Cognitive Psychology 159 (2025) 101745

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Lossy encoding of distributions in judgment under uncertainty☆

Tadeg Quillien [a],*, Neil Bramley [b], Christopher G. Lucas [a]

[a] School of Informatics, University of Edinburgh, United Kingdom
[b] Department of Psychology, University of Edinburgh, United Kingdom

ARTICLE INFO

Dataset link: https://osf.io/wz649

Keywords:
Computational modeling
Probability
Judgment under uncertainty
Social cognition
Information theory

ABSTRACT

People often make judgments about uncertain facts and events, for example 'Germany will win the world cup'. Judgment under uncertainty is often studied with reference to a normative ideal according to which people should make guesses that have a high probability of being correct. According to this normative ideal, you should say that Germany will win the world cup if you think that Germany is in fact likely to win. We argue that in many cases, judgment under uncertainty is instead best conceived of as an act of lossy compression, where the goal is to efficiently encode a probability distribution, rather than express the probability of a single outcome. We test formal computational models derived from our theory, showing in four experiments that they accurately predict how people make and interpret guesses. Our account naturally explains why people dislike vacuously-correct guesses (like 'Some country will win the world cup'), and sheds light on apparently sub-optimal patterns of judgment such as the conjunction fallacy.

1. Introduction

People often make judgments about uncertain facts and events. These judgments are often compared to a normative ideal derived from probability theory. For example, researchers explore how people's judgments conform to the standards of probability theory (Griffiths & Tenenbaum, 2006), and how they deviate from these standards (Kahneman et al., 1982). The challenge to explain the successes and failures of human probabilistic reasoning has given rise to a large body of productive research (e.g. Busemeyer et al., 2011; Cosmides & Tooby, 1996; Dasgupta et al., 2020; Fox & Rottenstreich, 2003; Gigerenzer, 1991; Griffiths & Tenenbaum, 2006; Johnson-Laird et al., 2015; Juslin et al., 2007; Koralus & Mascarenhas, 2013; Krynski & Tenenbaum, 2007; Oaksford & Chater, 2007; Quillien et al., 2023; Tversky & Kahneman, 1983; Zhu et al., 2020). This comparison to a normative standard is in line with the more general idea in cognitive science that it is useful to think about the abstract structure of the information-processing problems the mind is solving (Anderson, 1990; Cosmides & Tooby, 1994; Marr, 1982).

Work on judgment under uncertainty typically uses as a normative standard that we will call the *Probability-maximizing* criterion: you should endorse a statement in proportion to the probability that the statement is correct. Under *Probability-maximizing*, you should endorse 'Germany will win the world cup' if you think that Germany is likely to win the world cup. Consider for example the research on 'conjunction errors' (Tversky & Kahneman, 1983). According to *Probability-maximizing*, it is a mistake to endorse the statement 'Linda is a bank teller and is active in the feminist movement' to a greater extent than 'Linda is a bank teller', because the former statement cannot have higher probability of being true.

☆ The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article. All data, modeling and analysis scripts have been made available on the Open Science Framework at https://osf.io/wz649.

* Correspondence to: Department of Psychology, University of Edinburgh, United Kingdom.
E-mail address: tadeg.quillien@gmail.com (T. Quillien).

https://doi.org/10.1016/j.cogpsych.2025.101745
Received 23 October 2024; Received in revised form 24 May 2025; Accepted 29 May 2025

Available online 24 June 2025

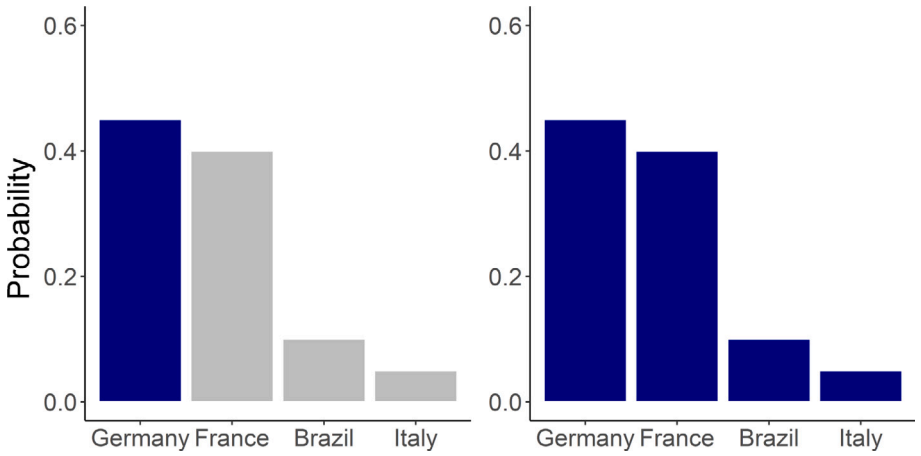

Fig. 1. Difference between the probability-maximizing standard (left) and our view (right). Under the probability-maximizing standard, the guess 'Germany will win' only encodes information about the probability that Germany will win. Under our account, it encodes information about the whole probability distribution.

In this paper, we suggest that there is another, complementary normative standard against which judgment under uncertainty can be evaluated. According to this standard, a statement about an uncertain fact or outcome (a 'guess') conveys information about a *probability distribution*, rather than the probability of a single outcome. In a sense that we will make precise later, a guess consists in a compressed encoding of a distribution.

The two different standards (*Probability-maximizing* and *Distribution-encoding*) are complementary in the sense that each specifies a different information-processing problem that the mind is faced with. *Probability-maximizing* is the right standard in cases like betting: when someone asks you which horse you would like to bet on, then all else being equal you should nominate the horse you think is more likely to win. *Distribution-encoding* is appropriate in cases where you need to efficiently encode the contents of your probabilistic beliefs; for example when you need to communicate about more than the probability of a single outcome.

It is helpful to make an analogy with the way scientists use probability. Sometimes scientists are looking to compute a particular probability, for example when they compute a *p*-value, or their statistical power to detect an effect. Other times they are trying to compute a summary of a distribution. For example when a researcher reports the mean and standard deviation of a measured variable, she is reporting a summary that captures some (incomplete) information about the empirical distribution of that variable. Or, when she computes the correlation between two variables, she is reporting a partial summary of the joint distribution over these variables. Clearly, both types of practice (computing a particular probability, and computing summaries of a distribution) are essential to the way scientists process information. By analogy, it makes sense that each should also play a role in our cognitive lives.

1.1. Scope of the work

We do not aim to provide an account of all phenomena in probabilistic reasoning. In many psychological tasks, people's judgments are already well-captured by assuming that people are approximately solving the *Probability-maximizing* problem, perhaps subject to noise, cognitive limitations, or sub-optimal processing (Beck et al., 2012; Costello & Watts, 2014, 2018; Griffiths & Tenenbaum, 2006; Zhu et al., 2020). We are interested in a different set of cases, where we think that people face the *Distribution-encoding* problem. A paradigmatic example would be replying to your friend who just asked who you think will win the next world cup. Arguably the problem your mind is facing in that situation is that of giving your friend an idea of your subjective probability distribution over the possible winners.

In such cases, *Distribution-encoding* makes sense of many intuitive judgments that are difficult to account for with *Probability-maximizing*. Suppose that your beliefs about who will win the next world cup are as in Fig. 1: you think that Germany, France, Brazil and Italy have 45%, 40%, 10% and 5% probabilities of winning, respectively. It seems that 'Germany or Italy will win' is somehow a worse guess than 'Germany will win'. Under the *Probability-maximizing* standard, this intuition is a mistake, since $Pr(\text{Germany} \vee \text{Italy}) > Pr(\text{Germany})$. More generally, the *Probability-maximizing* standard runs into the problem that it seems to entail *Vacuous-guessing*: the thesis that we should strive to always prefer statements with probability 1, like 'any country might win the world cup' (Holguin, 2022; Yaniv & Foster, 1995).

Distribution-encoding explains the intuitive lack of appeal of *Vacuous guessing*. If a guess functions as an attempt to efficiently summarize your distribution over possible outcomes, then vacuous guesses will typically be bad. 'Any country might win' misleadingly implies that you think every country has about the same chance of winning. Similarly, 'Germany or Italy will win' misleadingly implies that you think Italy has better chances than France.

In sum, our main goal is to account for people's judgments in situations where *Distribution-encoding* is plausibly the appropriate normative standard. However, we suggest that our perspective may also make sense of some of the mistakes – like conjunction

errors – that people make in tasks where they are asked about the probability of a single outcome. Researchers sometimes explain these mistakes by arguing that participants are replacing the question they are being asked with a different question (Kahneman & Frederick, 2002). In this spirit, we suggest that mistakes in probabilistic reasoning may often arise because people respond as if they were solving the *Distribution-encoding* problem even in tasks where they are explicitly asked for *Probability-maximizing*.

For example, guessing that Linda is 'a bank-teller involved in the feminist movement' can better convey your subjective distribution over Linda's likely features (depending on what you know about her) than guessing she is 'a bank teller'. We develop this argument in more detail in the General Discussion (see also Dorst & Mandelkern, 2021). There we also argue that *Distribution-encoding* is consistent with the related Disjunction fallacy (Bar-Hillel & Neter, 1993), with over-confidence in interval estimation (Juslin et al., 1999), and with people's intuitive conception of surprise and likelihood (Teigen et al., 2022; Teigen & Keren, 2003).

1.2. Relation to other perspectives

Our thesis is that *Distribution-encoding* can capture many facts about human probabilistic reasoning from an abstract, computational-level perspective (Anderson, 1990; Cosmides & Tooby, 1994; Marr, 1982). This view is compatible with other perspectives. In particular, the intuition that people avoid making misleading or uninformative guesses can be expressed in terms of concepts from linguistics, in particular in terms of *pragmatics*. People might avoid making vacuous guesses in order to conform to Gricean maxims of conversation, or to maximize relevance (Grice, 1975; Sperber & Wilson, 1986). A similar argument can be formulated in terms of *rational speech acts*, where the speaker optimizes the inference that the listener will derive from an utterance (Degen, 2023; Goodman & Frank, 2016; Shafto et al., 2014). A related idea is that people care about a statement's *specificity*, preferring to endorse guesses that mention a relatively small subset of the possible outcomes, because specificity is a useful proxy for informativeness. Under this proposal, judgment under uncertainty is governed by a trade-off between probability-maximizing and specificity (Dorst & Mandelkern, 2021; Yaniv & Foster, 1995).

We suggest that *Distribution-encoding* extracts, at an abstract level, the underlying logic that makes these proposals work. Consider for example the proposal that people favor specific guesses because they are a useful proxy for informativeness (Dorst & Mandelkern, 2021; Yaniv & Foster, 1995). We think that this intuition can be naturally explained in information-theoretic terms: more specific guesses typically provide better summaries of the relevant probability distribution.

In what follows we formally define our proposal, and report the results of experimental studies testing its fit to people's intuitions. A terminological note: we use the word 'guess' as a shortcut for 'statement about an uncertain fact or event', following its use in previous literature (Dorst & Mandelkern, 2021; Holguin, 2022). Our use of the term differs somewhat from everyday usage: for example, some statements count as 'guesses' under our use even if the speaker has very high confidence in them.

1.3. Efficient summaries of probability distributions

People often represent a subjective probability distribution over a set of relevant possible outcomes, or possible states of affairs.[1] You might for example think that among the possible answers to the question 'how many member states are in the European Union?', some possible answers (like 25) are more likely than others (2, or 60), even if you may not be able to explicitly verbalize your probability estimates.[2] This idea has been successful in many areas of cognition, from perception to high-level cognition and semantics (Fleming et al., 2012; Griffiths & Tenenbaum, 2006; Knill & Richards, 1996; Lassiter, 2011; Oaksford & Chater, 2007; Tenenbaum et al., 2011; Vul & Pashler, 2008).

Creating efficient summaries of a probability distribution is probably an important information-processing problem for the mind. For example, someone might ask you how many states you think are in the European Union, and you have to craft a verbal summary of your subjective distribution over possible answers, like 'between 20 and 30'. Or, you might want to remember your subjective distribution while minimizing the cost of storing this information in memory.

On this view, a guess can function as a lossy compression of an underlying probability distribution. We can make this proposal precise by using ideas from information theory (Gagie, 2006; MacKay, 2003; Sims, 2016). One can think of lossy compression as a process where an input (in our case, an agent's subjective probability distribution over relevant possible outcomes) is *encoded* in a compressed form, which can then be read by a *decoder* (see Fig. 2). The faithfulness of the encoding can be quantified as the extent to which the decoded output diverges from the original input (Berger, 2003; Sims, 2016).

Formally, we consider a reasoner that has a distribution P over some outcome of interest, and needs to create a summary representation g as a compressed version of this distribution. We assume the existence of an optimal decoder that can decode the representation g to re-construct a distribution Q_g. The encoding scheme should minimize the extent to which Q_g diverges from P,

[1] In line with Bayesian analyses of probability, a speaker's subjective probability distribution represents the degree of credence the speaker assigns to various possible outcomes or states of affairs (Jaynes, 2003). Subjective distributions can be shaped by statistical frequency, as when we reason that team A and B are equally likely to win a football game because each team has won 10 of their previous 20 encounters. Subjective distributions can also be shaped by many other factors, such as prior knowledge or causal reasoning; we may for instance adjust our probability estimate that team A will win the game if we learn that the best player in team B was injured.

[2] We are not committed to the idea that people always *explicitly* represent probability distributions, or that they do so in a perfectly coherent way. In many cases people may represent probabilities only in an implicit format, for instance within a generative model from which they draw samples (Chater & Oaksford, 2013; Icard, 2016; Vul et al., 2014). In these cases our proposal is that speakers make guesses that aim to communicate the probability distribution latent in the generative model.

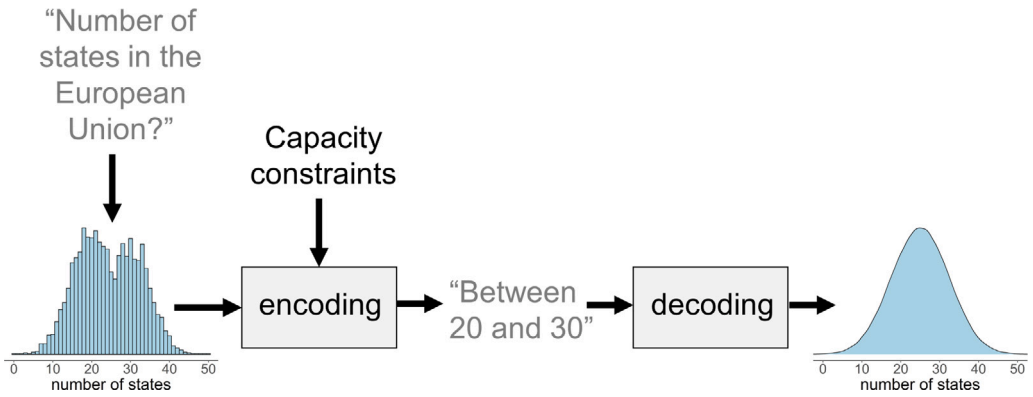

Fig. 2. Abstract characterization of our framework. The agent has a subjective probability distribution *(left)* over possible answers to a question *(top left)*. The guess (middle) provides a compressed encoding, which can later be decoded to yield an approximate re-construction of the original distribution (right). For a model of how people interpret interval estimates in particular, see Study 4.

subject to various constraints. We can define the faithfulness of Q_g as an approximate reconstruction of P by using a measure D of the 'divergence' between distributions Q_g and P. In our analyses we will use the Kullback–Leibler divergence of Q_g from P (Gagie, 2006; Kullback & Leibler, 1951)[3]:

$$D_{\mathrm{KL}}(P\|Q_g) = \int_i P(i) \log\left(\frac{P(i)}{Q_g(i)}\right) \tag{1}$$

The KL-divergence is widely used as a measure of the faithfulness of a decoded probability distribution (Gagie, 2006; Tishby et al., 2000). Note however that the measure is not well-behaved in some contexts where $Q_g(i) = 0$; in these contexts one could use more robust divergence measures.[4]

The value of the representation g (the 'guess') is inversely related to the divergence of Q_g from P. Of course, in the absence of constraints the best g is trivially just an identical copy of P (which yields $D_{\mathrm{KL}}(P\|Q_g) = 0$). There might however be constraints that force the reasoner to create a more coarse-grained representation. These constraints could in principle be formalized as a bound on the quantity of information that the reasoner can extract from P when creating g, following work in rate–distortion theory and related methods (Berger, 2003; Sims, 2016; Tishby et al., 2000). In the present work, we found that it was easier to make experimental predictions if we instead consider constraints on the possible guesses that the reasoner can make. That is, we assume that the reasoner has access to a menu G of possible representations, and has to choose the representation $g \in G$ that minimizes the divergence between Q_g and P. Formally, the reasoner is looking for the representation $g^\star$ such that:

$$g^\star = \arg\min_{g \in G} D_{\mathrm{KL}}(P\|Q_g) \tag{2}$$

In principle this framework is very general and can be applied in several different domains. Guessing might for example be an intra-personal process meant to compress information in memory, where the decoding stage corresponds to later memory retrieval (see e.g. Gershman, 2021). For concreteness in this paper we will often use as an example the case of verbal communication. In that context a *speaker* utters a guess g with the intent to get a *listener* to re-construct a good estimate Q_g of the speaker's probability distribution P. Although ideally the speaker would simply enumerate his full subjective distribution, several constraints prevent him from doing so, like limited time, computation or conceptual knowledge.

In the context of verbal communication, our framework implies some relatively strong, idealized assumptions. First, the speaker knows the inference Q_g that the listener will draw. Second, the listener's inference is approximately optimal. Third, the speaker and listener have a common understanding of the set of relevant possibilities over which the distribution is defined. There is of course a lot that could be say about the mechanisms that make these assumptions approximately appropriate (e.g. conventions, pragmatic inferences) and about when they might fail. Our purpose however is to use communication as an example illustrating our more abstract framework, not to provide a complete account of verbal communication of probabilistic beliefs (see also General Discussion). Our idealized assumptions may not always hold exactly, but we submit they are often close enough to help us capture key phenomena in judgment under uncertainty. The next section illustrates our account in a simple setting.

[3] One intuition for the use of the Kullback–Leibler divergence is the following. The $\log\left(\frac{P(i)}{Q_g(i)}\right)$ term measures the difference in the surprise (technically, the 'surprisal') experienced by an agent using P and agent using Q_g, when observing outcome i. The agent with access to P would like to minimize the expected value of this difference; he thinks that outcome i will occur with probability $P(i)$, and therefore wants to minimize $\int_i P(i) \log\left(\frac{P(i)}{Q_g(i)}\right)$. See for example Egré et al. (2023).

[4] We thank an anonymous reviewer for highlighting this point. In exploratory analyses we have verified that our model yields almost identical predictions when using one such measure, the Jensen–Shannon divergence. The Jensen–Shannon divergence is computed by first defining a 'mixture' distribution $M = (P + Q_g)/2$, and then computing the average of the KL divergences of M from P and from Q_g.

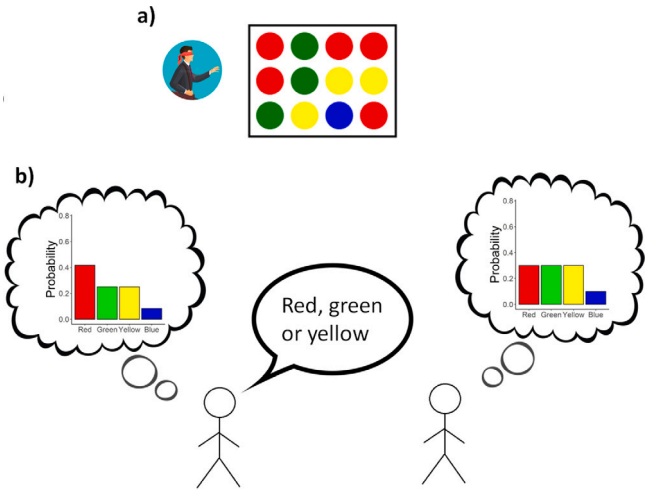

Fig. 3. **a**: Someone will randomly draw a ball from the box; the speaker must guess which color will come out. **b**: The speaker communicates his subjective probability distribution over the outcome of the draw (left) by making a guess (middle). The listener infers a probability distribution from the guess (right).

1.4. Case study: disjunctive guesses

Consider the box in Fig. 3, containing balls of different colors. If someone randomly draws a ball from the box, which color will it be? We study the case where the speaker is only allowed to make *disjunctive* guesses, for example 'Red', 'Blue or Green', 'Yellow or Red or Green', etc. His goal is to communicate his subjective distribution over the outcome of the draw (i.e. for each color, what is the probability that the ball drawn from the urn will be of that color?) to a listener. The listener knows that the box contains red, yellow, blue and green balls, but cannot see inside the box and so does not know the exact proportions of each color.

If you tell someone that the ball will be "red, green, or yellow", what can they infer? Remember that our general framework is not committed to a particular model of how the listener infers the probability distribution Q_g, but in the current setting we can obtain a relatively simple model of that inference. Specifically, the listener can infer that red, yellow, and green are more probable outcomes than blue, but she has no reason to think that any of the three colors (red, yellow, green) is more likely than the others. So, her best bet is to construct a probability distribution that looks like the one in Fig. 3b (right). As another example, if you tell her "it will be a red ball", her best bet is to infer a probability distribution over outcomes that looks like the one in Fig. 4b. More generally, the principle of indifference (Jaynes, 2003; Laplace, 1820) implies that listeners should infer that outcomes mentioned in the guess have equal probability, and have higher probability than outcomes not mentioned in the guess (see the methods section of Study 1 for details).[5]

The speaker makes a guess g that he expects will result in a low divergence between the distribution Q_g inferred by the listener and his own distribution P. For example, the guess 'Red, Green or Yellow' is a good candidate guess because the distribution over possible outcomes inferred by the listener (Fig. 3, right) does not diverge too much from the speaker's distribution (Fig. 3, left).

Our account delivers the intuitive result that the vacuous guess 'Red, Green, Yellow or Blue' is not necessarily the best, despite having probability 1. Including all possible outcomes in the guess would cause the listener to infer a flat distribution, which might be quite unlike the speaker's subjective distribution. We can also explain the intuition that good guesses respect 'clustering', in the sense that when two outcomes A and B have a similar probability, they should either be both included in the guess or both left out (Dorst & Mandelkern, 2021). If an urn has 4 red, 4 green, 2 yellow and 2 blue balls, for example, it seems strange to guess that the ball will be 'Red, Green or Yellow'. Guesses that include outcome A but not outcome B imply a distribution where A is much more likely than B, and this distribution is a bad approximation of the speaker's distribution if the speaker takes A and B to be equally likely.

We can also account for the fact that whether a guess is good seems to be relative to the question at hand. Consider again the urn in Fig. 3a, with 5 red balls out of 12. If someone asks 'which color will come out?', 'red' seems a good guess. But if someone asks 'will the ball be red or non-red?', 'non-red' seems better. The question 'red or non-red?' implies that the relevant probability distribution has support $\{\text{Red}, \neg\text{Red}\}$, such that the guess 'red' would misleadingly imply that $Pr(\text{Red}) > Pr(\neg\text{Red})$. In some contexts

[5] Here we are modeling how the listener constructs her probability distribution over possible outcomes of the random draw from the urn. That is, the listener assigns some probability to the random draw yielding a red ball, to the random draw yielding a blue ball, etc. Note that the listener might also have a probability distribution over the contents of the urn. That is, the listener might assign some probability to the possibility that the contents of the urn are [9 red balls, 1 green ball, 1 yellow ball, 1 blue ball], some probability to the possibility that the contents are [8 red balls, 2 green balls, 1 yellow ball, 1 blue ball], etc. The relationship between these two distributions is that if the listener is coherent, she should be able to derive the first probability distribution (over possible outcomes of the draw) from the second probability distribution (over possible contents of the urn), by marginalization. We explore this marginalization as a (slightly more complex) model of the listener's inference in the context of the RSA model we present in Appendix.

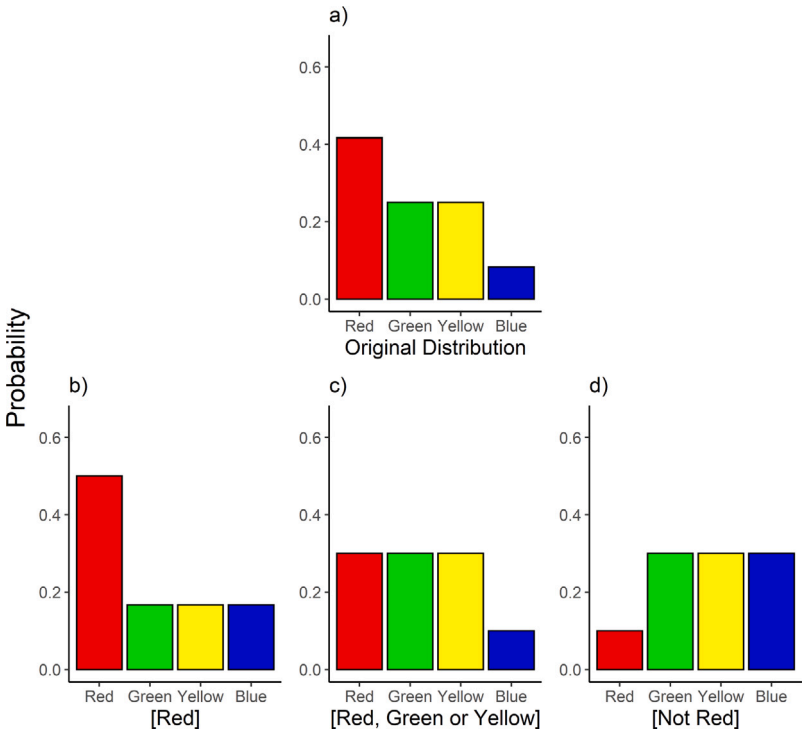

Fig. 4. **a**: Speaker's subjective probability distribution over possible outcomes. **b, c, d**: distributions over possible outcomes consistent with hearing the guesses "it will be a red ball", "it will be a red, yellow or green ball", "it won't be a red ball'. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

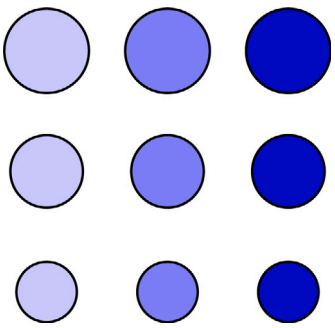

Fig. 5. Shapes varying along a similarity gradient.

the question under discussion may be implicit; we do not directly address how people infer it from the context, but we note that existing work in linguistics explores the issue (e.g. Roberts, 2012).

Note that the model of the listener's inference we gave in this section is specific to the example of our balls-and-urns setting. In general, the inference drawn by a listener will depend on the specifics of the task. Consider for example a task where you see a shape briefly displayed on a screen, and you have to guess which shape it was, by picking shapes from an array like the one in Fig. 5. Here the different shapes vary along a similarity gradient (color and size), so a listener hearing your guess should infer a probability distribution where the probability that a shape appeared on the screen is a function of its similarity to the shapes selected by the guesser.

1.5. Accuracy and specificity

Earlier we mentioned a theory according to which people like guesses that balance a trade-off between accuracy and specificity. According to this account, a good guess is likely to be correct but does not mention too many outcomes (Dorst & Mandelkern, 2021;

Goldsmith et al., 2002; Skipper, 2023; Yaniv & Foster, 1995).[6] The experiments we report below also give us the opportunity to test the predictions of this theory.

As we said earlier, the trade-off hypothesis is in principle consistent with our own hypothesis. One can interpret the trade-off model as a more descriptive, process-level theory than our own computational-level analysis. Under this interpretation, the two accounts sit at different levels of analysis and so are not in competition with each other. It might for example be that guesses function to encode a speaker's subjective probability distribution, but that people make guesses that are both likely and specific because this is a good enough heuristic to fulfill that function.

In general, the trade-off hypothesis makes very similar experimental predictions as our account, and therefore our experiments are not primarily designed to arbitrate between the two accounts. We wish however to highlight one interesting way that their predictions diverge. Consider the urn in Fig. 3 (with 5 red, 3 green, 3 yellow and 1 blue balls). When asked what color will come out, it might seem natural to say 'Red', and maybe also to say 'Red, Green or Yellow'. The guess 'Red or Green' might seem less natural. This pattern of intuition can be described as a U-shape in the relationship between the size of a guess (how many possible outcomes it mentions) and its quality: there is a guess of size 2 that seems less natural than both a size-1 and a size-3 guess.

It is of course an empirical question whether people's judgments actually display this sort of pattern. But before looking at the data, it is interesting to ask whether a given account predicts that such a U-shaped pattern is possible. In the Appendix we prove that the trade-off model by Dorst and Mandelkern (2021) predicts that people's judgments will never (except for noisy responding) exhibit such a U-shaped relationship between guess size and guess quality. Intuitively, a trade-off analysis of guess quality holds that, if someone prefers to say 'Red' instead of 'Red or Green', this means that they place a high weight on specificity relative to accuracy. Since the guess 'Red, Green or Yellow' is even less specific than 'Red or Green', the speaker is bound to prefer 'Red or Green' to 'Red, Green or Yellow'. In other words if someone prefers a size-1 guess to a size-2 guess then they will necessarily prefer the size-2 guess to a size-3 guess. By contrast, our information-theoretic account predicts that U-shaped patterns will be relatively common. Intuitively, the guess 'Red or Green' has the misleading implication that Green balls are more frequent than Yellow balls—so the speaker might convey a more accurate depiction of his subjective probability distribution by saying either 'Red' or 'Red, Green or Yellow' instead.

1.6. Overview of empirical tests

In what follows we report four experiments that test the quantitative predictions of computational models that implement our theory. Our tasks typically have no right or wrong answers; we ask participants to make guesses about an uncertain fact or outcome, or to judge the quality of possible guesses one could make, for example.

Studies 1 to 3 use the paradigm described above (disjunctive guesses about ball colors); in Study 1 participants rate the quality of different guesses one could make, in Study 2 participants compose their own guesses, and in Study 3 participants infer the contents of an urn on the basis of someone else's guess. In Study 4, we study how people evaluate the quality of a guess about a continuous quantity, when they know the correct answer. Data and R code (for modeling and data analysis) are available for all studies on the Open Science Framework at https://osf.io/wz649.

2. Study 1

In two studies (1a and 1b), we test our account in the context (described above) of disjunctive guesses in a simple urn scenario. Participants were shown urns containing balls of different colors (as in Fig. 3a), whose content we systematically varied in a within-subject design. We asked participants to rate the quality of different guesses that one could make about the outcome of a random draw from the urn. We compared their ratings with the predictions of our information-theoretic model (henceforth, compression model), the accuracy/specificity trade-off model, and a simple probability-maximizing model.

2.1. Materials and measures

Participants saw urns containing 12 balls of different colors (Red, Yellow, Blue, Green; there was at least one ball of each color in each urn). For each urn, we asked participants to rate the quality of four guesses about the outcome of a random draw from the urn, on a Likert scale from 1 (bad guess) to 9 (good guess). The guesses were of the form "The player will draw {·}", where {·} was a disjunction of possible colors (e.g. "a red ball or a yellow ball"). We call the number of colors in {·} the *size* of a guess. For example, {Red or Yellow} is a guess of size 2.

We constructed four guesses, of sizes 1, 2, 3 and 4, per urn, by first building a guess with the most frequent color, then a guess with the two most frequent colors, etc. For example, for the urn shown in Fig. 3a, we constructed the guesses {Red}, {Red or Yellow}, {Red, Yellow or Green} and {Red, Yellow, Green or Blue} (In cases where some colors have equal frequency we randomly imposed an artificial ordering on them when constructing guesses). All guesses for a given urn were presented alongside the urn on a single page, and the order of presentation of the guesses on the page was randomized. Different urns were presented on different pages, and the order of presentation of urns was randomized. No feedback was given.

[6] Some authors (Dorst & Mandelkern, 2021; Yaniv & Foster, 1995) use the term 'informative'. We use 'specific' following Skipper (2023), who points out that it is a more neutral and less theoretically loaded word.

We define the 'profile' of an urn as a list of four numbers, specifying the number of balls of the most frequent color, the number of balls of the second most frequent color, and so on. For example, the urn in Fig. 3a has profile [5,3,3,1]. We used 13 different profiles in Study 1a, and 10 in Study 1b. All participants saw one urn for each profile. The content of the urns was procedurally generated for each participant, by first randomly sampling one profile (without replacement), then randomly sampling a frequency ordering over colors, and randomizing the position of the balls in the urn.

2.2. Procedure

Participants were recruited on Prolific and completed the experiment on a web-based interface. We first asked participants to familiarize themselves with the setting by randomly drawing a few times from two different urns. Then they read a short set of instructions explaining the task. In the main phase of the study, participants rated the quality of four guesses per urn—each page featured a picture of a different urn, alongside four different guesses to rate. Participants then completed a short set of questions probing whether they understand how probability works in the current context (we do not analyze these reports here). Finally, they completed a few demographic questions and were redirected to Prolific for payment.

Studies 1a and 1b had essentially identical designs, with the following exceptions. Study 1b was shorter, with 10 instead of 13 different urns per participant. For exploratory purposes, we also varied whether the instructions framed the task as explicitly involving communication. In Study 1a we simply told participants that they were about to rate different possible guesses, while in Study 1b we asked them to imagine that they would be communicating with a friend who cannot see the contents of the box (but knows that boxes contain red, blue, green and yellow balls, in unknown proportion). Likert scales were labeled with 'bad guess' and 'good guess' in Study 1a, and 'bad answer' and 'good answer' in Study 1b. Interested readers can walk through the experiments at https://eco.ppls.ed.ac.uk/~tquillie/guesses/ and https://eco.ppls.ed.ac.uk/~tquillie/guesses-b/.

2.3. Participants

We recruited US residents from Prolific (in Study 1a, N=38, 24 female, 13 male, 1 other, mean age = 30.8, SD = 9.5; in Study 1b, N=39, 24 female, 14 male, 1 other, mean age = 30.7, SD = 9.4). Participants were compensated £1 for their participation (median completion time was about 8 min) and participation was restricted to Prolific users with a 90+% approval rate.

2.4. Computational modeling

2.4.1. Compression model

We defined our general framework for the compression model in the introduction. Here we specify the decoding function in the current setting. According to the principle of indifference (Jaynes, 2003; Laplace, 1820), an agent should assign the same probability to two outcomes if there is no reason to see one of them as more likely.[7] As such, the decoder infers that outcomes mentioned in the guess are more likely than un-mentioned outcomes, but makes no distinctions otherwise. So, the decoder infers that mentioned outcomes are γ times as likely as un-mentioned outcomes, where $\gamma > 1$ is a free parameter.[8] The guess 'Red or Green', for example, implies that Red and Green are each γ times as likely as Blue and Yellow.

Denote the probability of an outcome not mentioned in the guess as p. Then the probability of an outcome mentioned in the guess is γp. It follows that the decoder infers the following distribution Q_g:

$$Q_g(i) = \begin{cases} \frac{1}{n_{\neg g} + \gamma n_g} & \text{if } g(i) = 0 \\ \frac{\gamma}{n_{\neg g} + \gamma n_g} & \text{if } g(i) = 1 \end{cases} \tag{3}$$

where $g(i)$ denotes whether outcome i is mentioned in the guess, n_g is the number of outcomes mentioned in the guess, and $n_{\neg g}$ is the number of outcomes not mentioned in the guess (see Appendix for proof). For example, for $\gamma = 4$, the guess 'Red or Green' translates to $Q_g(\text{Red}) = Q_g(\text{Green}) = .4$, and $Q_g(\text{Blue}) = Q_g(\text{Yellow}) = .1$.

We also allow for the original probability distribution P to deviate from the normative distribution, for example because of perceptual or representational noise. Formally, we assume that the distribution P' from which the guess is constructed might be more spread out or more concentrated than the normative probability distribution P. We construct P' by applying the following transformation to each element i of P:

$$P'(i) = \frac{P(i)^\alpha}{Z} \tag{4}$$

where Z is a normalizing constant ensuring that all elements in P' sum to 1, and α is a free parameter which controls to what extent the distribution gets concentrated or spread out. For values of $\alpha < 1$, the probability distribution gets spread out; for $\alpha > 1$, it

[7] In principle the decoder might take into account the order in which colors are mentioned. We leave this possibility aside to keep the model simple.

[8] We assume that the decoder has access to basic information about the setup, i.e., that there are 12 balls in the box, that they can be red, yellow, green and blue, but does not know in which proportions these colors are represented. For simplicity, we consider settings where the decoder knows what combinations of colors are possible, but our approach is in principle compatible with situations where that is not the case.

gets concentrated (areas with a lot of probability mass get even more probability mass to the detriment of other areas). Low values of α result in guesses that mention more possible outcomes.

We can then compute the value of a guess as inversely related to the KL divergence of Q_g from P' (we add 1 to the denominator to keep the measure between 0 and 1):

$$V(g) = \frac{1}{1 + D_{\mathrm{KL}}(P' \| Q_g)} \tag{5}$$

2.4.2. Accuracy-specificity trade-off model

To implement the accuracy-specificity model, we used the equation provided in Dorst and Mandelkern (2021), where the value of a guess is:

$$V(g) = P'(g) J^{S(g)} \tag{6}$$

where $P'(g)$ is the *accuracy* of a guess, i.e., the probability that the guess is correct, $S(g) = n_{\neg g}/(n_{\neg g} + n_g)$ is its *specificity*: the proportion of possible outcomes (here, of possible colors) that it does not include. $J \geq 1$ is a free parameter that regulates how sensitive people are to specificity relative to accuracy (for $J = 1$ the speaker only cares about accuracy; higher values of J correspond to a higher weight for specificity).[9]

Finally, we also consider a naive model that simply computes the value of a guess as its probability. As for the compression model, when computing predictions for the trade-off and the probability-maximizing models we allow for the possibility that speakers use a slightly distorted distribution P', modulated by a free parameter α.[10]

2.4.3. Model evaluation

We fit each model both at the individual- and at the group level, by finding the parameter values that maximize the log-likelihood of the data under the model. We compute model fit using the AIC, a measure of model fit that penalizes overly complex models. For robustness we also perform additional individual-level model comparisons by computing Bayes Factors derived from marginal likelihoods.[11] To compute the likelihood, we assumed that each human rating is drawn from a truncated-discretized normal distribution with standard deviation σ and mean $1 + 8m^s$, where m is the model prediction (σ and s are free parameters we fit to the data).[12]

2.5. Results

Figs. 6 and 7 show the average ratings, along with model predictions for the compression and probability-maximizing model (fit at the group level), for Study 1a and 1b. Overall, participants' mean ratings for a guess tend to closely track the probability of that guess. As such, the probability-maximizing model has the best fit to the data at the group level, see Table 1. Ratings do not track probability *perfectly*, however. Consider for example the urn with profile [9,1,1,1], with nine balls of one color and one ball each for the other colors: participants rate a guess of size 1 (that mentions only the most frequent color, and has probability 9/12) as better than a guess of size 3 (with probability 11/12).

Looking at the data at the individual level reveals a richer picture: different participants appeared to use different strategies (see Fig. 9). While many participants were probability-maximizers, a substantial number of participants exhibit a more subtle pattern of judgments. Formally, about half of participants are best-fit by the naive probability-maximizing model, while among the remaining participants, about two-thirds are best-fit by the compression model (see Table 1). Fig. 9 displays the individual-level correlations between model and participant predictions, showing that the probability-maximizing model provides quite a bad fit for some participants.

The first set of participants rated the quality of a guess mostly on the basis of its probability. They gave highest ratings to guesses that mention all possible outcomes and therefore have probability 1. Fig. 8 (left panel) shows the ratings made by one such participant (in Study 1b). These participants are not easily accounted for by our compression model, because the probability of a guess is only an imperfect proxy of its quality as an encoding of the underlying distribution. In contrast, the probability-maximizing model (as well as the trade-off model, which reduces to the probability-maximizing model when setting $J = 1$) naturally explains

[9] Dorst and Mandelkern (2021) also suggest that people might use a different value of J across different questions. Briefly, their idea is that people might prefer to use a value of J that maximizes 'distinctiveness', i.e. the ratio between the value of the best guess and the value of the second best guess for the question at hand. In exploratory analyses, we implemented a variant of the model incorporating this idea (following footnote 22 in the original paper), but found that it did not improve model fit—optimizing distinctiveness tends to result in high values of J, predicting that people will have a strong preference for succinct guesses (R code for this analysis is available on our OSF page). Therefore in all the analyses we report here we simply assume that a participant uses the same value of J across urns.

[10] While the α parameter was not present in Dorst & Mandelkern's proposal, we find that including it improves model fit somewhat, even after accounting for the extra complexity.

[11] We computed the marginal likelihood by Monte Carlo simulation, taking 10^4 samples per participant and per model. We sampled model parameters from weakly informative priors, sampling α from an exponential distribution with rate 1/2, and s from an exponential distribution with rate 1. We sampled γ and J as $1 + V$, where V is exponentially distributed with rate 1/4, and σ as $.5 + W$, where W is exponentially distributed with rate 2. Bayes Factors are computed as $e^{LL_{model1}}/e^{LL_{model2}}$, where LL_{modelX} is the marginal log-likelihood for model X.

[12] The transformation $f(m) = 1 + 8m^s$ maps model predictions onto the 1–9 scale used by participants (see e.g. Griffiths & Tenenbaum, 2005). The trade-off model predictions are unbounded, so we first re-scale these predictions to the interval $[0, 1]$ before applying the transformation.

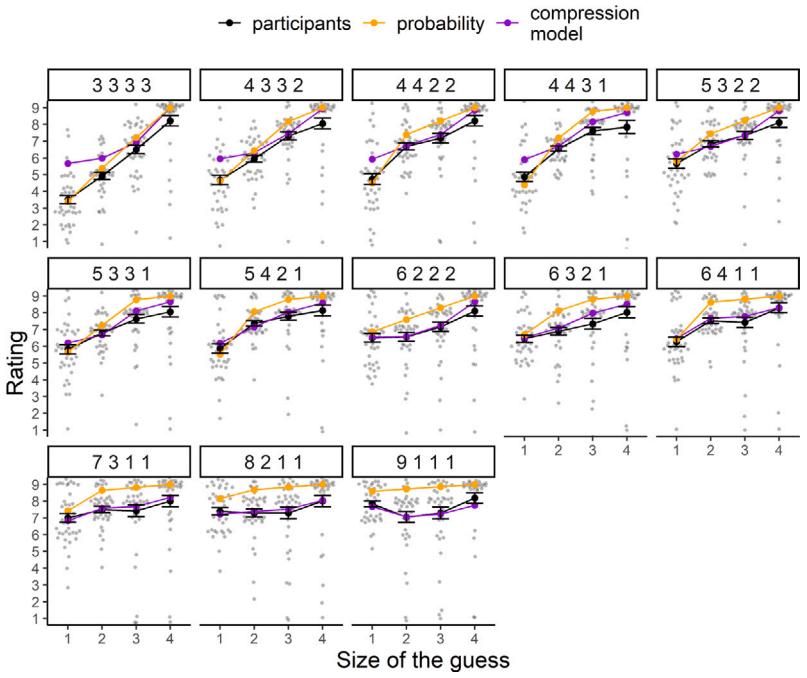

Fig. 6. Study 1a: Average participant ratings (black), along with predictions of the compression (purple) and probability-maximizing model (orange). The trade-off model makes the exact same predictions as the probability model here (best-fitting value of J at the group-level is $J = 1$). Error bars represent the standard error of the mean. Gray dots display individual ratings (jittered for visibility). Panel labels represent the profile of an urn: for example, an urn labeled [9,1,1,1] has 9 balls of one color, and one ball each of the other colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

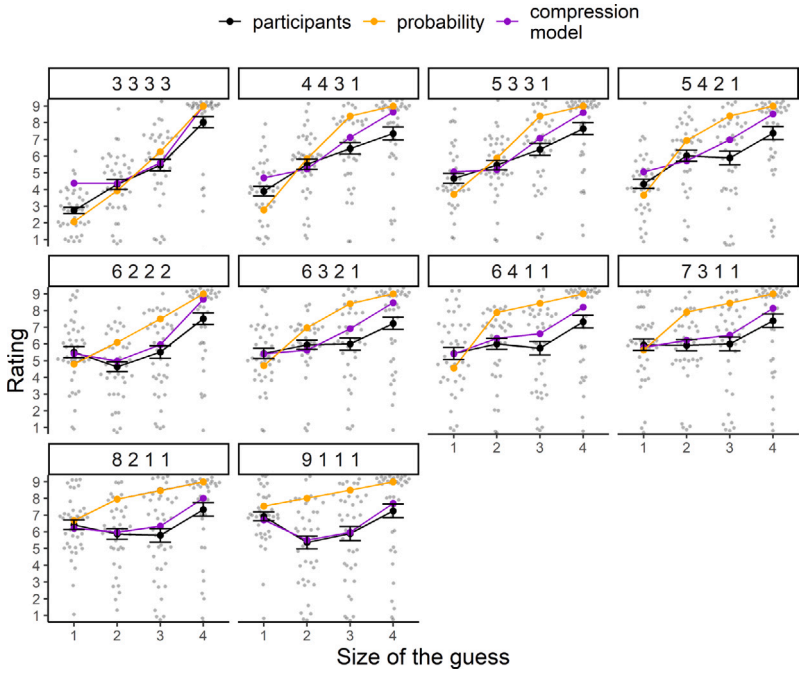

Fig. 7. Study 1b: Average participant ratings (black), along with predictions of the compression (purple) and probability-maximizing model (orange). The trade-off model makes the exact same predictions as the probability model here (best-fitting value of J at the group-level is $J = 1$). Error bars represent the standard error of the mean. Gray dots display individual ratings (jittered for visibility). Panel labels represent the profile of an urn: for example, an urn labeled [9,1,1,1] has 9 balls of one color, and one ball each of the other colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Model Fit, Study 1a and 1b. AIC: Akaike Information Criterion (lower values indicate better fit). *n* best fit: number of participants best fit by each model. We classify a participant as best fit by a model if that model has the lowest AIC, or if the model's Bayes Factor is larger than 3 against all other models.

Model	AIC (group-level fit)	*n* best fit (AIC)	*n* best fit (BF)
compression (Study 1a)	7059	11	8
trade-off (Study 1a)	6886	6	5
probability (Study 1a)	6884	21	22
compression (Study 1b)	6383	15	11
trade-off (Study 1b)	6385	7	7
probability (Study 1b)	6383	17	18

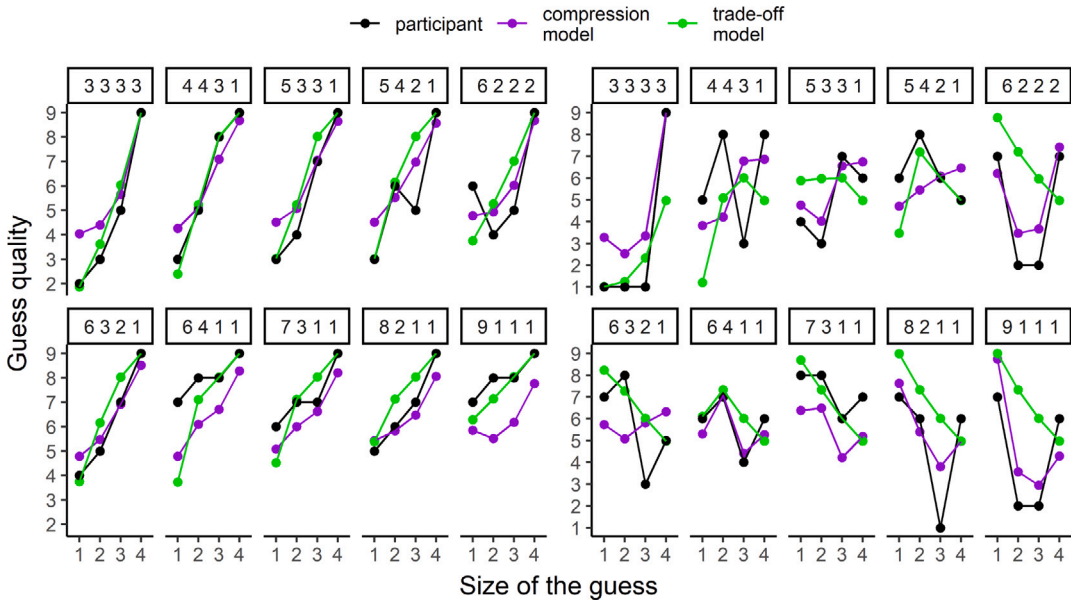

Fig. 8. Ratings from two representative participants in Study 1b, along with the predictions of the compression (purple) and trade-off (green) models, fitted to these participants' data. The participant on the left appears mostly sensitive to the probability of a guess, while the participant on the right has a more subtle pattern of judgments, sometimes preferring less likely, shorter guesses. Panel labels represent the profile of an urn: for example, an urn labeled [9,1,1,1] has 9 balls of one color, and one ball each of the other colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

these participants' judgments. The presence of probability-maximizing participants explains why the compression model has an overall poorer group-level fit than the other models.

There was nonetheless also a substantial number of participants (in both studies) who did something different than probability-maximizing—see for example the participant highlighted on the right of Fig. 8. These participants favored long guesses when colors are equally frequent (as in the urn with profile [3,3,3,3] which has 3 balls of each color), but they preferred shorter guesses for urns where one color was predominant. For example, for an urn with 9 red balls out of 12, these participants would favor the guess "The player will draw a red ball". For an urn with 6 yellow balls and 4 blue balls, many of them would favor the guess "The player will draw a yellow ball or a blue ball".

The judgments of these participants are naturally accounted for by the compression model. The model favors guesses that mention the most likely outcomes, because such guesses implicitly encode a distribution that is close to the speaker's probability distribution over possible outcomes. Therefore the model naturally favors short guesses when one or a few colors dominate (e.g. an urn with 9 red balls out of 12), and long guesses when all colors have the same frequency.

The trade-off model can also account for this pattern of judgments. The model values guesses that are both likely and specific. For an urn with 9 red balls out of 12, the guess "it will be red" is likely enough (it will come out true 75% of the time), and it is very specific because it rules out 3/4 of the possible outcomes. The model favors longer guesses (like "it can be any color") for urns with more equal color frequencies, as the gain in specificity from leaving out one color is not worth the decrease in probability.

Finally, we observe a U-shaped relationship between guess size and guess quality for some urn profiles. This U-shaped pattern is apparent at the group level, for example for urn profile [9,1,1,1], see Figs. 6 and 7. It can also be found at the individual level, especially among participants who are not best-fit by the probability-maximizing model—see for example the participant at the right

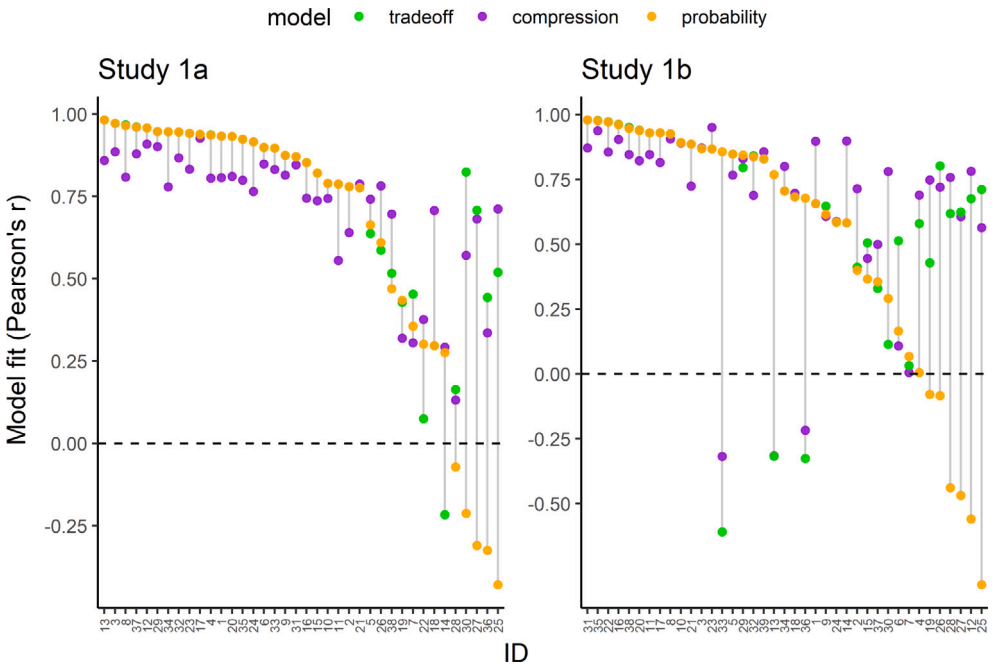

Fig. 9. Study 1: Individual model fits. Each point corresponds to the correlation between the judgments of one participant and the trade-off model (green), the compression model (purple) or the simple probability model (orange). Gray lines connect points belonging to the same participant. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of Fig. 8.[13] To give an example, when the urn has 9 red balls and 1 ball of each other color, a participant might rate a size-1 guess ('Red') and a size-4 guess ('Red, Green, Yellow or Blue') as both better than a size-2 guess ('Red or Green') or a size-3 guess ('Red, Green or Yellow'). As discussed in the introduction, the trade-off model cannot (even in principle) predict this pattern. In contrast, the compression model often exhibits a U-shaped pattern when participants' judgments do.

2.6. Discussion

Study 1 provides initial evidence for our rational analysis of guesses in terms of compression, and to some extent for Dorst and Mandelkern (2021)'s account in terms of an accuracy-specificity trade-off. A substantial number of participants made judgments that could not be accounted for by a pure probability-maximizing strategy, and were better fit by the compression and the trade-off models.

We nonetheless still find that a large number of participants simply responded in function of probability. This result might be a consequence of the relatively unnatural response format (rating the quality of a guess). Some participants may have been induced to rely on probability because this is the only measure of guess quality for which they have an explicit concept.

In Study 2, we make the task more natural, asking participants to compose their own guesses.

3. Study 2

Study 2 used the same setup as Study 1a, except that we let participants compose their own guesses. For each urn, participants had to complete the statement "The ball drawn from the box will probably be:", and could make any of 15 possible disjunctive guesses (for instance "Red or Blue or Green", "Blue", "Yellow or Green", etc.) by clicking on four buttons on the screen, one for each color (see Fig. 10). Clicking on a button added the color to the guess. Participants could also remove a color already in the guess by clicking on the button for that color again. The buttons were presented in a 2*2 array. The position of each color in that array was randomized across participants, but was the same across all trials for a given participant.

We also added two attention checks. During the instructions, participants were told to make a guess with two colors to get familiar with the interface (the two colors were randomly specified for each participant). Participants who did not include these two colors in their guess were excluded from analysis. Additionally, the last trial of the task contained an urn in which two colors were absent. Participants who included a color that was absent from the urn in their guess were excluded from analysis (we used

[13] Figures for all individual participants are available on the Open Science Framework at https://osf.io/wz649.

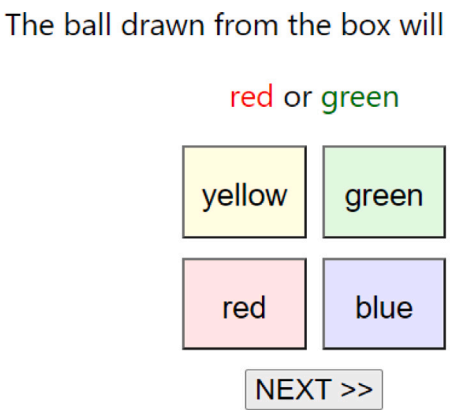

Fig. 10. Study 2: Partial screenshot of the experimental interface.

this trial purely as an attention check). The procedure was otherwise similar to Study 1a, and participants made guesses about 13 different urns.

3.1. Modeling

In addition to the compression, trade-off, and naive probability models, we also consider a simple heuristic model according to which participants include a color in a guess if the number of balls of that color is at or above a given threshold θ. For example, if $\theta = 2$, people include in their guess all colors that are present in at least two balls in the current urn – so, for the urn profile [6,3,2,1], people include the three most frequent colors in their guess (because there are three colors with 2 balls or more), but they only include one color for the urn profile [9,1,1,1]. See Appendix for complete model specification.[14]

To generate the probability that a given model would make a given guess, we passed model judgments through a soft-max function, such that the probability of making a given guess g_i is a function of its quality relative to all other possible guesses one could make about the current urn:

$$Pr(G = g_i) \propto e^{\beta V(g_i)} \tag{7}$$

where $V(g_i)$ is the value that the model assigns to guess g_i (for our main model, $V(g_i)$ is given by Eq. (5)), and β is an inverse temperature parameter controlling the stochasticity of choices (lower values of β correspond to more stochastic choices).

3.2. Participants

We recruited 98 US residents (72 female, 22 male, 4 other, mean age: 33, sd: 16) from Prolific. Participation was restricted to users with a more than 90% approval rate and who had completed between 50 and 1000 previous submissions on the platform. We excluded from analysis 34 participants who failed an attention check, yielding a final sample of 64 participants.

3.3. Results

We can visually inspect some properties of people's guesses by plotting the proportion of guesses of different sizes for each urn profile (Fig. 11). This reveals a lot of diversity in the guesses that participants make, even for the same urn profile. Looking at urn profile [5,3,3,1] for example, many participants made a guess mentioning a single color (i.e. a guess of size 1), many others made a guess of size 3, and a smaller proportion of participants made guesses of size 2.

Despite this diversity, are there systematic patterns in participants' guesses? We can first look at whether participants make guesses that are 'Pareto-optimal' in terms of accuracy (i.e. probability of being correct) and specificity (the proportion of possible outcomes they leave out). A guess is optimal in that sense if it is impossible to construct a guess that is more specific but not less accurate than the current guess, or more accurate but not less specific. In our context, a guess is Pareto-optimal if there is no other color in the current urn that is strictly more frequent than one of the colors mentioned in the guess.[15] We find that the overwhelming

[14] We included this model because it occurred to us as a salient alternative hypothesis when we piloted the task ourselves.

[15] What we call 'Pareto-optimality' has also been called 'cogency' by Holguin (2022), and 'filtering' by Dorst and Mandelkern (2021). Pareto-optimality along the accuracy-specificity axis is obviously a prediction of the trade-off model, but it is also predicted by our information-theoretic model. This is because if a guess g mentions color B but not color A, and A is more frequent than B in the current urn, then a guess g' that mentions A instead of B would a strictly better encoding (in information-theoretic terms) of the underlying probability distribution.

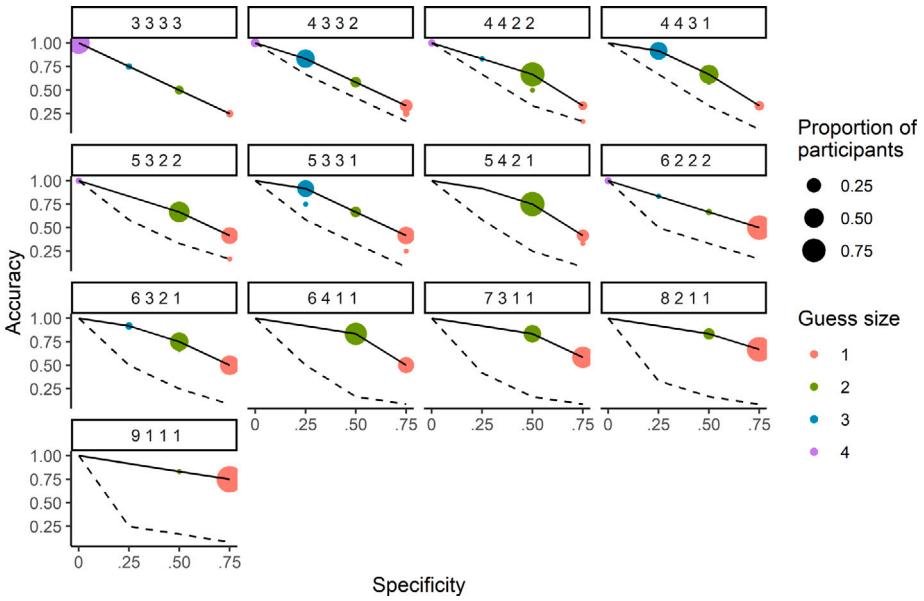

Fig. 11. Study 2: Proportion of participants making a given guess, as the function of the size, accuracy and specificity of the guess, for each urn profile. The size of a circle corresponds to the proportion of participants making a guess with the corresponding accuracy and specificity. Solid Black lines represent the Pareto frontier: guesses that cannot be made more specific without losing accuracy, or vice-versa. Dashed black lines represent the inefficiency frontier: guesses that cannot be made worse on one dimension without getting better on the other dimension.

Table 2
Fit of each model to the data, Study 2. Pearson's r indicates the correlation between the proportion of participants making a guess and the model probability of making that guess. AIC: Akaike Information Criterion — lower values indicate better fit. *n* best: number of participants for which the model has a Bayes Factor larger than 3 against all other models.

Model	Pearson's r	AIC	*n* best
Compression	.964	1676	15
Trade-off	.929	1960	27
Threshold	.871	2388	7
Naive probability	.428	3426	0
Random	NA	4506	0

majority (98%) of participants' guesses are Pareto-optimal (they lie along the black lines on Fig. 11), compared to an expected 41% for a random guesser.

Participants also appear sensitive to 'inflection points' in the exchange rate between accuracy and specificity. Consider the urn profile [6,4,1,1]. Its Pareto frontier has a relatively shallow slope between size-4 and size-2 guesses, and then a steep slope between size-2 and size-1 guesses. Most participants made a size-2 guess, as if trading accuracy for specificity up to the point where it was no longer efficient.

Because qualitative patterns like Pareto-optimality might also have simpler explanations, we now assess model fit quantitatively for a stricter test of our proposal.

3.3.1. Modeling results

We first fit each model at the group level, by finding the parameter values that maximize the log-likelihood of the data under the model. Table 2 describes the fit of each model to the data, and Table 3 shows the best-fitting parameter values for each model.

The compression model has a very good fit to the data. The correlation between the probability that the model makes a guess and the proportion of participants making that guess is very high, $r(193) = .964$, $p < .001$. This correlation is still very large even when restricting the analysis to the set of Pareto-optimal guesses, $r(78) = .957$, $p < .001$.

Thus, the compression model is able to accurately track how participants modulate the size of their guesses as a function of the urn profile. This can be seen more clearly in Fig. 12, where we plot the proportion of (Pareto-optimal) guesses of a given size made by the compression model and by participants, for each urn profile. Overall, participants tend to make guesses that are assigned high probability by the model. For instance, for the urn profile [4,3,3,2], most participants made a guess of size 3 (for instance, "it will probably be red or blue or green" for an urn containing 4 red balls, 3 blue balls, 3 green balls, and 2 yellow balls), and this is also the compression model's preferred guess.

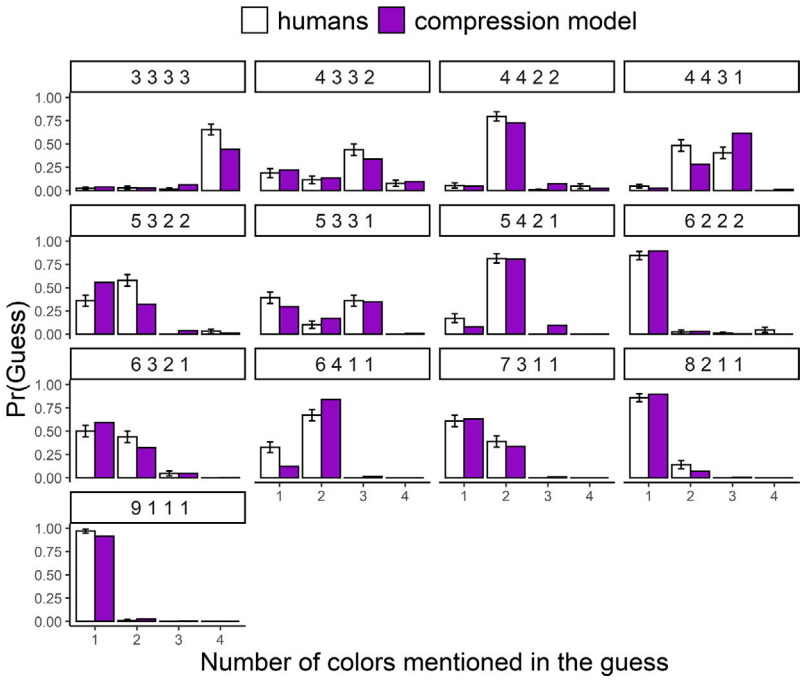

Fig. 12. Study 2: Proportion of human participants making a given guess, and model probability for that guess, as a function of urn profile and guess size, for guesses lying on the Pareto frontier. Note: for some urn profiles, several different guesses can correspond to the same guess size. When this is the case, we compute the average choice probability across all these guesses. Note that probabilities do not necessarily sum to 1, because guesses lying outside the Pareto frontier are not represented.

Table 3
Best-fitting values of the model parameters, for each model, Study 2.

Model	β	α	Param1
Compression	46.8	1.14	$\gamma = 1.99$
Tradeoff	8.11	2.11	$J = 3.01$
Threshold	5.30	3.41	$\theta = 3$
Naive probability	5.39	7.09	

The model can also explain the variability in participants' guesses. Guesses for urn profile [5,3,3,1], for example, show a U-shaped pattern: most participants made guesses of size 1 or 3, while a smaller proportion made size-2 guesses; this pattern is reflected in the probability mass that the model assigns to these options. By contrast, when most participants make the same guess (as for the urn profile [9,1,1,1], where almost all participants make a size-1 guess), the model also puts most of its probability mass on that guess.

Participants also appear to have a preference for 'clustering', in the sense that when two outcomes A and B have a similar probability, they rarely include one outcome in their guess but not the other. For the urn profiles [4,4,2,2] and [5,4,2,1], for instance, the vast majority of participants made a guess that included the two most frequent colors. This preference for clustering is also reflected in the compression model predictions.

The trade-off model has a slightly lower fit to the data than the compression model : its predictions are correlated with human choice proportions at r(193) = .929, p <.001, (see Fig. A.1 in the Appendix). The model is unable to explain some of the subtle features of the data, such as the U-shaped patterns described above. The model predicts that there is one optimal guess size for a given urn, and the quality of a guess diminishes monotonically as a function of its distance from the optimal guess size. The trade-off model also drastically under-estimates the proportion of participants who make a size-4 guess for an urn where all colors are equally frequent.

Next, we analyzed the data at the individual level: for each participant and each model, we computed the marginal likelihood of the data under that model.[16] Twenty-seven participants were best-fit by the trade-off model, while 15 participants were best fit by

[16] We computed the marginal likelihood by Monte Carlo simulation, taking 10^4 samples per participant and per model. We sampled model parameters from weakly informative priors, sampling α from an exponential distribution with rate 1/2, β from an exponential with rate 1/50, θ from a uniform distribution, and sampling γ and J as $1 + V$, where V is exponentially distributed with rate 1/4. We do not use AIC to perform model comparison at the individual level because the small number of trials (13 discrete choices) per participant led to identifiability issues. Specifically, we performed a model recovery analysis where

the compression model, and 7 participants by the threshold model (there were also 15 participants for which only weak evidence favored a given model). As such, although the compression model has the best fit to the data at the group level, the trade-off model is the best-fitting model for a larger number of individual participants. In particular, some participants adopted the policy of always making a size-1 guess, picking the most frequent color in the urn. This relatively low-effort policy is well-modeled by the trade-off model by setting J to a large number.

More generally, the trade-off model assumes that the length of a guess carries a direct cost (because longer guesses are less specific). This assumption hinders the model's ability to capture the group-level distribution of guesses (see above), but it can help it capture the fact that making longer guesses took more effort in our task (they required clicking on more buttons). So the model can to some extent capture the 'laziness' of some individual participants.

The other models we considered had a worse fit to human judgment than the models mentioned above. The naive probability model predicts that participants should have made the maximally-inclusive guess ("Red or Green or Blue or Yellow") every time, because that guess always has probability 1. Yet this was never the modal guess, except for the case where all colors are equally frequent. The simple threshold model is relatively effective at finding the modal guess for most urn profiles, but is unable to account for the variability in people's judgments, predicting that almost all participants will select the same guess for a given urn profile, see Fig. A.2 in the Appendix. In exploratory analyses, we find that the model has a relatively poor fit even when we allow its parameter values to vary from participant to participant, showing that the model cannot explain variability in the data by assuming that different participants have different thresholds.

3.4. Discussion

We asked people to make disjunctive guesses about the outcome of a simple game of chance. If participants were motivated to maximize their probability of being correct, they would have always made the guess that included all possible outcomes. Participants actually made much more variable guesses, which varied in a systematic way as a function of the relative frequency of colors in the urn.

Participants' judgments are well-explained by our information-theoretic model, according to which guesses encode an approximation of the speaker's distribution over possible outcomes. The model is able to explain subtle patterns in participants' judgments. For example, there were urns for which different participants made different guesses, and urns for which almost all participants made the same guess. This pattern is reflected in the model predictions, suggesting that participants vary in the guesses they make when the model sees these guesses as equally good.

Participants' guesses are also broadly consistent with an account of guessing at a complementary, more descriptive level, according to which guesses strike a trade-off between accuracy and specificity (Dorst & Mandelkern, 2021). However, while the trade-off model provides a good account of many individual participants, it has difficulty accounting for the specific shape of the group-level distribution of guesses. In the next two studies, we investigate whether *listeners* interpret guesses as implicitly encoding a probability distribution.

4. Study 3

Results from Studies 1 and 2 suggest that guesses encode a compressed representation of the speaker's subjective probability distribution. Can listeners decode this representation? To address this question, we run an 'inverted' version of Study 2: we show participants someone else's guess, and ask them to infer which urn the speaker was looking at. Specifically, in each trial, we show participants a speaker's guess as well as two urns A and B, and ask them to indicate which urn they think the speaker was looking at when he made the guess.

4.1. Computational modeling

The optimal Bayesian decoder for this task is given by:

$$Pr(\text{Urn X}|\text{Guess}) \propto Pr(\text{Guess}|\text{Urn X})Pr(\text{Urn X}) \tag{8}$$

where the likelihood $Pr(\text{Guess}|\text{Urn X})$ is the probability that a speaker looking at Urn X would make a given guess. We further assume a uniform prior over urns (i.e. $Pr(\text{Urn A}) = Pr(\text{Urn B})$), allowing us to re-write the above expression as:

$$Pr(\text{Urn B}|\text{Guess}) = \frac{Pr(\text{Guess}|\text{Urn B})}{Pr(\text{Guess}|\text{Urn A}) + Pr(\text{Guess}|\text{Urn B})} \tag{9}$$

We do not have direct access to the likelihood $Pr(\text{Guess}|\text{Urn X})$, but we can estimate it. We do so in two different ways. Our first approach makes no theoretical commitment about the speaker's behavior, but simply estimates the likelihood $Pr(\text{Guess}|\text{Urn X})$ as

we simulated the judgments of virtual participants, and fit these simulated data with the compression and trade-off model. When simulating data under the assumption that the compression model is the correct generative process, we find that 64% of simulated participants are nonetheless better-fit by the trade-off model (as assessed by AIC). When generating simulated participants using the trade-off model as the generative process, only 5% of simulated participants are incorrectly better-fit by the compression model. This suggests that the trade-off model is more prone to over-fitting at the individual level.

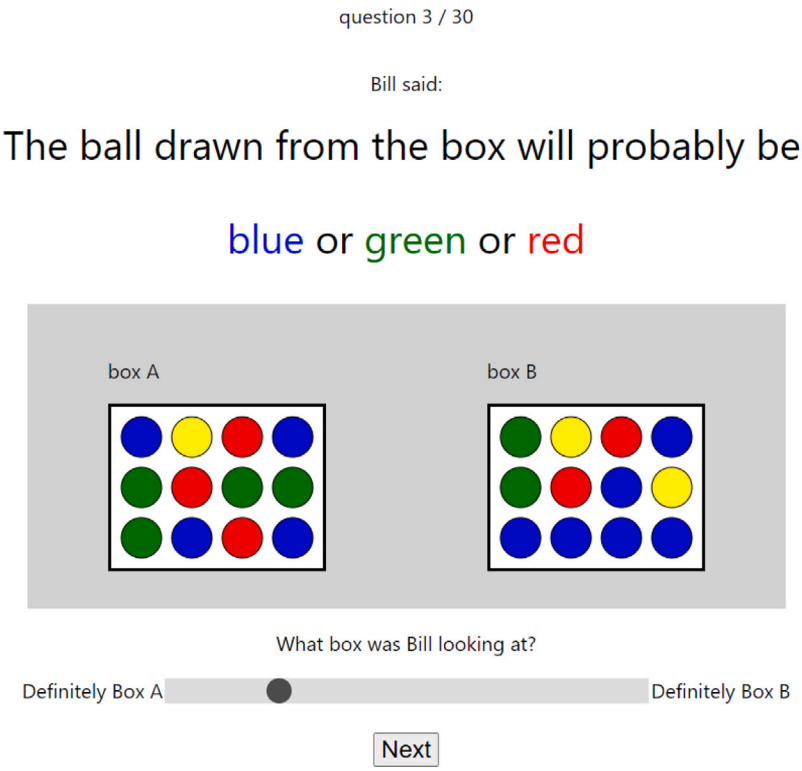

Fig. 13. Study 3: Partial screenshot of the experimental interface.

the proportion of participants in Study 2 who made that guess when looking at Urn X.[17] Our second approach uses the likelihood defined by a given computational model tested in Study 2 (for example the compression model), with the best-fitting parameters that we derived at the group level for that model in our analysis of Study 2.

Note that Eq. (8) describes a *pragmatic* listener (cf. Goodman & Frank, 2016), who can approximately model the way that speakers make guesses, and makes inferences by inverting this model.[18] Therefore we call the model a 'pragmatic listener', although we do not make strong process-level claims about the way participants complete the task.

4.2. Methods

4.2.1. Procedure

After signing a consent form and reading instructions (similar to the previous experiments), participants first completed four trials of the production task from Study 2, to get familiar with the setting. In the main task, we then asked participants to imagine that another person called Bill also had to make similar guesses. For each trial, we displayed two urns on the screen (labeled 'box A' and 'box B'), as well as the guess that Bill made, and asked participants which box they think he was looking at. We indicated the guess made by the speaker in the following format:

"Bill said:

The ball drawn from the box will probably be:

[guess]"

Where [guess] was a disjunction of colors, for example 'red', or 'green or yellow'. Below the boxes, we asked "What box was Bill looking at?", and participants answered using a slider scale ranging from 'Definitely Box A' to 'Definitely Box B', but otherwise unlabeled (internally the scale ranges from 1 to 100). See Fig. 13 for a partial screenshot of the experimental interface.[19]

To keep the task non-trivial, we only used pairs where each urn has the same frequency ordering over colors, and guesses that are Pareto-optimal in the sense defined in Study 2 (colors included in the guess are more or equally frequent, in both urns, than

[17] We add $\epsilon = 0.001$ to the proportion in cases where no participant in Study 2 made that guess, to avoid divide-by-0 errors in later computations.

[18] Note that the pragmatic listener in the current task is not the same as the 'literal' listener whose inferences are anticipated by the speaker in our model for Studies 1 and 2. There is no inconsistency with the current analysis, however: in Studies 1 and 2, the speaker only sees one urn, and so he cannot strategically adjust for the specific challenge faced by our listeners in the current task.

[19] Instructions also made clear that among the two urns on the screen in a given trial, one was a random urn that was not shown to the speaker. That is, the speaker does not have the opportunity to maximize how informative his guess is with respect to the task of discriminating A from B.

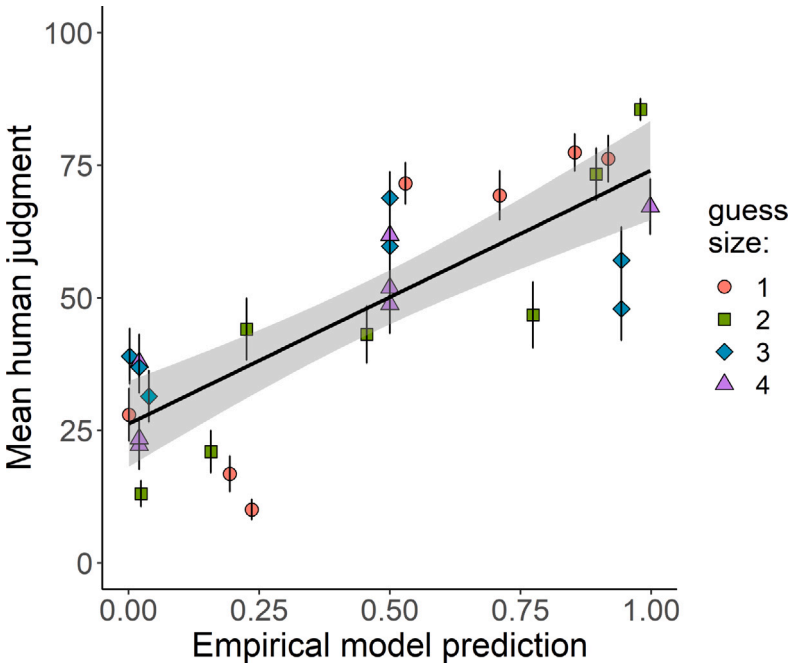

Fig. 14. Study 3: Mean human judgment (preference for box B) as a function of the predictions of the pragmatic listener model with an empirical likelihood (derived from production data in Study 2). Error bars represent the standard error of the mean.

colors not included in the guess). Thus participants could not solve the task simply by exploiting differences in frequency orderings across urns (there were never any trial where, for instance, the guess is 'yellow' and yellow is the most frequent color in urn A but the third most frequent color in urn B).

We designed stimuli by computing, for every possible pair of urn profiles and each possible guess that obey the criteria above, the prediction of our Listener model (calibrated with the compression-based likelihood). We then randomly sampled 7 trials for each guess size (from 1 to 4) that smoothly spanned the range of predicted probabilities (from 0 to 100% chance of box B), resulting in a total of 28 trials (see Table A.1).

We also generated two attention check trials with an obvious answer, where each urn has nine balls of a given color and one ball of each other color, but the dominant color is different in each urn, and the guess mentions the dominant color in urn A. Participants who did not give a rating of 50% of more for urn A in either trial were excluded from analysis. Trials were presented in randomized order, and the position of the urns within a pair (which urn was assigned to 'box A') was counter-balanced.

4.2.2. Participants

We recruited 49 participants (23 female, 1 other, mean age=45, sd=15) from Prolific. Participation was restricted to US residents with a 90%+ approval rate who had taken between 50 and 1000 previous studies on the platform. Participation took on average 10 min, and participants were compensated £1.20 for their participation. We excluded from analysis 13 participants who failed at least one attention check, for a final sample of 36 participants.[20]

4.3. Results

The model derived from the empirical production data in Study 2 provides a good account of the current data, with no free parameter. On average, the correlation between a participant's judgments and model predictions was r(26) = .50, inter-quartile range = .32 to .69. Aggregating across participants, the correlation between model predictions and mean human judgment was r(26) = .81, p <.001; see Fig. 14. Results for example trials are displayed in Fig. 15.

We then perform the same analysis for the model that uses the likelihood from the compression model. On average, the correlation between a participant's judgments and model predictions was r(26)=.48, inter-quartile range=.20 to .72. Aggregating across participants, the correlation between model predictions and mean human judgment was r(26) = .79, p <.001; see Fig. 16.

[20] The first attention check was the same as in Study 2, but participants in the current experiment failed that attention check at a higher rate than in Study 2. Excluding every one of these participants would have led us to discard more than half the sample (final N=23), so we adopted a softer criterion, retaining participants who mentioned at least one the requested colors. Analysis with the stricter exclusion criterion yields virtually identical results.

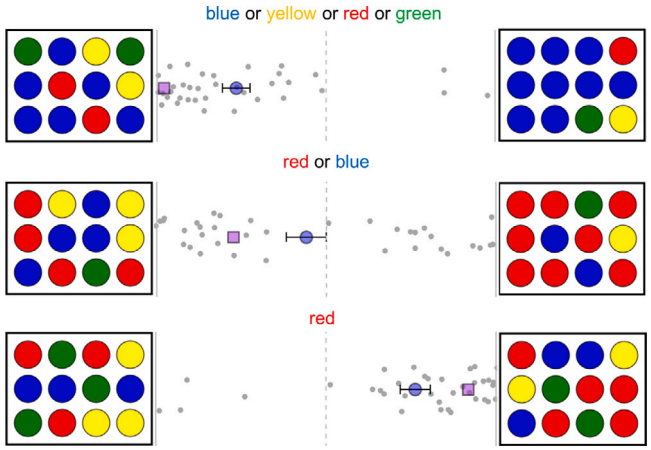

Fig. 15. Study 3: Results for a selection of trials (see Fig. A.4 for all trials). Gray dots represent individual ratings; predictions for the empirical model are in purple, and mean human ratings are in blue. Error bars represent the standard error of the mean. (The particular urns and guesses displayed here are meant to illustrate the abstract structure of a trial: during the experiment the ordering over colors and ball positions were procedurally generated for each participant.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

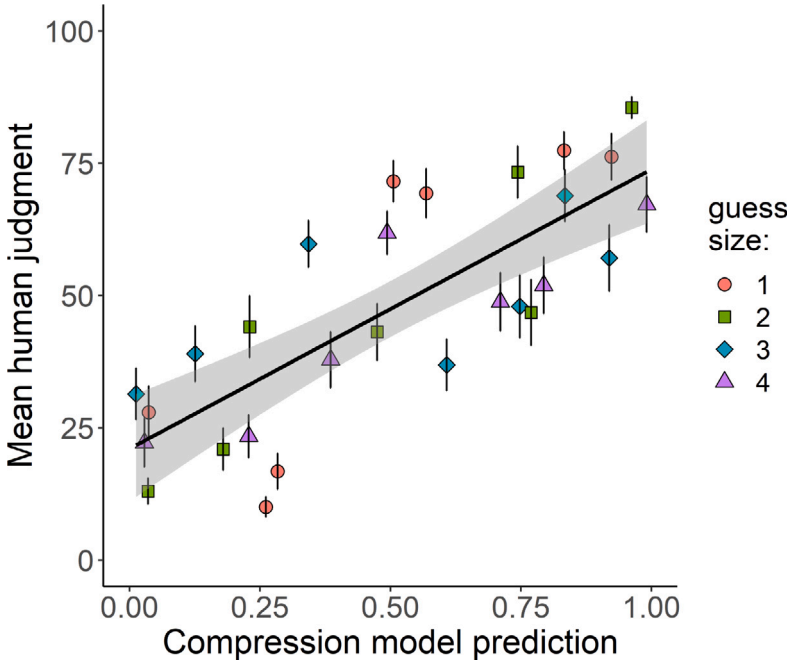

Fig. 16. Study 3: Mean human judgment (preference for box B) as a function of the predictions of the pragmatic listener model with the compression-based likelihood. Error bars represent the standard error of the mean.

We find a similarly good correlation when we use a likelihood derived from the naive probability model (mean individual-level correlation, $r(26) = .49$, inter-quartile range: .48 to .62; item-level correlation: $r(26) = .79$, $p < .001$).[21]

[21] We report this result for completeness, but given the poor performance of the naive-probability model in the previous study, we suspect that the good performance of the naive-probability likelihood is an artifact of the following property of the current task. To perform well in the task, a pragmatic listener does not necessarily need to correctly rank the probability of making a guess within a given urn; what matters is the relative likelihood of the guess *across urns*. So, even though the naive-probability likelihood incorrectly predicts that most people will make a size-4 guess for an urn with profile [9,1,1,1], it correctly predicts that the proportion of size-4 guesses will be higher for urn profile [3,3,3,3] than urn profile [9,1,1,1]. Therefore a pragmatic listener using a naive-probability likelihood can successfully infer that someone who made a size-4 guess was looking at the [3,3,3,3] urn.

Performing a similar analysis with likelihoods from the trade-off and threshold models (again with the parameters obtained in Study 2) yields slightly lower agreement with the data, r(26) = .73, p <.001 (trade-off model) and r(26) = .63, p <.001 (threshold model).

Could participants simply have chosen the urn for which the guess had the highest probability of being correct? Such a heuristic would lead people to be indifferent between the two urns when the speaker makes a size-4 guess, since a size-4 guess always has probability 1 of being correct, regardless of the urn contents. Participants actually drew strong inferences even for size-4 guesses—see for example the trial on top of Fig. 15.

4.4. Discussion

The current results suggest that people can decode the distributional information encoded in a guess, at least in the context of verbal communication. Participants were able to reverse-engineer which probability distribution the speaker had in mind (i.e. which urn he was looking at), on the basis of the speaker's guess. Specifically, their judgments were well-predicted by a Bayesian decoder calibrated with the production data from speakers in Study 2, without any free parameter.

The fact that people's judgments are well-predicted by a normative benchmark provides evidence that guesses perform their communicative function well. Future research could more deeply investigate the exact process by which listeners make their inferences. In the next study, we provide additional evidence that people extract distributional information from a guess.

5. Study 4

To test the generality of our framework, here we investigate its predictions in a different setting and in a different task. Specifically, we study how people evaluate guesses about continuous quantities, in a context where the correct answer is already known.

Study 4 is a conceptual replication of a classic experiment by Yaniv and Foster (1995). We ask participants to evaluate the quality of a guess relative to the ground truth. For instance, suppose that the speaker guessed that there are between 165 and 185 member states in the United Nations. Given that there are actually 193 member states in the UN, how good was the speaker's guess? We make the hypothesis that people treat the guess as implicitly encoding a probability distribution over the answer, and judge that a guess is good if the probability distribution they decode from the guess assigns a high probability to the correct answer (see Fig. 17).

5.1. Model

Our general framework assumes that listeners infer a probability distribution Q_g from a guess g. The current study focuses on testing this assumption. Specifically, we test the hypothesis that when people evaluate an interval guess of the form "x is between x_{low} and x_{high}" (for example, "There are between 165 and 185 member states in the United Nations"), they implicitly treat this guess as encoding a probability distribution.

Intuitively, such a guess conveys information both about the mean of the speaker's distribution (it is probably around the midpoint of the interval), and about its standard deviation (the speaker makes wider guesses the more uncertain he is). Therefore, we assume that the listener infers that the speaker's distribution has mean μ and standard deviation σ, where μ is the middle of the interval, and σ is proportional to the interval width:

$$\mu = x_{\text{low}} + \frac{x_{\text{high}} - x_{\text{low}}}{2} \tag{10}$$

$$\sigma = k(x_{\text{high}} - x_{\text{low}}) \tag{11}$$

where k is a free parameter. There are an infinity of possible distributions that obey these constraints, but we assume that the listener infers a normal distribution, with mean μ and variance σ^2. This choice is motivated on normative grounds: The normal distribution is the maximum entropy distribution for known mean and variance, meaning that if all we know about a distribution is its mean and its variance, a normal distribution is the representation that imports the fewest extra assumptions (Jaynes, 2003).[22] The quality of a guess is then the probability that this distribution assigns to the correct answer; see Fig. 17.

5.2. Accuracy-specificity model

We also consider whether people's judgments strike a trade-off between accuracy and specificity, following the original model that Yaniv and Foster (1995) used to model their data. According to this model, the quality of a guess is inversely related to the quantity:

$$L = \frac{|t - m|}{w} + \alpha \log(w) \tag{12}$$

where t is the correct answer, $m = x_{\text{low}} + (x_{\text{high}} - x_{\text{low}})/2$ is the midpoint of the participant's guess, $w = x_{\text{high}} - x_{\text{low}}$ is the width of the interval, and α is a free parameter controlling the weight that people assign to specificity relative to accuracy. Intuitively,

[22] Technically speaking, many of the quantities in our study lie on partially bounded intervals (for example a distance cannot be less than 0 kilometers) and thus normality is only an approximation of the maximum entropy distribution. To keep the model simple and intuitive we pass over this issue.

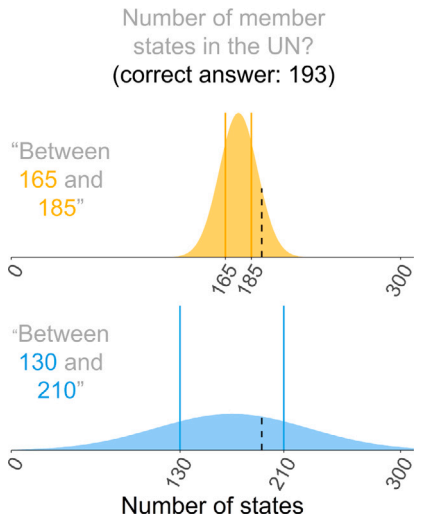

Fig. 17. Illustration of our model. Listeners infer a probability distribution whose mean and variance are determined by the center and the width of the interval, respectively. The dashed line indicates the probability assigned to the correct answer. The first interval (orange) should be seen as a better guess than the second interval (blue), despite not formally including the correct answer. Distributions were generated by multiplying the width of an interval by $k = .74$, the value that provides the best fit to our experimental data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the first term ($\frac{|t-m|}{w}$) is inversely related to the accuracy of the guess, while the second term ($\log(w)$) is inversely related to its informativeness.

5.3. Method

Participants were told to imagine that they were a researcher preparing for a presentation, and that they had asked two research assistants for their estimates about a given number (Yaniv & Foster, 1995). Participants were for example told that the two assistants were asked the question "what was the date of the first transatlantic flight?", and that one assistant responded "1930 to 1970" and another responded "1915 to 1923", while the correct answer was 1927. Participants were asked which of the two assistants gave a better answer. To prevent carry-over effects, they were also asked to imagine that the assistants were different in each scenario.

Each participant made a choice for 20 different trials (see Table A.3 in the Appendix). Each trial features a ground truth (the correct answer to the question) and two different interval guesses (one made by assistant A and one made by assistant B). Participants were asked "which estimate is better?", and had to select either A or B. Trials were presented in randomized order, and the identity of the assistants (whether assistant A and B made a given statement) was randomized across trials and participants. We also included as an attention check a trial for which one guess was unambiguously better, and excluded from analysis participants who failed to select that guess (data from this trial were not otherwise included in the main analysis).

5.3.1. Participants

We recruited 99 US residents (51 male, 46 female, 1 other, mean age = 34.4, sd = 12.2) from Prolific. Participation was restricted to users with a more than 90% approval rate and who had completed between 50 and 1000 previous submissions on the platform. We excluded from analysis one participant who failed an attention check, yielding a final sample of N=98.

5.4. Computational modeling

In addition to the compression and trade-off models, we tested simple heuristic models (following Yaniv & Foster, 1995). According to these models, the quality of a guess is determined by:

- 'Nearest-boundary' distance: (inverse of) distance between the ground truth and the interval boundary nearest to the ground truth.
- 'Farthest-boundary' distance: (inverse of) distance between the ground truth and the interval boundary farthest from the ground truth.
- 'Absolute error": (inverse of) distance between the ground truth and the midpoint of the interval.
- 'Normalized error': (inverse of) absolute error divided by the width of the interval.
- 'Interval width': (inverse of) interval width.
- 'Inclusion': a binary variable indicating whether the interval contains the ground truth.

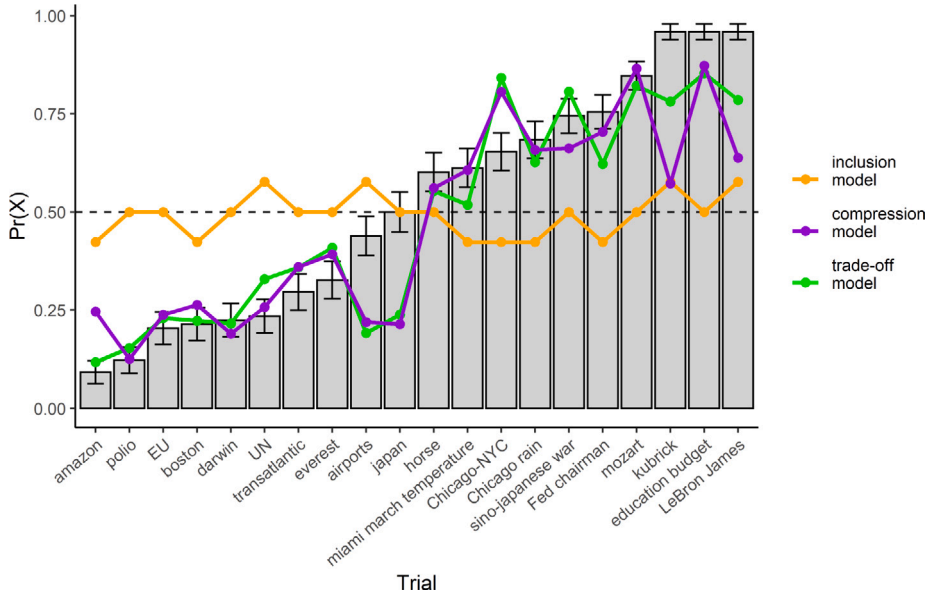

Fig. 18. Study 4: Proportion of participants preferring the guess by assistant X over the guess by assistant Y, along with predictions of the compression (purple) and trade-off (green) models. For comparison we also include a heuristic model according to which a guess is good if it contains the correct answer (orange)—see Table 4 for the fits of other heuristic models. Error bars represent standard errors. See Table A.1 for list of trials. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For each model and each trial, we first compute the quality of the guesses made by assistant A and B under the model. We then compute the probability of choosing a guess via a soft-max function over guess quality, with a parameter β controlling the stochasticity of answers (as in Study 2, see Eq. (7)). We fit the models to the data, both at the group- and the individual level, by finding the parameter values that maximize the log-likelihood of the data.

5.5. Results

For 18 out of 20 items, a statistically significant majority of participants chose the guess that the compression model (as well as the trade-off model) judged to be better. Human choice proportion for the 2 other items was not significantly different from 50%. All the heuristic models have lower classification accuracy (see Table 4).

The quantitative predictions of the compression model were highly correlated with the proportion of participants choosing a given guess, $r(18) = .865$, $p < .001$, see Fig. 18, although the best-fitting model was the trade-off model, $r(18) = .915$, $p < .001$. Remarkably, the predictions of the trade-off model and the compression model are highly correlated with each other, $r(18) = .951$, $p < .001$. All heuristic models had a lower fit to the data than the compression or the trade-off model.

To verify that the results are not an artifact of averaging or over-fitting, we also conducted individual-level model comparisons, both by computing the Akaike Information Criterion (AIC) for each participant and for each model, and by computing Bayes Factors derived from marginal likelihoods[23]; see Table 4. Again the compression and trade-off model have the best fit to the data, with the trade-off model fitting slightly better. We also note some heterogeneity in the individual data. Although a substantial proportion of participants are best fit by the compression or trade-off models, some participants are best fit by heuristic strategies; for example the data from 15 participants are best explained by the normalized error model. We note however that there are relatively few participants for which the evidence in favor of a model is strong (i.e. BF > 3 against all other models).

5.5.1. Heuristic models

Here we give some intuition for why the heuristic strategies fail to account for people's judgments.

The inclusion heuristic says that an interval that fails to include the correct answer should never be preferred to one that does.[24] For instance, to the question "how many countries are in the UN?" (correct answer: 193), the guess "40 to 300" should be preferred to "165 to 185" since only the former actually includes the correct answer. Yet 76% of participants chose the latter guess.

[23] We computed the marginal likelihoods by Monte Carlo simulation, drawing 10^4 samples per participant and per model. We used weakly informative priors, sampling each parameter from an exponential distribution with rate 1.5.

[24] Some accounts of the semantics of interval guesses (e.g. Egré et al., 2023) seem to make this prediction, since they posit that all the probability mass should fall within the interval.

Table 4
Fit of the different models to the data, Study 4. Classification accuracy is the proportion of items where the model's preferred guess is chosen by 50% or more participants. AIC: Akaike Information Criterion (lower values indicate better fit), calculated as the sum of individual-level AICs. N best (AIC): number of participants for which the model has the lowest AIC. N best (BF): Number of participants for which the model has a Bayes Factor larger than 3 against all other models. Pearson's correlation: correlation between model prediction and proportion of participants making a choice. β and param1 indicate the group-level best-fitting value of parameters for each model.

	Classification accuracy	AIC	N best (AIC)	N best (BF)	Pearsons' r	β	param1
Compression	1	1931	15	4	.865	.58	$k = .74$
Trade-off	1	1792	46	5	.915	1.59	$\alpha = .49$
Absolute error	.9	2279	6	0	.397	1.23	
Normalized error	.55	2540	15	3	.517	.31	
Nearest boundary	.55	2650	6	0	.405	2.47	
Farthest boundary	.65	2358	0	0	.348	.50	
Inclusion	.7	2653	9	3	.155	.23	
Interval Width	.6	2468	1	0	.230	.16	
Random	.5	2717	2	0	NA		

The normalized and absolute error heuristics both posit that good guesses are those whose midpoint is close to the ground truth, and that people are not penalized for giving overly large intervals. For example, to the question "what is the monthly salary of the Fed chairman?" (correct answer: $16900), the guesses "$18000 to $22000" and "$4000 to $35000" both have a midpoint about $20000, which predicts that people should be indifferent among them. In fact, 75% of participants preferred the first guess.

The nearest-boundary heuristic holds that people prefer intervals that have at least one boundary close to the correct answer. For example, to the question "what is the average gestation length of a horse? (correct answer: 11 months), the guess "12 to 30 months" has a boundary (12 months) which is very close to the correct answer, so people should prefer this guess to the guess "7 to 9 months". Yet 60% of participants preferred the latter guess. Similarly, the farthest boundary heuristic holds that people prefer guesses that minimize the distance between the correct answer and the boundary farthest from the correct answer. For example, to the question "what is the yearly budget of the US department of education ($68 billion), the guess "$75 to $120 billion" has a boundary ($120 billion) that is very far from the correct answer, so people should prefer the guess "$95 to $110 billion". In fact 96% of participants chose the first guess.

5.5.2. Original data from yaniv & foster (1995)

The original data for Yaniv & Foster (1995) have been lost (Yaniv, personal communication), but we can analyze data from the eight sample items displayed in their paper. We find that the compression model has a good fit to these data, $r(6) = .96$, $p < .001$, see Appendix for details.

5.6. Discussion

We conceptually replicated the results of Yaniv and Foster (1995): when people evaluate guesses relative to the correct answer, they prefer those that strike a trade-off between accuracy (being close to the correct answer) and specificity (not including too wide a range). We also find support for our account of *why* people have this preference: guesses that are both accurate and specific implicitly encode a probability distribution that assigns a high probability to the correct answer. Specifically, a model that infers the speaker's subjective probability distribution from the guess, and computes the probability that the inferred distribution assigns to the true value, accounted for the data almost as well as Yaniv & Foster's original model. Furthermore, the predictions of the two models were highly correlated with each other. Participants' judgments could by contrast not be explained by simple heuristics, such as preferring guesses whose interval contains the correct answer, or guesses whose midpoint is closer to the correct answer.

5.7. The production of interval guesses

The current results have implications for how *speakers* should make interval guesses: they should adjust the width of their interval in such a way that the listener accurately infers the uncertainty in the speaker's distribution. We find that listeners in our experiment are best-fit by $k = .74$: they interpret the width of the interval as equal to $\frac{1}{0.74}$ times the standard deviation of the underlying probability distribution. If speakers use a similar value of the scaling parameter k to modulate the width w of their guesses, then they should generate intervals that extend within about $.67\sigma$ from their subjective mean on each side, since $\sigma = kw$ implies $\frac{w}{2} = \frac{\sigma}{2k} = .67\sigma$.

In a normal distribution, about 50% of the probability density lies within 0.67 standard deviations of the mean, so speakers with $k = .74$ should make interval guesses that they see as about 50% likely to contain the correct answer. If we also assume that speakers are approximately well-calibrated (they can reliably estimate their uncertainty), this hypothesis predicts that speakers will offer intervals that contain the correct answer only about half the time. There is indeed a large literature supporting this prediction (Alpert & Raiffa, 1982; Cesarini et al., 2006; Juslin et al., 1999; Klayman et al., 1999; McKenzie et al., 2008; Moore et al., 2015; Russo & Shoemaker, 1992; Soll & Klayman, 2004; Teigen & Jørgensen, 2005; Yaniv & Foster, 1997). For example, Yaniv and Foster

(1997) asked participants to make interval guesses about a variety of real-world quantities, and found that the proportion of intervals that contained the correct answer was consistently slightly less than 50% (46%, 43% and 45% respectively in their three studies). Interestingly, in one of their studies they asked participants to give 95% confidence intervals, i.e. intervals that participants were 95% confident included the correct answer. Yet only about 45% of the intervals did actually include the correct answer. This rate was similar to the hit rate in two other studies where participants did not have to reach a specified target (Yaniv & Foster, 1997).

These results, and many others (e.g. Cesarini et al., 2006; Teigen & Jørgensen, 2005) suggest that when participants are asked to construct a 95% confidence interval, they largely disregard the overt instruction and instead construct a guess that is optimized for another purpose, i.e. give a good encoding of their subjective probability distribution. The interval they give is much too narrow for a 95% confidence interval, but it would have been remarkably effective for communicating their subjective distribution to the listeners in the current study.

6. General discussion

People often make judgments about uncertain facts or events. These judgments are often compared to a normative standard according to which one should endorse statements as a function of the probability that they are correct. We have suggested that it is often useful to instead consider people's judgments as good solutions to a *Distribution-encoding* problem: statements about uncertain facts or events often work as compressed summaries of a probability distribution.

This account makes successful quantitative predictions about what guesses people make (Study 2), what inferences listeners draw from a guess (Study 3), and what people judge to be a good guess both when they already know (Study 4) or do not know the correct answer (Study 1). Below we explore how our account sheds light on puzzling phenomena in probabilistic reasoning. Then, we discuss the scope of our work, some limitations, and directions for future research.

6.1. Accounting for puzzling phenomena in probabilistic reasoning

Here we apply our perspective to some already-documented patterns in probabilistic reasoning. Note that some of these patterns occur in tasks where *Probability-maximizing* seems to be the appropriate normative standard. These tasks require the participant to estimate the probability of a single outcome, for example. It is at first sight not obvious why our perspective would be relevant to understanding these phenomena, but we suggest that the human mind sometimes over-applies its tendency to solve for the *Distribution-encoding* problem. In other words, people might sometimes respond as if they were solving the *Distribution-encoding* problem even in tasks where they are explicitly asked for *Probability-maximizing*.

6.1.1. Surprise and likelihood

Whether we see an outcome as surprising does not only depend on the probability of that outcome. It also depends on the probability of other possible outcomes (Attneave, 1959; Kahneman & Tversky, 1982; Teigen & Keren, 2003). For example, an event that had a 10% probability is surprising if there was another event with probability 30%, but is less surprising if this was actually the most probable event (Teigen & Keren, 2003). Judgments of whether an event was 'likely' also depend on the probability of alternative possible events (Teigen, 1988; Windschitl & Wells, 1998, see also Lassiter, 2011; Yalcin, 2010). When estimating the value of a continuous quantity, people also judge that intervals that lie in the center of the relevant distribution are more likely than equally-probable larger intervals in the tails of the distribution (Teigen et al., 2022).

We suggest that when people make judgments of surprise and likelihood, they often do so with the goal of communicating their subjective probability distribution about the relevant outcome or state of affairs. From this perspective, people judge a possible outcome as likely in order to communicate that this outcome would be mentioned in a good guess about what will happen. Our model predicts that an outcome with probability p will sometimes be included and sometimes be left out of the optimal guess, depending on the probability of other outcomes. This naturally explains why our judgments of whether an outcome is likely depend on the probability of other outcomes. Our proposal also explains why interval estimates in the center of a distribution are judged as more likely than intervals in the tails: intervals in the tails of a distribution are poorer representatives of the distribution, and therefore make for poor guesses.

Similarly, expressing surprise at an event might communicate what the speaker's subjective distribution was like before the speaker observed the event. So, if the speaker assigned 10% probability to event A but 30% probability to event B, and A subsequently happens, the speaker expresses surprise as a way to convey that he would not have guessed that A would happen.

6.1.2. Overconfidence in interval estimation

People consistently produce over-confident confidence intervals. For example, when people are asked to estimate a numerical interval that they think is 95% likely to contain the correct answer to a question, they give an interval that contains the correct answer about 50% of the time (Alpert & Raiffa, 1982; Cesarini et al., 2006; Juslin et al., 1999; Klayman et al., 1999; McKenzie et al., 2008; Moore et al., 2015; Russo & Shoemaker, 1992; Soll & Klayman, 2004; Teigen & Jørgensen, 2005; Yaniv & Foster, 1997). Existing explanations of this pattern assume that it arises from cognitive limitations (e.g. Juslin et al., 2007; Moore, 2022; Zhu et al., 2023). As we argued in our discussion to Study 4, apparent overconfidence might also arise because speakers are trying to communicate about their subjective probability distribution. Participants in interval production studies might be disregarding or misunderstanding the experimenter's instructions (e.g., of producing a 95% confidence interval) and instead might be trying to communicate their uncertainty in a way that conforms to the expectations of an audience. Our experimental results in Study 4 suggest that the audience indeed expects interval estimates to represent 50% confidence intervals.

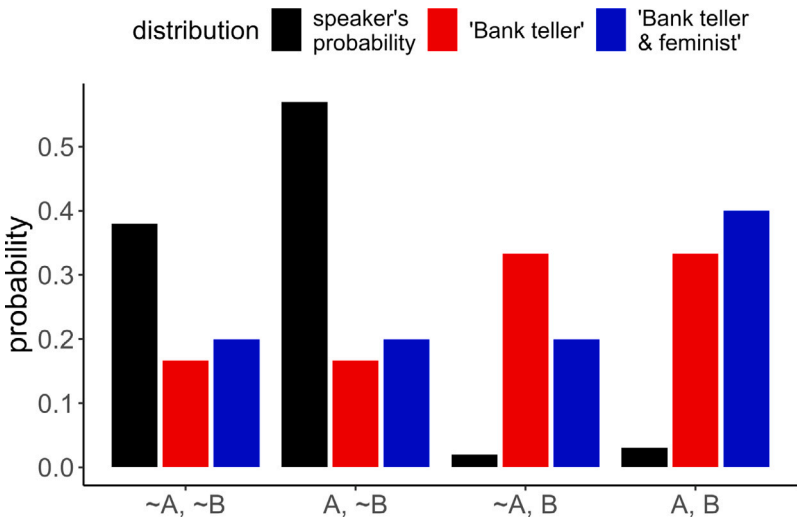

Fig. 19. Modeling Tversky & Kahneman's Linda case. Black bars represent the speaker's subjective distribution. A: Feminist, B: Bank teller. The guess 'Bank teller' suggests a probability distribution like the one in red. The guess 'Bank teller & Feminist' suggests a distribution like the one in blue. Although both guesses are poor approximations of the speaker's distribution, 'Bank teller & Feminist' is better than 'Bank teller', because it diverges less from the speaker's distribution (KL = .72 vs KL = .89). Here we used $Pr(A) = .6$, $Pr(B) = .05$, but the result is not sensitive to the specific parameters used, as long as Pr(Feminist) > Pr(Bank teller). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6.1.3. Extension fallacies

In a classic paper, Tversky and Kahneman (1983) documented a *conjunction fallacy* in intuitive judgment: people sometimes assign a higher probability to A&B than to A, in blatant violation of the extension rule of probability theory.[25] For example, Tversky and Kahneman (1983) gave participants the following description:

"Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations".

Participants tended to rate the statement 'Linda is a bank teller and is active in the feminist movement' as more probable than 'Linda is a bank teller'. We suggest that the *Distribution-encoding* standard might (at least partly) contribute to the conjunction fallacy. That is, people tend to make conjunction errors when the conjunction A&B is a better encoding of their subjective probability distribution than A. In the Linda case, 'Feminist and Bank Teller' (A&B) is a better guess than 'Bank Teller' (B), under plausible assumptions. We assume that the speaker implicitly considers a probability distribution over the following four possibilities:

- Linda is NOT a feminist and is NOT a bank teller. ($\neg A \neg B$)
- Linda is a feminist and is NOT a bank teller. ($A \neg B$)
- Linda is NOT a feminist and is a bank teller. ($\neg A B$)
- Linda is a feminist and is a bank teller. ($A B$)

Under our account, the guess 'Feminist and Bank teller' communicates that the fourth possibility (AB) is more likely than any of the other three possibilities. The guess 'Bank teller' communicates that the last two possibilities ($\neg AB$) and (AB) are more likely than the first two. For example, assuming that the speaker has subjective probabilities $Pr(A) = .6$ and $Pr(B) = .05$, and that $Pr(A)$ and $Pr(B)$ are independent, the speaker's distribution can be represented by the black bars in Fig. 19. Setting the γ parameter at $\gamma = 2$, the distributions that a listener would infer from 'Bank Teller' and 'Bank Teller and Feminist' are shown in red and blue.

In *absolute* terms, both statements make for bad guesses, because they imply that Linda is more likely than not to be a bank teller; a better guess would have been 'Linda is a feminist and is not a bank teller'. However, 'Bank teller and Feminist' is a good guess *relative* to 'Bank teller'; we suggest that this difference in the guess value of the two statements might play a key role in the corresponding difference in probability judgments.

People also violate the extension rule of probability theory when they judge that $Pr(A) > Pr(A \vee B)$, a mistake called the *disjunction* fallacy. Consider Danielle, a creative and introverted woman who enjoys reading. People judge that she is likely to be a Literature student, and rate this probability even higher than the probability that she is a Humanities student—even though membership in the first category entails membership in the second (Bar-Hillel & Neter, 1993).

Again, we suggest that the disjunction fallacy tends to arise when 'A' solves the *Distribution-encoding* problem better than 'A or B'. To illustrate, we consider a toy model of the Danielle case where the speaker entertains the following possibilities:

[25] The extension rule states that if S2 is a subset of S1, the probability of S2 cannot exceed that of S1.

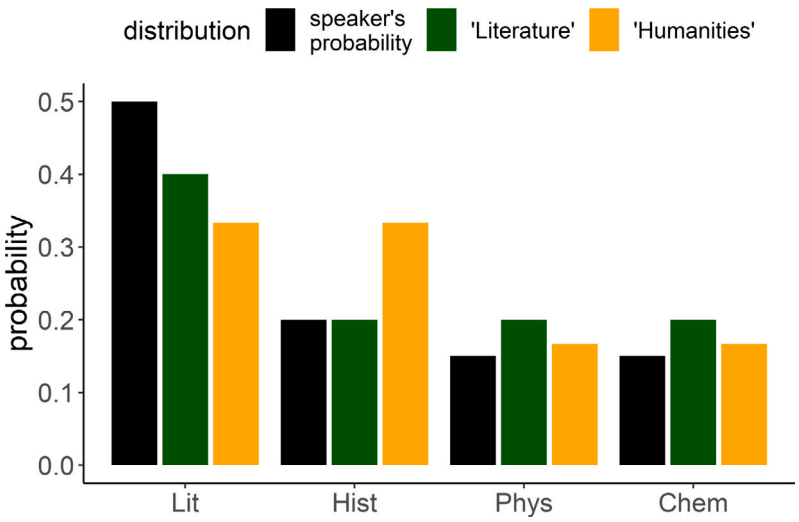

Fig. 20. Modeling the disjunction fallacy. The speaker' subjective distribution is in black. The guess 'Literature' suggests a probability distribution like the one in green. The guess 'Humanities' suggests a distribution like the one in orange. 'Literature' is a better approximation of the speaker's distribution than 'Humanities' (KL = .025 vs KL = .069). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- Danielle studies Literature
- Danielle studies History
- Danielle studies Engineering
- Danielle studies Chemistry

Guessing that Danielle is a Humanities student implies that she is more likely to be a Literature or History major than she is to be studying Engineering or Chemistry. Guessing that Danielle studies Literature implies that she is more likely to study Literature than any other major. If the speaker thinks that *Pr*(Literature) is high enough compared to *Pr*(History), the guess 'Literature' is a better approximation of his subjective probability distribution than 'Humanities', see Fig. 20.

There is of course already an extensive literature on the conjunction fallacy and related phenomena (e.g. Bar-Hillel & Neter, 1993; Busemeyer et al., 2011; Chung et al., 2023; Costello, 2009; Hertwig & Gigerenzer, 1999; Ludwin-Peery et al., 2020; Sablé-Meyer & Mascarenhas, 2022; Tentori et al., 2004; Tversky & Kahneman, 1983).[26] Conjunction errors probably have other causes besides those we are suggesting, and making a full case for the present proposal is beyond the scope of our paper.

We note that Dorst and Mandelkern (2021) present an account of the conjunction fallacy that is closely related to ours. They review the relevant empirical literature, and argue comprehensively that the pattern of people's conjunction errors is highly consistent with their accuracy-specificity trade-off model. Their arguments can also be used to support our own proposal. To give just one example (adapted to fit the details of our account), we predict that conjunction errors should substantially diminish when the question is presented in a *frequency format*. Suppose participants have to consider 100 people who fit Linda's description, and are asked to estimate the proportion of these people that are feminist bank tellers. This framing suggests that the relevant probability distribution is over the *number* of people who are feminist bank tellers (whereas the classical framing suggests that the relevant distribution is over Linda's *features*). A helpful guess in this context is one that is aligned with the speaker's estimate of the frequency of feminist bank tellers in the sample, and thus should be consistent with probability theory. If you think for example that out of 100 people that fit Linda's description, about 4 are likely to be feminist bank tellers, then '4' is a good guess. Indeed, frequency formats substantially diminish the rate of conjunction errors (Hertwig & Gigerenzer, 1999).

6.1.4. Implications for the rationality of judgment under uncertainty

There is a sense in which our perspective on judgment errors like the conjunction fallacy 'rationalizes' these errors: we suggest that these errors arise from the operation of a well-designed system. We attribute the error to a mis-application of this cognitive system outside of its proper domain. This mis-application is of course a mistake with potentially costly consequences. The conjunction fallacy is for example observed even in contexts where people have to make monetary bets on an outcome (Erceg & Galić, 2014; Nilsson & Andersson, 2010; Tversky & Kahneman, 1983). In this context, betting more money on A&B than A is clearly sub-optimal.

[26] Our proposal is perhaps most closely related to proposals according to which conjunction errors arise from conversational inferences. Our proposal is however distinct from previously proposed accounts along this line. For example, it has been suggested that participants infer that a speaker who says 'A' instead of 'A and B' means 'A and not B' (see Dulany & Hilton, 1991; Moro, 2009; Politzer & Noveck, 1991, for discussion). This is a different proposal than ours—for example on Fig. 19 the guess 'Bank teller' induces the listener to assign equal probability mass to both 'Bank teller and feminist' and 'Bank teller and not feminist'.

6.2. Scope of the work

Our argument is formulated at an abstract level—in David Marr's terms, it belongs to the computational level of analysis (Marr, 1982). We think that this abstraction is valuable, helping us focus on very general constraints on the shape of probabilistic reasoning. To re-iterate the core of our argument: Reasoning under uncertainty typically involves assessing the probability of many possible outcomes—reasoners must represent a *distribution*, rather than a single probability. As such, it makes sense that judgments about uncertain events would be designed to encode information about this distribution. These considerations apply, in principle, to any agent that must reason under uncertainty and create compressed representations of probability.

6.2.1. Verbal communication

At a less abstract level, one might be interested in how judgment under uncertainty works in more specific contexts, like verbal communication. We think that our proposal gives a useful starting point for thinking about these issues, although it does not provide a complete account.

In our information-theoretic framework, a guess provides an encoding of a distribution, which can then be decoded. In the context of communication a natural interpretation is that the speaker is encoding his subjective probability distribution in an utterance and the listener is doing the decoding. These kinds of dynamics can be formalized in models of *pragmatics*. For example the Rational Speech Acts framework models communication as recursive mindreading across a (virtual) hierarchy of speakers and listeners (RSA; Degen, 2023; Goodman & Frank, 2016; Zaslavsky et al., 2021). There is a deep connection between the RSA and information theory, as the RSA equations can be derived from information-theoretic principles (Wang et al., 2020; Zaslavsky et al., 2021). So, computational models of pragmatics like RSA can be seen as one possible implementation of our more general information-theoretic proposal.

The insight we provide is that for models of pragmatics to successful capture the way people talk about uncertainty, they will probably have to assume that the speaker's goal is to communicate his subjective probability distribution over the relevant facts or outcomes of interest. Indeed, existing applications of RSA to probabilistic reasoning make this assumption (e.g. Egré et al., 2023; Herbstritt & Franke, 2019; van Tiel et al., 2022).

To demonstrate the applicability of RSA to our phenomena of interest, we implemented an RSA model for our task in Study 2; see Appendix. The model assumes that the literal meaning of 'probably X' is about the probability of X (namely that $Pr(X)$ is above a certain threshold), but that the pragmatic speaker tries to communicate about his whole subjective probability distribution.[27] We find that the model is able to give a relatively good account of our data in Study 2, although its fit is not as good as that of our compression model.

We also find that the RSA model is more complicated to implement than our more abstract compression model, for example because it must represent distributions over distributions. This complexity can be a drawback, showing the comparative usefulness of our more abstract approach. At the same time, embracing that complexity will sometimes be necessary to model some situations that go beyond the assumptions of our simpler model. For example, speakers often aim to communicate their *belief* about a distribution, rather than the distribution itself (as our model assumes). Consider a student answering a teacher's question. The student's goal is not to change the teacher's belief about the correct answer, but to communicate something about his own beliefs.

Estimating the quality of a guess may also involve sophisticated causal inferences, for instance about whether the speaker possesses a good internal model of the relevant domain. Consider for instance the answer "8880 to 8885 meters" to the question "what is the height of Mount Everest?" (correct answer: 8849 m). If we interpret the guess as conveying a probability distribution narrowly centered on 8882.5 (the interval's midpoint), the guess assigns negligible probability to the correct answer. As such, the model we use in Study 4 would judge it a very bad guess. But it is tempting to say this is a good guess, for example because it is manifestly non-random: the guess would be unlikely to be in the correct ballpark if the speaker had absolutely no idea about the height of Mt Everest.

We have studied guesses in two different formats: disjunctions of possible outcomes (Studies 1 to 3) and numerical intervals (Study 4). There are of course many other ways that people express their uncertainty. They use for example vagueness, as in 'Around 30 people will show up at the party' (Egré et al., 2023). They also modulate the level of uncertainty in their guess, using words like *may*, *must*, *possibly* and *likely* (Herbstritt & Franke, 2019; Lassiter, 2010; Yalcin, 2007). Future work should investigate the applicability of our approach to these different formats, while integrating the insights of existing accounts of the semantics of uncertainty expressions (e.g. Alpert & Raiffa, 1982; Budescu et al., 2009; Budescu & Wallsten, 1995; Dhami & Mandel, 2022; Egré et al., 2023; Herbstritt & Franke, 2019; Kao et al., 2014; Lassiter, 2010; Meder et al., 2022; Wallsten & Budescu, 1983; Yalcin, 2007).

6.2.2. Computational cost

At an algorithmic level, computing the quality of a guess requires computing the target probability distribution, which is in many cases intractable. So people might need to approximate the distribution, using for example sampling-based methods (Bramley et al., 2017; Davis & Rehder, 2020; Vul et al., 2014; Zhu et al., 2020) or avoid this computation entirely and use heuristics.

[27] The assumption about the literal meaning of 'probably X' follows proposals from, e.g., Lassiter (2010), Yalcin (2007). We also make this choice to show that our key claim (that guesses ultimately are designed to encode a distribution rather than a single probability) is in principle consistent with a different assumption about the literal meaning of probability expressions.

6.3. Comparison with the accuracy/specificity trade-off hypothesis

Researchers have proposed that good guesses strike a trade-off between accuracy and specificity: they have a high probability but do not mention too many possible outcomes (Dorst & Mandelkern, 2021; Skipper, 2023; Yaniv & Foster, 1995). Here we provide the first (to our knowledge) empirical test of the formal model of the accuracy-specificity trade-off introduced by Dorst and Mandelkern (2021, see Studies 1 and 2). We also replicate an experiment by Yaniv and Foster (1995) that investigates if people prefer guesses that are accurate and specific. What are the implications of our results for the trade-off hypothesis?

On one hand, our data show that the trade-off hypothesis is generally a good descriptive account of people's judgments. In Study 4, we replicate Yaniv and Foster's finding that their formal model tracks people's intuitions closely. In Studies 1 and 2, Dorst and Mandelkern's formal model provides a good account of some of the general trends in the data, and in particular we find that many individual participants are best-fit by the model.

On the other hand, our data also suggest that the trade-off hypothesis paints at best an incomplete picture. In particular, the trade-off hypothesis predicts the impossibility of a U-shaped pattern between guess size and guess quality: for example if a short guess A is better than a longer guess B, then B should also be better than any longer guess C. But we observe systematic violations of this principle in Studies 1 and 2. People tend to exhibit U-shaped patterns of judgments, and they exhibit them in precisely those contexts where our information-theoretic account predicts that they should.

If the trade-off hypothesis is viewed as a descriptive account, it is not necessarily in tension with our computational-level theory. We expect that the function of guesses is to efficiently encode the speaker's subjective probability distribution, but the human mind is probably not implementing a fully optimal solution to this problem. Instead, people might use heuristics that provide a good enough approximation of that solution. It is possible that one such heuristic is to try to make guesses that are both accurate and specific: in general this strategy will lead to efficient communication of one's probabilistic beliefs. As such, our account can be seen as providing a rational explanation for why people might (sometimes) optimize an accuracy-specificity trade-off.

6.4. Related phenomena and future directions

One of the primary motivations for our proposal is that *Distribution-encoding* makes sense of many phenomena in probabilistic reasoning that defy explanation in terms of the *Probability-maximizing* standard. People's judgments might also diverge from *Probability-maximizing* for other reasons. First and most obviously, this divergence might sometimes be a symptom of genuine irrationality, or reflect the use of limited cognitive resources. Second, people sometimes guess in order to collect information, as in information-seeking games like Twenty Questions and Mastermind (Cheyette et al., 2023). In these contexts, guessing is equivalent to asking a question, and asking a question whose answer you already know is generally a bad strategy, so you should not make high-probability guesses.[28] Third, guesses that have low-probability are often considered impressive if successful, especially if we know something that others do not. For example, scientists often gain credit by making correct predictions that are surprising from the point of view of other existing theories, rather than correct-but-obvious predictions (Villarreal et al., 2023).

Our theory relies on pre-existing intuitions about what counts as a natural partition of the space of possibilities. In the urn example we use in Studies 1 to 3, the natural partition of the outcome space is {Red, Green, Yellow, Blue}. But if people viewed the relevant contrast to be {Red, not Red}, then our theory would make different predictions, for example judging that 'Not Red' is a better guess than 'Red' for the urn in Fig. 3a. One can think of the relevant partition as depending on an implicit 'question under discussion', suggesting connections between our approach and work on the semantics of questions (e.g. Koralus & Mascarenhas, 2013; Roberts, 2012).

Our modeling framework assumes a pre-existing, fixed set of constraints over the guesses that speakers can make. An interesting avenue for future research is to model agents that can choose at which level of precision to encode their probabilistic beliefs, and must navigate a trade-off between the costs and benefits of high-fidelity encoding (Berger, 2003; Sims, 2016).

Finally, our model often recommends making guesses that seem in some sense 'representative' of the probability distribution to be encoded. The idea that people use representativeness as a heuristic for probability judgment has a long history in cognitive science (Kahneman & Frederick, 2002; Kahneman & Tversky, 1972). Our framework offers a possible computational explanation for the pull that representative outcomes exert on probabilistic reasoning; future work could explore this possibility further.

7. Conclusion

To handle uncertainty about the world, the mind needs to represent probabilities. Probabilistic reasoning typically involves representing the probability of not only one, but many possible outcomes of an event: people must at some level represent *probability distributions*. As such, a natural idea is that when we think and talk about uncertain facts or events, we implicitly encode our subjective probability distribution over the relevant set of possible outcomes. We show that this natural hypothesis can qualitatively and quantitatively account for many aspects of judgment under uncertainty.

[28] Speculatively, the logic of question-asking could also explain why people tend to make interval guesses that are 50% likely to contain the correct answer. When asking yes/no questions, it is typically optimal to ask questions for which one is maximally uncertain.

CRediT authorship contribution statement

Tadeg Quillien: Writing – original draft, Formal analysis, Data curation, Conceptualization. **Neil Bramley:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition. **Christopher G. Lucas:** Writing – review & editing, Supervision, Resources, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

For helpful discussion and feedback, we thank Dan Lassiter, Hugo Mercier, Bonan Zhao, as well as audience members at the Evolution and Social Cognition Team at ENS Paris and at the meetings of the Cognitive Science and Human Behavior and Evolution Societies. We also thank the editor and two anonymous reviewers. This work was supported by EPSRC New Investigator Grant (EP/T033967/1) to NB and CGL.

Appendix

A.1. Listener model, studies 1-2

Denote the probability of an outcome not mentioned in the guess as p. Then the probability of an outcome mentioned in the guess is γp. Given that the probabilities of all possible outcomes must sum to 1, we have:

$$n_g \gamma p + n_{\neg g} p = 1$$

where n_g is the number of possible outcomes mentioned in the guess, and $n_{\neg g}$ the number of possible outcomes not mentioned in the guess. It follows that the probability of an outcome unmentioned in the guess is:

$$p = \frac{1}{n_{\neg g} + \gamma n_g}$$

from which we also get that the probability of an outcome mentioned in the guess is:

$$\gamma p = \frac{\gamma}{n_{\neg g} + \gamma n_g}$$

A.2. Proof that the trade-off model cannot predict U-shaped patterns

In Studies 1 and 2, we find that people's judgments sometimes exhibit a 'U-shaped' relationship between guess size and quality. For example, in Study 2 many more participants make a size-1 or size-3 than a size-2 guess when looking at an urn with profile [5,3,3,1]. The question arises whether the trade-off model could in principle account for this phenomenon. Here we show that it cannot.

Of course the model could trivially predict a U-shaped pattern if we consider guesses that are manifestly irrational: for example one can show that a size-1 and a size-3 guess are both better than a size-2 guess if the size-2 guess only mentions the two least frequent colors in the urn. As such we restrict our analysis to 'Pareto-optimal' guesses. A guess is optimal in that sense if it is impossible to construct a guess that is more specific but not less accurate than the current guess, or more accurate but not less specific. In the context of our urns-and-balls paradigm, a guess is Pareto-optimal if there is no other color in the current urn that is strictly more frequent than one of the colors mentioned in the guess. (Study 1 only featured Pareto-optimal guesses, and 98% of the guesses produced in Study 2 were Pareto-optimal.)

Remember that the trade-off model computes the value of a guess as:

$$V(g) = P(g)J^{S(g)} \tag{6}$$

Where the specificity $S(g)$ is the proportion of colors in the urn that are not mentioned in g. Consider three Pareto-optimal guesses a, b, c, where c mentions more colors than b and b mentions more colors than a. We will show that if $V(a) > V(b)$, then $V(b) > V(c)$. For example, if a size-1 guess is better than a size-2 guess, then a size-2 guess is necessarily better than a size-3 guess.

Proof. For conciseness we write $P_x = Pr(x)$ and $S_x = S(x)$. It will be useful to denote the difference in probability and specificity between guess a and guess b as follows:

$$\Delta P = P_b - P_a \tag{13}$$

$$\Delta S = S_b - S_a \tag{14}$$

We start with the assumption (to be relaxed later) that there is a gap of size 1 between the successive guesses, i.e. that a, b and c mention n, $n+1$ and $n+2$ colors respectively. It follows that the difference in specificity between them is equal: $S_c - S_b = S_b - S_a = \Delta S$. Given this assumption, $V(c)$ is maximized if the probability increase in going from b to c is the same as the probability increase in going from a to b, i.e. if $P_c - P_b = P_b - P_a = \Delta P$. This is because Pareto-optimality implies that $P_c - P_b$ cannot exceed ΔP.[29] Therefore we assume that $P_c - P_b = P_b - P_a = \Delta P$: If we can show that $V(b) > V(c)$ holds under this assumption, then it holds for all other possible values of $P_c - P_b$.

The inequality $V(b) > V(c)$ is equivalent to:

$$P_b J^{S_b} > P_c J^{S_c} \tag{15}$$

i.e.:

$$(P_a + \Delta P) J^{S_a + \Delta S} > (P_a + 2\Delta P) J^{S_a + 2\Delta S} \tag{16}$$

i.e.:

$$\frac{P_a + \Delta P}{P_a + 2\Delta P} > \frac{J^{S_a + 2\Delta S}}{J^{S_a + \Delta S}} \tag{17}$$

i.e.:

$$\frac{P_a + \Delta P}{P_a + 2\Delta P} > J^{\Delta S} \tag{18}$$

On the other hand, we already know, from $V(a) > V(b)$, that:

$$P_a J^{S_a} > P_b J^{S_b} \tag{19}$$

i.e.:

$$P_a J^{S_a} > (P_a + \Delta P) J^{S_a + \Delta S} \tag{20}$$

i.e.:

$$\frac{P_a}{P_a + \Delta P} > J^{\Delta S} \tag{21}$$

Putting Eqs. (18) and (21) together, we realize that $V(b) > V(c)$ is true as long as $\frac{P_a + \Delta P}{P_a + 2\Delta P} > \frac{P_a}{P_a + \Delta P}$, which the quotient rule of calculus shows to be always true.

We have just shown that if a, b and c mention n, $n+1$ and $n+2$ colors, then $V(a) > V(b)$ implies $V(b) > V(c)$. We now show that the result generalizes to any (a, b, c) triplet of Pareto-optimal guesses where a has more colors than b and b more colors than c.

Note first that if a guess g_n with n colors has higher value than a guess with $n+1$ colors, then it follows from the result derived above (by induction) that the sequence $V(g_n), V(g_{n+1}), V(g_{n+2}), \ldots$ is monotone decreasing.

Now we show that $V(a) > V(b)$ implies $V(b) > V(c)$. It follows from $V(a) > V(b)$ that every Pareto-optimal guess whose size is between a and b must have higher value than b.[30] Therefore the Pareto-optimal guess g_{b-1} whose size is 1 less than the size of b has $V(g_{b-1}) > V(b)$. Therefore the sequence $V(g_{b-1}), V(b), V(g_{b+1}), \ldots, V(c)$ is monotone decreasing, so $V(b) > V(c)$. □

A.3. RSA model in Study 2

We consider a model within the Rational Speech Act framework (RSA; Degen, 2023; Franke & Jäger, 2016; Goodman & Frank, 2016), inspired by an application of the framework to judgments of probability by Herbstritt and Franke (2019). The model relies on a threshold semantics for probability statements: the *literal* meaning of "probably X" is $Pr(X) \geq \theta$, where the threshold θ is a free parameter. The literal meaning of "the ball will probably be Red or Green" is for example $Pr(\text{Red} \vee \text{Green}) \geq \theta$. The model assumes that the listener interprets the guess according to its literal meaning, but that the speaker makes the guess that gets the listener to infer a probability distribution that is as close as possible to the speaker's subjective probability distribution over possible outcomes. In that respect, the RSA model is quite close in spirit to the compression model, and we see it essentially as another potential implementation of the general idea underlying our rational analysis.

When hearing the guess "the outcome will probably be X", the listener concludes that the speaker's distribution is such that $Pr(X) > \theta$, and uses this information to update her prior belief about the speaker's distribution, via Bayes' rule. The listener's prior

[29] Pareto-optimality implies that the colors in b are more or equally frequent as the colors not in b. Therefore the extra color in c is either as frequent or less frequent than the colors in b, so adding that color to the guess cannot result in a gain in probability greater than ΔP.

[30] Any other pattern would violate the result we derived earlier, because it would imply that there is a local U-turn somewhere along a and b. I.e. there would have to be three successive guesses g_1, g_2, g_3 (with size n, $n+1$ and $n+2$) somewhere between a and b with $V(g_1) > V(g_2)$ and $V(g_2) \leq V(g_3)$, which we saw is impossible.

belief about the speaker's distribution is a distribution over distributions (i.e. a meta-distribution). That is, the speaker might think that the frequencies of colors in the box are [Red: 5, Green: 3, Blue: 3, Yellow: 1] or he might think that they are [Red: 2, Green, 2, Blue: 2, Yellow: 6], or any other combination.

We assume that the listener has a flat prior over distributions, such that every possible combination of colors in the urn is equally likely (subject to the constraints that there are at least 1 ball of each color). When the listener hears the guess, she simply eliminates from her meta-distribution all those that are incompatible with the guess, i.e. those distributions where $Pr(X) < \theta$. For example if the guess is "probably Red or Green", then the listener eliminates distributions where $Pr(\text{Red}) + Pr(\text{Green}) < \theta$.

Finally, the listener computes her posterior belief over the probability that a randomly drawn ball would be of a certain color by marginalizing over all the remaining distributions in her meta-distribution.

The speaker can anticipate the inference that the listener will draw from a guess, and computes the quality of a guess as inversely related to the K-L divergence of the distribution inferred by the listener relative to the speaker's subjective probability distribution (Eq. (5)). Like the other models we consider in Study 2, the RSA model assumes that the speaker's distribution might be subject to slight distortions (Eq. (4)) and that the probability of guess production is determined by a softmax function (Eq. (7)).

A.3.1. Results

We fit the model to the human data in Study 2. We obtain $\alpha = 1.96$, $\beta = 14.93$, and $\theta = .75$ as best-fitting parameters. The RSA model has a good fit to the data : its predictions are correlated with human choice proportions at $r(193) = .907$, $p < .001$ (see Fig. A.3) and its AIC is 2115.

The model is however unable to explain some of the subtle features of the data. Similar to the trade-off model, it predicts that there is one optimal guess size for a given urn, and that the quality of a guess diminishes monotonically as a function of its distance from the optimal guess size—as such it cannot account for the U-shaped patterns in the data. However, we anticipate that the model might be able to capture these patterns if we allowed higher levels of recursive mindreading—for example by letting a Level-2 speaker anticipate the inferences of a pragmatic listener. We leave an exhaustive exploration of these possibilities for future research.

A.4. Threshold model in Study 2

According to the threshold model, participants include a color in a guess if the number of balls of that color is at or above a given threshold θ. For example, if $\theta = 2$, people include in their guess all colors that are present in at least two balls in the current urn – so, for the urn profile [6,3,2,1], people include the three most frequent colors in their guess (because there are three colors with 2 balls or more), but they only include one color for the urn profile [9,1,1,1].

To make the model stochastic, we assume that the quality of a guess is 1 if the guess includes all and only colors with θ balls or more, and $\frac{1}{1+L}$ otherwise, where L is the number of colors that are either included in the guess whereas they should not be (because $n_{\text{color}} < \theta$), or not included in the guess whereas they should be (because $n_{\text{color}} \geq \theta$). The probability of each guess is then given by passing these values to a soft-max function, as specified in the main text (Eq. (7)).

A.5. Figures for alternative models, Study 2

See Figs. A.1–A.3.

A.6. List of trials, Study 3

See Table A.1 and Fig. A.4.

A.7. Original data from Yaniv & Foster (1995)

Unfortunately, the original data for Yaniv and Foster (1995) have been lost (Yaniv, personal communication), but we can analyze data from the sample items displayed in the paper (reproduced in Table A.2). Specifically, there are available data for three items from the preliminary experiment (N=20), and for five items from Experiment 3 (N=30). For each item, we have the guesses made by assistants A and B, the correct answer, as well as the proportion of participants who picked assistant A as the best. We analyzed data from these eight items together, by finding the best-fitting values of k (the scaling constant that transforms the width of the guess into a standard deviation) and β (the temperature parameter) for the compression model. We find that for best-fitting values $k = 1$, $\beta = 1.3$, the correlation between the preference of the compression model and participants' choices is $r(6) = .96$, $p < .001$, see Fig. A.5. We also find that Yaniv & Foster's original accuracy-specificity trade-off model has a similarly good fit to these data ($r(6) = .98$, $p < .001$).

A.8. List of items from Yaniv & Foster (1995)

See Table A.2.

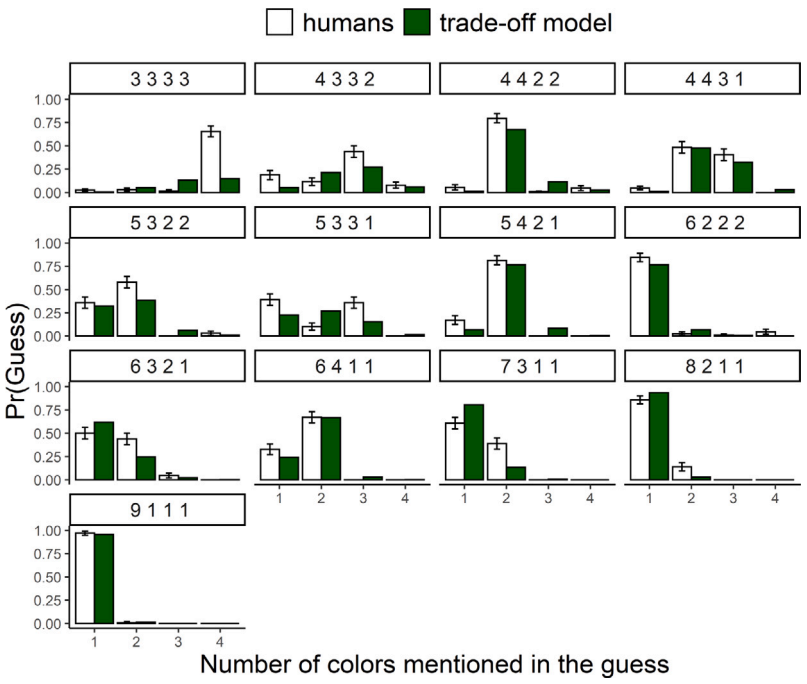

Fig. A.1. Study 2: Proportion of human participants making a given guess, and trade-off model probability for that guess, as a function of urn profile and guess size, for guesses lying on the Pareto frontier. Note: for some urn profiles, several different guesses can correspond to the same guess size. When this is the case, we compute the average choice probability across all these guesses. Note that probabilities do not necessarily sum to 1, because guesses lying outside the Pareto frontier are not represented.

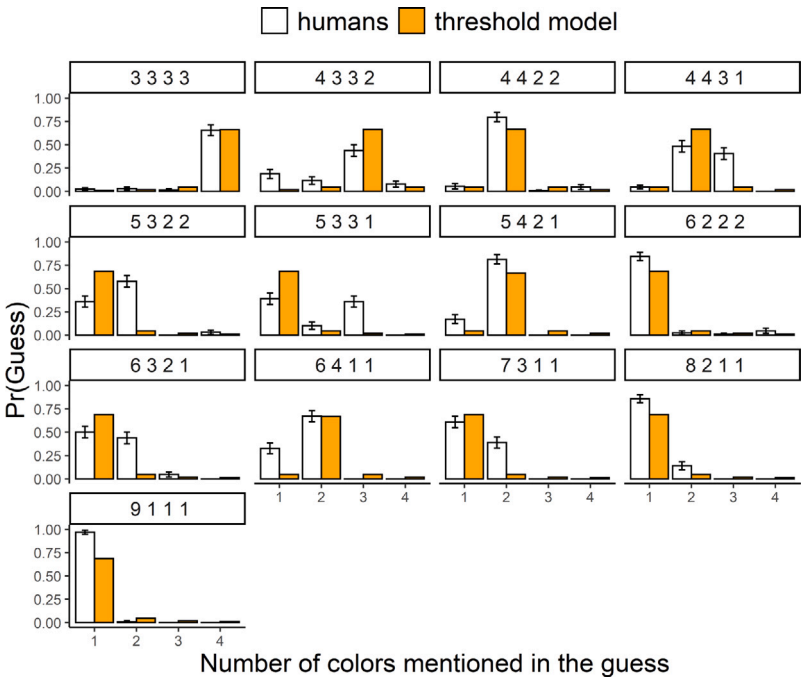

Fig. A.2. Study 2: Proportion of human participants making a given guess, and threshold model probability for that guess, as a function of urn profile and guess size, for guesses lying on the Pareto frontier. Note: for some urn profiles, several different guesses can correspond to the same guess size. When this is the case, we compute the average choice probability across all these guesses. Note that probabilities do not necessarily sum to 1, because guesses lying outside the Pareto frontier are not represented.

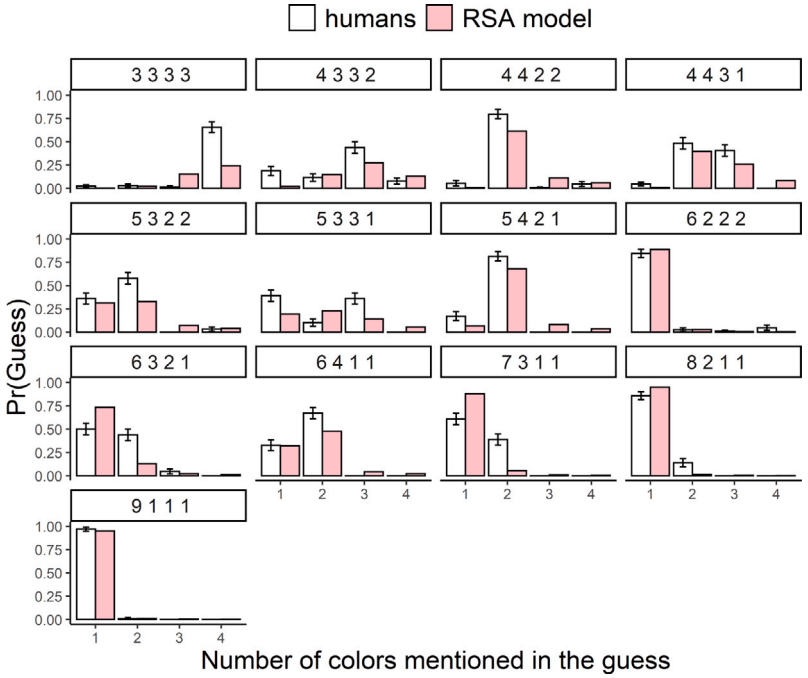

Fig. A.3. Study 2: Proportion of human participants making a given guess, and RSA model probability for that guess, as a function of urn profile and guess size, for guesses lying on the Pareto frontier. Note: for some urn profiles, several different guesses can correspond to the same guess size. When this is the case, we compute the average choice probability across all these guesses. Note that probabilities do not necessarily sum to 1, because guesses lying outside the Pareto frontier are not represented.

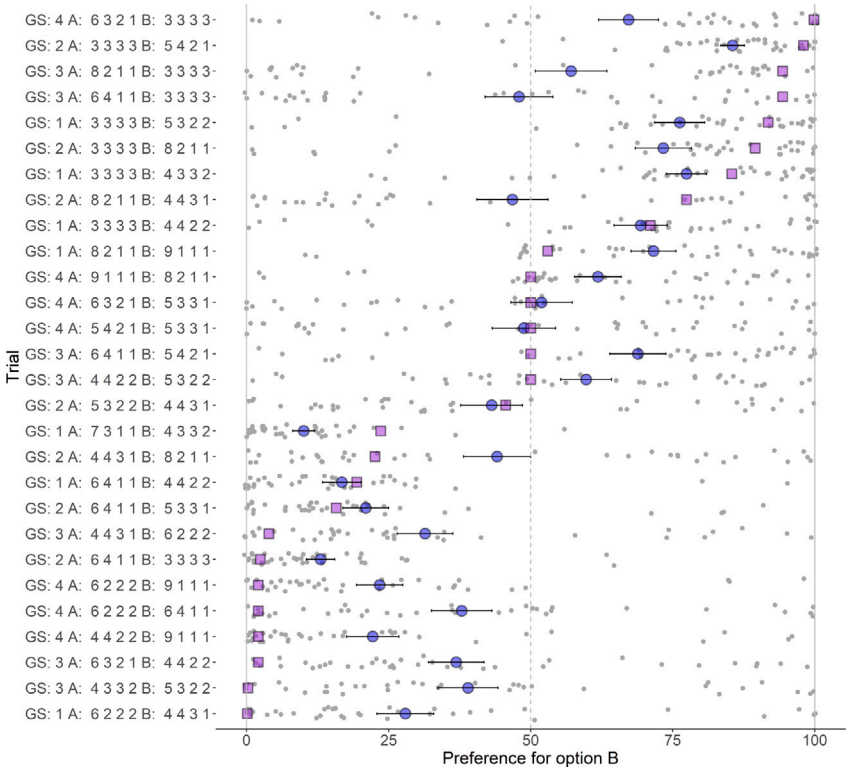

Fig. A.4. Detailed choice data for Study 3, along with predictions of the pragmatic listener model with empirically-derived likelihood. Gray dots represent individual ratings, blue circles represent mean human ratings, and purple squares are model predictions. Error bars represent the standard error of the mean. GS: guess size; A: profile for urn A; B: profile for urn B. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table A.1
List of trials, along with predictions of the pragmatic listener model (calibrated with empirically-derived likelihoods) and mean participant ratings, Study 3. Number lists represent the profile of an urn: for example, an urn labeled [9,1,1,1] has 9 balls of one color, and one ball each of the other colors. Note that which box was labeled as 'A' or 'B' was counter-balanced across participants.

urn A	urn B	guess size	$Pr(B)$, empirical model	$Pr(B)$, mean human rating
3 3 3 3	4 3 3 2	1	0.853907135	77.41666667
3 3 3 3	4 4 2 2	1	0.71043771	69.33333333
3 3 3 3	5 3 2 2	1	0.917860554	76.22222222
6 2 2 2	4 4 3 1	1	0.001182383	27.94444444
6 4 1 1	4 4 2 2	1	0.193815064	16.77777778
7 3 1 1	4 3 3 2	1	0.235956814	10.05555556
8 2 1 1	9 1 1 1	1	0.529881839	71.61111111
3 3 3 3	5 4 2 1	2	0.979972896	85.52777778
3 3 3 3	8 2 1 1	2	0.894944708	73.33333333
4 4 3 1	8 2 1 1	2	0.225877193	44.08333333
5 3 2 2	4 4 3 1	2	0.455965242	43.11111111
6 4 1 1	3 3 3 3	2	0.024111675	13.02777778
6 4 1 1	5 3 3 1	2	0.157721796	20.97222222
8 2 1 1	4 4 3 1	2	0.774122807	46.77777778
4 3 3 2	5 3 2 2	3	0.002275313	38.97222222
4 4 2 2	5 3 2 2	3	0.5	59.75
4 4 3 1	6 2 2 2	3	0.039221469	31.41666667
6 3 2 1	4 4 2 2	3	0.020460358	36.88888889
6 4 1 1	3 3 3 3	3	0.943262411	47.94444444
6 4 1 1	5 4 2 1	3	0.5	68.86111111
8 2 1 1	3 3 3 3	3	0.943262411	57.11111111
4 4 2 2	9 1 1 1	4	0.020460358	22.19444444
5 4 2 1	5 3 3 1	4	0.5	48.80555556
6 2 2 2	6 4 1 1	4	0.020460358	37.86111111
6 2 2 2	9 1 1 1	4	0.020460358	23.41666667
6 3 2 1	3 3 3 3	4	0.99848082	67.22222222
6 3 2 1	5 3 3 1	4	0.5	51.91666667
9 1 1 1	8 2 1 1	4	0.5	61.83333333

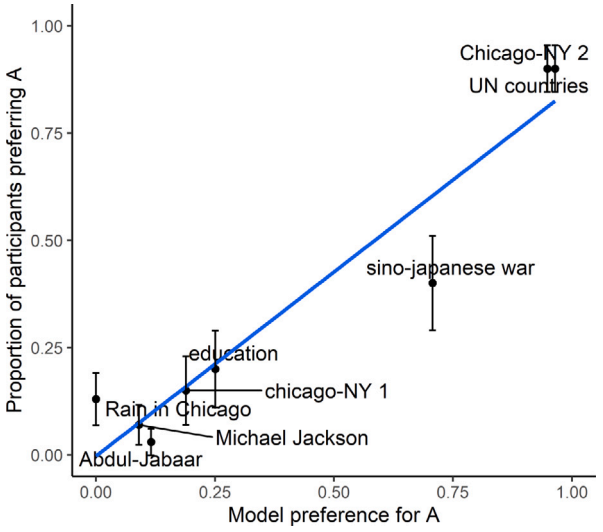

Fig. A.5. Proportion of participants preferring the guess from assistant A, as a function of the compression model's preference for A, in items from Yaniv and Foster (1995) for which data were available. Error bars represent standard errors. Regression line is shown in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table A.2
Experimental stimuli in Yaniv and Foster (1995). For example, to the request "Average number of rainy days in Chicago?", assistant A answered "160 to 165" while the other assistant answered "140 to 180".

Label	Text	Ground truth	Lower A	Upper A	Lower B	Upper B
Rain in Chicago	Average number of rainy days in Chicago?	130	160	165	140	180
Michael Jackson	Amount of money received by Michael Jackson in 1987 to star in a series of Pepsi commercials?	15 million	1 million	20 million	12 million	14 million
Abdul-Jabaar	Total number of points scored by Kareem Abdul-Jabaar in 19 years of playing basketball (as of 1987–1988 season)?	37 639	30 000	45 000	37 000	40 000
Chicago–NY 1	Air distance between Chicago and New York?	713	800	850	600	800
education	Amount of money spent on education by the US federal government in 1987?	22.5 billion	20 billion	40 billion	18 billion	20 billion
Sino–japanese war	Date the Sino–Japanese War began?	1894	1870	1890	1875	1925
UN countries	Number of United Nation member countries?	159	140	150	50	300
Chicago–NY 2	Air distance between Chicago and New York?	713	730	780	700	1500

Table A.3
Experimental stimuli in Study 4. For example, to the request "Year Charles Darwin was born?", one assistant answered "1825 to 1835" while the other assistant answered "1780 to 1800". Assistants were identified by the letters A and B (which of assistant X and Y got called A or B was randomly determined for each item and participant).

Label	Text	Ground truth	Lower X	Upper X	Lower Y	Upper Y
darwin	Year Charles Darwin was born?	1809	1825	1835	1780	1800
polio	Year the polio vaccine was invented?	1952	1970	1980	1920	1945
EU	Number of member states in the European Union?	27	15	20	30	40
UN	Number of member states in the United Nations?	193	40	300	165	185
boston	Number of people living in Boston?	684 000	610 000	640 000	550 000	700 000
japan	Number of people living in Japan?	126	100	110	135	170
airports	Number of public airports in the United States?	5217	5000	20 000	4000	4800
everest	Height of Mount Everest?	8849	9000	12 000	8000	8500
amazon	Length of the Amazon river?	6400	6600	12 000	6000	7000
transatlantic	Date of the first transatlantic flight?	1927	1930	1970	1915	1923
Sino–japanese war	Date the Sino–Japanese War began?	1894	1880	1910	1890	1970
Chicago rain	Average number of rainy days in Chicago?	130	140	160	10	145
miami march temperature	Average temperature in Miami in March?	22	25	30	15	45

(*continued on next page*)

Table A.3 (*continued*).

Label	Text	Ground truth	Lower X	Upper X	Lower Y	Upper Y
Chicago–NYC	Air distance between Chicago and New York City?	713	740	800	700	2000
education budget	Annual budget of the US department of education?	68	75	120	95	110
LeBron James	Number of points scored by LeBron James in his entire NBA career?	37 024	32 000	40 000	20 000	35 000
Fed chairman	Monthly salary of the chairman of the U.S. Federal Reserve?	16 900	18 000	22 000	4000	35 000
horse	Average gestation length of a horse?	11	7	9	12	30
mozart	Age of Mozart when he died?	35	20	30	44	49
kubrick	Number of movies directed by Stanley Kubrick?	13	10	16	7	11

A.9. List of items in Study 4

See Table A.3.

Data availability

All data, modeling and analysis scripts have been made available on the Open Science Framework at https://osf.io/wz649.

References

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In *Judgment under uncertainty: heuristics and biases*.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Attneave, F. (1959). *Applications of information theory to psychology: A summary of basic concepts, methods, and results*. Henry Holt.

Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments.. *Journal of Personality and Social Psychology*, *65*(6), 1119.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Berger, T. (2003). Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301.

Budescu, D. V., Broomell, S., & Por, H.-H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science*, *20*(3), 299–308.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In *Psychology of learning and motivation*: *vol. 32*, (pp. 275–318). Elsevier.

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*(2), 193.

Cesarini, D., Sandewall, Ö., & Johannesson, M. (2006). Confidence interval estimation tasks and the economics of overconfidence. *Journal of Economic Behavior and Organization*, *61*(3), 453–470.

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, *37*(6), 1171–1191.

Cheyette, S. J., Callaway, F., Bramley, N. R., Nelson, J. D., & Tenenbaum, J. (2023). People seek easily interpretable information. In *Proceedings of the annual meeting of the cognitive science society*: *vol. 45*, (45).

Chung, W., Dorst, K., Mandelkern, M., & Mascarenhas, S. (2023). The conjunction fallacy: Confirmation or relevance?.

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. *Cognition*, *50*(1–3), 41–77.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*(1), 1–73.

Costello, F. (2009). How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, *22*(3), 213–234.

Costello, F., & Watts, P. (2014). Surprisingly rational: probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463.

Costello, F., & Watts, P. (2018). Invariants in probabilistic reasoning. *Cognitive Psychology*, *100*, 1–16.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412.

Davis, Z. J., & Rehder, B. (2020). A process model of causal reasoning. *Cognitive Science*, *44*(5), Article e12839.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540.

Dhami, M. K., & Mandel, D. R. (2022). Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences*.

Dorst, K., & Mandelkern, M. (2021). Good guesses. *Philosophy and Phenomenological Research*, http://dx.doi.org/10.1111/phpr.12831.

Dulany, D. E., & Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, *9*(1), 85–110.

Egré, P., Spector, B., Mortier, A., & Verheyen, S. (2023). On the optimality of vagueness: "around", "between" and the gricean maxims. *Linguistics and Philosophy*, 1–56.
Erceg, N., & Galić, Z. (2014). Overconfidence bias and conjunction fallacy in predicting outcomes of football matches. *Journal of Economic Psychology*, *42*, 52–62.
Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *367*(1594), 1280–1286.
Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, *14*(3), 195–200.
Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift Für Sprachwissenschaft*, *35*(1), 3–44.
Gagie, T. (2006). Compressing probability distributions. *Information Processing Letters*, 133–137.
Gershman, S. J. (2021). The rational analysis of memory. In *Oxford Handbook of Human Memory*. Oxford University Press Oxford, UK.
Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, *2*(1), 83–115.
Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size memory reporting.. *Journal of Experimental Psychology: General*, *131*(1), 73.
Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.
Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.
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. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.
Herbstritt, M., & Franke, M. (2019). Complex probability expressions & higher-order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data. *Cognition*, *186*, 50–71.
Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy'revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*(4), 275–305.
Holguin, B. (2022). Thinking, guessing and believing. *Philosopher's Imprint*, 1–34.
Icard, T. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, *7*, 863–903.
Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge University Press.
Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*(4), 201–214.
Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *25*(4), 1038.
Juslin, P., Winman, A., & Hansson, P. (2007). The naive intuitive statistician: A naive sampling model of intuitive confidence intervals. *Psychological Review*, *114*(3), 678.
Kahneman, D., & Frederick, S. (2002). *Representativeness revisited: Attribute substitution in intuitive judgment*: *vol. 49*, (49–81), (p. 74).
Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge.
Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.
Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*(2), 143–157.
Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.
Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216–247.
Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.
Koralus, P., & Mascarenhas, S. (2013). The erotetic theory of reasoning: bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives*, *27*, 312–365.
Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty.. *Journal of Experimental Psychology: General*, *136*(3), 430.
Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.
Laplace, P. S. (1820). *Théorie analytique des probabilités*: *vol. 7*, Courcier.
Lassiter, D. (2010). Gradable epistemic modals, probability, and scale structure. In *Semantics and linguistic theory*: *vol. 20*, (pp. 197–215).
Lassiter, D. (2011). *Measurement and modality: The scalar basis of modal semantics* (Ph.D. thesis), New York University.
Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, *31*(12), 1602–1611.
MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.
McKenzie, C. R., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes*, *107*(2), 179–191.
Meder, B., Mayrhofer, R., & Ruggeri, A. (2022). Developmental trajectories in the understanding of everyday uncertainty terms. *Topics in Cognitive Science*, *14*(2), 258–281.
Moore, D. A. (2022). Overprecision is a property of thinking systems. *Psychological Review*.
Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. In *The wiley blackwell handbook of judgment and decision making*: *vol. 2*, (pp. 182–209). Wiley Online Library.
Moro, R. (2009). On the nature of the conjunction fallacy. *Synthese*, *171*, 1–24.
Nilsson, H., & Andersson, P. (2010). Making the seemingly impossible appear possible: Effects of conjunction fallacies in evaluations of bets on football games. *Journal of Economic Psychology*, *31*(2), 172–180.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.
Politzer, G., & Noveck, I. A. (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research*, *20*, 83–103.
Quillien, T., Tooby, J., & Cosmides, L. (2023). Rational inferences about social valuation.. *Cognition*.
Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*, 1–69.
Russo, J., & Shoemaker, P. (1992). Managing overconfidence. *Management Review*, 7–17.
Sablé-Meyer, M., & Mascarenhas, S. (2022). Indirect illusory inferences from disjunction: a new bridge between deductive inference and representativeness. *Review of Philosophy and Psychology*, *13*(3), 567–592.
Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.
Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.
Skipper, M. (2023). Good guesses as accuracy-specificity tradeoffs. *Philosophical Studies*.
Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(2), 299.
Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*: *vol. 142*, Citeseer.
Teigen, K. H. (1988). When are low-probability events judged to be 'probable'? Effects of outcome-set characteristics on verbal probability estimates. *Acta Psychologica*, *67*(2), 157–174.
Teigen, K. H., & Jørgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *19*(4), 455–475.

Teigen, K. H., Juanchich, M., & Løhre, E. (2022). What is a "likely" amount? Representative (modal) values are considered likely even when their probabilities are low. *Organizational Behavior and Human Decision Processes*, *171*, Article 104166.
Teigen, K. H., & Keren, G. (2003). Surprises: Low probabilities or high contrasts? *Cognition*, *87*(2), 55–71.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.
Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science*, *28*(3), 467–477.
van Tiel, B., Sauerland, U., & Franke, M. (2022). Meaning and use in the expression of estimative probability. *Open Mind*, *6*, 250–263.
Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. arXiv preprint.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.. *Psychological Review*, *90*(4), 293.
Villarreal, M., Etz, A., & Lee, M. D. (2023). Evaluating the complexity and falsifiability of psychological models. *Psychological Review*.
Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.
Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.
Wallsten, T. S., & Budescu, D. V. (1983). State of the art—Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, *29*(2), 151–173.
Wang, P., Wang, J., Paranamana, P., & Shafto, P. (2020). A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, *33*, 17582–17593.
Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology*, *75*(6), 1411.
Yalcin, S. (2007). Epistemic modals. *Mind*, *116*(464), 983–1026.
Yalcin, S. (2010). Probability operators. *Philosophy Compass*, *5*(11), 916–937.
Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, *124*(4), 424.
Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, *10*(1), 21–32.
Zaslavsky, N., Hu, J., & Levy, R. (2021). A rate–distortion view of human pragmatic reasoning. In *Proceedings of the society for computation in linguistics 2021* (pp. 347–348).
Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719.
Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. N. (2023). The autocorrelated Bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychological Review*.