written form under each display. The visual displays were the same as in the eye-tracking task, only that the target object appeared inside a grey frame (so that participants would evaluate each description in relation to that object).

Procedure

For the eye-tracking task, displays and descriptions were distributed in two lists using a Latin square design, and participants were evenly allocated to either list. Each trial lasted 3,000 ms. In between trials, participants had to click on a cross in the center of the screen. This was done to ensure that the cursor was equidistant from the four objects at the start of each trial. Item presentation was individually randomized.

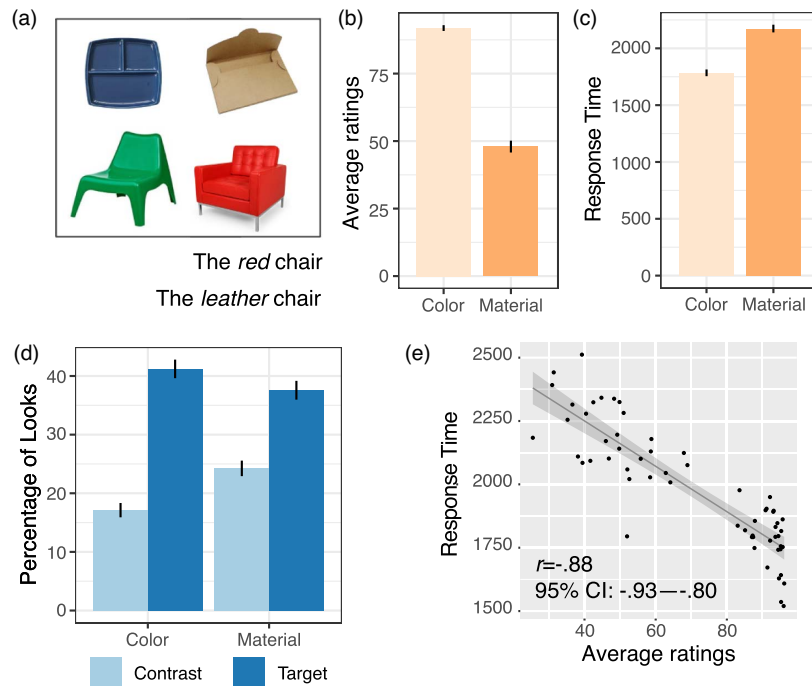
In the description-rating task, participants were asked to rate the color versus material descriptions of each target on a scale from 0 to 100 according to how likely they would be to use each description in that particular context. Item presentation was individually randomized.

Experiment data and analyses files are available at <https://osf.io/bezua/>. This study was not preregistered.

Results

In the eye-tracking task, net dwell time was calculated separately for the target and the contrast object for the duration of each trial. Overall, participants directed 41.23%, $CI_{95\%}$ [39.61, 42.80], of fixations to the target and 17.14%, $CI_{95\%}$ [15.79, 18.39], of fixations to the contrast in the color trials. This effect was attenuated for material trials. In these trials, participants directed 37.69%, $CI_{95\%}$ [36.13, 39.16], of fixations to the target and 24.23%, $CI_{95\%}$ [15.89, 18.45], of fixations to the contrast. Consistent with this, participants showed a slower response time for material relative to color descriptions, mean ratings: 1.78 s for color versus 2.19 s for material; $t(29) = -11.98$; $p < .0001$.

We ran three linear mixed effects models in which we modeled Adjective Type as a predictor of (a) percentage of net dwell time on the target, (b) percentage of net dwell time on the contrast object, and (c) response times (measured from the onset of the description until the participant clicked on an object). In addition to the fixed effect for Adjective Type, the models included random slopes for Adjective Type for subject and item and random intercepts for subject and item (i.e., the maximum random effect structure; see Supplemental Materials for details).

Figure 2*Sample Trial and Results From the Test of Our Model Assumptions*

Note. (a) Example trial. The four objects in the sample display have been enlarged for illustrative purposes. Each trial had a target object (the red leather chair, in this case) with a competitor (the green plastic chair, in this case). (b) Average ratings for the two descriptions of each target (i.e., how likely participants would be to use the color or the material description to refer to the target) in the description-rating task. Participants showed a reliable preference for color descriptions over material descriptions. (c) Average response time as a function of adjective in the instructions in the visual search task. Participants were significantly faster to click on the correct target when they heard a color adjective (e.g., “the red chair”) relative to a material adjective (e.g., “the leather chair”). (d) Percentage of fixations on the Target and the Contrast object (e.g., the red leather chair vs. the green plastic chair) in the visual search task. (e) Average rating in the description-rating task (x -axis) against response time (y -axis), showing that descriptions rated as more likely to be produced were the ones that led listeners to identify the referent faster. Error bars indicate 95% bootstrapped confidence intervals. See the online article for the color version of this figure.

The first model revealed that participants fixated more on the target in the Color Adjective condition than in the Material Adjective condition ($\beta = 3.91$; $t = 3.42$), confirming that color is indeed more salient than material. The second model showed that participants fixated more on the contrast object in the Material Adjective condition than in the Color Adjective condition, which suggests that searching by material created more hesitation between the two potential targets (e.g., the leather chair vs. the plastic chair) than searching by color (e.g., distinguishing the red chair from the green chair; $\beta = -7.04$; $t = -5.80$). These looking patterns are in line with the distinction that our visual search model makes between properties that can be identified from the periphery (such as color) and properties that require directly fixating on the object (such as material). Finally, the third model revealed that participants were also significantly faster to click on the target object in the Color condition than in the Material condition ($\beta = -402.37$; $t = -9.67$). It is important to bear in mind that the visual displays were the same in

the two conditions: what changed was the description of the target (e.g., “The red chair” vs. “The leather chair”), which in turn had an effect on the efficiency of the listener’s visual search for the object.

In the description-rating task, participants rated color descriptions almost twice as high as material descriptions, mean ratings: 92 for color versus 49 for material on a scale from 0 to 100; $t(29) = 17.67$; $p < .0001$. These results support our second assumption: speakers are sensitive to the differential visual salience of color and material properties, preferring to use color descriptions than material descriptions to refer to the targets in our displays.

More importantly for the test of our model assumptions, when comparing speakers’ description ratings with listeners’ response times in the eye-tracking task, there was a strong, negative correlation between the average description rating and the average RT for each color and material description, $r = -0.88$; $CI_{95\%} [-0.93, -0.80]$, confirming that speakers preferred to use those descriptions that led listeners to faster target identification. Moreover, the

correlation between average rating and average RT remained significant for each adjective type separately, Color: $r = -0.45$, $CI_{95\%} [-0.70, -0.11]$; Material: $r = -0.54$, $CI_{95\%} [-0.76, -0.22]$.

Can the ICE Model Explain Referential Over-Specification?

To evaluate the ICE model's capacity to explain reference production, we began by testing whether it could reproduce known qualitative phenomena in referential over-specification. Specifically, previous work has found that (a) speakers are more likely to use redundant color words in visual displays with more objects (Gatt et al., 2017; Koolen et al., 2016; Paraboni et al., 2007; Rubio-Fernandez, 2019); (b) this propensity, however, decreases as a function of the number of objects that share the same color as the target referent (Koolen et al., 2013; Long et al., 2020; Rubio-Fernandez, 2016, 2019); and (c) in identical visual displays with the same target, English speakers (who position adjectives before nouns) are more likely to use redundant color words relative to Spanish speakers (who position adjectives after nouns; Rubio-Fernandez, 2016, 2019; Rubio-Fernandez et al., 2020). These phenomena show that people have a nuanced preference towards producing color words, which is shaped by display density, color gamut, and word order—as the incremental efficiency theory predicts.

To evaluate if our model could reproduce these effects, we considered a simplified version of our efficient search model that only considered bare definite descriptions (e.g., “the star”) and color-modified descriptions (e.g., “the pink star”), as these phenomena have been primarily established with paradigms in which material information is not available (e.g., using simple geometrical shapes of different colors). Throughout, we used the preset model parameters and did not change them as a function of the phenomena or data set that we evaluate.

Redundancy as a Function of Display Density

To test the first effect—speakers are more likely to use redundant color words in visual displays with more objects—we created visual displays that ranged from 3 to 16 objects, all with a unique target of a unique color (e.g., a single blue star in an array of geometrical shapes, none of which were blue or stars). Because in all cases, the bare description is sufficiently informative, the alternative brevity model predicts that speakers should never use color words, regardless of the objects in the display. Figure 3a shows the results from this analysis. Like people, our model's preference for redundant color words increases as a function of the number of objects in the scene.

Our ICE model replicates this effect, because the bare definite description (e.g., “the star”) has a higher search cost in scenes with more objects (i.e., the more objects in the scene, the longer the listener must spend searching for the referent, since finding an object by its category requires fixating on the actual objects). By contrast, the color-modified description has a low visual search cost, independent of scene size, because, in these displays, color is a unique property of the referent (e.g., if there is only one blue object, then “the blue star” always helps the listener fixate on the referent, because they can identify color from their visual periphery). Therefore, adding more objects to a scene creates a larger utility difference

between these two expressions, leading to a stronger preference for using color words.

Redundancy as a Function of Color Variation

We next tested whether our model's propensity to use redundant color words decreased when all objects were the same color. To test this possibility, we created arrays of nine objects, all with a unique target (e.g., a single star among a set of geometrical shapes). We then manipulated the display's monochromaticity, ranging from 1/9 (the target has a unique color) to 9/9 (all objects share the same color).

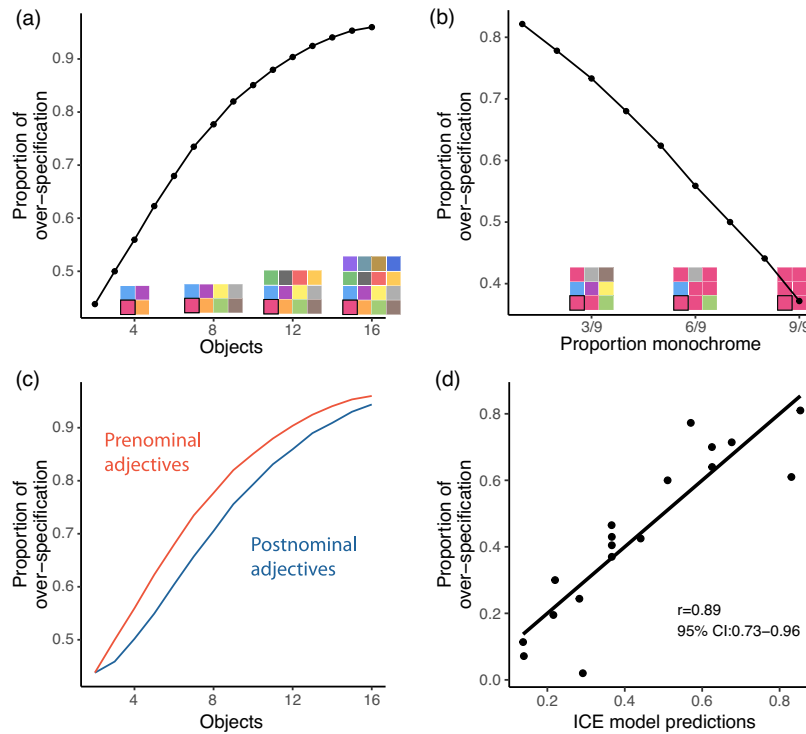
Figure 3b shows the results from this analysis. Once again, our model reproduced the empirical pattern, where monochromaticity reduces people's propensity to use redundant color words. Our model replicates this effect, because the search cost associated with the bare definite description (e.g., “the star”) is the same for all trials (as the number of objects in the scene does not change). By contrast, the color-modified description (e.g., “the pink star”) reduces visual search cost in polychrome displays (because color can be identified from the periphery), whereas in monochrome displays, the color-modified description *hinders* the listener's visual search (because color is not informative, so it delays the listener's ability to resolve reference until they hear the noun). Therefore, the inclusion of a color word increases the utility in polychrome displays (helping the listener) and decreases the utility in monochrome displays (delaying the listener). Importantly, our model predicted a gradual decrease in the use of redundant color adjectives between these two extreme cases (i.e., the referent color is unique in the display vs. the referent color is uniform across the display), replicating the pragmatic sensitivity observed in the human data.

Note that, once again, because the target could be uniquely specified by the noun, a brevity-based model, such as our alternative model, would always favor the bare description, independent of the distribution of colors in the display.

Cross-Linguistic Variation

We next sought to test whether our model could reproduce cross-linguistic phenomena. When describing the same target in equivalent displays, English speakers—who position adjectives prenominally—are more likely to use redundant color words relative to Spanish speakers—who position adjectives postnominally. To test this, we created arrays of 3–16 objects, all of them with a unique target of a unique color (e.g., as with the first analysis, a single blue star in an array of geometrical shapes, none of which were blue or stars). We then ran our original model along with a modification to our model that positions adjectives postnominally.

Figure 3c shows how our model reproduced this cross-linguistic effect. Our model's preference for redundant color words increased as a function of the number of objects in the display with both language structures. However, the preference for redundant color words was higher when adjectives were positioned prenominally (as in English) relative to postnominally (as in Spanish), as reported empirically in Rubio-Fernandez et al. (2020). Our ICE model replicates this effect, because the utility advantage of the color-modified description (e.g., “the blue star”) over the bare definite description (e.g., “the star”) is larger in languages with prenominal modification, because the color word is made available to the listener earlier in processing, enabling them to identify the referent

Figure 3*Model Simulations and Fits to Documented Effects on Redundant Color Use*

Note. (a) Number of objects in a visual scene (x-axis) against model probability of producing redundant color words (y-axis). Consistent with Rubio-Fernandez (2019), our model is more likely to use redundant color words in denser (polychromatic) visual scenes. (b) Proportion of objects of the same color in a scene with nine objects (x-axis) against model probability of producing redundant color words (y-axis). Consistent with Long et al. (2020) and Rubio-Fernandez (2021), our model decreases its propensity to use color words as a function of monochromaticity. (c) Number of objects in a visual scene (x-axis) against model probability of producing redundant color words (y-axis) based on adjective positioning. Consistent with Rubio-Fernandez et al. (2020), our model is more likely to use redundant color words when it positions adjectives prenominally (as in English) relative to postnominally (as in Spanish). (d) Model predictions (x-axis) against empirical data (y-axis) from previous studies. See the online article for the color version of this figure.

by its color (and hence from their visual periphery). By contrast, in languages with postnominal modification, the listener's visual search is initially guided by the noun (which requires fixating on different shapes to find the star), with color speeding up visual search only later in processing.

Quantitative Model Fit to Previous Data

Having found that our model reproduces the qualitative phenomena behind the incremental efficiency theory, we then tested whether it could capture these phenomena quantitatively, by comparing its predictions to the referential expressions the people produce in interactive communicative tasks conducted in person. To achieve this, we combined the following four data sets consisting of production data elicited in interactive laboratory tasks.

The first data set consisted of speaker-generated referential expressions over polychrome displays of increasing density (2, 4, and 8 shapes) from Rubio-Fernandez (2019), showing that speakers

increase redundant color use in denser displays. The second data set consisted of speaker-generated referential expressions over two types of four-shape displays, one monochrome (all objects of the same color) and one polychrome (all objects of different colors), used to show a reduced use of redundant color words in monochrome displays, from Long et al. (2020).

The third data set shows the same effect, but using a graded set of conditions, where all displays had nine objects and the proportion of monochrome objects varied from one to nine, from (Rubio-Fernandez, 2021). Finally, our fourth data set consisted of speaker-generated referential expressions over four- and 16-item polychrome displays with production data from Spanish and English speakers, from Rubio-Fernandez et al. (2020).

These four data sets capture the three phenomena in color over-specification that our model simulations capture. Because each data set includes only a few key contrasts, we combined all data (for a total of $n = 18$ conditions tested) and compared averaged judgments against our model, using once again our preset model parameters.

People's propensity to use redundant color words is affected by a multitude of factors, including color discriminability (Viethen et al., 2017), perceptual grouping (Koolen, 2019), and color typicality (Rubio-Fernandez, 2016; Sedivy, 2003; Westerbeek et al., 2015), among others. We, therefore, expected our model to capture quantitative graded phenomena across conditions, but not necessarily the baseline levels of redundancy, which are likely influenced by additional factors. Thus, to make our results easier to interpret, we used a linear regression to predict judgments as a function of model predictions. Note that correlations are invariant to linear transformations, and the application of a linear regression, therefore, does not affect any of our statistical results, and instead, it only helps show how quantitative changes in the model relate to quantitative changes in people's propensity to over-specify.

Figure 3d shows the results from this analysis, with model predictions on the x -axis and over-specification in these papers on the y -axis. Our model showed a correlation of $r = 0.89$, $CI_{95\%} [0.73, -0.96]$. This high correlation shows strong evidence for the ICE model, given that these data were obtained from previous in-person interactive studies (i.e., with a listener present when the participant produces referential expressions), using different paradigms and without fitting any parameters. Because, in all of these data sets, the target can always be uniquely specified with the noun alone, the alternative Brevity model treats all of these events as equivalent and produces no variance throughout (thus making it impossible to compute a correlation).

Can the ICE Model Explain Graded Acceptability Judgments?

Our results so far show that the ICE model captures patterns of reference production and over-specification in interactive laboratory tasks. Nonetheless, production tasks can hide away people's intuitions about expressions that they choose not to produce. Thus, to evaluate our model in a more holistic way, we also designed a graded acceptability task in which we asked participants to rate how likely they were to use each description, allowing us to evaluate our model not only based on its preferred expression, but also on the full distribution of expressions that it produces.

Method

Participants

Forty-one participants ($M_{\text{age}} = 34.2$) from the U.S. (as determined by their IP address) were recruited through Amazon's Mechanical Turk framework for our main study. Additional 41 participants ($M_{\text{age}} = 30.90$) were recruited for a control condition through the Prolific testing platform.³ Twenty-two additional participants were recruited, but excluded in the study, because they failed at least one of the four catch trial (see Results).

Stimuli

Stimuli consisted of 30 critical trials, each consisting of a grid of objects like those in Figure 4. Objects came in nine colors: black, blue, brown, green, orange, pink, purple, red, and yellow, and in nine materials: cardboard, ceramic, glass, leather, metal, paper, plastic, wood, and wool. The target object appeared inside a grey

frame and its position was counterbalanced across trials. The 30 trials consisted of three different density levels: three objects (8 trials), six objects (10 trials), and nine objects (12 trials). In three of these trials (one of each density), the target object was of a unique category, color and material, making modification unnecessary (e.g., the minimal description "the basket" would be sufficiently informative). The remaining 27 trials included a contrast object of the same category as the target (e.g., a second chair) and varied depending on whether the contrast object was of a different color than the target (6 trials, 2 of each density); a different material (6 trials, 2 of each density), or differed along both dimensions (15 trials, 3 with 3 objects, 5 with 6 objects, and 7 with 9 objects). To vary the relative efficiency of color and material adjectives for visual search, one or more objects in the display acted as "competitors" and were the same color as the target and the contrast object (1 color competitor in 6 trials, 2 of each density; 2 color competitors in 2 trials, 1 with 6 objects and 1 with 9 objects; and 3 color competitors in 1 trial with 9 objects) or the same material as the target and the contrast object (1 material competitor in 6 trials, 2 of each density; 2 material competitors in 2 trials, 1 with 6 objects and 1 with 9 objects; and 3 color competitors in 1 trial with 9 objects). See Supplemental Materials for full stimuli. In addition to the 30 critical trials, we included four catch trials, designed to ensure that participants were paying attention. In each of these trials, only one of the four descriptions matched the target, thus allowing us to ensure that participants were reading the expressions when rating them.

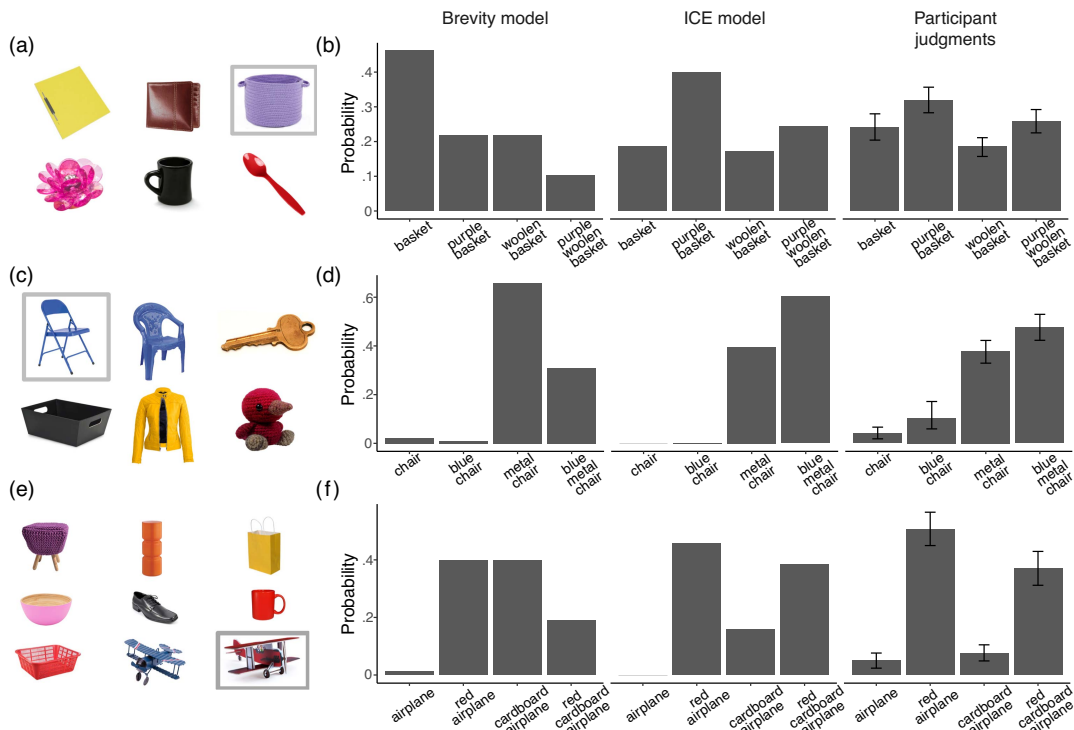
Procedure

Participants read a brief tutorial and then completed all trials in a randomized order. In each of the test trials, participants rated on a scale from 0 to 100 how likely they were to refer to the target in the display using each of four descriptions (0 = *Very unlikely*, 50 = *Somewhat likely* and 100 = *Very likely*), and the sliding button appeared at 50. Descriptions included no modification, a color adjective, a material adjective or both types of adjectives. The order of these descriptions was randomized individually across trials. The four catch trials were identical with the difference that only one of the descriptions matched the target object, and the other three described objects that were not present in the display.

The control condition was identical to our test condition, with one critical difference: participants rated the same utterances with the same visual displays, but now reported how likely they were to use each description when speaking on the phone. The instructions explained that participants were calling a friend to tell them about different objects that they had ordered from a catalogue, but their friend did not have a copy of the catalogue on the other side of the line. This condition, therefore, enables us to test whether our model explains people's preference for referential expressions only when speakers assume physical copresence with their listener.

Model predictions, experiment data, and analyses files are available at <https://osf.io/bezua/>. This study was not preregistered.

³ This control condition was run to address a reviewer concern. Since the original dataset was collected, attention levels in Mechanical Turk workers have decreased, and have been found to be lower relative to the Prolific platform (Arechar & Rand, 2021). We therefore used this new platform to ensure that a low effect (as we predict) could not be due to decreased attention in Mechanical Turk.

Figure 4*Example Trials From Our Main Experiment Along With Model Predictions and Participant Judgments*

Note. (a) Display where the target—the basket—can be identified without any modifiers. (b) The alternative Brevity model favors the shortest utterance ("the basket") and decays as a function of utterance length. By contrast, the ICE model, like participants, dis-prefers the shortest utterance (despite it being sufficiently informative) and instead introduces a redundant color word. (c) Trial where the target must be disambiguated by material rather than color. (d) The alternative Brevity model dis-prefers color-modified descriptions, while the ICE model and participants use both color and material. (e) A situation, where either color or material can disambiguate the target. (f) The alternative Brevity model favors expressions with a single modifier. The ICE model and participants prefer utterances that include color and material. ICE = incremental collaborative efficiency. See the online article for the color version of this figure.

Results

Each participant's judgments were normalized within trials, so that they added up to 1. Participants who did not place at least 80% of the probability on the correct description on any of the four catch trials were excluded from the study ($n = 22$). Judgments were then averaged across all participants (but separately for the test and control conditions).

Test Condition

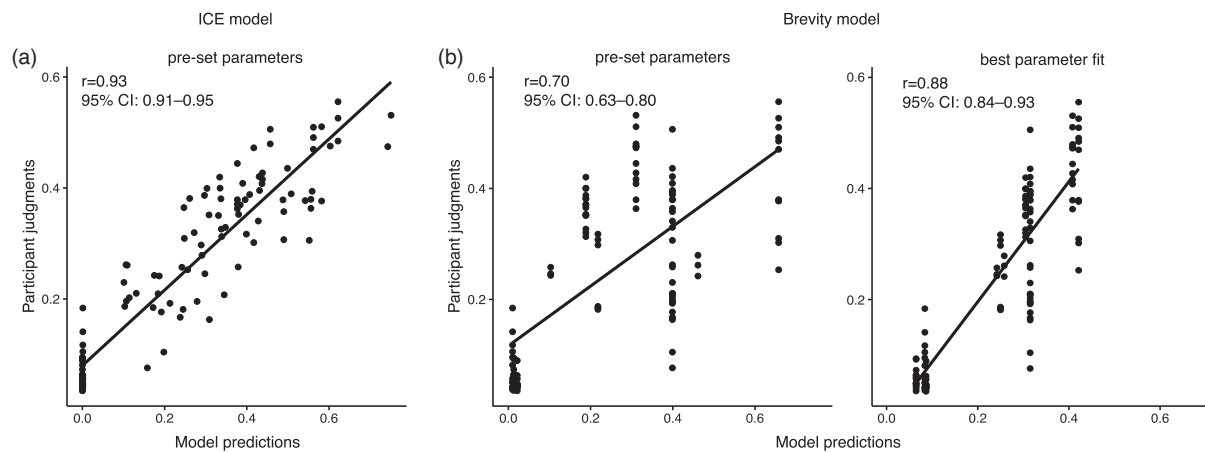
For our test condition, our main (ICE) model showed a correlation of $r = 0.93$, $CI_{95\%} [0.91, 0.95]$, while the alternative (brevity) model showed a lower correlation of $r = 0.70$, $CI_{95\%} [0.63, 0.80]$. Moreover, our ICE model showed a significantly higher correlation relative to the alternative model, $\Delta r = 0.22$; $CI_{95\%} [0.14, 0.29]$.

As noted in the computational framework, the parameters for our ICE model (fixation rate = 1 with $\tau = 2$) were set prior to data collection and the alternative Brevity model's parameters were selected so as to make the two models comparable (cost = 0.09 with $\tau = 0.12$; see Computational Framework for explanation). To test whether our results were affected by this initial choice, we next

fit each model's parameter to participant data to maximize their performance (as determined by the correlation; see Supplemental Materials for details). Fitting the Brevity model's parameters to participant data (cost = 0.01 with $\tau = 0.31$) increased the model's correlation to $r = 0.88$, $CI_{95\%} [0.84, 0.93]$. While this correlation was substantially higher than the one we obtained with the preset parameters ($\Delta = .18$ increase), fitting the parameters increased the correlation without capturing participant data in a nuanced manner. As shown in Figure 5b, the fitted Brevity model achieved a higher correlation by *losing* sensitivity (correlations increase when prediction errors are balanced on both directions of the linear model, a property that the Brevity model exploits). Moreover, this correlation was still reliably lower than the correlation from our ICE model with preset parameters, $\Delta r = 0.04$; $CI_{95\%} [0.006, 0.08]$. Thus, the ICE model with preset parameters outperformed even a Brevity model fit to participant judgments.

The parameters that best fit our ICE model to participant data (fixation rate = 1 with $\tau = 2.9$) increased the correlation to 0.94, $CI_{95\%} [0.92, 0.96]$, which was also reliably higher than the best Brevity model fit, $\Delta r = 0.06$; $CI_{95\%} [0.02, 0.09]$. Critically, fitting the parameters to our model had little effect on our predictions. Our preset model had a correlation of $r = 0.99$, $CI_{95\%} [0.99, 0.99]$, with

Figure 5
Experiment Results



Note. Each dot represents a referential expression (four expressions per trial). The *x*-axis shows the model predictions and the *y*-axis shows average participant judgments. Diagonal black lines show best linear fit. (a) ICE model results using pre-set parameters. (b) Brevity model results using pre-set parameters, and results using parameters fit to participant data. ICE = incremental collaborative efficiency.

the fitted model, giving further evidence of the small impact that the parameters have over the ICE model (see Supplemental Materials for additional details).

Figure 4 shows three example trials that illustrate the differences between the two models (using the preset model fits). The event in Figure 4a consists of six objects with a target basket. Because there is only one basket, the Brevity model selects “basket” as the best utterance (because it is sufficiently informative, and carries the minimal cost), followed by “purple basket” and “woolen basket,” as they both use one unnecessary word, and ending with “purple woolen basket” which uses two unnecessary words (Figure 4b). By contrast, the ICE model mimics participant judgments, favoring “purple basket” and “purple woolen basket,” because locating a purple object is easier than finding a basket (Figure 4b). Figure 4c shows a situation, where mentioning the object’s material is necessary. Consequently, the Brevity model prefers “metal chair” over “blue metal chair” (Figure 4d). Our ICE model and participants, however, prefer “blue metal chair” over “metal chair” (Figure 4d), which help the listener first identify the blue objects, and find the metal one between those two (rather than relative to the entire visual scene). Finally, Figure 4e shows a case, where the referent can be identified through color or material. The Brevity model assigns equal probability to “red airplane” and “cardboard airplane” and a lower probability to “red cardboard airplane” (Figure 4f). Our ICE model and participants instead assign the highest probability to “red airplane,” followed by “red cardboard airplane,” and disprefer “cardboard airplane,” despite it being sufficiently informative (Figure 4f).

The Brevity model with the best parameter fits also reveals how this model fails to explain participant judgments. The fitting process selected the lowest possible WC available in our search space (0.01; see Supplemental Materials for detail), suggesting that the Brevity model performed better when the egocentric WC was minimized, further suggesting that WCs had minimal explanatory power.

Control Condition

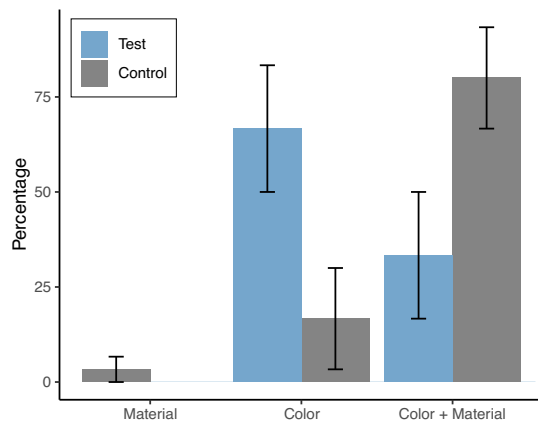
The results so far show that people use and prefer referential expressions designed to reduce the listener’s time and effort resolving the referent (as revealed by our modeling of four interactive communication data sets, and the test condition in this experiment). Is it possible that people use this strategy in an inflexible way, even in situations, where speakers’ choice of words cannot help the listener’s visual search? Our control condition tested this possibility by running the same task with a different cover story, where participants had the task of describing an object by telephone to a listener who did not have visual access to the objects.

Participant judgments now showed a correlation of $r = 0.41$, $CI_{95\%} [0.27, 0.56]$, with our ICE model and a correlation of $r = 0.23$, $CI_{95\%} [0.10, 0.37]$, with the Brevity model, suggesting that participants were no longer attempting to support listeners’ visual search, or to establish reference with as few words as possible. Instead, participants now appeared to favor being more descriptive than they would be otherwise (Figure 6). Overall, participants favored the most descriptive utterance (color and material plus noun) on 80% of the trials ($n = 24$ out of 30), and included either color or material on the remaining 20% of the trials ($n = 5$ mentioning color, and $n = 1$ mentioning material). By contrast, participants in the test condition favored the most descriptive utterance (color and material plus noun) on 33.33% of the trials ($n = 10$ out of 30), and included color alone on the remaining 66.6% of the trials ($n = 20$ out of 30).

Discussion

Here, we introduced the ICE framework, which aims to capture referential communication based on the idea that speakers build reference designed to minimize listeners’ *time* and *effort*. To achieve this, our computational model built referential expressions using a nuanced model of listeners’ real-time cognitive processing, as they search for the referent. In support of our framework, we found that

Figure 6
Distribution of Preferred Description Type Across Conditions



Note. The x-axis shows descriptions modified by material (e.g., “the plastic chair”), descriptions modified by color (e.g., “the blue chair”), and descriptions modified by color and material (e.g., “the blue plastic chair”). The y-axis shows the percentage of trials, where a description of that type obtained the highest score, as a function of condition. Bare definite descriptions are not shown, because these were never selected as the most popular in any of the trials. See the online article for the color version of this figure.

our computational model naturally produced three separate empirical phenomena associated with referential over-specification, and it showed a quantitative fit to four different published data sets of interactive referential communication (Can the ICE Model Explain Referential Over-Specification? section). Moreover, our computational model was also able to quantitatively match people’s acceptability judgments, including speakers’ beliefs about the relative value of utterances that people avoid using (Can the ICE Model Explain Graded Acceptability Judgments? section). Finally, we also found an empirical effect, where people’s preference for different referential expressions showed a strong negative correlation with the relative speed with which listeners identify the referent in hearing such expressions. Altogether, our work thus suggests that the principles by which speakers generate referential expressions for physical objects are aligned with how listeners process these expressions and search for the corresponding objects.

Our work advances an interpretation of the Gricean Maxim of Quantity (Grice, 1975) that is in line with the Principle of Least Collaborative Effort (Brennan & Clark, 1996; Clark, 1996; Clark et al., 1983; Clark & Marshall, 1981; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986; Horton & Brennan, 2016), whereby speakers do not attempt to simply be unambiguous and succinct, and are instead also motivated to ensure that they ease the listeners’ effort in resolving the referent. The idea that speakers might introduce redundant adjectives to support listener visual search has received recent attention and has been proposed in a qualitative manner (Rubio-Fernandez, 2016, 2021; Rubio-Fernandez et al., 2020). Our work adds to this literature by presenting a precise computational formalization of this idea, grounded in estimates of visual search difficulty.

Our computational framework is also consistent with an increased attention to the importance of real-time communication for pragmatics (Davies & Richardson, 2021; Eberhard et al., 1995; Rubio-Fernandez

et al., 2020; Rubio-Fernandez & Jara-Ettinger, 2020; Sedivy et al., 1999), including computational models that, like ours, consider how listeners process words incrementally as a message unfolds over time (Dale & Reiter, 1995; Waldon & Degen, 2021). Our work is consistent with these models, but diverges in that we focused on presenting a computationally precise account for what it means for a speaker to be motivated to support listener real-time visual search.

Our findings directly speak to a much debated question in the pragmatics literature: whether, and to what extent, speakers tailor their utterances to their interlocutors (Clark & Krych, 2004; Clark & Murphy, 1982; Horton & Keysar, 1996). This is known as *audience design*. Referential communication studies have shown that speakers adjust their referential expressions to the needs of their interlocutors, distinguishing experts from novices, for instance (Heller et al., 2012; Isaacs & Clark, 1987). Recent work has also found that computational frameworks centered around audience design—where utterances aim to maximize the chance that listeners recover the intended message—successfully capture a wide range of linguistic phenomena, from scalar implicature to hyperbole (Goodman & Frank, 2016). Our work advances this view, showing the nuance in audience design: Speakers produce utterances tailored to help listeners efficiently recover the intended message in real time, showing sensitivity to the visual context and to their interlocutors’ visual search. Importantly, our control condition reveals that speakers’ sensitivity changes when the visual context is not shared between interlocutors.

Implications of the Structure and Nature of Referential Costs

Most directly, our results show that allocentric costs are critical for explaining how speakers produce referential expressions. In addition, we also found that a model with a purely egocentric brevity cost function failed to capture both qualitative and quantitative patterns of reference. Even when the parameters of this alternative Brevity model were fit to participant data, the best fit underperformed relative to our ICE model. Interestingly, the best Brevity model obtained through parameter fitting was the one that used the smallest possible production cost available in our parameter space. These results suggest that utterance-length penalization had minimal explanatory power in our task, to the point that the model performed better by giving the lowest possible weight. These results are consistent with related work that has also found that egocentric production costs do not confer significant benefits to reference models (Degen et al., 2020). While this finding might seem surprising in the context of a simple interpretation of the Gricean Maxim of Quantity, we believe that these results are intuitive—in everyday conversations, slightly longer utterances rarely feel more costly, and people often say more than what would be considered necessary by a rigid brevity analysis.

At the same time, our results do not imply that reference is guided purely by allocentric costs. Effective collaborative communication requires interlocutors to minimize the joint communicative effort, which must involve comparing and combining egocentric and allocentric costs. Related work shows that linguistic preferences in speakers are also shaped by egocentric costs, such as the effort associated with preempting an ambiguity (Ferreira et al., 2005) or adopting the listener’s visual perspective (Nadig & Sedivy, 2002). Thus, our work adds to these previous findings by showing that speaker referential

costs include allocentric components, guided by a sensitivity to listeners' visual search. While numerous psycholinguistic studies have investigated audience design, the space of costs that shape reference production remain poorly understood (for a review of so-called "speaker-internal" and "addressee-oriented processes" in reference production, see Arnold, 2008). Nonetheless, our work shows that the space of these costs is conceptually rich, and must include allocentric components that depend on a consideration of our listeners, and not only on an expectation of brevity.

Our work also reveals an interesting conceptual distinction between production and selection costs. That is, some cost functions represent which referential expressions come to mind, while other cost functions capture information that speakers compute when deciding which expression to produce. Consider, for example, a speaker who can describe an object as "the sofa" or "the chaise." In a situation like this, the speaker might be more likely to select the first expression, because the word "sofa" is easier to retrieve from memory than the word "chaise." This preference can be captured by a production cost that represents the difficulty of retrieving different words from memory (and is, therefore, not a cost that the speaker calculates). Suppose, however, that the speaker is a furniture salesman, who has little problem retrieving both expressions from memory. In this case, the speaker might, nonetheless, be more likely to use the expression "the sofa," because the listener will have an easier time processing it, but they might also select "the chaise" if their customer seems to be looking for something exclusive. These different preferences can be captured by a selection cost that represents information that the speaker is calculating about different expressions when deciding which one to use.

This distinction suggests that cost functions in reference can both represent the difficulty of generating different expressions (production costs) and the value that speakers assign to expressions that they can easily generate (selection costs). Intuitively, allocentric costs are typically of the selection type (they are estimated to decide what a listener will find useful), while egocentric costs are typically of the production type (they are forces that shape what comes to speakers' minds). While similar distinctions have been made in psycholinguistic studies of audience design (see Arnold, 2008), to the best of our knowledge, it remains an open question the extent to which different types of costs are production-based or selection-based.

Referential Communication Beyond a Computational Level of Analysis

Our framework was designed to capture reference production at a computational level of analysis (Chater & Oaksford, 1999; Marr, 1982). That is, our model sought to formalize the computational goals behind people's choice of words during reference production. Our results provide evidence that the computations underlying reference production are shaped for collaborative communication, rather than by an attempt to identify the shortest unambiguous utterance. Do our model's representations and algorithms resemble the underlying implementations that instantiate these computations in the human mind?

At the highest level, our computational model uses a mental model of listeners to estimate the difficulty of real-time visual search. Research in Theory of Mind has long argued that it is critical to distinguish our own mental states, such as beliefs and desires, from those of others, because these have great variability

across agents (Repacholi & Gopnik, 1997; Wimmer & Perner, 1983). Visual systems, however, are largely homogeneous. In addition, while agents often differ in what they can see from their physical position in space, the principles of what we find visually salient (such as color) are the same for everyone. Thus, it is possible that speakers rely on their own visual system to estimate the listener's difficulty in identifying the referent, rather than using separate representations.

Even if speakers produce efficient referential expressions by relying on their own visual system, this process would still require some amount of perspective taking. Recent work on social neuroscience has argued that, even when people share the same perspective, they must monitor the extent to which others' mental states match or mismatch their own (Deschrijver & Palmer, 2020). Thus, at a minimum, a speaker in a situation of *copresence* (Clark & Marshall, 1981) needs to be aware that they are sharing a physical space with their listener and can, therefore, rely on *mutual salience* (Rubio-Fernandez, 2019, 2021). People's sensitivity to copresence is evidenced by our control condition, showing that people cease to prefer referential expressions designed for efficient visual search when they are told the information will be communicated to a listener who cannot see the objects (see also Hawkins et al., 2021 for related evidence that speakers monitor copresence). Therefore, monitoring whether we can or we cannot rely on mutual salience is a form of perspective taking that is necessary for successful, efficient referential communication. From this standpoint, it is possible that speakers use a mental model of the process of visual search, but relying on their own visual system's estimates of what is visually salient.

At the same time, there are two components of our model that are unlikely to hold at an algorithmic level of analysis. First, our model estimated visual difficulty by quantifying the number of fixations. However, there are alternate measures, such as time, that correlate with fixation count and that people might use to estimate the burden that referential expressions impose on listeners. Second, we estimated cost via Monte Carlo simulations that averaged over different potential fixation patterns, but such method can be rather inefficient (although see Hamrick et al., 2015 for some evidence of simulation as a plausible algorithmic implementation in other domains of cognition). In these cases, while the algorithms humans use likely diverge from the ones that we used, our work helps constrain the space of possibilities, as our results establish that the underlying algorithm must correlate with the output of our model, given the tight correspondence between our model predictions and participant judgments.

One additional important clue towards the underlying algorithmic implementation comes from the work of Gatt et al. (2017), who found that the time that it takes for people to generate a referential expression increases with (a) the complexity of the visual scene and (b) the complexity of the generated expression. An appropriate algorithmic-level theory of reference production must replicate these effects. The ICE model that we presented here requires more computation as a function of visual scene complexity, therefore, explaining the first effect (see Supplemental Materials for details). However, our implementation could only replicate the second effect if it is explicitly implemented to consider the value of referential expressions in order of complexity (comparing them against a general baseline expectation of referential efficiency). While this assumption appears reasonable (and more likely than,

e.g., assuming that speakers first consider the most complex expressions before considering simpler ones), this also implies that our model could generate any pattern of data based on the order in which it considers the space of possible expressions. Therefore, our model's potential to fit this data is not necessarily algorithmic evidence for our account, as it would also be able to fit *alternative* patterns of data that did not occur (see Roberts & Pashler, 2000 for a discussion of this point for theory evaluation).

Finally, our work also did not implement a key component that is likely to affect referential communication at an algorithmic level: resource rationality (Lieder & Griffiths, 2020). In more complex situations, effective referential communication requires not only a sensitivity to visual search, but also a representation of others' knowledge and beliefs (Sperber & Wilson, 1986), and these social computations are thought to be costly for interlocutors to compute (Horton & Keysar, 1996; Keysar et al., 2000). It is, therefore, likely that people, at an algorithmic level, flexibly decide to devote more or less computational resources to the problem depending on the task complexity. Consistent with this, Hawkins et al. (2021) recently advanced a resource-rational view of reference that helps explain reference in situations, where knowledge is asymmetrical, and which we believe will be crucial for a full algorithmic understanding of reference.

Evidence From Cognitive Neuroscience

Recent work in cognitive neuroscience also provides some evidence about the neural implementation supporting social reference production, and here, we briefly discuss how these findings relate to our model.

First, our computational model represented the value of utterances through utility functions inspired by those used to model human Theory of Mind and pragmatics (Baker et al., 2017; Goodman & Frank, 2016; Jara-Ettinger et al., 2020). While these utility-based frameworks enjoy the most support at the computational level, recent work has found evidence that utility-based algorithms for social reasoning are implemented in the dorsomedial prefrontal cortex (dmPFC; Collette et al., 2017), one of the key regions responsible for representing others' mental states (Jamali et al., 2021; Saxe & Powell, 2006). While these results provide some initial algorithmic and implementation evidence for utility functions in social reasoning, these studies have focused on social learning, and more research is needed to understand if these algorithms extend to communicative interactions.

Beyond utility functions, our computational model also posits that speakers use a mental model of their listener. Recent work has also found evidence that people encode communicative expectations through a mental model of the speaker, which appear to be implemented in frontal–striatal circuits and in the ventromedial prefrontal cortex; Mi et al., 2021). In the context of our work, these results provide some initial evidence for the use of mental models in communication, but leave open the question of whether, at an algorithmic level, speakers also use mental models of the listener as they speak.

Collaborative Reference in More Complex Situations

Our computational model implemented the incremental efficiency theory using the standard test-bed for empirical and computational work on reference: referential communication about objects

presented in controlled visual displays. This enabled us to formalize our theory in precise computational terms and evaluate it in a quantitative manner against controlled data sets and alternative accounts (e.g., Clarke et al., 2013; Koolen et al., 2013; Nadig & Sedivy, 2002). While our work included model comparisons to data obtained from live interactive communicative tasks (Quantitative Model Fit to Previous Data section), these experiments still consisted of simple events, where speakers had to identify a single object. However, field studies suggest that real-world interactive referential communication is substantially more complex than what these experimental paradigms capture (e.g., Clark & Schaefer, 1989; Schober & Clark, 1989; Wilkes-Gibbs & Clark, 1992), and a complete theory of reference must account for these phenomena. In particular, because reference is one of the most fundamental functions of communication, humans routinely communicate about a wide space of referents that go beyond physically observable objects, employing a wide variety of communicative devices to convey reference—from an ostensive glance or a pointing gesture to complex noun phrases (e.g., “Did you see the old couple who just walked past?”). How might the ICE framework apply to these situations?

The first challenge is that real-world referential communication often occurs in complex natural scenes that can easily involve hundreds of potential referents. In cases like these, the cognitive effort required for the speaker to encode every object in the scene to design a useful referential expression may be prohibitively expensive. From a collaborative efficiency perspective, this situation licenses people to speak, as they think (rather than to think before they speak); as the benefit of producing the best referential expression is outweighed by the time it would take to construct such expression. These phenomena may lead to *referential repair*, where speakers produce suboptimal referential expressions which they then correct, as they recognize their failure to communicate (Fusaroli et al., 2017). While the model that we presented here does not yet accommodate these effects (but see van Arkel et al., 2020), we believe that these phenomena provide further evidence for the communicative principles instantiated in the ICE framework, as they reflect a collaborative effort to communicate efficiently, with a sensitivity to the real-time dynamics of communication. In future work, we hope to explore these effects by including a sensitivity to speaker real-time production effort.

Related research also shows that people employ a variety of solutions to the problem of reference in complex environments, such as generating expressions that include salient features of the object to preempt any potential ambiguity that the speaker might be unaware of (Clarke et al., 2013; Gatt et al., 2017; Hawkins et al., 2021), or using other referential devices, such as looking at the object and/or pointing at it (Bangerter & Chevalley, 2005). In the context of the ICE framework, such effects may be incorporated by extending the action space to include combinations of physical actions and utterances, and provide speakers with uncertainty over the potential space of referents. In this case, the ICE framework predicts that people's decision to over-specify should be modulated by their beliefs about undetected objects (e.g., what may be found on a kitchen top vs. a desk) and the accuracy of pointing gestures depending on the interlocutors' distance to the objects or their visual salience. We hope to test these predictions in future work.

Other phenomena that emerge in more complex communicative interactions are the interlocutors' reliance on referential pacts

(e.g., implicitly agreeing to call a shoe “the loafer”; Brennan & Clark, 1996) and common ground (e.g., calling one of two mugs “the mug” once it has been identified in the conversation). Recent research suggests that common ground can also affect the use of adjectives: speakers are more likely to overspecify color when the target is only known to them, than when it is in common ground with their listener (Rubio-Fernandez, 2019). These phenomena may be incorporated into the ICE framework by giving our model an even richer mentalistic representation of the event, whereby speakers are aware that they can both refer to co-present objects, or to objects in common ground. From this standpoint, when discussing past events in common ground or when referring to abstract ideas, listeners’ visual search is no longer the guiding metric of efficient communication. In cases like these, speakers may be more inclined to begin with memorable properties over non-memorable ones (i.e., relying on salience against common ground rather than perceptual salience; Clark et al., 1983). For example, if “the tall man” and “the mean man” refer to the same person, the first expression might be more efficient when helping the listener identify the man in a crowd, while the second expression might be more efficient when asking the listener to recall someone that they previously encountered. In related research, we have found that people can build nuanced representations of common ground from single linguistic events, suggesting that these representations are readily available for communicative purposes (Jara-Ettinger & Rubio-Fernandez, 2021b).

While these broader predictions of the incremental efficiency theory remain to be tested empirically and developed computationally, the ICE framework provides a first step towards modeling more complex types of referential communication. Under our framework, this would require using an allocentric cost function that models our intuitive theory of “the mind’s eye” and how it brings information in memory to bear. This is a direction that we hope to explore in future work, helping further test the scope of the incremental efficiency theory in a computational manner, and to reveal the extent to which people can use rich intuitive models of their listener’s mind to compute allocentric costs in communicative interactions.

Conclusions

Overall, our model and experimental study support the view that communication is best understood as a social activity, where interlocutors share a joint goal to transmit their intended messages efficiently. Our work shows that, even in the simplest linguistic activities, such as deciding what to call an object, speakers construct reference with their listeners in mind, adapting what they say not only to how much information their message contains, but also to how their listeners will extract the meaning that is conveyed. Our results, in turn, advance an interpretation of the Gricean Maxim of Quantity, where efficiency is not quantified through an egocentric bias to be brief, but through our nuanced understanding of other people’s minds.

References

- Adams, R. C., & Chambers, C. D. (2012). Mapping the timecourse of goal-directed attention to location and colour in human vision. *Acta Psychologica*, 139(3), 515–523. <https://doi.org/10.1016/j.actpsy.2012.01.014>
- Arechar, A. A., & Rand, D. G. (2021). Turing in the time of COVID. *Behavior research methods*, 53, 2591–2595. <https://doi.org/10.3758/s13428-021-01588-4>
- Arnold, J. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, 23(4), 495–527. <https://doi.org/10.1080/01690960801920099>
- Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1), 361–374. <https://doi.org/10.1016/j.pragma.2010.07.013>
- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1, Article 0064. <https://doi.org/10.1038/s41562-017-0064>
- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3), 329–349. <https://doi.org/10.1016/j.cognition.2009.07.005>
- Bangerter, A., & Chevalley, E. (2005). Pointing and describing in referential communication. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 27). Erlbaum.
- Belke, E. (2006). Visual determinants of preferred adjective order. *Visual Cognition*, 14(3), 261–294. <https://doi.org/10.1080/13506280500260484>
- Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during “same”-“different” decisions. *European Journal of Cognitive Psychology*, 14(2), 237–266. <https://doi.org/10.1080/09541440143000050>
- Bramão, I., Reis, A., Petersson, K. M., & Faisca, L. (2011). The role of color information on object recognition: A review and meta-analysis. *Acta Psychologica*, 138(1), 244–253. <https://doi.org/10.1016/j.actpsy.2011.06.010>
- Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6), 1482–1493. <https://doi.org/10.1037/0278-7393.22.6.1482>
- Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, 3(2), 57–65. [https://doi.org/10.1016/s1364-6613\(98\)01273-x](https://doi.org/10.1016/s1364-6613(98)01273-x)
- Clark, H. H. (1996). *Using language*. Cambridge University Press.
- Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62–81. <https://doi.org/10.1016/j.jml.2003.08.004>
- Clark, H. H., & Marshall, C. R. (1981). Definite knowledge and mutual knowledge. In A. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge University Press.
- Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J.-F. Le Ny & W. Kintsch (Eds.), *Advances in psychology* (Vol. 9, pp. 287–299). North-Holland.
- Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13(2), 259–294. https://doi.org/10.1207/s15516709cog1302_7
- Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22(2), 245–258. [https://doi.org/10.1016/S0022-5371\(83\)90189-5](https://doi.org/10.1016/S0022-5371(83)90189-5)
- Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1), 1–39. [https://doi.org/10.1016/0010-0277\(86\)90010-7](https://doi.org/10.1016/0010-0277(86)90010-7)
- Clarke, A. D., Elsner, M., & Rohde, H. (2013). Where’s wally: The influence of visual salience on referring expression generation. *Frontiers in Psychology*, 4, Article 329. <https://doi.org/10.3389/fpsyg.2013.00329>
- Collette, S., Pauli, W. M., Bossaerts, P., & O’Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *eLife*, 6, Article e29718. <https://doi.org/10.7554/eLife.29718>
- Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2), 233–263. [https://doi.org/10.1016/0364-0213\(95\)90018-7](https://doi.org/10.1016/0364-0213(95)90018-7)
- Davidoff, J. (1991). *Cognition through color*. MIT Press.
- Davidoff, J. (2001). Language and perceptual categorisation. *Trends in Cognitive Sciences*, 5(9), 382–387. [https://doi.org/10.1016/S1364-6613\(00\)01726-5](https://doi.org/10.1016/S1364-6613(00)01726-5)

- Davies, C., & Richardson, A. (2021). Semantic as well as referential relevance facilitates the processing of referring expressions. *Journal of Pragmatics*, 178, 258–269. <https://doi.org/10.1016/j.pragma.2021.03.024>
- Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to “overinformative” referring expressions. *Psychological Review*, 127(4), 591–621. <https://doi.org/10.1037/rev0000186>
- Deschrijver, E., & Palmer, C. (2020). Reframing social cognition: Relational versus representational mentalizing. *Psychological Bulletin*, 146(11), 941–969. <https://doi.org/10.1037/bul0000302>
- Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24(6), 409–436. <https://doi.org/10.1007/BF02143160>
- Engelhardt, P. E., Bailey, K., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54(4), 554–573. <https://doi.org/10.1016/j.jml.2005.12.009>
- Engelhardt, P. E., Demiral, S. B., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, 77(2), 304–314. <https://doi.org/10.1016/j.bandc.2011.07.004>
- Ferreira, V. S., Slevc, L. R., & Rogers, E. S. (2005). How do speakers avoid ambiguous linguistic expressions? *Cognition*, 96(3), 263–284. <https://doi.org/10.1016/j.cognition.2004.09.002>
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), Article 998. <https://doi.org/10.1126/science.1218633>
- Fukumura, K., & Carminati, M. N. (2021). Over-specification and incremental referential processing: An eye-tracking study. *Journal of Experimental Psychology: Learning, memory and cognition*. Advance online publication. <https://doi.org/10.1037/xlm0001015>
- Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). *Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions* [Conference session]. 39th Annual Conference of the Cognitive Science Society, Austin, Texas, United States.
- Gatt, A., Krahmer, E., van Deemter, K., & van Gompel, R. P. (2014). Models and empirical data for the production of referring expressions. *Cognitive Science*, 29(8), 899–911. <https://doi.org/10.1080/23273798.2014.933242>
- Gatt, A., Krahmer, E., van Deemter, K., & van Gompel, R. P. (2017). Reference production as search: The impact of domain size on the production of distinguishing descriptions. *Cognitive Science*, 41(S6), 1457–1492. <https://doi.org/10.1111/cogs.12375>
- Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11), 818–829. <https://doi.org/10.1016/j.tics.2016.08.005>
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics*, Vol. 3, *Speech Acts* (pp. 41–58). Academic Press.
- Gwendolyn, R., Cullimore, R. A., Henderson, J. M., & Ferreira, F. (2021). When more is more: Redundant modifiers can facilitate visual search. *Cognitive Research*, 6(1), Article 10. <https://doi.org/10.1186/s41235-021-00275-4>
- Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). *Think again? The amount of mental simulation tracks uncertainty in the outcome* [Conference session]. 37th Annual Conference of the Cognitive Science Society, Austin, Texas, United States.
- Hawkins, R. D., Gweon, H., & Goodman, N. D. (2021). The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive Science*, 45(3), Article e12926. <https://doi.org/10.1111/cogs.12926>
- Heller, D., Gorman, K. S., & Tanenhaus, M. K. (2012). To name or to describe: Shared knowledge affects referential form. *Topics in Cognitive Science*, 4(2), 290–305. <https://doi.org/10.1111/j.1756-8765.2012.01182.x>
- Horton, W. S., & Brennan, S. E. (2016). The role of metarepresentation in the production and resolution of referring expressions. *Frontiers in Psychology*, 7, Article 1111. <https://doi.org/10.3389/fpsyg.2016.01111>
- Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, 59(1), 91–117. [https://doi.org/10.1016/0010-0277\(96\)81418-1](https://doi.org/10.1016/0010-0277(96)81418-1)
- Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1), 26–37. <https://doi.org/10.1037/0096-3445.116.1.26>
- Jamali, M., Grannan, B. L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., & Williams, Z. M. (2021). Single-neuronal predictions of others’ beliefs in humans. *Nature*, 591(7851), 610–614. <https://doi.org/10.1038/s41586-021-03184-0>
- Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29, 105–110. <https://doi.org/10.1016/j.cobeha.2019.04.010>
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20(8), 589–604. <https://doi.org/10.1016/j.tics.2016.05.011>
- Jara-Ettinger, J., & Rubio-Fernandez, P. (2021a). *Incremental communicative efficiency model*. <http://osf.io/bezua>
- Jara-Ettinger, J., & Rubio-Fernandez, P. (2021b). Quantitative mental state attributions in language understanding. *Science Advances*, 7(47). <https://doi.org/10.1126/sciadv.abj0970>
- Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The Naïve Utility Calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123, Article 101334. <https://doi.org/10.1016/j.cogpsych.2020.101334>
- Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people’s preferences through inverse decision-making. *Cognition*, 168, 46–64. <https://doi.org/10.1016/j.cognition.2017.06.017>
- 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(33), 12002–12007. <https://doi.org/10.1073/pnas.1407479111>
- 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(1), 32–38. <https://doi.org/10.1111/1467-9280.00211>
- Koolen, R. (2019). On visually-grounded reference production: Testing the effects of perceptual grouping and 2D/3D presentation mode. *Frontiers in Psychology*, 10, Article 2247. <https://doi.org/10.3389/fpsyg.2019.02247>
- Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, 37(2), 395–411. <https://doi.org/10.1111/cogs.12019>
- Koolen, R., Krahmer, E., & Swerts, M. (2016). How distractor objects trigger referential overspecification: Testing the effects of visual clutter and distractor distance. *Cognitive Science*, 40(7), 1607–1647. <https://doi.org/10.1111/cogs.12297>
- Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1), 173–218. https://doi.org/10.1162/COLI_a_00088
- Kursat, L., & Degen, J. (2021). Perceptual difficulty differences predict asymmetry in redundant modification with color and material adjectives. *Proceedings of the Linguistic Society of America*, 6(1), 676–688. <https://doi.org/10.3765/plsa.v6i1.5003>
- 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, Article e1. <https://doi.org/10.1017/S0140525X1900061X>
- Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366), 1038–1041. <https://doi.org/10.1126/science.aag2132>

- Long, M., Moore, I., Mollica, F., & Rubio-Fernandez, P. (2021). Contrast perception as a visual heuristic in the formulation of referential expressions. *Cognition*, 217. <https://doi.org/10.1016/j.cognition.2021.104879>
- Long, M., Rohde, H., & Rubio-Fernandez, P. (2020). The pressure to communicate efficiently continues to shape language use later in life. *Scientific Reports*, 10(1), 1–13. <https://doi.org/10.1038/s41598-020-64475-6>
- Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., Markson, L., & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PLOS ONE*, 9(3), Article e92160. <https://doi.org/10.1371/journal.pone.0092160>
- Mangold, R., & Pobel, R. (1988). Informativeness and instrumentality in referential communication. *Journal of Language and Social Psychology*, 7(3–4), 181–191. <https://doi.org/10.1177/0261927X8800700403>
- Marr, D. (1982). *Vision*. MIT Press.
- Mi, Q., Wang, C., Camerer, C. F., & Zhu, L. (2021). Reading between the lines: Listener's vmPFC simulates speaker cooperative choices in communication games. *Science Advances*, 7(10), Article eabe6276. <https://doi.org/10.1126/sciadv.abe6276>
- Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, 13(4), 329–336. <https://doi.org/10.1111/j.0956-7976.2002.00460.x>
- Paraboni, I., & van Deemter, K. (2014). Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience*, 29(8), 1002–1017. <https://doi.org/10.1080/01690965.2013.805796>
- Paraboni, I., van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2), 229–254. <https://doi.org/10.1162/coli.2007.33.2.229>
- Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27, 89–110. <https://doi.org/10.1515/ling.1989.27.1.89>
- Repacholi, B., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, 33(1), 12–20. <https://doi.org/10.1037//0012-1649.33.1.12>
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2), 358–367. <https://doi.org/10.1037/0033-295x.107.2.358>
- Rubio-Fernandez, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*, 7, Article 153. <https://doi.org/10.3389/fpsyg.2016.00153>
- Rubio-Fernandez, P. (2019). Overinformative speakers are cooperative: Revisiting the Gricean maxim of quantity. *Cognitive Science*, 43(11), Article e12797. <https://doi.org/10.1111/cogs.12797>
- Rubio-Fernandez, P. (2021). Color discriminability makes over-specification efficient: Theoretical analysis and empirical evidence. *Humanities and Social Sciences Communications*, 8, Article 147. <https://doi.org/10.1057/s41599-021-00818-6>
- Rubio-Fernandez, P., & Jara-Ettinger, J. (2020). Incrementality and efficiency shape pragmatics across languages. *Proceedings of the National Academy of Sciences*, 117(24), 13399–13404. <https://doi.org/10.1073/pnas.1922067117>
- Rubio-Fernandez, P., Mollica, F., & Jara-Ettinger, J. (2020). Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology: General*, 150(3), 583–594. <https://doi.org/10.1037/xge0000963>
- Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17(8), 692–699. <https://doi.org/10.1111/j.1467-9280.2006.01768.x>
- Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2), 211–232. [https://doi.org/10.1016/0010-0285\(89\)90008-X](https://doi.org/10.1016/0010-0285(89)90008-X)
- Sedivy, J., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71(2), 109–147. [https://doi.org/10.1016/s0010-0277\(99\)00025-6](https://doi.org/10.1016/s0010-0277(99)00025-6)
- Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32(1), 3–23. <https://doi.org/10.1023/a:1021928914454>
- Sedivy, J. C. (2005). Evaluating explanations for referential context effects: Evidence for cricene mechanisms in online language interpretation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 345–364). MIT Press.
- Sonnenschein, S., & Whitehurst, G. (1982). The effects of redundant communications on the behavior of listeners: Does a picture need a thousand words? *Journal of Psycholinguistic Research*, 11(2), 115–125. <https://doi.org/10.1007/BF01068215>
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Harvard University Press.
- Tourtour, E., Delogu, F., Sikos, L., & Crocker, M. (2019). Rational over-specification in visually-situated comprehension and production. *Journal of Cultural Cognitive Science*, 3, 175–202. <https://doi.org/10.1007/s41809-019-00032-6>
- Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 1874–1882). NIPS Foundation.
- van Arkel, J., Woensdregt, M., Dingemanse, M., & Blokpoel, M. (2020). A simple repair mechanism can alleviate computational demands of pragmatic reasoning: Simulations and complexity analysis [Conference session]. 24th Conference on Computational Natural Language Learning, Stroudsburg, Pennsylvania, United States.
- van Deemter, K., Gatt, A., van Gompel, R. P., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4(2), 166–183. <https://doi.org/10.1111/j.1756-8765.2012.01187.x>
- van Gompel, R. P., van Deemter, K., Gatt, A., Snoeren, R., & Krahmer, E. J. (2019). Conceptualization in reference production: Probabilistic modeling and experimental testing. *Psychological Review*, 126(3), 345–373. <https://doi.org/10.1037/rev0000138>
- Viethen, J., van Vessel, T., Goudbeek, M., & Krahmer, E. (2017). Color in reference production: The role of color similarity and color codability. *Cognitive Science*, 41(Suppl. 6), 1493–1514. <https://doi.org/10.1111/cogs.12387>
- Waldon, B., & Degen, J. (2021). Modeling cross-linguistic production of referring expressions. *Proceedings of the Society for Computational Linguistics*, 4, 206–215. <https://doi.org/10.7275/vsfn-t057>
- Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*, 6, Article 935. <https://doi.org/10.3389/fpsyg.2015.00935>
- Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2), 183–194. [https://doi.org/10.1016/0749-596X\(92\)90010-U](https://doi.org/10.1016/0749-596X(92)90010-U)
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103–128. [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)
- Wu, S. A., & Gibson, E. (2021). Word order predicts cross-linguistic differences in the production of redundant color and number modifiers. *Cognitive Science*, 45(1), Article e12934. <https://doi.org/10.1111/cogs.12934>

Received March 22, 2021

Revision received October 15, 2021

Accepted October 17, 2021 ■