# Guesses as compressed probability distributions

**Tadeg Quillien**

School of Informatics

**Neil Bramley**

Department of Psychology

**Christopher Lucas**

School of Informatics

University of Edinburgh

## Author Note

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/?view_only=8d7019ee2b8d456d8c3c9b29049b75aa.

Correspondence concerning this article should be addressed to Tadeg Quillien, School of Informatics, University of Edinburgh. E-mail: tadeg.quillien@gmail.com

**Abstract**

People often make judgments about uncertain facts and events, for example 'Germany will win the world cup'. Here we present a rational analysis of these judgments: we argue that a guess functions as a compressed encoding of the speaker's *subjective probability distribution* over relevant possibilities. So, a statement like 'X will happen' encodes information not only about the probability of X but also, implicitly, about the probability of other possible outcomes. 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 it might shed light on apparently sub-optimal patterns of judgment such as the conjunction fallacy.

*Keywords:* Social cognition; computational modeling; probability; judgment under uncertainty; information theory

## Guesses as compressed probability distributions

### Introduction

People often make judgments about uncertain facts and events. Following Holguin (2022), we will call these judgments 'guesses'.[1] Some guesses are intuitively better than others. For instance, in the following pairs, (a) is better than (b):

(1) a. This fair coin will land on Heads.

b. This fair coin will land on its Edge.

(2) a. (looking at a small bird through binoculars) This might be a sparrow.

b. (looking at a small bird through binoculars) This might be a sparrow or an ostrich.

In both cases, a speaker would be more likely to utter (a) than (b), and a listener who hears (b) would find it less natural, or more misleading than (a). This suggests that there are implicit normative criteria that govern guessing. In this paper we are interested in these criteria: what makes a guess good?[2]

We are interested in these criteria at a fairly abstract level. There is a large literature on the cognitive processes involved in judgments under uncertainty (e.g. Tversky

[1] Note that our use of the term differs somewhat from its commonsense usage: for example a judgment might count as a guess under our framework even if the speaker is highly confident that it is the correct answer.

[2] To clarify the scope of this paper, it is helpful to contrast two distinct contexts in which guessing occurs. In many decision-making contexts—for instance when betting on a horse race or answering questions in a game show—there is a clearly-defined goal, like earning money or scoring points. In these cases the normative criteria for guessing are well-established: you should make the bet that maximizes your subjective expected utility (Savage, 1954). But there are many contexts—like telling your friend how many guests might show up at the party—without a clearly-defined metric for what constitutes success. In this paper we are interested in the norms that regulate guessing in the latter kind of context.

& Kahneman, 1983; Gigerenzer, 1991; Cosmides & Tooby, 1996; Griffiths & Tenenbaum, 2006; Oaksford & Chater, 2007; Juslin et al., 2007; Johnson-Laird et al., 2015; Zhu et al., 2020), and how these judgments are expressed linguistically (e.g. Budescu & Wallsten, 1995; Yalcin, 2007; Herbstritt & Franke, 2019; Meder et al., 2022; Lassiter, 2010; Alpert & Raiffa, 1982; Wallsten & Budescu, 1983; Budescu et al., 2009; Kahneman & Tversky, 1982; Cesarini et al., 2006; Dhami & Mandel, 2022). Here we abstract over many questions that arise in this literature — questions about the reasoning heuristics that people use, or the way that particular words modulate the meaning of a guess, for example. Rather, we are interested about whether anything general can be said about the logic of guessing.

Consider first a very simple hypothesis: You should make guesses that have a high chance of being correct. Under this view, you should guess that 'Germany is likely to win the world cup' if your subjective probability that Germany will win is high. Probability theory is widely seen as a normative framework for reasoning under uncertainty, so it makes sense that the norms of guessing would be related to probability. The simple probability-maximizing account sometimes makes incorrect predictions, however. Remember our second example above:

(2) a. (looking at a small bird through binoculars) This might be a sparrow.

b. (looking at a small bird through binoculars) This might be a sparrow or an ostrich.

Strictly speaking, the probability that a bird is 'a sparrow or an ostrich' is at least as high as the probability that it is a sparrow. Therefore, a probability-maximizing observer seeing a small bird should prefer 'It might be a sparrow or an ostrich' to 'It might be a sparrow'—but this seems absurd (Holguin, 2022; Yaniv & Foster, 1995; Dorst & Mandelkern, 2021). As another example, the probability-maximizing account predicts that the guess 'Some country will win the world cup' is a great guess, because it is almost guaranteed to be true. In general, the simple probability-maximizing account predicts that

we should make maximally vacuous guesses, that leave every possible option open and so have probability 1.

At this point, one might make the following suggestion for how to explain these counter-examples to the simple account. It might be that people want to make guesses that are likely to be correct, but they also care about other things beyond probability. People might want to make guesses that are relevant or carry the right amount of information (e.g. Sperber & Wilson, 1986; Grice, 1975). Or the extra ingredient might be *specificity*: people do not like guesses that vacuously mention too many of the possible answers to a question (Yaniv & Foster, 1995; Dorst & Mandelkern, 2021; Skipper, 2023).

Here we suggest that we can actually understand the psychology of guessing without having to venture outside of probability theory. Probability theory suggests a very natural account of guessing: the guess 'X will happen' is evaluated with respect to the entire *probability distribution* over the relevant outcome. The guess communicates not only that $Pr(X)$ is high, but also implicitly says something about the probability of the possible alternatives to X. In a sense that we will make more precise later, a guess can function as an *approximation* of the reasoner's subjective probability distribution.

Figure 1 illustrates how our view differs from the simple probability-maximizing account. Four countries are competing in the final phase of the world cup. You think that France, Germany, Brazil and Italy have 45%, 40%, 10% and 5% probabilities of winning, respectively. Under the simple account, whether 'Germany will win' is a good guess depends only on your belief about $Pr$(Germany wins); your belief about the other countries do not matter (see Left panel). Under our account, by contrast, the probability of each possible outcome is relevant to the guess you should make (Right). So, you may prefer guessing that 'Germany or France will win' rather than 'Germany will win', because the first guess provides a better approximation of your subjective probability distribution over all possible outcomes.

Our proposal naturally accounts for the intuition that guesses with higher

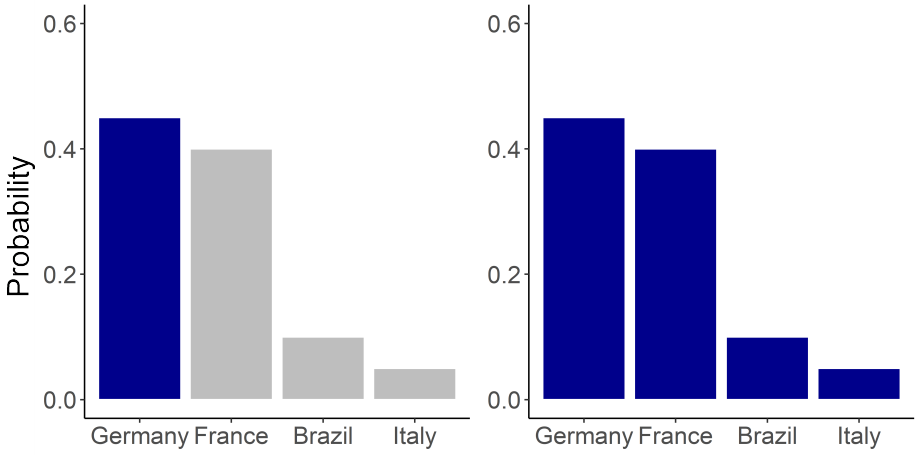


**Figure 1**

***Difference between a simple probability-maximising account (left) and our view (right)***. *Under the simple account, 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.*

probability are not always better. For example, 'Germany or Italy will win' is a worse guess than 'Germany will win', despite the fact that the former has higher probability (50% vs 45%). This is because the guess 'Germany or Italy will win' implies that Germany and Italy are the two most likely winners, and it is therefore a poor approximation of your subjective probability distribution (since you actually think that France and Brazil both have better odds than Italy). Coming back to our opening example, 'it might be a sparrow' is a better guess than 'it might be a sparrow or an ostrich' for similar reasons.

In this paper, we defend the idea that guesses encode an approximation of a reasoner's subjective probability distribution over the relevant possible outcomes. This idea can explain the characteristics of good guesses at a qualitative level, and, as we show in four experiments, also accounts for quantitative patterns in how people make and interpret guesses. We now present our account in more detail.

## A rational analysis of guesses

Our theory is derived from a rational task analysis of guessing, in the spirit of adaptationist and computational-level theories of cognition (Cosmides & Tooby, 1994; Marr, 1982; Anderson, 1990). We start with the assumption that people often represent (implicitly or explicitly) a subjective probability distribution over a set of relevant possible outcomes, or possible states of affairs. 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 exact probability estimates. In our view, a guess like 'probably between 20 and 30 states' functions as a lossy compressed encoding of this subjective probability distribution.

We draw on ideas from information theory (MacKay, 2003; Gagie, 2006; 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 Figure 2). The faithfulness of the encoding can be quantified as the extent to which the decoded output diverges from the original input (see e.g. Sims, 2016, for details on the mathematics of lossy compression, and their relevance to cognitive science).

Our claim is that people make guesses that provide good encodings of their underlying subjective probability distribution. In principle this framework is very general and can be applied in several different domains. Guessing could, 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 focus on the case of verbal communication, where a *speaker* communicates information to get a *listener* to re-construct a good estimate of the speaker's probability distribution.

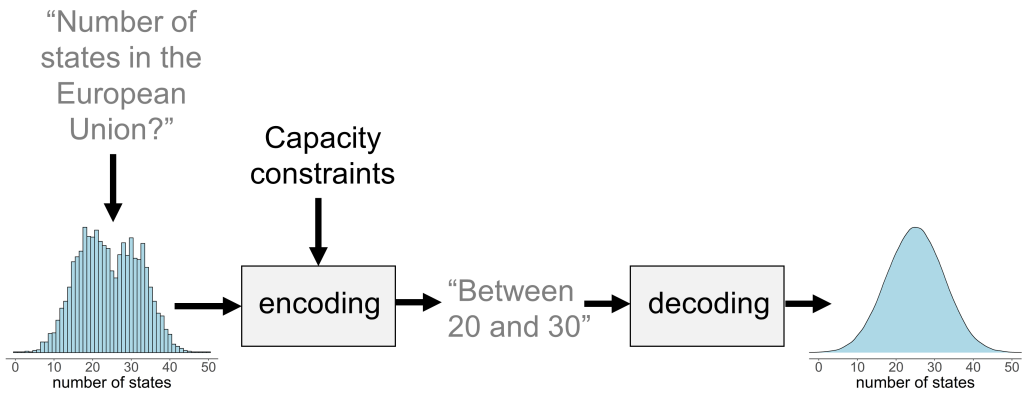


**Figure 2**

***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.*

### *Subjective probability distributions*

Our proposal is motivated by the general idea that the mind handles uncertainty by representing probability distributions over possible outcomes, or possible states of affairs. This idea has been successful in many areas of cognition, from perception to high-level cognition and semantics (Griffiths & Tenenbaum, 2006; Oaksford & Chater, 2007; Knill & Richards, 1996; Vul & Pashler, 2008; Tenenbaum et al., 2011; Fleming et al., 2012; Lassiter, 2011). 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.

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; Vul et al., 2014; Icard, 2016). In these cases our proposal is that speakers make guesses that aim to communicate the probability distribution latent in the generative model.

### *Constraints on the communication of subjective distributions*

Ideally, the speaker would be able to explicitly enumerate his full subjective probability distribution over the relevant possible outcomes. In practice, several constraints prevent this. For example:

- **Time**. Enumerating the probability of each possible outcome can take a lot of time if there are many different possible outcomes.
- **Conceptual knowledge**. Statistical concepts, like the normal distribution, are recent inventions, so people may lack the linguistic means to express complex statistical information. Someone's belief about the number of people living in the EU might be well-approximated by a normal distribution with some mean and variance, but they would be unable to say so explicitly.
- **Computation**. Explicitly computing a probability distribution, for example when computing a posterior, is either impossible (because there is no closed-form solution) or computationally expensive (MacKay, 2003). In many cases people might need to use approximation methods such as Markov Chain Monte Carlo when computing probabilities (Zhu et al., 2020; Bramley et al., 2017; Davis & Rehder, 2020). If people take a small number of samples, their estimates would be too coarse-grained to be useful.

These examples are not meant to form an exhaustive list, and they are not mutually exclusive. In this paper, we remain agnostic about the exact nature of the constraints that prevent people from full enumeration of their subjective distribution. We simply assume that explicit enumeration is impossible or impractical in most cases, and therefore people are skilled at communicating their subjective distributions in more compressed forms.

In our experiments, we will artificially specify the constraints ourselves: for example the speaker is only allowed to make guesses like 'Red', 'Blue or Green', etc, and cannot use other words or numbers. We view guessing as a problem of optimization under constraints: among the possible guesses the speaker could make, he must make the one that will most effectively convey his subjective probability distribution to a listener. We now state this problem more formally.

***Modeling framework***

We consider a speaker who has a subjective probability distribution $P$ over some outcome of interest, and can select a guess $g$ among a set of possible guesses, in order to convey information about $P$ to a listener. The speaker anticipates that the listener will infer a probability distribution $Q_g$ over possible outcomes on the basis of $g$.[3] The speaker's goal is to make the guess that causes the listener to infer a distribution $Q_g$ that is as 'close' as possible to the speaker's subjective probability distribution $P$ over possible outcomes. We can formalize this intuitive notion of 'closeness' between distributions using tools from information theory. Specifically, we say that a guess $g$ is a good encoding of the original distribution $P$ if the Kullback-Leibler divergence (KL-D)[4] of $Q_g$ from $P$ is low (Kullback

[3] Note that the listener could in principle also infer a probability distribution over the speaker's *beliefs*. For simplicity, we focus on the inference that the listener makes about the world (Modeling inferences about beliefs would require modeling a probability distribution over probability distributions).

[4] 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 the speaker and the listener when observing outcome $i$. The speaker would like to minimize the expected value of this difference; he thinks that outcome $i$ will occur will probability $P(i)$, and therefore wants to minimize

and Leibler, 1951, see Gagie, 2006), where the KL-D is computed as:

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

Therefore we assume that the value of a guess is inversely proportional to the KL divergence between the recovered distribution $Q_g$ and the speaker's subjective distribution $P$ (We add 1 to the denominator so that guess quality can range from 0 to 1):

$$V(g) = \frac{1}{1 + \mathrm{KL}(P||Q_g)} \tag{2}$$

In principle the listener might adjust the way she computes $Q_g$ as a function of how she thinks the speaker will behave, which can induce circularities (since the speaker anticipates how the listener computes $Q_g$). Instead for simplicity we follow recent work in computational modeling of social cognition (Frank et al., 2009; Shafto et al., 2014; Goodman & Frank, 2016) and hold the listener's update rule fixed while allowing the speaker's behavior to be an approximate best-reply to that fixed update rule.[5]

This framework is very general, and we are not strongly committed to any particular theory of how the listener interprets guesses, i.e. how she computes $Q_g$. In our case studies below we will rely on plausible assumptions about how people compute $Q_g$ in a given particular context, but our focus is on gathering evidence for our general framework rather than these particular assumptions.

---

$\int_i P(i) \log \left( \frac{P(i)}{Q_g(i)} \right)$. See Egré et al. (e.g. 2023) for a longer explanation.

[5] One possible theory of the computation of $Q_g$ is given by the Rational Speech Act Framework, which would define a recursive hierarchy of speakers and listeners, where the Level-0 listener interprets the guess according to some 'literal' meaning, a Level-1 speaker tailors his guess to the Level-0 listener, the Level-1 listener interprets the guess by taking the strategic intentions of the Level-1 speaker into account, and so on...(Goodman & Frank, 2016; Franke & Jäger, 2016; Degen, 2023). Models of this class are a special case of our general framework, where the 'fixed update rule' corresponds the update rule of the top-level listener (who by definition does not anticipate the behavior of the top-level speaker).

In the next section we provide a concrete illustration of our theory, and explain how it can account for the idiosyncratic nature of guesses. Then we report a test of this particular implementation of the theory. In a later experiment we test our framework in a different context, studying how people interpret guesses about continuous quantities.

**Case study: disjunctive guesses**

Consider the box in Figure 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 to a listener who 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", she 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 Figure 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 Figure 4b. More generally, we make the plausible assumption that listeners 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).

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

We can illustrate the comparative value of our account by showing that it can handle cases that are problematic for the simple probability-maximizing account. Guesses

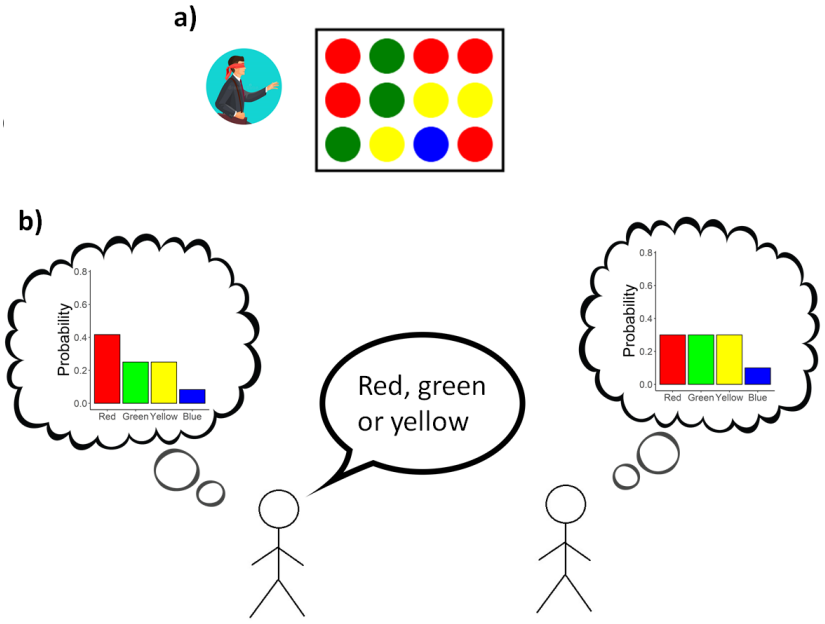


**Figure 3**

***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 (left) by making a guess (middle). The listener infers a probability distribution from the guess (right).*

that mention every possible outcome ('Red, Green, Yellow or Blue') are intuitively bad, despite having probability 1. This prediction follows naturally from our account, because 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.

To give another example, the guess 'it will not be red' is an intuitively worse guess than 'it will be red', despite the fact that $Pr(\neg\text{Red}) > Pr(\text{Red})$. 'Not red' assigns a low probability to Red, which is actually the highest-probability outcome in the speaker's mind, and therefore it defines a distribution that does not look anything like the one in the speaker's mind (see Figure 4d).

Our account also explains the intuition that good guesses respect 'clustering', in the

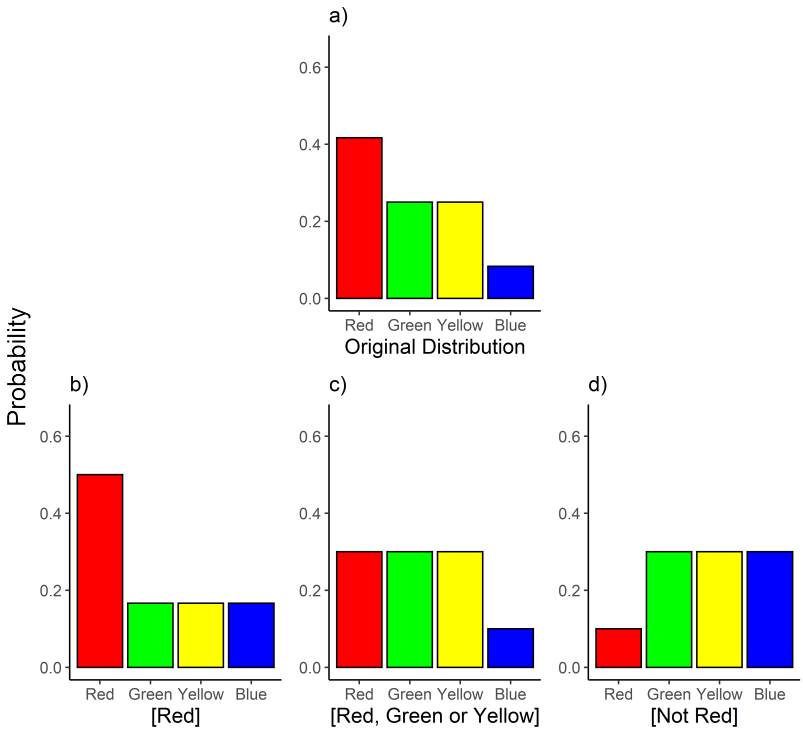


**Figure 4**

***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'.*

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.

### Accuracy and specificity

The experiments we report below also give us the opportunity to test the predictions of another account of the psychology of guessing. According to some

researchers, good guesses strike a trade-off between accuracy and specificity: a good guess is likely to be correct but does not mention too many outcomes (Yaniv & Foster, 1995; Dorst & Mandelkern, 2021; Skipper, 2023; Goldsmith et al., 2002)[6]. For example, Dorst and Mandelkern (2021) model the quality of a guess as:

$$V(g) = Pr(g)J^{S(g)} \tag{3}$$

where $Pr(g)$ is the *accuracy* of a guess, i.e. the probability that the guess is correct, and $S(g)$ is its *specificity*: the proportion of possible outcomes that it does not include. In the context of our urn-and-balls example above, a guess is specific to the extent that it mentions few colors (for example 'Red' is more specific than 'Red or Blue'). The parameter $J$ regulates how much people prioritize specificity relative to accuracy (higher values of $J$ correspond to higher weight for specificity)[7].

Is the accuracy-specificity trade-off framework in competition with our own analysis? It depends on the way we interpret it. If we interpret the trade-off framework as providing a computational-level hypothesis about the function of guessing, then it is an alternative to our own computational-level hypothesis. But we can also interpret the trade-off framework as giving a more descriptive, process-level hypothesis about how people make and interpret guesses. 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.

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

[7] To our knowledge, this formal model has not yet been tested empirically—in Studies 1 and 2 we take the opportunity to conduct such a test. In Study 4, we test the predictions of another formal implementation of the trade-off hypothesis, by Yaniv and Foster (1995).

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 Figure 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.

**Overview of empirical tests**

In what follows we report four experiments that test the quantitative predictions of computational models that implement our theory. 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 the 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/?view_only=8d7019ee2b8d456d8c3c9b29049b75aa.

## Study 1

In two studies (1a and 1b), we test our account in the context (described just above) of disjunctive guesses in a simple urn scenario. Participants were shown urns containing balls of different colors (as in Figure 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.

### 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 figure 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.

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 Figure 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.

**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 labelled 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 [Omitted for blind review].

**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) from Prolific. Participants were compensated £1 for their participation (median completion time was about 8 minutes) and participation was restricted to Prolific users with a 90+% approval rate.

**Computational modeling**

***Compression model***

In the introduction we defined our framework for the compression model, but did not commit to a particular theory of how listeners re-construct a distribution on the basis of a guess, i.e. how they compute $Q_g$. Here we offer plausible assumptions about the computation of $Q_g$ in the context of the current task.

We assume that when the listener hears a guess, she infers that outcomes mentioned in the guess are $\gamma$ times as likely as outcomes not mentioned in the guess, where $\gamma > 1$ is a free parameter.[8] The guess 'Red or Green', for example, implies that Red and Green are each $\gamma$ times as likely as Blue and Yellow. Note that this construction implies that all outcomes mentioned in the guess have the same probability as each other, and all outcomes

[8] We assume that the listener knows 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. Note that for simplicity here we consider settings where the audience knows what combinations of colors are possible, but our approach is in principle compatible with situations where that is not the case.

not mentioned in the guess have the same probability as each other. This is consistent with the principle of indifference (Jaynes, 2003; Laplace, 1820), according to which an agent should assign the same probability to two outcomes if there is no reason to see one of them as more likely.[9]

Denote the probability of an outcome not mentioned in the guess as $p$. Then the probability of an outcome mentioned in the guess is $\gamma p$. It follows that the listener infers the following distribution $Q_g$:

$$Q_g(i) = \begin{cases} \frac{1}{n_{\neg g} + \gamma n_g} & \text{if } g(i) = 0 \\ \frac{\gamma}{n_{\neg g} + \gamma n_g} & \text{if } g(i) = 1 \end{cases} \tag{4}$$

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 speaker's probability distribution 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 $P$:

$$P'(i) = \frac{P(i)^\alpha}{Z} \tag{5}$$

where $Z$ is a normalizing constant ensuring that all elements in $P'$ sum to 1, and $\alpha$ 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 gets concentrated (areas with a lot of probability mass get even more probability mass to

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

the detriment of other areas). Low values of $\alpha$ result in guesses that mention more possible outcomes.

We can then compute the value of a guess, following Equation 2, as:

$$V(g) = \frac{1}{1 + \mathrm{KL}(P' || Q_g)} \tag{2}$$

***Accuracy-specificity 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).[10]

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 $\alpha$.[11]

[10] 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. However, the idea is difficult to implement in practice, because distinctiveness can be trivially maximized by setting $J$ to infinity and making a maximally specific guess. Therefore we simply assume that a participant uses the same value of $J$ across urns.

[11] While the $\alpha$ 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.

### *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. To compute the log-likelihood, we assumed that each human rating is drawn from a truncated-discretized normal distribution with standard deviation $\sigma$ and mean $1 + 8m^s$, where $m$ is the model prediction ($\sigma$ and $s$ are free parameters we fit to the data).[12]

## Results

Figures 5 and 6 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 Figure 8). 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). Figure 8 displays the individual-level correlations between model and participant predictions, showing that the probability-maximizing model provides quite a bad fit for some participants.

[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.

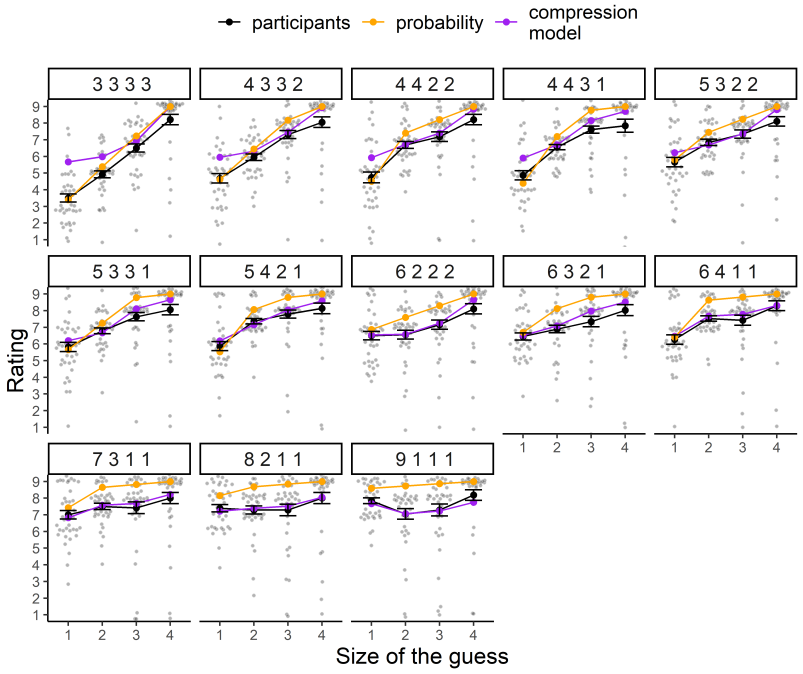


**Figure 5**

*Average participant ratings (black), along with predictions of the compression (purple) and probability-maximizing model (orange), in Study 1a. 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. Grey dots display individual ratings (jittered for visibility). Panel labels represent the profile of an urn: for example, an urn labelled [9,1,1,1] has 9 balls of one color, and one ball each of the other colors.*

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. Figure 7 (left panel) shows the ratings made by one such participant (in study 1b). 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 Figure 7. 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

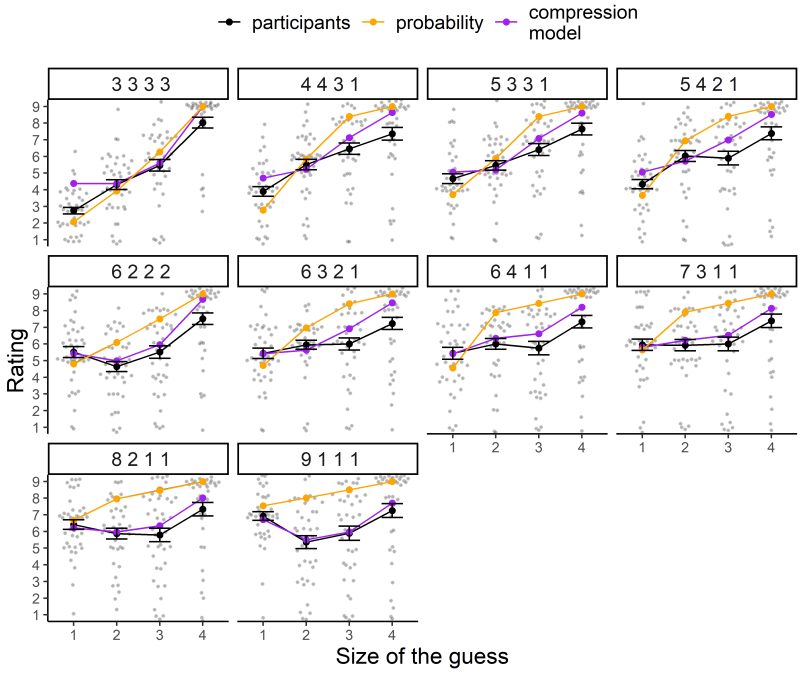


**Figure 6**

*Average participant ratings (black), along with predictions of the compression (purple) and probability-maximizing model (orange), in Study 1b. 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. Grey dots display individual ratings (jittered for visibility). Panel labels represent the profile of an urn: for example, an urn labelled [9,1,1,1] has 9 balls of one color, and one ball each of the other colors.*

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

| Model | AIC (group-level fit) | $n$ best fit |
|---|---|---|
| compression (Study 1a) | 7059 | 11 |
| trade-off (Study 1a) | 6886 | 6 |
| probability (Study 1a) | 6884 | 21 |
| compression (Study 1b) | 6383 | 15 |
| trade-off (Study 1b) | 6385 | 7 |
| probability (Study 1b) | 6383 | 17 |

**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, as assessed by AIC.*

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 Figures 5 and 6. 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 of Figure 7.[13] To give an example, when the urn has 9

[13] Figures for all individual participants are available on the Open Science Framework at

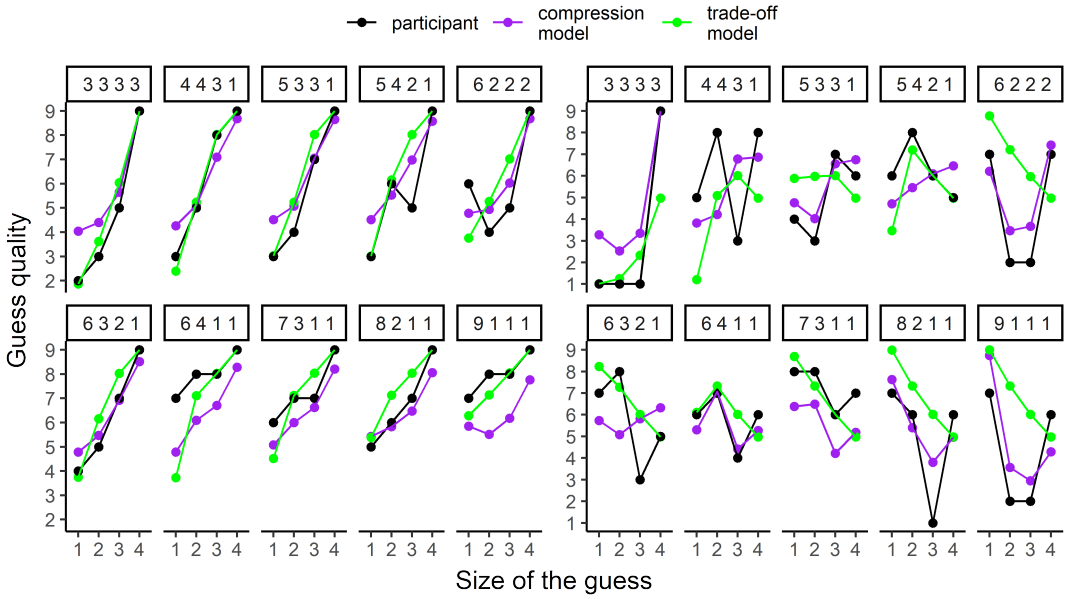


**Figure 7**

*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 labelled [9,1,1,1] has 9 balls of one color, and one ball each of the other colors.*

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.

## 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

---

https://osf.io/wz649/?view_only=8d7019ee2b8d456d8c3c9b29049b75aa.

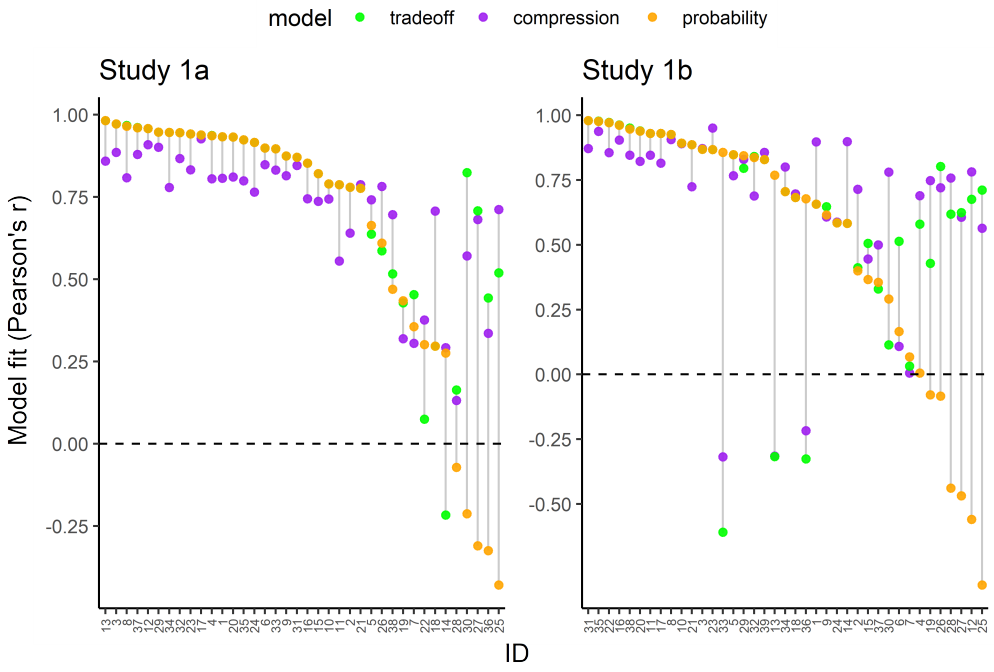


**Figure 8**

*Individual model fits, study 1. 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.*

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.

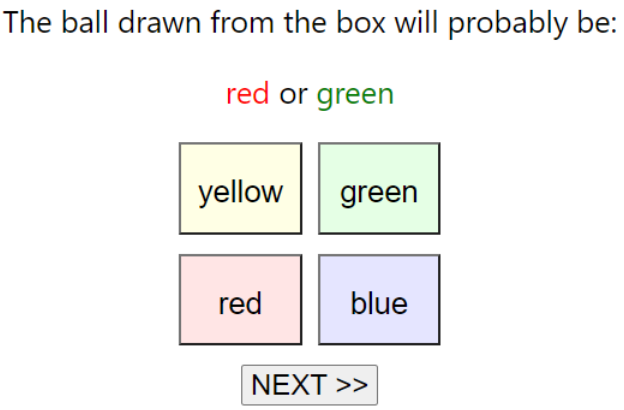


**Figure 9**

*Partial screenshot of the experimental interface, Study 2.*

## 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 Figure 9). 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 this trial purely as an

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

**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 $\theta$. 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 equation 2), and $\beta$ is an inverse temperature parameter controlling the stochasticity of choices (lower values of $\beta$ correspond to more stochastic choices).

**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

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

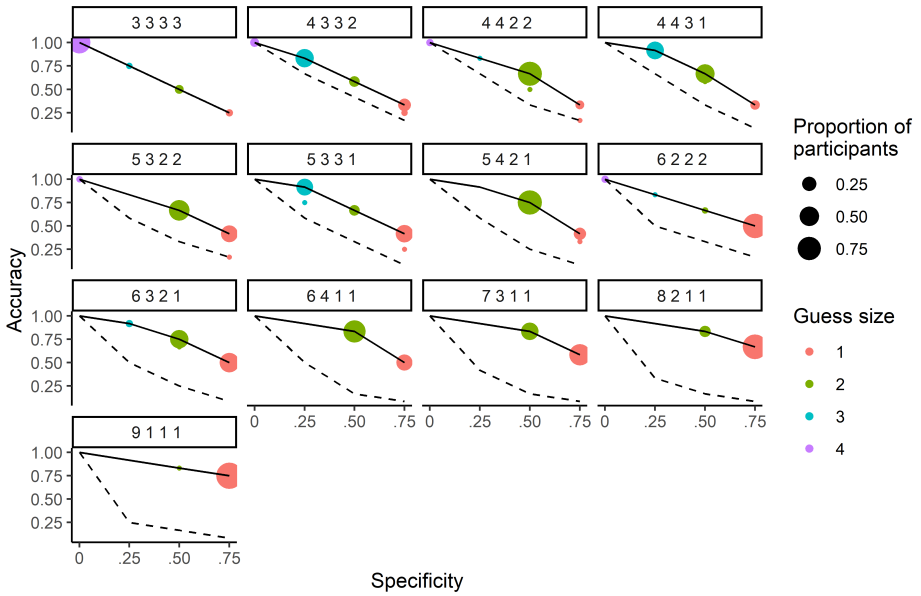


**Figure 10**

*Proportion of participants making a given guess, as the function of the size, accuracy and specificity of the guess, for each urn profile, Study 2. 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 can't be made more specific without losing accuracy, or vice-versa. Dashed black lines represent the inefficiency frontier: guesses that can't be made worse on one dimension without getting better on the other dimension.*

sample of 64 participants.

## Results

We can visually inspect some properties of people's guesses by plotting the proportion of guesses of different sizes for each urn profile (Figure 10). 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 majority (98%) of participants' guesses are Pareto-optimal (they lie along the black lines on Figure 10), 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.

***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

[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.

| Model | Pearson's r | AIC | $n$ best fit |
|---|---|---|---|
| Compression | .964 | 1676 | 19 |
| Trade-off | .929 | 1960 | 36 |
| Threshold | .871 | 2388 | 9 |
| Naive probability | .428 | 3426 | 0 |
| Random | NA | 4506 | 0 |

**Table 2**

*Fit of each model to the data. 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.*

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 Figure 11, 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.

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

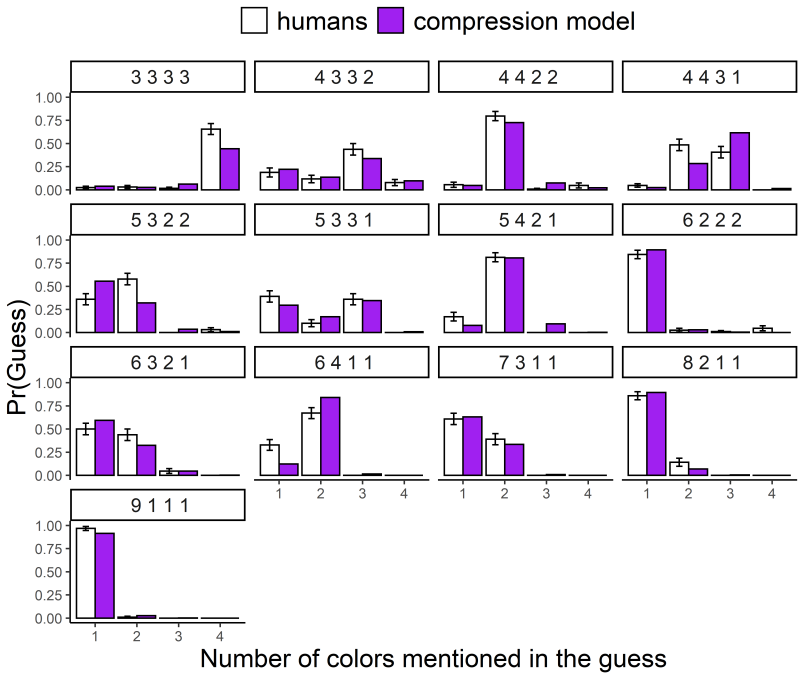


**Figure 11**

*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.*

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 $r(193) = .929$, $p < .001$,

| Model | $\beta$ | $\alpha$ | 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 | |

**Table 3**

*Best-fitting values of the model parameters, for each model.*

(see Figure A1 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 monotonously 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] Thirty-six participants were best-fit by the trade-off model, while 19 participants were best fit by the

[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 $\alpha$ from an exponential distribution with mean 2, $\gamma$ and $J$ from an exponential with mean 5 (truncated at 1), $\beta$ from an exponential with mean 50, and $\theta$ from a uniform distribution. 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 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.

compression model, and 9 participants by the threshold 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 Figure A2 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.

**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.

## 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.

**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 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 Equation 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.

[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.

## Methods

### *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 Figure 12 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 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).

[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.

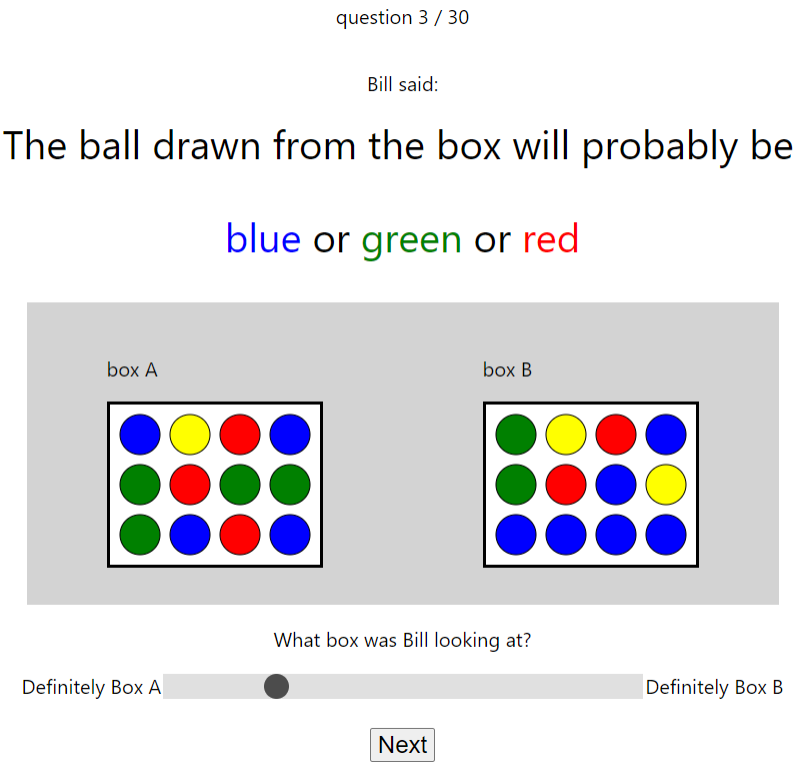


**Figure 12**

*Partial screenshot of the experimental interface, Study 3.*

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.

### *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 minutes, 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].

## Results

The model derived from the empirical production data in Study 2 provides a good fit to 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 Figure 13. Results for example trials are displayed in Figure 14.

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 Figure 15. We find a similarly good fit 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]

[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.

[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

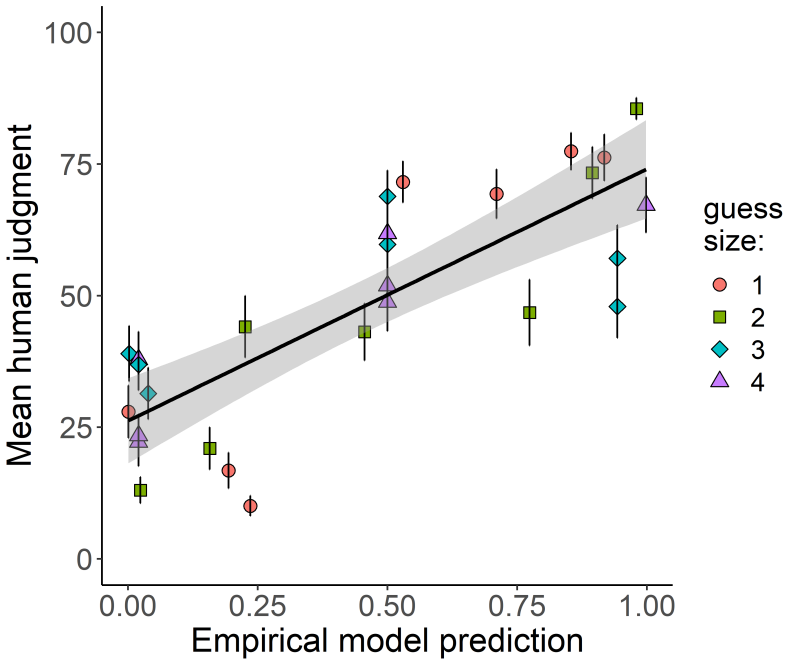


**Figure 13**

*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), Study 3. Error bars represent the standard error of the mean.*

Performing a similar analysis with likelihoods from the trade-off and threshold models (again with the parameters obtained in Study 2) yields a slightly lower fit to 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

---

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.

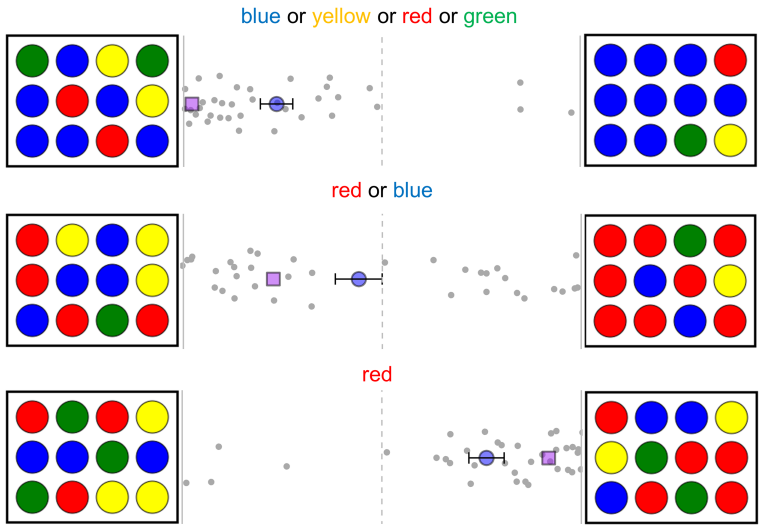


**Figure 14**

*Results for a selection of trials in Study 3 (see Figure A4 for all trials). Grey 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.)*

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 Figure 14.

**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

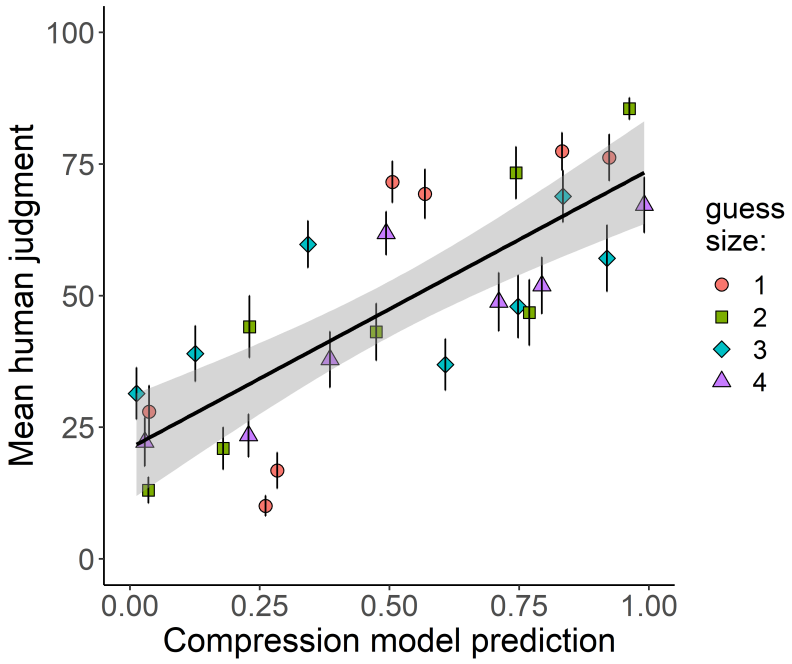


**Figure 15**

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

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.

## 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 Stations. 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 Figure 16).

**Model**

Formally, we model how people interpret interval guesses, of the form "$x$ is between $x_{\text{low}}$ and $x_{\text{high}}$" (for example, "There are between 165 and 185 member states in the United Nations"). 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 $\mu$ and standard deviation $\sigma$, where $\mu$ is the middle of the interval, and $\sigma$ 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 $\mu$ and variance $\sigma^2$. 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

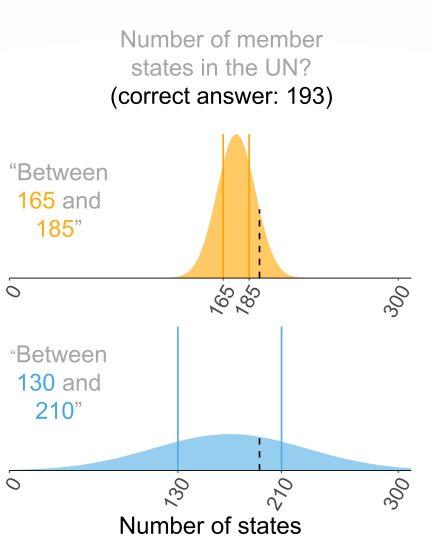


**Figure 16**

***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.*

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 Figure 16.

### 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:

[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.

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

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 $\alpha$ is a free parameter controlling the weight that people assign to specificity relative to accuracy. Intuitively, 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.

**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).

***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.

**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.

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 $\beta$ controlling the stochasticity of answers (as in Study 2, see equation 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.

**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 Figure 17, 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 computed the Akaike Information Criterion (AIC) for each participant and for each model. See Table 4 for the sum of all participants' AICs for each model. 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.

***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[23]. For instance, to the question "how many

[23] Some accounts of the semantics of interval guesses (e.g. Egré et al., 2023) seem to make this prediction,

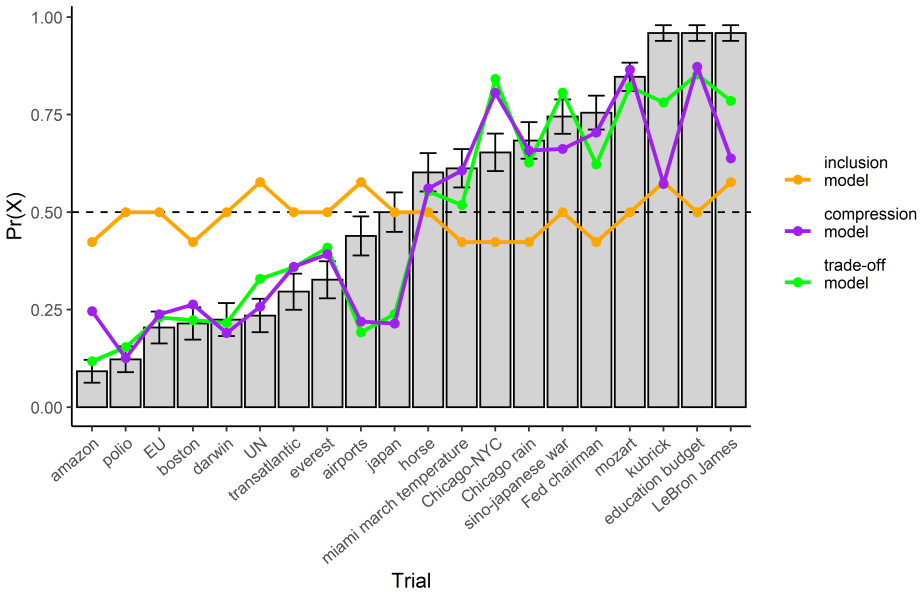


**Figure 17**

*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 A1 for list of trials.*

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.

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.

---

since they posit that all the probability mass should fall within the interval.

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

**Table 4**

*Fit of the different models to the data. 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). N best fit: number of participants best fit by the model, as assessed by AIC. Pearson's correlation: correlation between model prediction and proportion of participants making a choice. $\beta$ and param1 indicate the group-level best-fitting value of parameters for each model.*

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.

***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.

**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.

**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; McKenzie et al., 2008; Yaniv & Foster, 1997; Teigen & Jørgensen, 2005; Soll & Klayman, 2004; Klayman et al., 1999; Juslin et al., 1999; Moore et al., 2015; Russo & Shoemaker, 1992). 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.

## General Discussion

People often makes judgments about uncertain facts or events. Some of these 'guesses' strike us as intuitively better than others. What are the normative criteria that govern guessing? In this paper we explain the norms of guessing with a rational analysis of the problem that guessing solves. We suggest that speakers hold subjective probability distributions over possible answers to a question, but explicitly communicating the full distribution is difficult. Therefore they offer guesses that function as efficient (but lossy) encodings of their subjective distribution. This account explains the norms of guessing at a qualitative level, and 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 don't know the correct answer (Study 1). Below we explore some implications of this work for reasoning under uncertainty more generally. Then we discuss the scope of our work, some limitations, and directions for future research.

### Broader implications

In this more speculative section, we explore the possibility that guessing is, in some sense, the 'default mode' of human reasoning under uncertainty. That is, even in tasks where the correct thing to do is to reason about the probability of only one outcome, people favor answers that are informative about their whole subjective probability distribution—answers that make for good guesses. Below we briefly review relevant existing empirical findings.

#### *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 (Kahneman & Tversky, 1982; Attneave, 1959; 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 and Wells, 1998, see also discussion in Lassiter, 2011). 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 an outcome is judged as surprising if it would have been a bad guess prior to learning about the outcome. Conversely, people judge an outcome as likely if it is a good guess, or would have been included in a good guess. Under our account, 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 or surprising depend on the probability of other outcomes.

Our account 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.

### *Overconfidence in interval estimation*

People consistently produce over-confident confidence intervals. For example, if asked to estimate a numerical interval that they think is 95% likely to contain the correct answer to a question, that interval will contain the correct answer about 50% of the time (Alpert & Raiffa, 1982; Cesarini et al., 2006; McKenzie et al., 2008; Yaniv & Foster, 1997; Teigen & Jørgensen, 2005; Soll & Klayman, 2004; Klayman et al., 1999; Juslin et al., 1999; Moore et al., 2015; Russo & Shoemaker, 1992). Existing explanations of this pattern assume that it arises from cognitive limitations (e.g. Juslin et al., 2007; Zhu et al., 2023; Moore, 2022). As we argued in our discussion to Study 4, apparent overconfidence might also arise because speakers are trying to make good guesses. Participants in interval production studies might be disregarding 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.

***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[24]. 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'.

Following Dorst and Mandelkern (2021), we suggest that intuitions about what makes for a good guess 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 guess* than A. In the Linda case, 'Feminist and Bank Teller' is a better guess than 'Bank Teller', under plausible assumptions. We assume that the speaker implicitly considers a probability distribution over the following four possibilities:

- Linda is NOT a bank teller and is NOT a feminist. ($\neg T \neg F$)
- Linda is NOT a bank teller and is a feminist. ($\neg TF$)
- Linda is a bank teller and is NOT a feminist. ($T \neg F$)
- Linda is a bank teller and is a feminist. ($TF$)

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

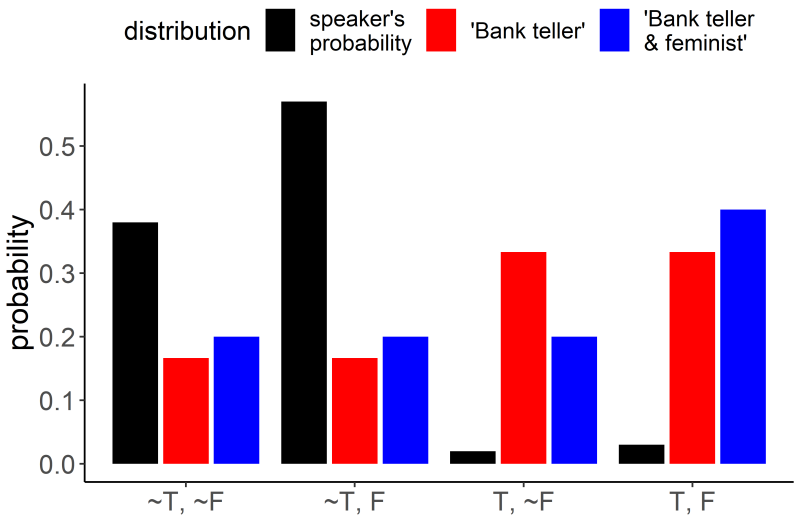


**Figure 18**

*Modeling Tversky & Kahneman's Linda case. Black bars represent the speaker's subjective distribution. 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 $Pr(T) = .05$, $Pr(F) = .6$, but the result is not sensitive to the specific parameters used, as long as $Pr(Feminist) > Pr(Bank\ teller)$.*

Under our account, the guess 'Feminist and Bank teller' communicates that the fourth possibility ($TF$) is more likely than any of the other three possibilities. The guess 'Bank teller' communicates that the last two possibilities ($T\neg F$) and ($TF$) are more likely than the first two. For example, assuming that the speaker has the subjective probabilities $Pr(T) = .05$, and $Pr(F) = .6$, and that $Pr(T)$ and $Pr(F)$ are independent, the speaker's distribution can be represented by the black bars in Figure 18. Setting the $\gamma$ 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 '$A$' is a better guess than '$A$ or $B$'. To illustrate, we consider a toy model of the Danielle case where the speaker entertains the following possibilities:

- 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 Figure 19.

There is of course already an extensive literature on the conjunction fallacy and related phenomena (e.g. Tversky & Kahneman, 1983; Bar-Hillel & Neter, 1993; Hertwig &

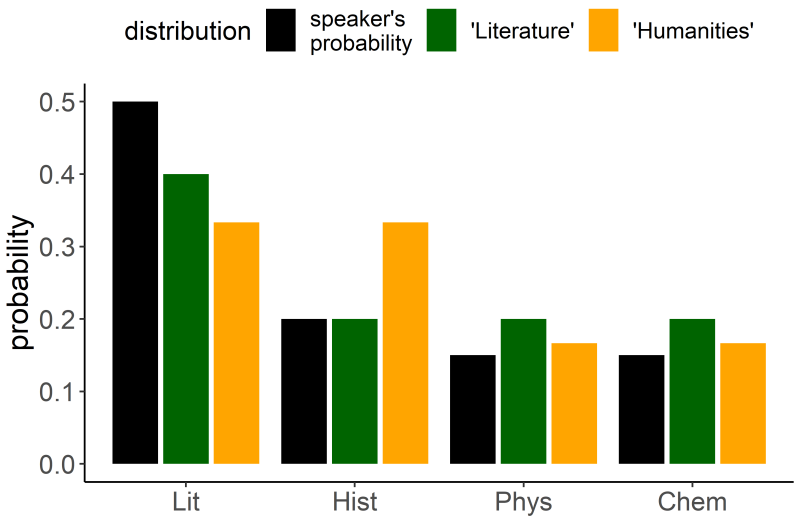


**Figure 19**

*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).*

Gigerenzer, 1999; Costello, 2009; Tentori et al., 2004; Busemeyer et al., 2011; Ludwin-Peery et al., 2020; Sablé-Meyer & Mascarenhas, 2022; Chung et al., 2023). Conjunction errors probably have other causes besides those we are suggesting, and making a full case for the current proposal is beyond the scope of the current paper. We note that Dorst and Mandelkern (2021) present an account of the conjunction fallacy that is closely related to ours: like us, they argue that conjunction errors arise when 'A&B' is a better guess than A, although their account of guess quality is slightly different than ours. Dorst and Mandelkern (2021) review the relevant empirical literature, and argue comprehensively that the pattern of people's conjunction errors is highly consistent with a guess-based account.[25]

[25] To give just one example (adapted to fit the details of our account), a guess-based account predicts 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

### *Implications for the rationality of judgment under uncertainty*

The patterns of judgment reviewed above seem like mistakes if we think that judgments like 'A is likely' should simply reflect the probability of A. These judgments appear more rational if they are viewed through a different lens, where guesses function as approximations of the whole probability distribution.

This conclusion must be qualified somewhat, however. Even if our account is correct, it still makes sense to view certain patterns of judgment as mistakes. The conjunction fallacy is for example observed even in contexts where people have to make monetary bets on an outcome (Tversky & Kahneman, 1983). In this context, betting more money on A&B than A is clearly sub-optimal. One possible interpretation of these effects is that the 'default mode' of judgment under uncertainty is to make guesses that provide a good encoding of one's subjective distribution, and non-normative judgment arises when people over-apply this tendency.

## 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 the core reason why guesses take the form that they do. 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 a distribution. It is

---

proportion of them 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).

impractical for speakers to recite their entire probability distribution out loud, and therefore they often make guesses that mention only one or a few possible outcomes. These considerations apply, in principle, to any agent that must reason under uncertainty and create compressed representations of probability.

At a less abstract level, and given our emphasis on verbal communication, the question arises of how to think of guesses in terms of concepts from linguistics. Linguists often distinguish between the *literal meaning* of an expression, and the meaning it actually conveys within a particular context (its *pragmatic* meaning, Grice, 1975; Goodman and Frank, 2016). So the following question might arise: does a statement like 'X will probably happen' encode distributional information in virtue of its literal meaning, or does it do so because of pragmatics?

Our account is agnostic with respect to this question—we are not committed to any position about the literal meaning of particular words. For example, the literal meaning of 'probably X' might not itself be about a probability distribution; instead it might mean that the probability of X is above some threshold (Yalcin, 2007; Lassiter, 2010). But the speaker might use this literal meaning strategically, in order to communicate his subjective probability distribution.

To illustrate, consider our urns-and-balls paradigm. Imagine that the speaker says 'The ball will probably be red', and the literal meaning of 'probably X' is 'Pr(X) > .5'. When the listener hears the guess, accommodating the fact that Pr(Red) > .5 requires her to update her whole probability distribution over the outcome: she has to decrease her subjective probability that the ball will be green, yellow or blue. If the speaker anticipates this inference well enough, he can select the guess strategically to align the listener's distribution close to its own.

This kind of dynamics can be formally modeled within the Rational Speech Acts framework (RSA; Goodman & Frank, 2016; Degen, 2023; Zaslavsky et al., 2021). The RSA framework models communication as recursive mindreading across a hierarchy of speakers

and listeners; at the lower level is a 'literal listener' who interprets the utterance according to its literal meaning. Researchers have designed RSA models of how people communicate about probabilistic information (e.g. Herbstritt & Franke, 2019; Egré et al., 2023; van Tiel et al., 2022). In the Appendix, we derive an RSA model for our task in Study 2. The model assumes that the literal meaning of 'probably X' is about the probability of X, but that the pragmatic speaker tries to communicate about his whole subjective probability distribution. We find that the model is able to give a relatively good account of our data in Study 2.

Ultimately, our core claim is that guesses encode information about the speaker's probability distribution over possible outcomes. We think that computational levels of pragmatics are one potential way to implement this hypothesis. However, the explanatory power of such a model comes from its assumption that the speaker is communicating his subjective probability distribution over possible outcomes; it does not come intrinsically from the model's emphasis on pragmatics. The more abstract approach that we take, embodied in our main model, has the advantage of focusing on this key assumption while abstracting away from specific implementation details.

More generally, there are many details for which our computational-level theory does not make strong commitments. At an algorithmic level, the question arises of how speakers access their subjective probability distribution. Because explicit enumeration of a distribution is in many cases intractable, speakers might need to engage in sampling-based approximation (Zhu et al., 2020; Vul et al., 2014; Bramley et al., 2017; Davis & Rehder, 2020).

### *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 (Yaniv & Foster, 1995; Dorst & Mandelkern, 2021; Skipper, 2023).

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 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 of the psychology of guesses. 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.

We note that the trade-off hypothesis, when viewed as a descriptive account, 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.

**Limitations and future directions**

Our account involves some over-simplifications. To make modeling more tractable, we have assumed that the speaker aims to change the listener's belief about the world. But

guesses can also aim to change the listener's meta-representations: i.e. change the listener's belief about the speaker's belief. In many cases this might be their primary function. 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.

Beyond communicating the speaker's distribution, guesses may also contain more general information about the speaker's knowledge. Consider for instance the answer "8880 to 8885 meters" to the question "what is the height of Mount Everest?" (correct answer: 8849 meters). If strictly interpreted as 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. It is however 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. Estimating a guess' quality may therefore involve sophisticated causal inferences, for instance inferences about whether the speaker possesses a good internal model of the relevant domain.

More generally, we recognize that guessing often performs other functions beyond encoding the speaker's beliefs. For example, betting in a horse race can be construed as a guess ('I guess this horse will win'), but that guess is governed by standard decision-theoretic considerations (people should make the bet that maximizes their subjective expected utility). In information-seeking games like Twenty Questions and Mastermind, a guess serves the function of collecting information, and players make guesses that help them narrow down the set of possible answers (Cheyette et al., 2023). In the current paper we focused on situations for which there is not an already obvious normative criterion—like expected utility or expected information gain—for what constitutes a good guess.

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 Figure 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. Roberts, 2012; Koralus & Mascarenhas, 2013).

Finally, our modeling framework assumes a pre-existing, fixed set of constraints over the guesses that speakers can make. An interesting avenue for future research could be to model speakers that can choose at which level of precision to communicate their probabilistic beliefs, and must navigate a trade-off between the cost of high-fidelity communication and the cost of misunderstanding.

### *The language of uncertainty*

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* (Yalcin, 2007; Herbstritt & Franke, 2019; Lassiter, 2010). Our approach can in principle account for the way people make guesses in many different formats, provided we have appropriate models of the way listeners infer a probability distribution from a given type of guess. Building these more specific models will require integrating the insights of existing accounts of the semantics of uncertainty expressions (e.g. Budescu & Wallsten, 1995; Yalcin, 2007; Herbstritt & Franke, 2019; Lassiter, 2010; Egré et al., 2023; Kao et al., 2014).

## 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 talk about uncertain facts or events, we implicitly communicate about our subjective probability distribution over the relevant set of possible outcomes. Here we have argued that this natural hypothesis can qualitatively and quantitatively account for the way that people make and interpret guesses.

## Appendix

### 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 $\gamma p$. 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:

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

### 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 $\Delta P$.[26] Therefore we assume that $P_c - P_b = P_b - P_a = \Delta P$: If we can show

[26] Pareto-optimality implies that the colors in $b$ are more or equally frequent as the colors not in $b$.

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.:

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

---

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 can not result in a gain in probability greater than $\Delta P$.

Putting Equations 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 with $n$ colors has higher value than a guess with $n+1$ colors, then it follows from the result derived above (by induction) that it has higher value than all guesses with $n+2$ or more colors.

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 $b$.[27] Therefore the Pareto-optimal guess $g_{b-1}$ whose size is 1 less than the size of $b$ has $V(g_{b-1}) > V(b)$. It follows (from the result in the paragraph above) that every guess of size greater than $b$ has lower value than $b$, so $V(b) > V(c)$. □

**RSA model in Study 2**

We consider a model within the Rational Speech Act framework (RSA; Goodman & Frank, 2016; Franke & Jäger, 2016; Degen, 2023), 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 $\theta$ 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,

[27] 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.

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 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 (Equation 2). Like the other models we consider in Study 2, the RSA model assumes that the speaker's distribution might be subject to slight distortions (Equation 5) and that the probability of guess production is determined by a softmax function (Equation 7).

***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 Figure A3) 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 monotonously 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.

**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 $\theta$. 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 $\theta$ 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 shouldn't 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 (Equation 7).

**Figures for alternative models, Study 2**

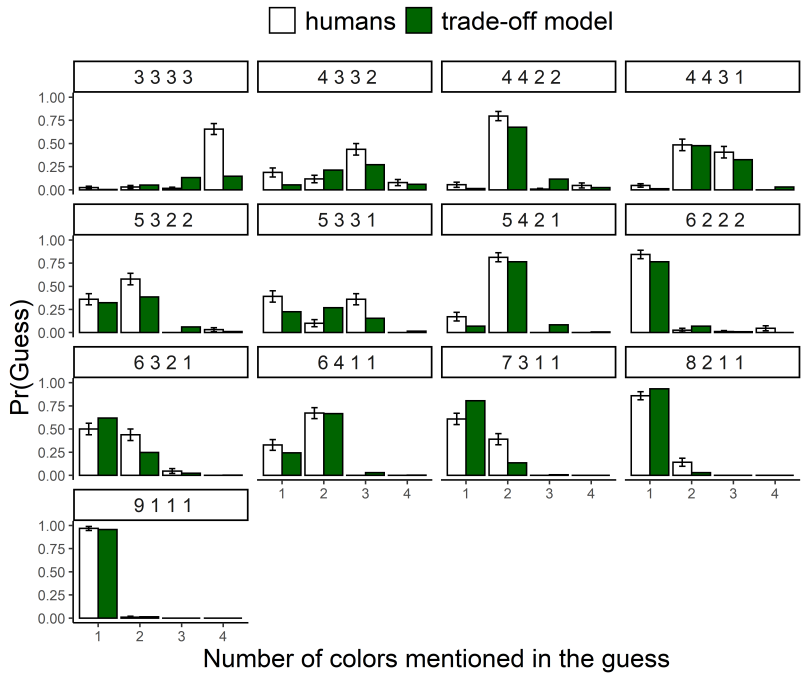


**Figure A1**

*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.*

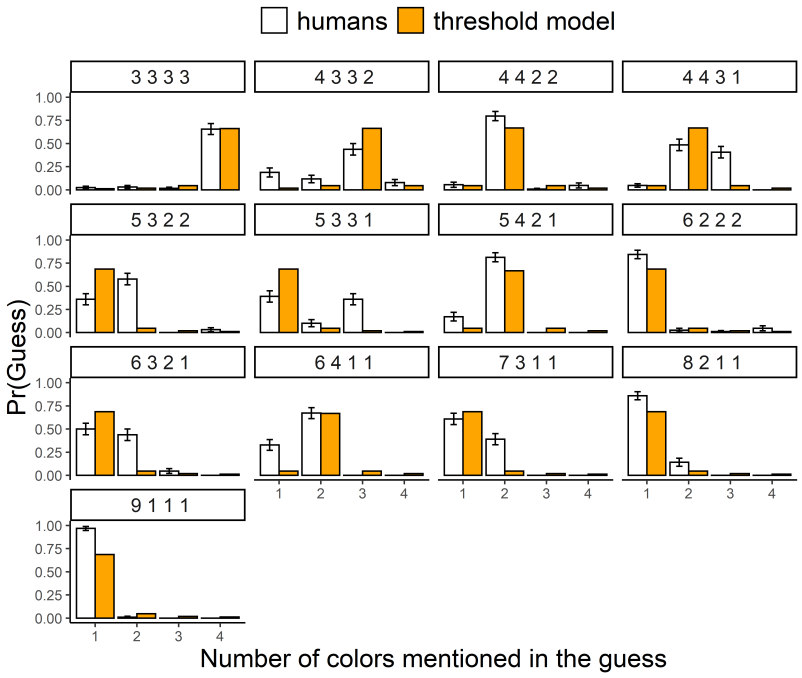


**Figure A2**

*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.*

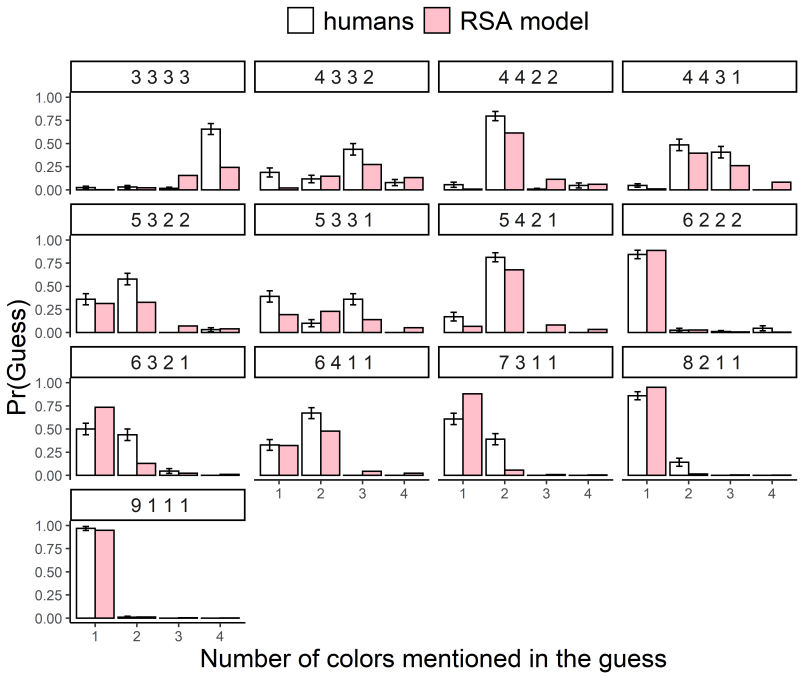


**Figure A3**

*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.*

**List of trials, Study 3**

**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 labelled [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 |

Continued on next page

**Table A.1 – continued from previous page**

| **urn A** | **urn B** | **guess size** | **empirical model** | **mean human rating** |
|---|---|---|---|---|
| 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 |

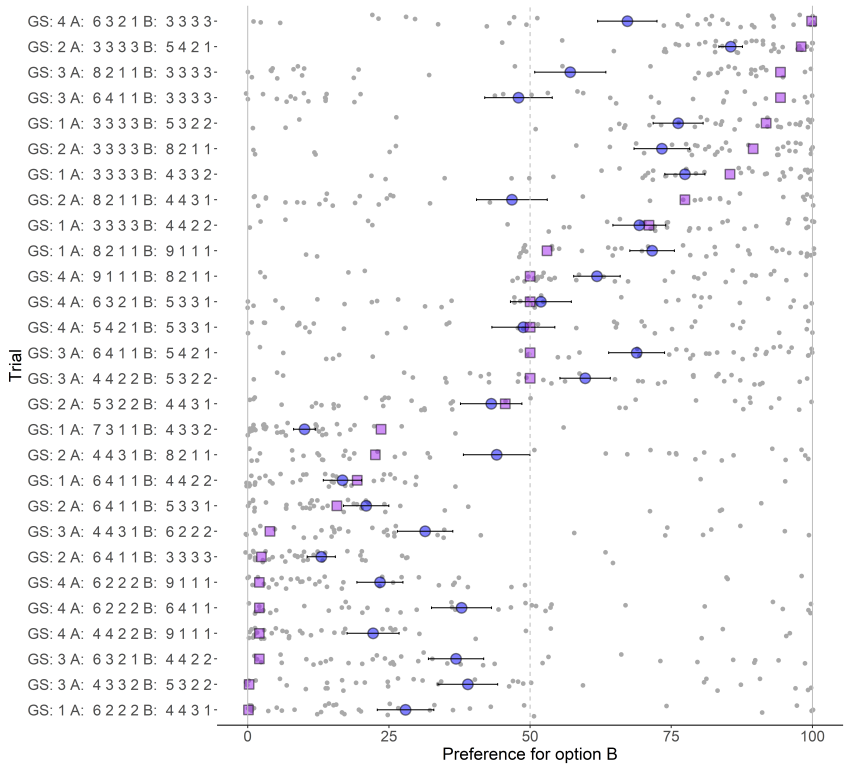


**Figure A4**

*Detailed choice data for Study 3, along with predictions of the pragmatic listener model with empirically-derived likelihood. Grey 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.*

### 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

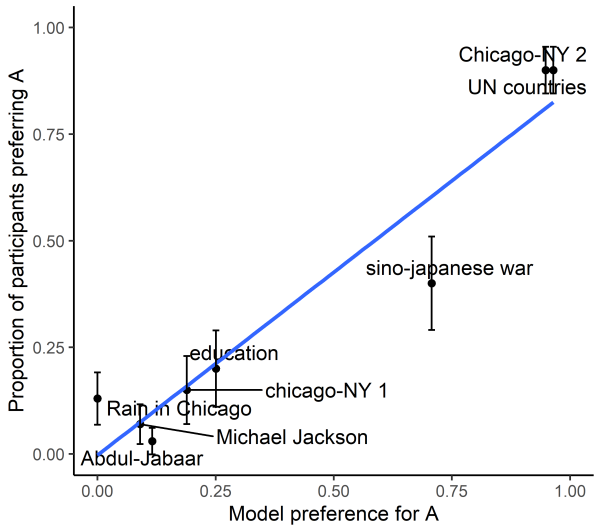


**Figure A5**

*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.*

transforms the width of the guess into a standard deviation) and $\beta$ (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 Figure A5. We also find that Yaniv & Foster's original accuracy-specificity trade-off model has a similarly good fit to these data ($r(6) = .98$, $p < .001$).

**List of items from Yaniv & Foster (1995)**

**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)? | 37639 | 30000 | 45000 | 37000 | 40000 |

Continued on next page

**Table A.2 – continued from previous page**

| **Label** | **Text** | **Ground truth** | **Lower A** | **Upper A** | **Lower B** | **Upper B** |
|---|---|---|---|---|---|---|
| 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 |

**List of items in Study 4**

**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 |

Continued on next page

**Table A.3 – continued from previous page**

| **Label** | **Text** | **Ground truth** | **Lower X** | **Upper X** | **Lower Y** | **Upper Y** |
|---|---|---|---|---|---|---|
| boston | Number of people living in Boston? | 684000 | 610000 | 640000 | 550000 | 700000 |
| japan | Number of people living in Japan? | 126 | 100 | 110 | 135 | 170 |
| airports | Number of public airports in the United States? | 5217 | 5000 | 20000 | 4000 | 4800 |
| everest | Height of Mount Everest? | 8849 | 9000 | 12000 | 8000 | 8500 |
| amazon | Length of the Amazon river? | 6400 | 6600 | 12000 | 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 |

Continued on next page

**Table A.3 – continued from previous page**

| **Label** | **Text** | **Ground truth** | **Lower X** | **Upper X** | **Lower Y** | **Upper Y** |
|---|---|---|---|---|---|---|
| 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 |
| 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? | 37024 | 32000 | 40000 | 20000 | 35000 |

Continued on next page

**Table A.3 – continued from previous page**

| **Label** | **Text** | **Ground truth** | **Lower X** | **Upper X** | **Lower Y** | **Upper Y** |
|---|---|---|---|---|---|---|
| Fed chairman | Monthly salary of the chairman of the U.S. Federal Reserve? | 16900 | 18000 | 22000 | 4000 | 35000 |
| 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 |

## 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.

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.

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* (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 & 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. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *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. J. (2009). How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, *22*(3), 213–234.

Davis, Z. J., & Rehder, B. (2020). A process model of causal reasoning. *Cognitive Science*, *44*(5), 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*. https://doi.org/10.1111/phpr.12831

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.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–1286.

Frank, M., Goodman, N., Lai, P., & Tenenbaum, J. (2009). Informative communication in word production and word learning. *Proceedings of the annual meeting of the cognitive science society*, *31*(31).

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. *Oxford handbook of human memory*.

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., & 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.

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. *Semantics and Linguistic Theory*, *20*, 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* (pp. 182–209). Wiley Online Library.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, *5*, 6–1.

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.

Savage, L. J. (1954). *The foundations of statistics*. Courier Corporation.

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*, 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.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293.

van Tiel, B., Sauerland, U., & Franke, M. (2022). Meaning and use in the expression of estimative probability. *Open Mind*, *6*, 250–263.

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.

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.

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. *Proceedings of the Society for Computation in Linguistics 2021*, 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*.