



# Shades of confusion: Lexical uncertainty modulates ad hoc coordination in an interactive communication task

Sonia K. Murthy<sup>a,b</sup>, Thomas L. Griffiths<sup>a,c</sup>, Robert D. Hawkins<sup>a,\*</sup>

<sup>a</sup> Department of Psychology, Princeton University, Princeton, NJ, United States of America

<sup>b</sup> Allen Institute for Artificial Intelligence, Seattle, WA, United States of America

<sup>c</sup> Department of Computer Science, Princeton University, Princeton, NJ, United States of America

## ARTICLE INFO

### Keywords:

Communication  
Concepts  
Social representations  
Word-color associations

## ABSTRACT

There is substantial variability in the expectations that communication partners bring into interactions, creating the potential for misunderstandings. To directly probe these gaps and our ability to overcome them, we propose a communication task based on color-concept associations. In Experiment 1, we establish several key properties of the mental representations of these expectations, or *lexical priors*, based on recent probabilistic theories. Associations are more variable for abstract concepts, variability is represented as uncertainty within each individual, and uncertainty enables accurate predictions about whether others are likely to share the same association. In Experiment 2, we then examine the downstream consequences of these representations for communication. Accuracy is initially low when communicating about concepts with more variable associations, but rapidly increases as participants form ad hoc conventions. Together, our findings suggest that people cope with variability by maintaining well-calibrated uncertainty about their partner and appropriately adaptable representations of their own.

## 1. Introduction

From jargon-filled scientific communication (Anderson-Cook, 2010; Bullock, Colón Amill, Shulman, & Dixon, 2019; Martínez & Mammola, 2021) and medical consultations (Castro, Wilson, Wang, & Schillinger, 2007; Korsch, Gozzi, & Francis, 1968; McCabe & Healey, 2018) to the linguistic battlefields of political discourse (Lakoff, 2006; Wodak, 1989), it can often feel as if we are speaking different languages. Speakers not only bring different perspectives, expertise, and background knowledge into a conversation, they may even have different expectations about what words mean (Labov, 1973; Marti, Piantadosi, & Kidd, 2019; McCloskey & Glucksberg, 1978). For example, cognitive psychologists, doctors, and lawyers all commonly use the word “trial” in professional settings but refer to different concepts (a stimulus presentation, a clinical study, and a court appearance, respectively). Such lexical variability is a troubling and pervasive challenge, setting the stage for misunderstandings and other communication breakdowns. How do we manage to understand each other when we cannot be sure that we’re starting on the same page?

Modern theories have addressed this challenge by viewing communication not as a unitary act of transmission but as an ongoing collaborative process where interlocutors must coordinate to reach mutual understanding

over time (Clark, 1996; Davidson, 1986; Krauss & Fussell, 1996; Reddy, 1979; van Arkel, Woensdregt, Dingemanse, & Blokpoel, 2020). These theories acknowledge idiosyncrasy, misunderstanding, and variation in literal meaning across partners and communities as a fundamental and unavoidable aspect of language use (Clark, 1998; Elman, 2004; Schuster & Degen, 2020; Wilson & Carston, 2007). Still, a central open question has concerned the underlying mental representations of signal meaning that support this collaborative process. One hypothesis raised by recent probabilistic theories of communication is that the ability to anticipate and flexibly overcome misunderstandings depends on each speaker’s initial *lexical uncertainty*, reflecting prior expectations about what messages may or may not mean to one’s partner (Bergen, Levy, & Goodman, 2016; Brochhagen, 2020; Hawkins et al., 2022; Potts & Levy, 2015). This idea builds on the classical construct of a mental lexicon containing signal-meaning mappings, allowing words to be grounded in external referents. For example, a speaker would consult their lexicon to evaluate the extent to which a given word like “dog” applies to a given animal they’ve encountered in the world. Lexical uncertainty replaces a fixed dictionary of mappings with a probability distribution over possible mappings that different partners may be using. For example, there may be more uncertainty about whether a given partner will share a given meaning for some

\* Corresponding author.

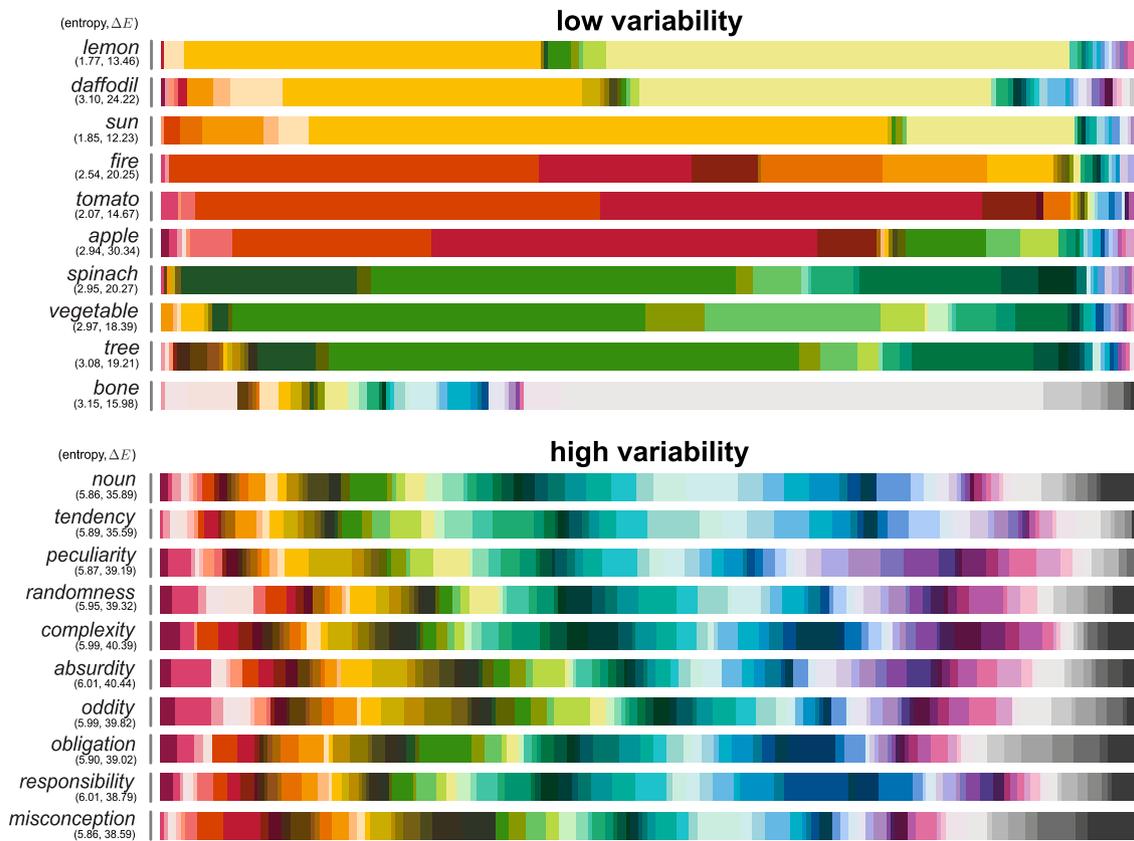
E-mail address: [rdhawkins@princeton.edu](mailto:rdhawkins@princeton.edu) (R.D. Hawkins).

<https://doi.org/10.1016/j.cognition.2022.105152>

Received 13 May 2021; Received in revised form 7 February 2022; Accepted 26 April 2022

Available online 20 May 2022

0010-0277/© 2022 Elsevier B.V. All rights reserved.



**Fig. 1.** Example color associations. We depict the response distribution of the 10 concepts with lowest and highest variability. The width of each color bar corresponds to the proportion of a particular color response for a given word. Colors are presented in a fixed order across bars but colors that received no responses are not visible in the given bar. We report two estimates of variability for each concept (see Methods; in both cases, higher is more variable).

technical jargon than whether they will share a given meaning for “dog.”<sup>1</sup>

While lexical priors have played a key explanatory role in computational models of communication, they have been challenging to measure and manipulate directly in experimental work. Classical studies of coordination and communication have typically been restricted to stimuli like ambiguous tangram shapes (Clark & Wilkes-Gibbs, 1986), line drawings (Krauss & Weinheimer, 1964), or complex scenes (Weber & Camerer, 2003) where lexical priors fall in a narrow, carefully-calibrated band of uncertainty between completely random (e.g., white noise) and completely universal (e.g., a photograph of a prototypical dog). Despite the narrow range of these stimulus spaces, it has been possible to isolate certain effects that are consistent with accounts of lexical uncertainty. For example, there is evidence that the *codeability* of tangrams (the number of distinct descriptions elicited; Hupet, Seron, & Chantraine, 1991), and the shared expertise of speakers (whether participants are equally familiar with New York landmarks; Isaacs & Clark, 1987) both affect the time it takes for speakers to reach mutual understanding. Still, there are many reasons to explore lexical uncertainty in richer stimulus spaces and other communication modalities. Richer stimulus spaces make it possible to observe and manipulate lexical priors spanning a broader range of uncertainty, from the most

idiosyncratic to the most universal. Other communication modalities present a further opportunity to overcome inherent challenges associated with natural language, where probability distributions must be estimated from sparse samples and poor coverage over the full (infinite) spaces of possible utterances and meanings. For example, measures of codeability for referential meaning (Hupet et al., 1991) are based on a single description from each participant, where most descriptions appear only once in the data set.

### 1.1. Color-concept associations as a window onto lexical uncertainty in communication

To proceed in the face of these methodological challenges, we propose a communication paradigm based on color-concept associations. While color has commonly been used as the *target* of reference in communication tasks (Caldwell & Smith, 2012; Monroe, Hawkins, Goodman, & Potts, 2017; Morin, Müller, Morisseau, & Winters, 2022; Winters & Morin, 2019), we instead ask participants to use a set of color chips as their *communication modality* (e.g., Roberts & Clark, 2018) to communicate the identity of a target concept in a context of distractors (e.g. *lemon*, *happiness*). While these color chips clearly differ in important ways from natural language utterances as a communication modality (see General Discussion), we argue that color-concept associations nevertheless provide a number of advantages for examining the consequences of lexical uncertainty. First, color-concept associations naturally span a wide range of possible priors (Fig. 1). For example, some concepts, like *lemon*, are expected to have strong, nearly universal color associations, reflected in tight priors  $P(\text{color}|\text{lemon})$ . Meanwhile, other concepts, like *fairness* are expected to have more idiosyncratic and distributed associations, reflected in looser and more spread-out priors

<sup>1</sup> While lexical uncertainty is typically defined as a joint distribution over the full set of signal-meaning tuples  $P(m,s)$  (see Appendix A), it can also be expressed as the conditional uncertainty over meanings for a given signal,  $P(m|s)$ , or as the conditional uncertainty over signals for a given meaning  $P(s|m)$ , since the baseline probabilities  $P(m)$  and  $P(s)$  are of lesser interest. We will loosely use *lexical prior* and *lexical uncertainty* to talk about this whole family of expressions, but the final conditional expression  $(s|m)$  is particularly useful for eliciting priors.

(Hutchings, 2004). Prior studies eliciting such associations have found systematic differences across different abstract concepts (Barchard, Grob, & Roe, 2017; Guilbeault et al., 2020; Mohammad, 2011; Rathore, Leggon, Lessard, & Schloss, 2019; Volkova, Dolan, & Wilson, 2012), which may also vary cross-culturally (Hupka, Zaleski, Otto, Reidl, & Tarabrina, 1997; Tham et al., 2020). We build on the elicitation methods developed in prior work while also introducing several methodological innovations to answer novel questions arising in the context of communication. Second, while natural language utterances are typically understood to be embedded in a complex and high-dimensional semantic space (Jones & Mewhort, 2007; Pennington, Socher, & Manning, 2014), color signals are embedded in a much lower-dimensional space with better-validated psychometric structure, which allows for denser sampling and explicit measurements of variation. Finally, color remains an important modality of natural communication in its own right (Riley, 1995), as evidenced by the deliberate choice of color palettes in graphic design (Marcus, 1982) and marketing (Labrecque & Milne, 2011), or our everyday metaphorical appeals to color when trying to convey complex emotional states that are challenging to describe with words (Lakoff & Johnson, 1980; Meier & Robinson, 2005; Van Leeuwen, 2011).

### 1.2. Three foundational questions about lexical uncertainty

We use the domain of color-concept associations to evaluate three foundational hypotheses about lexical uncertainty raised by recent theories of communication. First, and most fundamentally, do individuals actually maintain an internal probability distribution representing their uncertainty about associations (Fig. 2, top row), or do they only represent a point estimate giving their strongest association (Fig. 2, bottom row)? Second, we ask: is the population relatively homogeneous, composed of individuals sharing similar representations (Fig. 2, right column) or is the population actually more heterogeneous and

idiosyncratic (Fig. 2, left column)? Third, when it comes time to use these representations in a communicative context, is a given individual's representation purely *egocentric* or do they maintain well-calibrated expectations about whether their representation will be shared by *other agents*?

While we unpack these hypotheses and operationalize their predictions more thoroughly in subsequent sections, it is worth noting here that there is theoretical precedent for these questions not only in the communication literature but also in the literature on color-concept associations. For example, the *color inference* framework introduced by Schloss (2018) proposes that individuals store and continually update their color-concept associations from their experiences in the world. Under this framework, every concept has a corresponding association space with some weight placed on each possible color. These weights effectively give rise to an internal probability distribution that can, for example, be used to generate appropriately discriminative colors to convey different meanings (Mukherjee, Yin, Sherman, Lessard, & Schloss, 2022). Such resonances between models of meaning across the domains of natural language communication and of color-concept associations provides a further theoretical motivation for using the color domain to explore representations of lexical uncertainty.

In Experiment 1, we begin to explore basic theoretical properties of lexical uncertainty by eliciting color associations for a variety of concepts. By eliciting multiple responses from each participant, we were able to compare the *population-level* variability of responses to the *internal* variability within each individual. And by asking participants to estimate the extent to which others will share the same associations, we were able to probe the extent to which any given individual represents the population distribution, thus forming a basis for successful communication. Next, in Experiment 2, we directly measure the downstream effects of lexical priors on ad hoc coordination using an interactive communication task. Pairs of participants were asked to

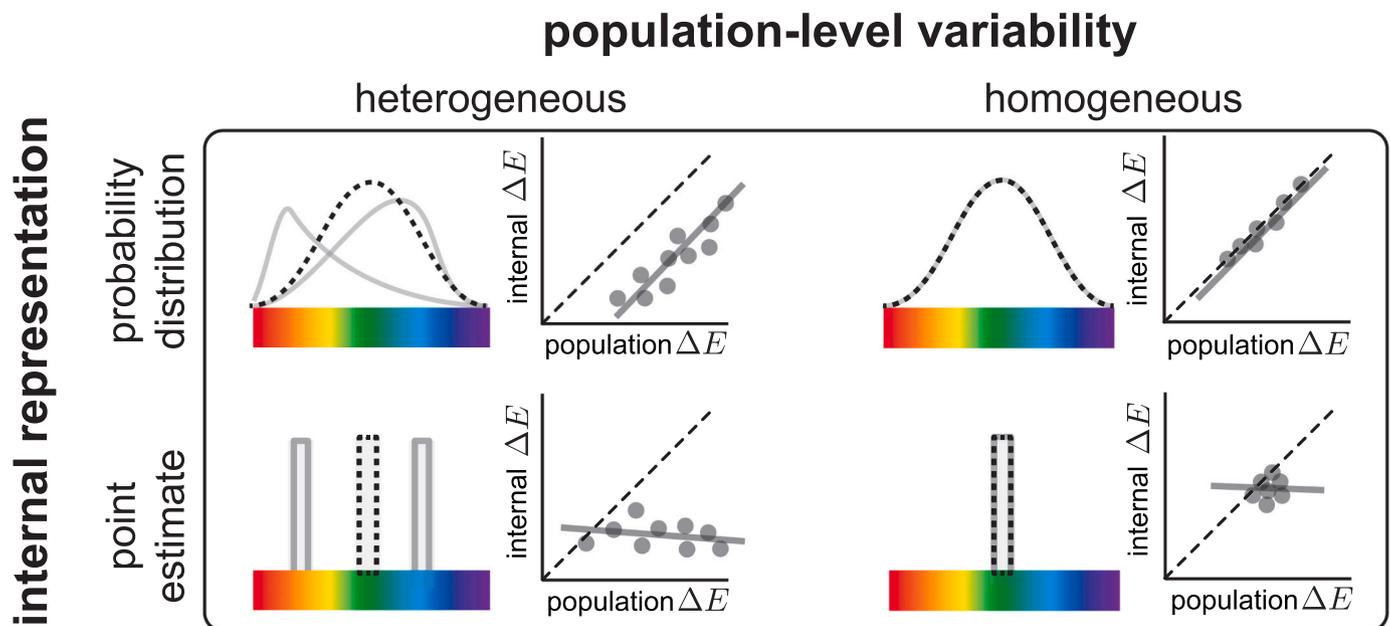


Fig. 2. Schematic of candidate hypotheses and corresponding predictions. We explore a  $2 \times 2$  space of proposals about how color-concept associations are internally represented by individuals (as a fully calibrated probability distribution representing uncertainty, or as a sparser point estimate) and how much these associations differ across individuals in the population (from a fully homogeneous population to a more heterogeneous population). These hypotheses make distinct predictions about the relationship between the variability of associations measured within an individual and across a population, which we evaluate in Experiment 1.

communicate about sets of concepts using color chips as signals. Critically, we used the corpus of associations we elicited in our first experiment to construct contexts that *manipulated* participants' initial priors: some associations were strongly shared across the population while others were more idiosyncratic. Taken together, our work suggests that people maintain well-calibrated prior expectations about the potential for miscommunication and use these flexible priors to rapidly adapt to their partner.

## 2. Experiment 1: Eliciting color-concept association priors

We begin by eliciting color association priors for concepts spanning a wide range of population variability. While there are a number of existing datasets that may be used to assess variability at the population level (e.g. [Mohammad, 2011](#); [Tham et al., 2020](#); [Volkova et al., 2012](#)), there were three specific desiderata for our work that were not satisfied by these existing datasets. First, we required participants to not only choose the color they themselves most associated with a concept, but also to predict whether that association would be shared by others. Such an explicit query about expected agreement was not included in the protocol of previous studies. Second, while previous studies revealed differences in population-level variability, it remains unclear whether such variability arises within individuals (e.g. from internal probability distributions) or only at the population-level (e.g. from individuals with different point estimates). To distinguish between these possibilities, we required multiple blocks of judgements for each participant; previous variants of the task only collected a single best exemplar per participant. Finally, we required a set of stimuli that would span a wide range of expected association variability from strong (e.g. "lemon") to weak (e.g. "fairness"). Previous studies have focused on smaller sets of concepts (e.g. [Tham et al., 2020](#), used only 59 abstract concepts, screening out words that corresponded to concrete objects like "lemon") or only collected coarse-grained responses for basic-level colors (e.g. [Mohammad, 2011](#)) asked participants to choose from the 11 color terms from [Berlin & Kay, 1969](#)).

### 2.1. Methods

#### 2.1.1. Participants

We recruited 733 participants from Amazon Mechanical Turk, restricting location to the United States. After implementing our pre-registered exclusions criteria, we were left with data from 485 of these participants. 129 participants were excluded for failing one of four attention checks that were interspersed throughout the experiment. These attention checks asked participants to provide color associations for basic-level color words (e.g., "red", "orange", "yellow"). The first two attention checks were embedded in a block of initial practice trials while the other two were inserted randomly into the experiment (one in the first half, and one in the second half). Participants who did not provide a color within a (relatively permissive) set of valid responses were immediately removed from the experiment for the base payment. We also removed 5 additional participants who provided the same response for more than three trials in a row, or consistently responded in less than 1000 ms, both of which indicated blind guessing without reading the stimulus prompt. Finally, an additional 65 participants were excluded due to colorblindness. To test color vision, we presented participants with three Ishihara plates that detected common red-green deficiencies or more extreme colorblindness. We excluded only those participants whose responses indicated more extreme colorblindness (or inattention).

#### 2.1.2. Stimuli

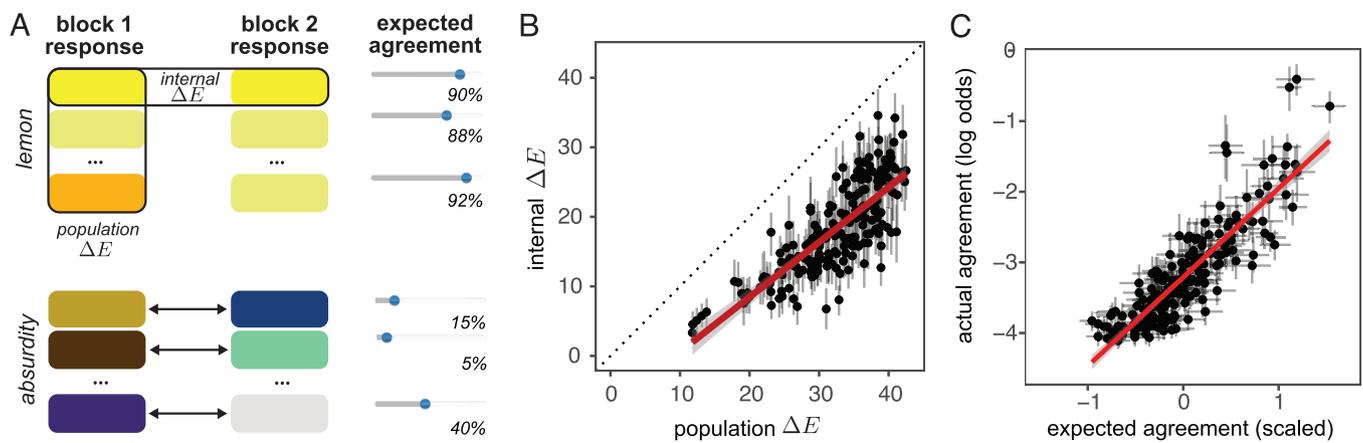
We considered two factors when selecting our stimulus set. First, we required a relatively large number of concepts to control for possible item effects. Second, we needed these concepts to span a wide range of different priors (i.e. different color associations, with different levels of variability), but were unable to know these priors beforehand. We thus considered a set of 5500 candidate concepts drawn from the Glasgow Norms dataset ([Scott, Keitel, Becirspahic, Yao, & Sereno, 2019](#)), which provides ratings for words on 9 different scales, including properties like imageability and concreteness. Concreteness represents the degree to which something can be experienced by the senses, while imageability represents the extent to which a word invokes a mental image.<sup>2</sup> We used these measures as rough proxies for the level of variability in color associations we could expect for a concept and selected a balanced set of 200 concepts from this candidate set using the following procedure. We began by imposing a familiarity threshold (familiarity  $\geq 4.0$ ) to ensure that the majority of participants were likely to know the concept that each stimulus word referred to. Next, we selected the 500 words with the highest concreteness ratings (the *concrete* set) and the 500 words with the lowest concreteness ratings (the *abstract* set). We then sampled 100 words from each set to obtain a roughly uniform distribution of imageability. Lastly, we conducted a manual pass over the resulting 200 words to ensure that they were consistent in part of speech (e.g., converting adjectives to their noun form) and to replace any that were offensive, confusing, or redundant with one another. These 200 words were randomly divided into 5 subsets, each containing 20 words from the *abstract* set and 20 words from the *concrete* set.

#### 2.1.3. Task, design, & procedure

Each participant was assigned one of the 5 distinct word sets. On each trial, a single word was presented with a set of 88 (virtual) Munsell chips sampling a wide range of color space.<sup>3</sup> Participants were instructed to click the color they most associated with the target word. To control for differences in individuals' color displays, participants were instructed to take the experiment on a desktop or laptop computer and ensure that their screens were set to their default brightness and color temperature (e.g. to turn off programs like Flux). Participants were also screened through a pre-test asking them to select color swatches for words like 'blue' and 'red', ensuring that any differences in color displays lay within tolerance of color boundaries (see Supplemental Figs. S3-S5). To estimate internal variability, we presented these words in a blocked sequence. After providing responses for all 40 words in the set, presented in randomized order, participants repeated the task a second time ([Fig. 3A](#)). On the second block, participants were also asked: "How strongly do you expect others to share your color association for this word?" We presented a slider ranging from "not at all" (most people will have a different color association than I do), to "very strongly" (most people will have the same color association as I do) with a midpoint labeled "somewhat" (roughly the same number of people will have the same or different color associations as I do). This question allowed us to compare the true proportion of shared responses to each participants' expectations.

<sup>2</sup> Though these two measures are highly correlated with each other ( $r = 0.93$ ), they are considered distinct aspects of a word's semantics ([Paivio, Yuille, & Madigan, 1968](#); [Richardson, 1975](#)). For example, emotion words like *anger* may be rated low on concreteness but high on imageability; conversely, some scientific or medical concepts like *diabetes* may be high on concreteness but low on imageability.

<sup>3</sup> The World Color Survey ([Kay, Berlin, Maffi, Merrifield, & Cook, 2009](#)) used the set of 320 chips (40 evenly spaced hues crossed with 8 levels of lightness) proposed by [Lenneberg and Roberts \(1956\)](#). Later studies down-sampled this set to 160 ([Heider, 1972](#)) by removing hues at intermediate levels of 2.5 and 7.5, and then further down-sampled to 80 ([Gibson et al., 2017](#); [Zaslavsky, Kemp, Tishby, & Regier, 2019](#)). Our 88-chip set was derived by adding 8 achromatic chips to the 80-chip set from [Gibson et al. \(2017\)](#), allowing participants to select grayscale values.



**Fig. 3.** Design and results of Experiment 1. (A) We elicit two associations from each participant for a given word (1st block and 2nd block). On the second block, participants additionally estimate the probability that others would provide the same association. Internal  $\Delta E$  is measured as the similarity between the two responses provided by the same individual, while population  $\Delta E$  is measured between different individuals. (B) Internal  $\Delta E$  of color associations for each concept tracks population variability  $\Delta E$ , although a single participants' responses are somewhat closer together than expected from sampling from the population distribution. (C) Expectations are well-calibrated to the true agreement odds. Error bars for each concept are bootstrapped 95% CIs.

#### 2.1.4. Evaluation Metrics

We measured variability in two different ways: (1) the entropy of the discrete response distribution and (2) the average pairwise similarity between responses in color space.<sup>4</sup> Our entropy measure was computed on the distribution of response counts over the 88 Munsell chips, using the Schurmann-Grassberger estimator (i.e. adding pseudo-counts of  $1/88$  to each bin). Entropy is expected to be high when responses are spread out across many different color chips, and low when participants all concentrate their responses on a small number of colors. Our pairwise distance measure is computed by taking the perceptual similarity  $\Delta E$  (where  $E$  stands for *Empfindung*, German for “sensation”) between colors in the psychometrically-validated CIELAB color space. It is close to 0 when colors are perceptually similar (colors with  $\Delta E < 1$  are not able to be discriminated by the human eye), and reaches values close to 100 for extremely dissimilar colors. We use the CIE2000 definition of  $\Delta E$ , which accounts for distortions in perceptual uniformity (Sharma, Wu, & Dalal, 2005). For each word, we derived a measure of internal variability, *internal*  $\Delta E$ , by taking the  $\Delta E$  between the color chosen by a given participant in the first block and the second block. We also obtained a word-level measure of population variability, *population*  $\Delta E$ , by taking the average pairwise  $\Delta E$  among every pair of color responses provided by different participants.<sup>5</sup>

<sup>4</sup> We also pre-registered a procedure to estimate parametric variability by fitting a multi-dimensional Gaussian distribution over color space, but chose to replace this measure by the  $\Delta E$  measure. It is highly correlated with the Gaussian measure,  $r = 0.861$ , but better accounts for metric distortions in color space and the existence of multi-modal response distributions.

<sup>5</sup> We may define population  $\Delta E$  in several ways, depending on where the different participants' responses are drawn from. For example, block 1  $\Delta E$  compares responses from different participants in the first block ( $\Delta(c_{1i}, c_{1j})$ ), block 2  $\Delta E$  restricts to the second block, ( $\Delta(c_{2i}, c_{2j})$ ), and a “cross-block”  $\Delta E$  compares responses in one block to those obtained from other participants in the other block, ( $\Delta(c_{1i}, c_{2j})$ ). These different ways of measuring population  $\Delta E$  are highly correlated ( $r = 0.994$  between the block 1 and cross-block versions;  $r = 0.993$  between the block 2 and cross-block versions;  $r = 0.973$  between the block 1 and block 2 versions), and results are invariant to this choice. While Fig. 3A shows the block 1 variant to make it clear that the axis of comparison is across different participants, we report results using the version that pools together all responses across both blocks to get the most highly powered estimate.

## 2.2. Results

### 2.2.1. Associations are more variable for abstract words

We begin by examining our use of concreteness and imageability as proxies for constructing stimuli that elicit a wide range of associations. These proxies were motivated by recent work suggesting that the extent to which people share associations for a concept may be related to the extent that they share the same *sensory experiences* for that concept, such that abstract concepts have more variable associations than concrete and imageable concepts. For example, recent analyses of Google Image search results have found that raw color distributions found in the top images retrieved for abstract terms are indeed more variable on average than those measured for concrete terms (Desikan et al., 2020). At the same time, however, the relationship between these measures is likely more complex: recent work has also found coherent relationships among abstract concepts (Guilbeault et al., 2020) and some abstract words may nevertheless have strong associations (e.g. *anger* is red and *sadness* is blue). To assess this relationship in our dataset, we constructed two regression models predicting our two population variability measures (entropy and  $\Delta E$ ) from the corresponding concreteness and imageability ratings reported by (Scott et al., 2019) for each word (see Appendix B, Fig. S1).

We found that both concreteness and imageability are independently correlated with entropy (Spearman rank correlation  $\rho = -0.51$ ,  $p < 0.001$  and  $\rho = -0.54$ ,  $p < 0.001$ , respectively), suggesting that color associations for more abstract words were much more variable. However, in a combined model including both predictors we only found a main effect of imageability ( $b = -0.31$ ,  $t(191) = -3.19$ ,  $p = 0.002$ ), with no independent contribution of concreteness ( $b = 0.03$ ,  $p = 0.74$ ), suggesting that the information provided by concreteness may be redundant. Similarly, both concreteness and imageability are independently correlated with  $\Delta E$  ( $\rho = -0.53$ ,  $p < 0.001$  and  $\rho = -0.52$ ,  $p < 0.001$ , respectively), but we found the opposite relationship in the combined model: concreteness was a weak but significant predictor of  $\Delta E$  ( $b = -1.22$ ,  $t(191) = -2.07$ ,  $p = 0.039$ ), while imageability was not ( $b = -0.8$ ,  $p = 0.26$ ). Together, these findings support a strong negative relationship between abstractness and variability, regardless of how we measure them, but we were unable to distinguish the unique contributions of concreteness from those of imageability due to their high collinearity in the combined model. Importantly, while they were useful for choosing stimuli, neither measure was a particularly good proxy for population variability in absolute terms. We therefore use our own direct estimates of variability (i.e. entropy and  $\Delta E$ ) to construct high and

low variability conditions in Experiment 2, rather than relying on coarser concreteness or imageability norms.

### 2.2.2. Population variability reflects individual uncertainty

Given that we successfully elicited associations with different levels of variability in the population, we now proceed to ask how this variability is related to the internal representation of color-concept associations within each individual. First, we ask whether individuals maintain a single *point estimate* representing a single strongest association or instead maintain a full *internal probability distribution* over different associations. Second, we ask whether the population is relatively *homogeneous*, where all individuals maintain roughly the same representations, or whether it is more *heterogeneous*, containing individuals with somewhat differing representations (see Fig. 2). To evaluate this space of candidate hypotheses, we leveraged our blocked design to compare *internal variability* against *population variability*.

In particular, we draw on the notion of the “crowd within” (Fiechter & Kornell, 2021; Hourihan & Benjamin, 2010; Rauhut & Lorenz, 2011; Steegen, Dewitte, Tuerlinckx, & Vanpaemel, 2014; van Dolder & van den Assem, 2018; Vul & Pashler, 2008) that has recently been proposed for phenomena like forecasting and estimation in the decision-making literature (see Herzog & Hertwig, 2014, for a review). The “crowd within” views a judgement as a sample from an (implicit) probability distribution, explaining why the same participant may make different judgements at different times, and why averaging together multiple judgements from the *same* participant may yield better predictions (an analogue of the “wisdom of the crowd” where judgements are averaged across different participants). Thus, while it is unrealistic to perfectly reconstruct any single individual’s prior distribution from only two responses per concept, a key insight of this literature is that we do not need to reconstruct the *full* distribution maintained by any single participant. It is statistically sufficient to compare the distance between a small number of samples obtained from a given participant for a given concept (internal  $\Delta E$ ) against the expected null distribution of distances derived from different participants for that concept (population  $\Delta E$ ; see Fig. 3A).

Each hypothesis predicts a different pattern of relationships between internal  $\Delta E$  and population  $\Delta E$  (see Fig. 2). The distinction between different internal representations concerns whether internal and population variability are related across concepts. The *point estimates* representation allows for some response noise but assumes such noise is identical across all concepts, hence internal variability should be independent of population variability and we would expect similar internal  $\Delta E$  across blocks for all concepts.<sup>6</sup> Meanwhile, the *internal probability distribution* account predicts that individuals maintain different distributions for different concepts, which vary in their internal variability, hence there should be a systematic relationship between internal  $\Delta E$  and population  $\Delta E$  (regardless of whether the population is more homogeneous or heterogeneous). The distinction between population homogeneity and heterogeneity, on the other hand, concerns the extent to which population variability is higher or lower overall than internal variability. In the extreme case where every participant  $i$  shared the same distribution of color associations  $P(r|c)$  in a perfectly homogeneous population, then the entire dataset of responses  $r_{ij}$  would be drawn from the same distribution  $r_{ij} \sim P(r|c)$  and the distances between two samples taken from the same participant  $i$ ,  $\Delta(r_{i1}, r_{i2})$ , would be the same, in expectation, as the distances between samples taken from distinct participants,  $\Delta(r_{1j}, r_{2j})$ . In other words, we may obtain a distribution of the expected variation in any pair of sampled responses, under the null hypothesis of a homogeneous population, and compare the extent to which the actual pairs of samples we obtained from the same participant

<sup>6</sup> More formally, under this hypothesis, we could suppose an individual  $i$ ’s responses  $j$  are expected to be drawn from  $r_{ij} \sim \mathcal{N}(c_i, \epsilon)$  where  $c_i$  is their point estimate for the concept and  $\epsilon$  captures a fixed probability that they may click on nearby color chips rather than the precise value  $c_i$ .

are more or less similar than expected under this null.<sup>7</sup>

Our findings are shown in Fig. 3B. We found that the average  $\Delta E$  between an individual participant’s responses was strongly correlated with the overall population’s  $\Delta E$ ,  $r = 0.81$ ,  $t(192) = 19.04$ ,  $p < 0.001$ . That is, for concepts where the population as a whole most disagreed with one another, each individuals’ own responses across blocks also tended to disagree with one another. This pattern was only predicted by the *internal distribution* account and not the *point estimates* account, which cannot explain why there would be such systematic differences in internal  $\Delta E$  across concepts. At the same time, we found that internal variability was lower than population variability overall; the estimated intercept in a linear regression significantly differed from zero,  $b = -17.3$ ,  $t(192) = 20.8$ ,  $p < 0.001$ , consistent with a more heterogeneous account where not all participants shared exactly the same internal distribution. A homogeneous population would be closer to the line of unity.

### 2.2.3. Expectations about others are well-calibrated

Our results so far support the view that each individual implicitly maintains a full probability distribution or *prior* for each concept, which tracks the true population variability but which tends to be somewhat narrower on average. However, it is unclear whether this representation has a social component. In principle, communicative success depends on each individual’s expectations about how their *partner* will understand (or misunderstand) a given message, not just whether individuals represent similar meanings. This distinction is crucial for understanding the consequences of lexical priors for resolving misunderstandings in communication. If individuals are unable to represent whether *others* will share the same prior (whether it is a point estimate or a distribution), they might over- or under-estimate agreement and inhibit successful communication.<sup>8</sup> To assess whether participants maintain well-calibrated social expectations, we turned to the expected agreement measure collected on the second block. For each expected agreement rating in our dataset, we calculated the log odds that other participants’ color responses actually matched that participants’ reported color association,  $\log(p/(1 - p))$ , representing ground-truth agreement. We found a strong correlation between expected agreement and true agreement (Pearson’s correlation  $r = 0.88$ ,  $t(197) = 25.5$ ,  $p < 0.001$  and Spearman’s rank correlation,  $\rho = 0.87$ ,  $p < 0.001$ ), suggesting that individuals’ expectations about the others’ were remarkably well-calibrated to the true statistics (see Fig. 3C).<sup>9</sup>

<sup>7</sup> There exist other possible explanations for a smaller internal  $\Delta E$ , although we have taken measures to minimize them; for example, if participants’ response was strongly influenced by their response on the first block, then their samples may violate the assumption of independent draws from the distribution. However, such influence is unlikely given the number of intervening stimuli in each block and the relatively long delay between samples. We return to this concern in the discussion.

<sup>8</sup> Of course, there are multiple ways for well-calibrated social expectations to be achieved. One possibility is that individuals maintain their own idiosyncratic associations, based on their unique experiences, but also track the degree of divergence from the overall population and correct for it in social settings using theory of mind (i.e. they are aware that their associations are idiosyncratic). Another possibility is that agents are egocentric (i.e. maintain a single internal distribution of associations) but have tuned their own internal distribution over time to match the population distribution. Distinguishing between these representational possibilities is outside the scope of this paper.

<sup>9</sup> While the log-odds linking function is orthogonal to our question of interest, we note that it is consistent with previous observations about how participants spontaneously map sliders to logarithmic rather than linear scales (e.g. Griffiths & Tenenbaum, 2005, 2007; Landy, Silbert, & Goldin, 2013).

### 3. Experiment 2: The consequences of lexical priors for communication

Experiment 1 established several key properties of word-color associations: more abstract words have more variable associations, this variability is represented within individuals, and individuals can accurately predict whether an association is likely to be shared with others. These properties are precisely those that probabilistic models of communication have associated with *lexical uncertainty*, the recognition that particular word meanings or concepts may or may not be shared with others. Such uncertainty may allow for more flexible adaptation than simple point estimates: individuals may be able to anticipate potential confusions ahead of time, and possess rich enough knowledge about likely alternative associations to rapidly adjust their expectations on the fly. Here, we examine the downstream effects of these priors in a Pictionary-like communication task where participants sent color swatches as messages that allowed their partner to guess a target word. We hypothesized that target words with nearly universal color associations, reflected in strong, tightly overlapping priors, would provide a common foundation and allow for instant communicative success. Meanwhile, words with more variable or uncertain color associations would be more difficult to communicate about. In either case, individual pairs should still be able to adaptively coordinate on mutually agreeable solutions given the flexibility provided by their initial uncertainty.

#### 3.1. Methods

##### 3.1.1. Participants

We recruited 234 participants on Amazon Mechanical Turk and paired them up to form 117 dyads. After removing 6 dyads that disconnected before completing the task and 4 additional dyads where at least one participant failed our attention checks, we were left with 107 dyads in our sample. Participants were screened for comprehension and color vision before being paired with a partner (see Supplemental Fig. S6).

##### 3.1.2. Stimuli

We constructed 100 four-word contexts from the 200 words we used in Experiment 1. We designed these contexts to span a broad range of priors based on the population variability we estimated in the prior

experiment. We iteratively sampled contexts to satisfy to two main constraints, both derived from possible pragmatic context effects which may mask the effects of priors: (1) we aimed to group words with similar variability while (2) minimizing the extent to which different words in the same context have overlapping priors. To satisfy these constraints, we first ordered words by their estimated response entropy and greedily sampled from the list of words to build an initial context. To check the extent to which these words had overlapping priors, we computed the Jensen-Shannon (JS) divergence between the Experiment 1 response distributions for each pair of words in the proposed context. We imposed a minimum divergence threshold of 0.3: when the context exceeded this threshold for a word, we replaced it with the next in the list until a satisfactory set was formed. We repeated this procedure to obtain 50 contexts where each word appeared in exactly one context and overlap within each context was low. To obtain a distinct alternative set of contexts, we repeated the same procedure with the additional criterion of rejecting pairs of words that had previously appeared together in the first set. These 50 contexts were divided into a “high prior variability” condition (the top 25) and a “low prior variability” condition (the bottom 25).

##### 3.1.3. Procedure

Participants were paired into dyads to play an interactive reference game using color as the communication medium (Fig. 4). Participants were randomly assigned to *speaker* and *listener* roles and placed in an environment containing a context of 4 concept words and 88 Munsell chips, both shared in common ground. At the beginning of each trial, one of the 4 words was privately shown to the speaker as the target word. The speaker was then instructed to choose a color chip from the set of Munsell chips that would best allow the listener to select that target from the distractors. After the listener received the speaker’s message and clicked on one of the words, both participants received feedback: the listener was shown the true target and the speaker was shown the listener’s selection. Participants were awarded a performance bonus of \$0.03 for each correct response.

##### 3.1.4. Design

We tested the effect of lexical priors by manipulating the target words in a within-dyad design: each dyad was assigned two 4-word contexts, one “high prior variability” context and one “low prior

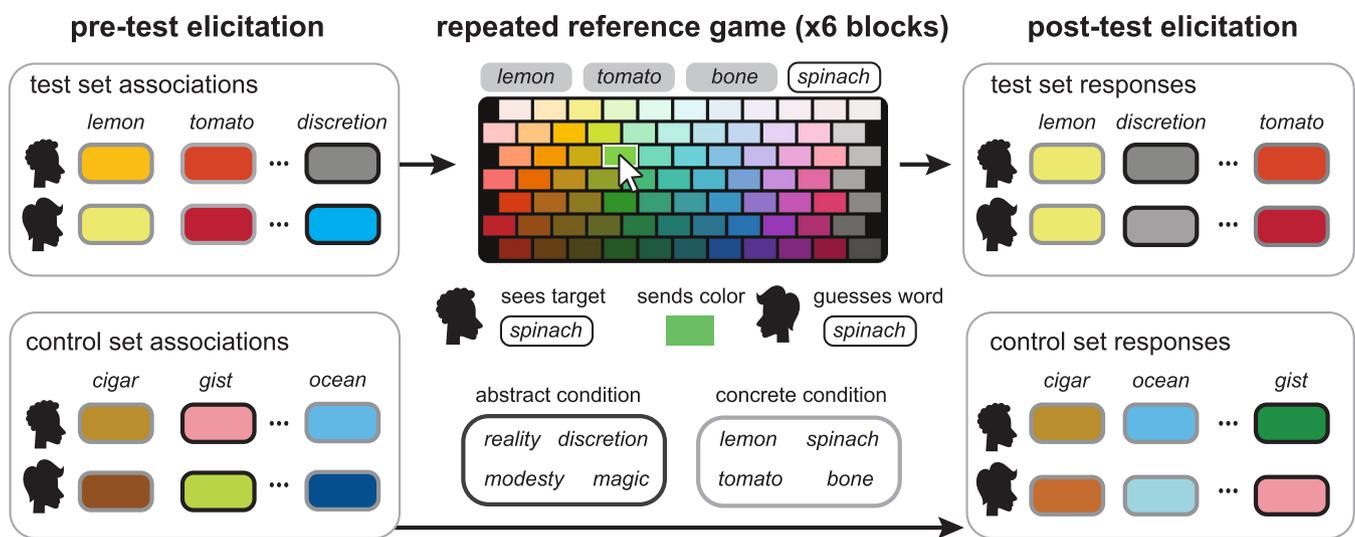
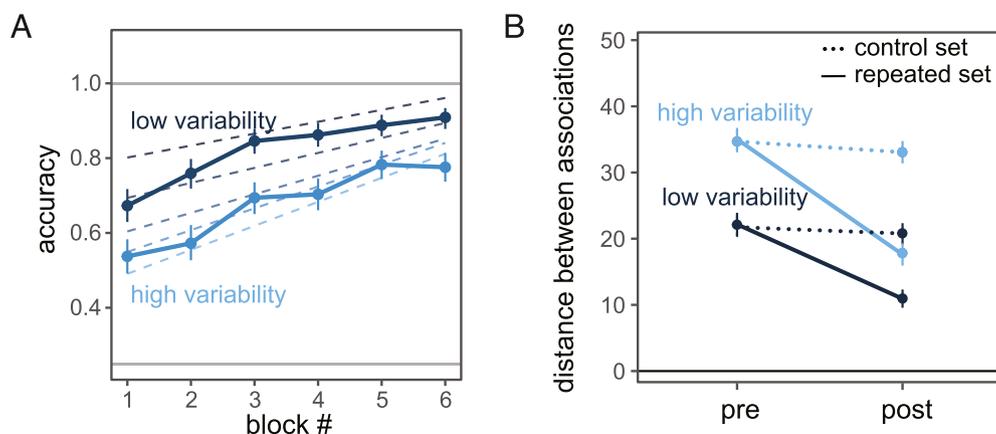


Fig. 4. Design of Experiment 2. Participants are paired into dyads and assigned to speaker and listener roles. Both participants are shown the full set of Munsell chips and the same context of 4 words. The speaker is additionally shown which one of these 4 words is the target word and asked to select a color to send to the listener. The listener then guesses the word they believe is the target. Participants independently complete a color association elicitation task before and after the reference game for the test set of eight words appearing in the communication game and also a control set of eight additional words.



**Fig. 5.** Results of Experiment 2. (A) Communicative success increases across repeated interaction. Dashed lines represent quintiles of initial similarity between partners' associations in the pre-test (a finer-grained measure than condition). (B) Speakers' associations begin closer in color space for low prior variability words than for high prior variability words, but converge to closer associations in the post-test, relative to a control set that did not appear in the reference game. Error bars are bootstrapped 95% CIs.

variability" context. The trial sequence was constructed from 6 repetition blocks, allowing us to observe the trajectory of behavior as each target is referred to multiple times. The four target words from each context were randomly interleaved in each block, for a total of 48 trials (six blocks of eight words). Participants switched roles between repetition blocks. Finally, we included two blocks of the prior elicitation task used in Experiment 1. At the beginning and end of the reference game, we asked participants to provide associations for the 8 *test* words in the contexts they were assigned as well as 8 *control* words not encountered during the reference game task – four with high prior variability and with low prior variability.

## 3.2. Results

### 3.2.1. Shared priors facilitate communicative success

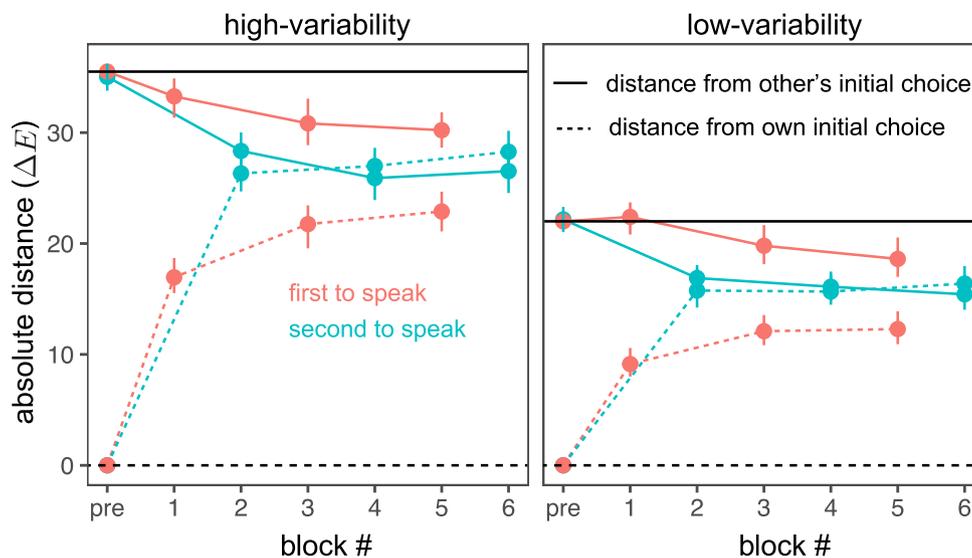
Our first prediction concerned the effect of priors on communicative success, which we operationalized as the probability that the listener correctly selects the target (chance is 25%). We hypothesized that participants would initially struggle more to communicate in the high prior variability condition, compared to the low prior variability condition, where they could take advantage of stronger expectations and more closely overlapping priors. At the same time, we expected participants to improve through interaction, as they built common ground across repeated appearances of the same targets. To test these hypotheses, we constructed a logistic regression to predict correctness at the trial-by-trial level, including fixed effects of condition (high prior variability context vs. low prior variability context), repetition block number, and their interaction, with the maximal random-effect structure that converged (random intercepts and both main effects at the dyad-level). We found a significant main effect of condition ( $b = 0.44, z = -4.8, p < 0.001$ ) with higher accuracy for the low prior variability condition throughout the task. We also found a significant improvement in accuracy across the task in both conditions ( $b = 0.43, z = 12.0, p < 0.001$ ), reflecting ad hoc coordination. Finally, these effects were clarified by a significant interaction ( $b = 0.06, z = -2.6, p = 0.009$ ), likely reflecting ceiling effects for the low-variability conditions. On the first round, dyads were at approximately 54% accuracy in the high variability condition, compared to 67% accuracy in the low variability condition. By the final round, they achieved approximately 78% and 91%

accuracy, respectively.<sup>10</sup> Because each condition contained a wide variety of items spanning different priors (reflected in high variances for random effect estimates), we also probed these effects using a more fine-grained, individualized measure. We computed the  $\Delta E$  distance between the two participants' pre-test responses for each word in each game, and found that quintiles of this continuous measure followed the same trend (dashed lines in Fig. 5A).

### 3.2.2. Rapid convergence to shared meanings

To evaluate the extent to which participants flexibly shifted their associations for target concepts over the course of interaction, we compared pre-test and post-test responses. We operationalized the similarity between partners' associations as the  $\Delta E$  between their responses at each phase. For example, as a manipulation check, we found that participants' responses were indeed more similar in the pre-test for words in the low variability condition than the high variability condition,  $d = 12.78, t(117.3) = 9.1, p < 0.001$ , implying that we successfully constructed separable context sets from the population variability estimates obtained in Experiment 1. Critically, however, we hypothesized that participants' responses in the post-test would become significantly closer for the words that were repeated in the reference game, compared to a comparable set of control words that only appeared in the pre-test and post-test. We tested these predictions in a mixed-effects linear regression model including fixed effects of phase (pre-test vs. post-test), word set (repeated vs. control), and condition (high vs. low variability) as well as all interactions, including random intercepts and slopes for all main effects). All variables were effect-coded to facilitate interpretation of interactions. Because the three-way interaction in this model is complex to reason about, we begin by considering a sub-model restricted to the repeated set only. We found a significant main effect of condition,  $b = 0.25, t(110.8) = 10.2, p < 0.001$ , with responses to low prior variability words being more similar at both phases, as well as a main effect of phase  $b = -0.38, t(120.0) = 16.49, p < 0.001$ , with more similar responses in the post-test for all words. We also find a significant interaction,  $b = -0.08, t(355.1) = -3.87, p = 0.0001$ , with high prior variability words experiencing an even larger shift, likely reflecting a floor effect. We evaluate the null hypothesis that this convergence simply reflects additional practice with the task using our control words, which appeared only in the pre-test and post-test. We found evidence of

<sup>10</sup> A Bayesian logistic regression model with the maximal random-effect structure at both the dyad- and item-level yielded similar effects (see Appendix Figure S2 for full results)



**Fig. 6.** Dynamics of adaptation in Experiment 2. Dotted lines represent distance from one's own initial choice while solid lines represent the distance from one's partner's initial choice. The red and blue lines track each participant's identity as they alternate roles. Both participants tend to shift away from their initial associations over time (dotted lines rise) while also moving closer to their partner's initial associations (solid lines fall). However, there is significant asymmetry in the directionality of adaptation: the first participant to take the role of the speaker shifts from their initial association to a lesser extent than the second participant that takes the role of the speaker. Horizontal black lines are provided for easier comparison to pre-test distances; error bars represent bootstrapped 95% CIs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a three-way interaction, where the interaction between phase and condition reported above is significantly different from the relationship found for control words, where similarity between partners remained relatively unchanged between the pre- and post-test (Fig. 5B).

### 3.2.3. The dynamics of adaptation

While we established strong effects of adaptation from the pre-test to the post-test, it remains unclear how, exactly, adaptation unfolds throughout the communicative interaction. Our final exploratory analysis examines how each participant's choices as speaker change from round to round as they observe one another's behavior. We took advantage of our experimental design, which required participants to alternate between the speaker and listener roles at the outset of each repetition block. We measured two distances at each block  $i$  (see Fig. 6): (1) the distance between the chosen color chip chosen by the speaker on that block and their own initial choice as measured in the pre-test (which we call "distance from own initial") and (2) the distance between the speaker's chosen color chip and their partner's initial choice (which we call "distance from other's initial"; note that the partner's initial choice is unobserved). We tested the effects of speaker turn (first vs. second), initial variability condition (high vs. low), and block number in mixed-effects regression models predicting these distances, including random intercepts at the subject-level and item-level as well as random effects for speaker turn at the subject level.

First, overall, participants tend to shift away from their initial associations over time,  $b = 1.2$ ,  $t(4713) = 5.5$ ,  $p < 0.001$ , while also moving closer to their partner's initial associations,  $b = -1.1$ ,  $t(4702) = -4.7$ ,  $p < 0.001$ . These shift suggests that partners tend to settle somewhere in the middle of their own initial associations and their partner's initial associations. Second, however, we find a significant asymmetry in how much each participant adapts. The first participant to take the role of the speaker shifts away from their initial association to a lesser extent than the second participant that takes the role of the speaker,  $b = -4.9$ ,  $t(111.7) = -6.03$ ,  $p < 0.001$  and also shifts closer to their partner to a lesser extent,  $b = 3.7$ ,  $t(113.8) = 5.1$ ,  $p < 0.001$ . This asymmetry is consistent with a "first mover advantage" where the first participant in the speaker role only has their own *lexical prior* to go on, while the second participant has additional information from observing those initial choices and is able to use that information to anchor their choices closer to their partner's. The asymmetry created by this alternation is solidified over the remainder of the game: the second participant actually ends up adopting a convention that is closer to their partner's initial association than to their own. Third, we find that the initial variability condition appears to simply shift this overall pattern up or down without

qualitatively changing the effect. In the low-variability condition, participants begin closer together but adaptation proceeds similarly,  $b = -10.0$ ,  $t(211.4) = -10.8$ ,  $p < 0.001$  and  $b = -10.3$ ,  $t(216.9) = -9.7$ ,  $p < 0.001$  for one's own associations and one's partner's associations, respectively.

## 4. Discussion

Communication is a continual challenge. Even when we speak the same language, our vocabularies are full of words that may mean different things to different partners. In this paper, we investigated the mental representations of meaning that people use to overcome this challenge. Using the domain of color-concept associations, where population variability can be carefully measured and manipulated, we first established that variability does not take people by surprise: individuals represent internal uncertainty about associations rather than point estimates and these agreement were well-calibrated to the actual population-wide statistics. We then used these elicited distributions to systematically manipulate the degree of variability in the lexical priors of partners in an interactive communication game. Although communication was initially difficult for words with more variable associations, participants were able to quickly adapt their expectations based on common ground accumulated within the game, leading their priors to become more similar over time. Taken together, these findings suggest that partners enter communicative settings with well-calibrated but flexible priors about the likely difficulty of communication. Our work provides new support for recent probabilistic accounts of *lexical uncertainty* in communication and raises a number of new directions.

First, most prominently, translating our findings from the relatively low-dimensional domain of colors back to the much larger lexical priors for natural language expressions will require further methodological advances. While the gap between these domains may seem to limit generalization and any potential differences in signaling medium need to be evaluated empirically, we suggest that colors may have more in common with words as a signaling medium than is initially apparent (Schloss, 2018). Just as the meanings of discrete word "tokens" are typically thought to lie in a continuous space (such that we can meaningfully talk about the semantic similarity between words or sentences), participants were presented with a discrete set of 88 color "tokens" whose semantic similarity is also grounded in a continuous underlying space (i.e. LAB space). Just as natural-language speakers cannot directly convey meanings from the continuous semantic space and must pass through the bottleneck of discrete tokens, our participants could not

directly access the continuous color space; they had to pass through the bottleneck of discrete chips. Perhaps the most significant difference between domains is not discreteness but that the underlying similarity of the chips is exposed visually, whereas the semantic similarity of word tokens is typically not exposed visually (e.g. words that are nearby in meaning often appear very different when written out orthographically). Especially because we organized the discrete set of color tokens according to their visual similarity in the task interface, we may have set up an easier pathway for speakers to compare similar signals. But we expect that this difference would primarily affect search and retrieval from the space of discrete tokens rather than the representation of uncertainty about a given token's meaning or the choice to use that token.

Second, our evidence for lexical uncertainty raises deeper mechanistic questions about how, exactly, variability is encoded and learned by individuals. There are a number of different computational models for probabilistic meanings that have arisen in natural language processing (NLP), including Gaussian embeddings and Gaussian mixtures (Athiwaratun, Wilson, & Anandkumar, 2018; Brazinskas, Havrylov, & Titov, 2018; Vilnis & McCallum, 2015) as well as a number of different proposals for how these distributions may be learned over time depending on an individual's own idiosyncratic experiences (Johns & Jamieson, 2018; Johns, Jones, & Mewhort, 2019; Kleinschmidt, 2019; Kraljic, Samuel, & Brennan, 2008). A related direction is to better characterize the mechanistic processes allowing participants to align their associations to communicate better (as we found in Experiment 2), and whether updates to an individual's associations are long-lasting or transient. One lower-level explanation for alignment is that participants are simply priming one another and making certain associations more salient or accessible (Pickering & Garrod, 2006). A higher-level explanation, not mutually exclusive with priming, is that speakers update their beliefs to form partner-specific common ground (Hawkins et al., 2022). This hierarchical account predicts that the extent to which local adaptation will persist in longer-term updates depends on sustained use of that association over time, across different partners. For example, as a slang term is more consistently and widely used throughout a language community, it may eventually supplant whatever initial associations individuals had with that word and persist as a longer-term update. However, such persistence would likely require more than a single interaction.

A third direction for future work is to examine how sources of variability in color associations arise in the first place. While we confirmed a coarse relationship between the abstractness of a concept and the variability of its associations, it remains unclear how this relationship arises. On one hand, some portion of this relationship may be driven by sensory aspects of word representations. We are all exposed to roughly the same visual imagery statistics for concrete concepts such as "tree," suggesting strong associations with greens and browns. Meanwhile, abstract concepts such as "justice" lack concrete referents and associations may be driven by more idiosyncratic semantic properties that vary depending on each individual's own history with the concept. These terms may be more susceptible to cultural variation across different languages and latent social groups (Tham et al., 2020), which would be interesting to measure in broader cross-cultural samples.

It is worth noting several potential limitations of our data. For one, there are intrinsic challenges associated with accurately sampling responses from color space. While we evenly sampled color chips from the Munsell color chart (Landa & Fairchild, 2005; Munsell, 1905), following standard practice for color elicitation (Berlin & Kay, 1969; Brown & Lenneberg, 1954; Gibson et al., 2017; Sturges & Whitfield, 1995), there are known distortions created by this set of chips (Zaslavsky, Kemp, Regier, & Tishby, 2018). Most noticeably, discrepancies in the relative number of green-blue and red-pink chips compared to yellow-ish chips may have biased responses towards better-represented hues. We expected this distortion to have the biggest effect on *entropy-based* measures of variability, which is more sensitive to the support of the response distribution than our  $\Delta E$  measure. Second, our online data collection setup prevented us from ensuring perfectly consistent color

calibration across participants' screens, raising concerns that variability in associations is simply due to presentation noise. While we did our best to minimize such sources of variability and bias, our primary comparisons were importantly *relative* comparisons between different words (e.g. between words with more or less variability). Because the display of colors on a given participant's screen was likely to be fixed across the experiment, any noise or bias arising from the color display should contribute equally across all words and conditions. Third, it is possible that our use of within-participant designs, both when eliciting multiple responses across blocks in Experiment 1 and when interleaving high and low variability contexts in Experiment 2, may have resulted in spillover or "self-priming" effects that reduce our estimates of internal variability. Generally, these possible effects would work against our hypotheses: in Experiment 1, self-priming would have favored the point estimate hypothesis rather than the internal uncertainty hypothesis, and in Experiment 2, it would have reduced our estimate of differences across conditions. More broadly, it will be important to reproduce our findings using longer delays between blocks.

More broadly, color associations are of substantial interest for communication in their own right. The very properties that we highlighted in Experiment 1 may be responsible for the prevalence and usefulness of color in communicating about abstractions, relative to more concrete modalities (Gass, 1975; Johansson, Anikin, & Aseyev, 2020; Schloss, Witzel, & Lai, 2020; Winter, 2019). When a speaker says "love is blue" they draw attention to different semantic dimensions than "love is bright red," which may be difficult to reach with other metaphorical expressions. Thus, characterizing uncertainty in communication with color associations is not only useful for practical *visual communication* like the choice of color in data visualization (Lin, Fortuna, Kulkarni, Stone, & Heer, 2013; Schloss, Leggon, & Lessard, 2021; Setlur & Stone, 2015), or for design elements like signage (Mahnke, 1996; Schloss, Lessard, Walmsley, & Foley, 2018), but is core to a more general theory of multi-modal communication. The meanings of colors may have much in common with the meanings of words we use in everyday conversations. Rather than rigid dictionaries, well-calibrated probability distributions allow us to better anticipate misunderstandings and coordinate with one another to achieve mutual understanding.

## Author contributions

All authors conceived of the project and contributed to the study design. SM and RH implemented the experiments, collected data, performed all analyses, and wrote the manuscript. TG provided feedback on the manuscript. All authors approved the final version of the manuscript for submission.

## Credit Authorship Contribution Statement

Sonia Murthy: Conceptualization, Methodology, Investigation, Formal analysis, Software, Writing - Original Draft, Visualization.

Thomas Griffiths: Conceptualization, Writing- Reviewing and Editing.

Robert Hawkins: Conceptualization, Methodology, Investigation, Formal analysis, Software, Writing - Original Draft, Writing- Reviewing and Editing, Visualization, Supervision, Project administration.

## Author Note

Materials and code for reproducing all behavioral experiments and analyses are open and available online at <https://github.com/hawkrobe/color-ref>. The pre-registration for Experiment 1 is at <https://osf.io/v86xz> and the pre-registration for Experiment 2 is at <https://osf.io/gmqsv>. Correspondence should be addressed to Robert Hawkins ([rdhawkins@princeton.edu](mailto:rdhawkins@princeton.edu)).

## Acknowledgements

Christiane Fellbaum, Josh Peterson, and Ken Norman. This work was supported by NSF Grant #1911835 to RH.

We are grateful for helpful conversations with Bill Thompson,

## Appendix A. What is a lexical prior?

While the present work focuses on basic *qualitative* properties of lexical priors, and a formal model is not strictly necessary to interpret our findings, an overview of previous formal definitions nonetheless helps make this construct more precise (see Bergen et al., 2016; Hawkins et al., 2022, for a more extensive treatment). We begin by defining a *lexicon*  $\mathcal{L}$ , which specifies a specific set of form-meaning mappings, assign semantic meanings to all utterances in a language.

**Definition 1.** A lexicon  $\mathcal{L} : (u, m) \rightarrow \{0, 1\}$  is a function assigning a Boolean truth value<sup>11</sup> to every pair of utterances  $u \in \mathcal{U}$  and meanings  $m \in \mathcal{M}$ .

Traditionally, all speakers of a language are assumed to learn a single fixed lexicon specifying the meaning of every utterance in the language.

**Example 1.** Consider a simple referential language game where there are 2 possible utterances  $u \in \{u_1, u_2\}$  and 2 possible meanings  $m \in \{m_1, m_2\}$ . Then a single lexicon  $\mathcal{L}$  can be represented as a binary  $|\mathcal{U}| \times |\mathcal{M}|$  matrix with 2 rows and 2 columns, where each entry represents whether utterance  $u$  has meaning  $m$  or not, and  $\mathcal{L}(u, w)$  simply looks up the specified entry:

$$\mathcal{L} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

Now, it is straightforward to define a lexical prior as a probability distribution over such matrices, representing an agent's uncertainty over exactly which lexicon is being used by their partner.

**Definition 2.** A lexical prior is a probability distribution over the support of possible lexicons, denoted by  $P(\mathcal{L})$ .

Note that the lexical prior  $P(\mathcal{L})$  is distinct from  $P(u)$  or  $P(m)$ , which simply represent the background probability of a given utterance or meaning independently popping up in the environment.

**Example 2.** Let  $\mathcal{L}_{ij}$  be the matrix entry representing whether utterance  $u_i$  has meaning  $m_j$ . Then

$$P(\mathcal{L}_{ij}) \sim \text{Bernoulli}(5)$$

defines a maximally uninformative prior where every utterance is equally likely a priori to have every meaning.

**Example 3.** If we additionally assume that lexicons satisfy the constraint that entries sum to one for each column of the matrix (a weak form of mutual exclusivity, where every meaning is assumed to have exactly utterance that expresses it), then the lexical prior may be specified even more compactly in conditional form as

$$P(\mathcal{L}_{ij}) = P(u_i | m_j) \sim \text{Categorical}(\theta_j),$$

where every meaning  $m_j$  is associated with a vector  $\theta_j$  giving a distribution over utterances. We are now prepared to consider how these theoretical constructs correspond our setting of color-concept signaling games, where the meanings  $m$  are a discrete set of concept words (e.g. *lemon*, *randomness*) and the utterances  $u$  are a discrete set of 88 color chips. We could in principle try to elicit the full latent object  $\mathcal{L}$  by showing participants every color-concept pair  $(u, m)$  and collecting a slider response or 2AFC judgement about whether that participant (or that participant's partner) would endorse that specific mapping. This kind of elicitation procedure might be the most direct instantiation of the formal definition but is expensive to collect (e.g. requiring  $50 \times 88 = 4400$  judgements from a single participant to cover just a quarter of the concepts we use). Instead, we elicit the prior by taking samples from the conditional distribution  $P(u|m)$ , which also happens to be a well-vetted method for eliciting color associations. In other words, we query a column of the lexicon (i.e. conditioning on a given concept) and ask participants to draw a sample from the induced distribution over color chips. For this reason, it is convenient to define the lexical prior *for a given concept*  $m$  as the elicited distribution over colors given  $m$ , although it is a slight abuse of the term.

## Appendix B. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cognition.2022.105152>.

## References

Anderson-Cook, C. M. (2010). Hidden jargon: Everyday words with meanings specific to statistics. In C. Reading (Ed.), *Proceedings of the 8th International Conference on Teaching Statistics (ICOTS8)*.

Athiwaratkun, B., Wilson, A. G., & Anandkumar, A. *Probabilistic fasttext for multi-sense word embeddings*. (2018). Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) <https://aclanthology.org/P18-1001>.

<sup>11</sup> This definition can be generalized to a real-valued function  $\mathcal{L} : (u, m) \rightarrow \mathbb{R}$  (Degen, Hawkins, Graf, Kreiss, & Goodman, 2020), yielding a graded or "fuzzy" semantics, but we use the traditional truth-conditional for simplicity.

- Barchard, K. A., Grob, K. E., & Roe, M. J. (2017). Is sadness blue? The problem of using figurative language for emotions on psychological tests. *Behavior Research Methods*, 49(2), 443–456.
- Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9.
- Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley, CA: UC Press.
- Brazinskas, A., Havrylov, S., & Titov, I. *Embedding words as distributions with a Bayesian skip-gram model*, *Proceedings of the 27th International Conference on Computational Linguistics*. (2017). <https://aclanthology.org/C18-1151>, 1775–1789.
- Brochhagen, T. (2020). Signalling under uncertainty: Interpretative alignment without a common prior. *The British Journal for the Philosophy of Science*, 71(2), 471–496.
- Brown, R. W., & Lenneberg, E. H. (1954). A study in language and cognition. *The Journal of Abnormal and Social Psychology*, 49(3), 454–462.
- Bullock, O. M., Colón Amill, D., Shulman, H. C., & Dixon, G. N. (2019). Jargon as a barrier to effective science communication: Evidence from metacognition. *Public Understanding of Science*, 28(7), 845–853.
- Caldwell, C. A., & Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLoS One*, 7(8), Article e43807.
- Castro, C. M., Wilson, C., Wang, F., & Schilling, D. (2007). Babel babble: Physicians' use of unclarified medical jargon with patients. *American Journal of Health Behavior*, 31(1), S85–S95.
- Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.
- Clark, H. H. (1998). Communal lexicons. In K. Malmkjær, & J. Williams (Eds.), *Context in language learning and language understanding* (pp. 63–87).
- Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1), 1–39.
- Davidson, D. (1986). A nice derangement of epitaphs. *Philosophical grounds of rationality: Intentions, categories, ends*, 4, 157–174.
- 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.
- Desikan, B. S., Hull, T., Nadler, E., Guilbeault, D., Kar, A. A., Chu, M., & Sardo, D. R. L. (2020). Comp-syn: Perceptually grounded word embeddings with color. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1744–1751).
- van Arkel, J., Woensdregt, M., Dingemans, M., & Blokpoel, M. (2020). A simple repair mechanism can alleviate computational demands of pragmatic reasoning: Simulations and complexity analysis. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 177–194). Association for Computational Linguistics.
- van Dolder, D., & van den Assem, M. J. (2018). The wisdom of the inner crowd in three large natural experiments. *Nature Human Behaviour*, 2(1), 21–26.
- Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, 8(7), 301–306.
- Fiechter, J. L., & Kornell, N. (2021). How the wisdom of crowds, and of the crowd within, are affected by expertise. *Cognitive Research: Principles and Implications*, 6(1), 1–7.
- Gass, W. H. (1975). *On being blue: A philosophical inquiry*. Boston: D. R. Godine.
- Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., & Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40), 10785–10790.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51(4), 334–384.
- Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, 103(2), 180–226.
- Guilbeault, D., Nadler, E. O., Chu, M., Lo Sardo, D. R., Kar, A. A., & Desikan, B. S. (2020). Color associations in abstract semantic domains. *Cognition*, 201, Article 104306.
- Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A., Smith, K., Griffiths, T. L., & Goodman, N. D. (2022). From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review*. In press.
- Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, 93(1), 10.
- Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, 18(10), 504–506.
- Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(4), 1068.
- Hupet, M., Seron, X., & Chantraine, Y. (1991). The effects of the codability and discriminability of the referents on the collaborative referring procedure. *British Journal of Psychology*, 82(4), 449–462.
- Hupka, R. B., Zaleski, Z., Otto, J., Reidl, L., & Tarabrina, N. V. (1997). The colors of anger, envy, fear, and jealousy: A cross-cultural study. *Journal of Cross-Cultural Psychology*, 28(2), 156–171.
- Hutchings, J. (2004). Colour in folklore and tradition—The principles. *Color Research & Application*, 29(1), 57–66.
- Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1), 26.
- Johansson, N., Anikin, A., & Aseyev, N. (2020). Color sound symbolism in natural languages. *Language and Cognition*, 12(1), 56–83.
- Johns, B. T., & Jamieson, R. K. (2018). A large-scale analysis of variance in written language. *Cognitive Science*, 42(4), 1360–1374.
- Johns, B. T., Jones, M. N., & Mewhort, D. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review*, 26(1), 103–126.
- Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1), 1.
- Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey*. CA: CSLI Publications Stanford.
- Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, 34(1), 43–68.
- Korsch, B. M., Gozzi, E. K., & Francis, V. (1968). Gaps in doctor-patient communication: I. Doctor-patient interaction and patient satisfaction. *Pediatrics*, 42(5), 855–871.
- Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, 19(4), 332–338.
- Krauss, R. M., & Fussell, S. R. (1996). Social psychological models of interpersonal communication. In E. Higgins, & A. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 655–701). New York: Guilford Press.
- Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(1–12), 113–114.
- Labov, W. (1973). *The boundaries of words and their meanings* (New ways of analyzing variation in English).
- Labrecque, L., & Milne, G. (2011). Exciting red and competent blue: The importance of color in marketing. *Journal of the Academy of Marketing Science*, 40(5), 711–727.
- Lakoff, G. (2006). *Whose freedom? The battle over America's most important idea*. New York: Farrar, Straus and Giroux.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.
- Landa, E., & Fairchild, M. (2005). Charting color from the eye of the beholder. *American Scientist*, 93, 436–443.
- Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. *Cognitive Science*, 37(5), 775–799.
- Lenneberg, E. H., & Roberts, J. M. (1956). *The language of experience: A study in methodology*. Baltimore: Waverly Press.
- Lin, S., Fortuna, J., Kulkarni, C., Stone, M., & Heer, J. (2013). Selecting semantically resonant colors for data visualization. In , 32. *Computer graphics forum* (pp. 401–410).
- Mahnke, F. H. (1996). *Color, environment, and human response: An interdisciplinary understanding of color and its use as a beneficial element in the design of the architectural environment*. New York: John Wiley & Sons.
- Marcus, A. (1982). Color: A tool for computer graphics communication. *The Computer Image*, 76–90.
- Marti, L., Piantadosi, S. T., & Kidd, C. (2019). Same words, same context, different meanings: People are unaware their own concepts are not always shared. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 2296–2302).
- Martínez, A., & Mammola, S. (2021). Specialized terminology reduces the number of citations of scientific papers. *Proceedings of the Royal Society B*, 288(1948), 20202581.
- McCabe, R., & Healey, P. G. (2018). Miscommunication in doctor–patient communication. *Topics in Cognitive Science*, 10(2), 409–424.
- McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, 6(4), 462–472.
- Meier, B. P., & Robinson, M. D. (2005). The metaphorical representation of affect. *Metaphor and Symbol*, 20(4), 239–257.
- Mohammad, S. (2011, June). Even the abstract have color: Consensus in word-colour associations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 368–373). Portland, Oregon, USA: Association for Computational Linguistics.
- Monroe, W., Hawkins, R. D., Goodman, N. D., & Potts, C. (2017). Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5, 325–338.
- Morin, O., Müller, T. F., Morisseau, T., & Winters, J. (2022). Cultural evolution of precise and agreed-upon semantic conventions in a multiplayer gaming app. *Cognitive Science*, 46(2), e13113.
- Mukherjee, K., Yin, B., Sherman, B. E., Lessard, L., & Schloss, K. B. (2022). Context matters: A theory of semantic discriminability for perceptual encoding systems. *IEEE Transactions on Visualization and Computer Graphics*, 28(1), 697–706.
- Munsell, A. H. (1905). *A color notation*. Boston: Geo. H. Ellis Company.
- Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76(1p2), 1.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). (EMNLP).
- Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, 4(2), 203–228.
- Potts, C., & Levy, R. (2015). Negotiating lexical uncertainty and speaker expertise with disjunction. In , 41. *Proceedings of the annual meeting of the Berkeley linguistics society*.
- Rathore, R., Leggon, Z., Lessard, L., & Schloss, K. B. (2019). Estimating color-concept associations from image statistics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1), 1226–1235.
- Rauhut, H., & Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology*, 55(2), 191–197.
- Reddy, M. (1979). The conduit metaphor. *Metaphor and thought*, 2, 285–324.
- Richardson, J. T. E. (1975). Concreteness and imageability. *Quarterly Journal of Experimental Psychology*, 27(2), 235–249.
- Riley, C. A. (1995). *Color codes: Modern theories of color in philosophy, painting and architecture, literature, music, and psychology*. Lebanon, NH: University Press of New England.

- Roberts, G., & Clark, R. (2018). Emergence of vowel-like organization in a color-based communication system. In C. Kalish, M. A. Rau, X. J. Zhu, & T. T. Rogers (Eds.), *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 954–959).
- Schloss, K. B. (2018). A color inference framework. In G. MacDonald, & C. Biggam (Eds.), *Progress in colour studies: Cognition, language, and beyond*. John Benjamins.
- Schloss, K. B., Leggon, Z., & Lessard, L. (2021). Semantic discriminability for visual communication. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 1022–1031.
- Schloss, K. B., Lessard, L., Walmsley, C. S., & Foley, K. (2018). Color inference in visual communication: The meaning of colors in recycling. *Cognitive Research: Principles and Implications*, 3(1), 1–17.
- Schloss, K. B., Witzel, C., & Lai, L. Y. (2020). Blue hues don't bring the blues: Questioning conventional notions of color–emotion associations. *Journal of the Optical Society of America A*, 37(5), 813–824.
- Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, 203, Article 104285.
- Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51(3), 1258–1270.
- Setlur, V., & Stone, M. C. (2015). A linguistic approach to categorical color assignment for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), 698–707.
- Sharma, G., Wu, W., & Dalal, E. N. (2005). The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1), 21–30.
- Steege, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (2014). Measuring the crowd within again: A pre-registered replication study. *Frontiers in Psychology*, 5, 786.
- Sturges, J., & Whitfield, T. A. (1995). Locating basic colours in the Munsell space. *Color Research & Application*, 20(6), 364–376.
- Tham, D. S. Y., Sowden, P. T., Grandison, A., Franklin, A., Lee, A. K. W., & Ng, M.,... Zhao, J. (2020). A systematic investigation of conceptual color associations. *Journal of Experimental Psychology: General*, 149(7), 1311.
- Van Leeuwen, T. (2011). *The language of colour: An introduction*. London: Routledge.
- Vilnis, L., & McCallum, A. (2015). Word representations via Gaussian embedding. In Y. Bengio, & Y. LeCun (Eds.), *Proceedings of the 3rd International Conference on Learning Representations*.
- Volkova, S., Dolan, W. B., & Wilson, T. (2012). CLex: A lexicon for exploring color, concept and emotion associations in language. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 306–314).
- Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7), 645–647.
- Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science*, 49(4), 400–415.
- Wilson, D., & Carston, R. (2007). A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In N. Burton-Roberts (Ed.), *Pragmatics* (pp. 230–259). New York: Palgrave Macmillan.
- Winter, B. (2019). *Sensory linguistics: Language, perception and metaphor*. Philadelphia: John Benjamins Publishing Company.
- Winters, J., & Morin, O. (2019). From context to code: Information transfer constrains the emergence of graphic codes. *Cognitive Science*, 43(3), Article e12722.
- Wodak, R. (1989). *Language, power and ideology: Studies in political discourse*. Philadelphia: John Benjamins Publishing Company.
- Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31), 7937–7942.
- Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2019). Color naming reflects both perceptual structure and communicative need. *Topics in Cognitive Science*, 11(1), 207–219.