Memory and Representativeness

Pedro Bordalo Katherine Coffman Nicola Gennaioli
Frederik Schwerter Andrei Shleifer*

December 17, 2019

Abstract

We explore the idea that judgment by representativeness reflects the workings of episodic memory. In our model, subjects assess the probability of a hypothesis given data in terms of the ease of retrieval of instances of that hypothesis when cued with the data. Retrieval is driven by a measure of similarity which exhibits contextual interference: a data-cue is less likely to retrieve a hypothesis that occurs in other data. As a result, retrieval and probability assessments are context dependent. In a new laboratory experiment on cued recall, participants are shown two groups of images with different distributions of colors. In line with the model's predictions, we find that i) decreasing the frequency of a given color in one group significantly increases the recalled frequency of that color in the other group, ii) cueing different dimensions for the same set of images entails different interference patterns and different probabilistic assessments. A calibration of the model yields a good quantitative fit with the data, highlighting the central role of contextual interference.

Keywords: cued recall; interference; similarity; probabilistic judgments; heuristics and biases

*Bordalo: University of Oxford; pedro.bordalo@sbs.ox.ac.uk. Coffman: Harvard Business School; kcoffman@hbs.edu. Gennaioli: Bocconi University and IGIER; nicola.gennaioli@unibocconi.it. Schwerter: University of Cologne; frederik.schwerter@uni-koeln.de. Shleifer: Harvard University; shleifer@fas.harvard.edu. We thank Rahul Bhui, Ben Enke, Paul Fontanier, Sam Gershman, Thomas Graeber, Spencer Kwon, Joshua Schwartzstein, seminar participants at Harvard University and MIT Sloan, and four anonymous referees for helpful comments. Gennaioli thanks the European Research Council (GA 647782) for financial support. Schwerter and Shleifer thank the Sloan Foundation for financial support.

1 Introduction

A vast literature in psychology shows that individuals' probabilistic judgments are vulnerable to systematic errors (see Benjamin 2018 for a review). An important body of evidence, ranging from base rate neglect to the disjunction fallacy, focuses on the representativeness heuristic, the tendency to judge as likely events that are merely representative. Representativeness captures "the degree to which [an event] is similar in essential characteristics to its parent population" (Kahneman and Tversky, KT 1972). In a well-known example, subjects are given a short description of an introverted man and are asked to rank, in order of likelihood, several different occupations (Tversky and Kahneman, TK 1974). Subjects tend to state that the introverted man is more likely to be a librarian than a salesman, even though there are vastly more salesmen than librarians. Being an "introvert", the reasoning goes, makes the man more similar to a librarian than to a salesman, which causes an inflated assessment of likelihood.

But why do subjects use similarity in making probability judgments? Why should similarity sometimes cause them to think about unlikely events, so that judgments are incorrect? Why among introverted men do we think about the unlikely librarian instead of the more likely salesman? Why do we exaggerate the likelihood that an Irish person has red hair even though only 10% of them do (Bordalo et al., BCGS 2016)?

We address these questions starting from the idea that probability judgments are formed by selectively retrieving information from memory, cued by the problem at hand. In our model, a decision maker assesses the likelihood of different hypotheses in light of data d. In line with memory research (Kahana 2012), recall of a hypothesis h is driven by its similarity with the data d, and is subject to interference, so that hypotheses more similar to the data prevent less similar ones from coming to mind. When told someone is Irish – that is, when cued with the data $d =$ Irish – the decision maker selectively recalls the hair color h, in this case red, that is most similar to this data, and overestimates its likelihood.

Our main theoretical contribution is to formalize similarity between the data-cue and the hypothesis assessed, which constitute distinct and overlapping sets. Our formulation of similarity is inspired by Tversky's (1977) similarity metric, and captures the idea that two sets are more similar when the statistical correlation among their attributes is higher. In probabilistic judgments, this implies that a hypothesis h is more easily recalled when it is more likely to occur with the data d, but also when it is less likely to occur with other data $-d$. Similarity-based recall is then context dependent, in that recall about data d depends on the comparison data $-d$, sometimes causing neglect of likely events. For instance, when thinking about $d =$ Irish, we find it harder to recall the modal hair color $h =$ dark because it occurs very often in other nationalities $-d$, which interferes with its

recall in the Irish group. We refer to this mechanism as contextual interference, since interference with hypothesis h is determined by the context provided by the alternative data $-d$.[1]

Our second contribution is to present a novel experimental framework and a set of experimental results that provide evidence for this mechanism. The model yields testable predictions for how biases in probability judgments depend on the statistical associations between the data and hypotheses. To test these predictions, we run three (main) experiments with the following basic structure. Participants observe a sequence of 25 numbers and 25 decoy images (either shapes or words), which come in different colors. Among numbers, 15 are blue and 10 are orange. In one treatment, decoy images are all gray shapes so there is no color overlap with numbers. In another treatment, decoys are all blue words, which is also the modal color among numbers. Participants are then asked to: i) recall the numbers they saw, and ii) estimate the likelihood of different colors among numbers. We think of data as d = numbers, and the possible hypotheses, $h \in H$, as the colors.

In Study 1 we test contextual interference: we predict that the hypothesis h = orange is deemed more likely when the decoy data $-d$ are blue words than when they are gray shapes, because blue words interfere with the recall of blue numbers. For the same reason, fewer blue numbers should be recalled in this case. In Study 2 we vary the strength of interference by adding orange words to the decoy data $-d$. As blue words are replaced with orange ones, retrieval of blue numbers is less interfered with and retrieval of orange numbers is more interfered with. Our model predicts that the assessed likelihood of hypothesis h = orange should decrease with the number of orange words.

In Study 3 we test a key implication of our model of memory-based probabilistic judgments: descriptions of the judgment task that shape perceptions of similarity affect memory retrieval and thus probability judgments, even if they are normatively irrevelant. Specifically, we keep the color distribution of numbers and words fixed, but allow the images to differ along font size: either small or large. Crucially, all blue numbers are small, while all other items (orange numbers and blue words) are large. This ensures that, given d = numbers, the hypothesis h = orange is equivalent to the hypothesis h = large. However, our model predicts that participants should assess the size distribution

[1]The literature on similarity usually focuses on judgments of similarity between two objects of the same type, say similarity between faces (Tversky 1977, Kahana 2012). In contrast, we argue that to understand probabilistic judgments we must consider similarity between two sets: the data and the hypothesis. An example may help illustrate the difference. Consider the data "things that fly". A standard approach would consider the similarity among two specific airplanes depending on their differences along certain characteristics. This is the approach followed also in probabilistic judgment models such as PROBEX. In our approach, instead, we ask how similar is the broad set "airplanes" to the broad set "things that fly", which also includes birds and clouds, and thus how likely is "airplane" to be retrieved when cued with "things that fly".

of numbers differently from its color distribution, even though the two coincide. The reason is that when thinking about size, contextual interference by the decoy data penalizes large numbers (which are orange), but when thinking about color it penalizes blue numbers (which are small).

We find strong and robust evidence consistent with these predictions. To assess how our model quantitatively accounts for the observed variation in beliefs about the frequency of orange numbers. We use the data from Study 1 and from a subset of treatments in Study 2 to calibrate the model. We then use the calibrated model to predict, out of sample, the other results in Studies 2 and 3, as well as in additional experiments we ran. This exercise confirms that contextual interference is a quantitatively important determinant of beliefs and that our model offers a good quantitative fit.

In our approach, the representativeness heuristic derives from a basic cognitive mechanism: context-dependent, similarity-based recall. As such, it differs from existing models of faulty probability judgments. One strand of the literature invokes limited memory, including sampling models (Sanborn and Chater 2016), the Minerva MD model (Dougherty et al. 1999) or the exemplar PROBEX model (Nilsson, Olsson, and Juslin 2005). The first two models are based on noisy recall, which yields some biases but not the overestimation of unlikely events. In the PROBEX model, the exemplar hypothesis of data d does not depend on different data $-d$, which makes this model unsuitable to account for how probability judgments depend on the decoy group.[2]

Bhatia (2017) shows that social stereotypes, such as introverted male librarians and red-haired Irish, can be spread through the media and personal interactions, through selective repetition of the more representative cases. This semantic association and repetition of unlikely traits becomes a source of accessibility in memory that may cause their overestimation. While this mechanism certainly contributes to biases in social contexts, it does not explain the source of these biases (why does media focus on these specific traits), and why biases are widespread even in abstract judgments such as those in our experiments.

[2]Other models of memory are motivated not by biases in probabilistic assessments but by evidence on recall, specifically by the fan effect (Anderson 1974) and a large body of evidence on interference going back to the early 20th century (Jenkins and Dallenbach 1924, Whitely 1927, McGeoch 1932, Underwood 1957, Keppel 1968). Anderson (1974) shows that concepts associated with more items are more difficult to remember in response to any specific cue, as evidenced by slower response times and a larger error rate in a recognition task. Similarly, in word-pair recall tasks, associating the same cue word with different target words in different lists reduces recall of both the pair learned first and the pair learned later (see Kahana 2012 for a review). Two broad approaches to this evidence have been proposed: associative models such as the Adaptive Control of Thought - Rational (ACT-R, Anderson and Reder 1999), and inhibition models such as inhibitory control in retrieval (Anderson and Spellman 1995). As in most existing models within the similarity framework, interference here occurs across types associated with a cue, which in our model corresponds to interference across different hypotheses h associated with data d. Instead, our model and experiments highlights the importance of contextual interference, namely the association of hypotesis h with some other data $-d$.

Several papers explicitly engage with the representativeness heuristic. Tenenbaum and Griffiths (2001) formalize representativeness based on Tversky and Kahneman (1983), but do not link it to similarity-based recall and interference. Our analysis provides a foundation for the representativeness heuristic in terms of interference in recall. More broadly, we show that interference in recall generates beliefs that amplify true differences between distributions, a property introduced in Gennaioli and Shleifer (GS 2010) and BCGS (2016) that we refer to as the kernel of truth.

Our recent work shows that the kernel of truth helps account for a wide range of experimental and field evidence on beliefs in several domains: GS (2010) illustrate its explanatory power for several puzzles in probabilistic judgments, such as the conjunction fallacy and base rate neglect, that were originally attributed to representativeness (KT 1972). BCGS (2016) show that the same principle helps explain measured beliefs about political groups as well as social stereotypes, while BCGS (2019) show that its predictive power extends to beliefs about own ability, tainted by group stereotypes. The kernel of truth also helps account both qualitatively and quantitatively for measured expectations of market participants in macroeconomics and finance (Bordalo, Gennaioli, Shleifer 2018, Bordalo, Gennaioli, Ma, Shleifer 2018, Bordalo, Gennaioli, La Porta, Shleifer 2019).

The results in this paper link the kernel of truth to memory mechanisms, deepen our understanding of these biases, and shed light on how beliefs are shaped by contextual cues.The model does not explain several other judgment biases that have been documented by KT (1972) and others, such as the law of small numbers or biases due the anchoring and to availability heuristics. We return to these in the conclusion.[3]

The paper is organized as follows. Section 2 describes the model and the experimental framework. Section 3 describes our first set of experiments in which representativeness is controlled by varying the comparison group. Section 4 presents the experiment in which the database is the same for all participants but recall is cued along different dimensions. Section 5 discusses our findings and Section 6 concludes.

[3]Growing literatures in both Neuroscience and Economics explore the role of memory for valuation and decision making, particularly the role of cues in selectively retrieving information. Bornstein et al (2017) find that choices among risky alternatives can be altered by showing subjects images that co-occurred with payoffs in the past, even if these images are totally uninformative. The authors argue that the subjects' behavior is consistent with a cued recall mechanism: whether the subjects respond to the cue is commensurate with the reinstatement in the brain's visual areas of patterns associated with the cue. Enke, Schwerter and Zimmerman (2019) give experimental subjects news about a hypothetical asset, associating certain images with good news and others with bad news. When assessing the asset's value, subjects overreact to good news associated with images that co-occurred with past good news. This result is in line with the model of memory-based choice of Bordalo, Gennaioli, and Shleifer (2019) in which context (an image) cues recall of past experiences associated with it. Several papers document that individuals' decisions are shaped by recall of past experiences: assessments of expected inflation are based on goods bought frequently (Cavallo, Cruces, Perez-Truglia 2017), and willingness to participate in the stock market depends on own past returns (Malmendier and Nagel 2011, Wachter and Kahana 2019).

2 Model and Experimental Framework

2.1 The Model

Our model captures the idea that probability judgments, such as assessing the probability distribution of hair color among the Irish, are formed by selectively retrieving relevant instances from memory. Retrieval of different hypotheses (hair colors) is cued by the data (the Irish), and is driven by their similarity. We show that Tversky's (1977) model of similarity among sets yields a form of interference in recall that is critical to generating representativeness effects and several other predictions explored below.[4]

The memory database is described by a probability space with event space Ω and probability measure P, which summarizes a person's experiences including their frequencies. Two random variables are defined on this space: H, which we think of as the hypotheses the DM seeks to assess, and D, which we think of as data. The task of the decision maker is to assess the probability of hypothesis $h \in H$ given data $d \in D$. In our running example, Ω is the universe of people, $D =$ nationalities and $H = \{dark, light, red\}$ are different hair colors.

A Bayesian retrieves all possible hypotheses h given d and computes the true conditional probability $P(h|d)$ using the measure P.[5] Relative to this benchmark, in our model the data d cues retrieval of more similar hypotheses from memory, facilitating recall of certain hypotheses relative to others, and causing their probability in d to be inflated.

To formalize the similarity between data d and a hypothesis h we rely on Tversky's (1977) similarity metric, which applies to objects characterized as collections of features. Given two objects a and b with sets of features A and B, respectively, the similarity of a to b is defined as:

$$S(a,b) = \alpha f(A \cap B) - \beta f(A/B) - \gamma f(B/A), \tag{1}$$

where $A \cap B$ is the subset of features that are common to both objects, A/B and B/A are the features that are present in one object but not in the other, $f(.)$ is an increasing function, and α, β, γ are non-negative coefficients. According to Equation (1), the similarity between a and b is ceteris paribus higher when they have more features in common, and when each set has fewer distinct features not shared with the other.

[4] Tversky's (1977) model of similarity has also been used to study the problem of categorization, which is distinct from what we consider here. Kemp, Bernstein, and Tenenbaum (2005) and Gershman (2017) show that testing whether two objects have the same data generating process reduces, under some assumptions, to a measure of non-Euclidean similarity similar to Tversky's model.

[5] Formally, the decision maker computes the conditional probability by retrieving all elementary events ω consistent with each hypothesis h and data d, and by using the probability measure P to compute: $P(H = h|D = d) = \frac{\int_{\omega \in \Omega: H(\omega)=h, D(\omega)=d} dP(\omega)}{\int_{\omega \in \Omega: D(\omega)=d} dP(\omega)}$.

Equation (1) can be applied to the context of probabilistic assessments by noting that data d and each possible hypothesis h are also characterized by features, namely the different possible values of the random variables under consideration. For instance, $d =$ Irish is the set of people who are Irish but can have different hair color (red, light, dark), while $h = red$ is the set of people who have red hair and may have different nationalities (Irish, Spanish,...).

In this interpretation of Equation (1), the similarity between the set of red haired people and the set of Irish people depends on the size of three subsets. Similarity $S(d,h)$ between d and h increases in the size $P(d,h)$ of the set of people who are both Irish and have red hair. But similarity between d and h decreases in the size $P(d,-h)$ of the set of people who are Irish but do not have red hair, as well as in the size $P(-d,h)$ of the set of people who have red hair but are not Irish. Formally, the similarity between hypothesis h and data d is given by:

$$S(d,h) = \alpha f(P(d,h)) - \beta f(P(d,-h)) - \gamma f(P(-d,h)). \tag{2}$$

Similarity between hypothesis h and data d increases in their correlation: the more h and d occur together. In the previous example, for a given size of the red-haired Irish population $P(d,h)$, similarity between Irish and red hair is maximized when all Irish are red haired, $P(d,-h) = 0$, and when red hair occurs only among the Irish, $P(-d,h) = 0$.

Similarity is linked to retrieval through Luce's rule. That is, the probability of recalling h when cued with d is given by:

$$\widetilde{P}(h|d) = \frac{e^{S(d,h)}}{\sum_{h' \in H} e^{S(d,h')}}. \tag{3}$$

When cued with d, a hypothesis h more similar to d is more easily retrieved. Retrieval is normalized, so that recall probabilities add to one. Normalization captures a form of interference whereby recall of more similar items inhibits that of less similar ones. The decision maker estimates the probability of h in d by measuring the frequency with which h comes to mind among all retrieved d members. Through the law of large numbers, Equation (3) also yields the probability that subjects on average attach to hypothesis h given data d.

According to (3), recall of h, and thus its assessed probability, follow the well known law of frequency. The more people are both Irish and have red hair (higher $P(d,h)$), the more likely is $d =$ Irish to cue retrieval of $h =$ red hair. This effect is common to most memory-based models of probability assessments (Anderson and Reder 1999, Dougherty, Gettys, Ogden 1999, Kahana 2012).

The law of frequency is however subject to two kinds of interference. The first is

interference across different hypotheses. A person who has seen many dark haired Irish (high $P(d,-h)$) is less likely to recall red hair. This is due to both the β term in the similarity function (2) and the normalization of recall in Luce's rule (3). Many memory models, including Minerva DM and sampling, display this form of interference, which tends to produce underestimation of unlikely traits.

The second, and critical, form of interference occurs across different data. For a decision maker who has seen few non-Irish people with red hair (which implies a low $P(-d,h)$), Irish is more associated with red hair, and retrieval of red haired Irish is increased. Conversely, when thinking about dark hair among $d = Irish$, Spaniards or Italians may come to mind because this hair color is much more common among these other nationalities $-d$. This interferes with recalling dark haired Irish. This form of interference flows back from hypothesis h to different data $-d$.

This second mechanism only operates when $\gamma > 0$, and implies that *distinctive* traits are overestimated, even if they are unlikely. We call this the "kernel of truth" property. It leads subjects to overestimate a hypothesis h that is more associated with the current data d relative to other data $-d$, even though the latter association is normatively irrelevant. Our experiment tests this prediction precisely by varying the normatively irrelevant "decoy" distribution $P(-d,h)$ while holding the relevant probabilities $P(d,h)$ and $P(d,-h)$ constant.[6]

Consider a convenient formulation in which $f(x) = \ln(c+x)$, where c is a non-negative constant, and $\beta = 0$.[7] The assessed probability $\widetilde{P}(h|d)$ then becomes:

$$\widetilde{P}(h|d) \propto \frac{[c + P(h|d)P(d)]^{\alpha}}{[c + P(h|-d)P(-d)]^{\gamma}}. \tag{4}$$

Parameter α captures the strength of frequency-based recall. When α is high, likely hypotheses are oversampled from memory and their probability is inflated. For low α, the reverse holds: sampling depends less on $P(h|d)$, so that low likelihood hypotheses are oversampled across the board.

Parameter γ instead captures contextual interference, or the kernel of truth. If γ is high, when cued by data d we oversample hypotheses that are unlikely to occur with other data, i.e. with low $P(h|-d)$. The decision maker then inflates the probability of such distinctive hypotheses. The parameter c pins down a baseline probability of recall-

[6]Another difference from existing memory models is that in our setup encoding is perfect, not noisy as in the Minerva-DM of Dougherty, Gettys, Ogden (1999). Noisy encoding could be added to our model, but we stress that interference across groups relies on fairly accurate memories, in the sense that the co-occurrence of types and groups is correctly recorded. This is critical to obtain belief distortions that depend on the true features of the data.

[7]In our two-hypothesis setting, both β and α modulate the extent to which similarity $S(d,h)$ increases in the likelihood $P(d,h)$, so β is redundant.

ing a hypothesis, regardless of its frequency, which caps distortions when a hypothesis does not occur at all given the decoy data $-d$.

Equation (4) flexibly nests different kinds of judgments, both rational and biased. When $\alpha = 1$ and $c = \gamma = 0$, the model collapses to Bayesian inference. When instead $\alpha > 1$, and $c = \gamma = 0$ the model generates a form of over-reaction to data, in the sense of overestimation of hypotheses that are more likely given d, reminiscent of overconfidence (Moore and Healy 2008). Conversely, when $\alpha < 1$ the model generates a form of under-reaction to data. In the extreme case of $\alpha = 0$, all hypotheses are recalled with the same probability, so that unlikely ones are mechanically oversampled.

When $\gamma = 0$, the model fails to yield context-dependent beliefs and overestimation of distinctive hypotheses. This occurs only when $\gamma > 0$, so that assessements about d depend on the decoy distribution $-d$. This mechanism generates representativeness effects. In fact, if $\alpha = 1 + \gamma$ and $c = 0$, Equation (4) becomes:

$$\widetilde{P}(h|d) \propto P(h|d)\left[\frac{P(h|d)}{P(h|-d)}\right]^{\gamma}. \tag{5}$$

which exactly captures Kahneman and Tversky's (1983) definition of representativeness: "an attribute is representative of a class if it is very diagnostic; that is, the relative frequency of this attribute is much higher in that class than in a relevant reference class." In this case, our model yields the BCGS (2016) definition of representativeness and diagnostic expectations (BGS 2018), which direcly follow KT's definition, as special cases of similarity based recall.

Our experiments test the predictions that follow from $\gamma > 0$ by varying the decoy data $-d$ across treatments. Our experiments then allow us to quantify the parameters of the model, and hence to assess Equation (4) quantitatively.

2.2 Experimental Design

We now describe our main experiments, and the model's predictions for each of them.

Study 1. Participants are shown a sequence of 50 abstract images that vary along two dimensions, content and color. They are randomly assigned to one of two treatments: a condition in which they see 10 orange numbers, 15 blue numbers and 25 gray shapes ($gray$ treatment); or a condition in which they see the same 10 orange numbers and 15 blue numbers, but also 25 blue words (the $blue$ treatment). The memory databases are described below.

After observing the sequence of 50 images, participants are asked several questions. On the first screen, they are asked:

Table 1: Databases of Study 1

Image database in the *gray* treatment

		Hypotheses:		
		*o*range	*b*lue	*g*ray
Data:	*n*umbers	10	15	0
	decoy (*s*hapes)	0	0	25

Image database in the *blue* treatment

		Hypotheses:		
		*o*range	*b*lue	*g*ray
Data:	*n*umbers	10	15	0
	decoy (*w*ords)	0	25	0

(Q1) "An image was randomly drawn from the images that were just shown to you. The chosen image showed a number. What is the likely color of the chosen image?"

On the following screen, participants are given the same scenario but then asked:

(Q2) "What is the probability the number is orange?"

Finally, on the next screen, participants are asked:

(Q3) "How many orange numbers were shown to you?", and

(Q4) "'How many blue numbers were shown to you?"

In both treatments, subjects assess the distribution of the colors of the numbers, but the decoy distribution varies across treatments. All our experiments build on this basic design, which extends the preliminary experiments in BCGS (2016).

In Q1 and Q2 subjects are cued with the data $d =$ number and are asked to assess the probability of different colors. In Q3 and Q4 they are asked to recall how many orange and blue numbers were shown to them. Our model implies that assessments of likelihood and recall go hand in hand, and are shaped by the critical contextual interference parameter γ in Equation (4). Formally, Appendix A shows that Equation (3) yields the following predictions.

Prediction 1. *If $\gamma > 0$, the blue treatment reduces the assessed likelihood that the randomly drawn number is blue in Q1 and Q2 relative to the gray treatment. If $\gamma = 0$ there is no treatment effect.*

and

Prediction 2. *If $\gamma > 0$, the blue treatment reduces the share of blue numbers recalled in Q3 and Q4 relative to the gray treatment. There is also a positive correlation at the individual level between the share of blue numbers recalled from Q3 and Q4 and the assessed likelihood of the number being blue in Q1 and Q2. If $\gamma = 0$ there is no treatment effect.*

If $\gamma > 0$, in the blue treatment the retrieval of blue words interferes with the retrieval of blue numbers from the data d = number. Thus, blue numbers are undersampled relative to the gray treatment, in which interference in the retrieval of numbers of any color is not at work. As a result, the probability of the less likely hypothesis – that a number is orange – is overestimated in the $blue$ treatment. If instead $\gamma = 0$, the decoy distribution is irrelevant so there is no treatment effect.

Study 2. In this experiment, we manipulate the strength of contextual interference by creating five new variants of the decoy data (the distribution of numbers is always kept the same). We denote by $blue_k$ a treatment in which we replace k blue words with orange words. We allow for six such treatments varying the frequency of orange words $k \in \{0, 1, 3, 6, 10, 22\}$ (where $k = 0$ coincides with the blue treatment in Study 1). The model generates the following prediction.

Prediction 3. *If $\gamma > 0$, the assessed probability that a random number is orange in Q1 and Q2 and the share of recalled orange numbers in Q3 and Q4 decrease with the number of orange words k in the decoy distribution. If $\gamma = 0$ there is no treatment effect.*

Prediction 3 is a more stringent test of the kernel of truth. According to Equation (4), the probability of retrieving orange given d = number and thus the assessed likelihood of orange numbers both decrease in the share of orange words, because orange words interfere with the recall of orange numbers.

Study 3. In our final experiment, we test a key implication of the model: features of the task that shape perceptions of similarity affect memory retrieval and thus probability judgments, even if they are normatively irrevelant.

To explore this mechanism, we build on Study 1 where images are characterized by G (content) and T (color). We now let images vary also in a third attribute, specifically font size, $T' \in \{small, large\}$. The database for all participants is now:

All orange numbers are large and all blue numbers are small, while all words are blue and large.

We then run two treatments that vary only in the questions asked. In the Color treatment, we ask "What is the likely color of a randomly drawn number?" (Q1) and "What

Table 2: Database in Study 3

10 orange, large numbers
15 blue, small numbers
25 blue, large words

is the probability of a randomly drawn number being orange?" (Q2). In the Size treatment, we instead ask "What is the likely font size of a randomly drawn number?" and "What is the probability of a randomly drawn number being large?".

The point of this experiment is that the similarity of the data d = number to a given color, say orange, is different from the similarity of the same data d to the corresponding size, namely large. This implies that asking subjects to assess color will produce different judgments than asking them to assess size, even though the two tasks are normatively identical.[8] The model yields the following prediction.

Prediction 4. *If $\gamma > 0$, the share of recalled orange/large numbers and the assessment of the probability that a random number is orange/large is higher in the color treatment than in the size treatment. If $\gamma = 0$ there is no treatment effect.*

In the color treatment, contextual interference inhibits the recall of blue numbers due to the presence of blue words. In the size treatment, contextual interference inhibits the recall of large numbers due to the presence of large words. But because large numbers are orange, the size treatment inhibits the recall of orange numbers and their assessed likelihood relative to the color treatment. This effect of course only arises if $\gamma > 0$.

2.3 Discussion

Our model makes three distinctive predictions relative to the existing models of biased probabilistic assessments, and our experimental design allows us to test them.

First, our approach is based on selective recall. We measure recall directly in our experiment in Q3 and Q4. Some alternative approaches do not rely on basic cognitive functions, such as memory. Rather, they assume specific distortions without exploring their cognitive basis. For instance some papers assume that subjects compute $P(h|d)$ using $P(d|h)$ (Gigerenzer and Hoffrage 1995, Kahneman and Frederick 2002, Villejoubert

[8]Because in this experiment the distribution of images seen during the viewing stage is *identical* across treatments, there is no reason for a participant in one treatment to attend to or encode images differently from a participant in another treatment. This helps rule out the possibility that differential attention during encoding is driving our results. Although a precise distinction between selective encoding and selective recall is beyond the scope of this paper, it is useful to distinguish a process in which the instability of probabilistic assessments arises from the selectivity of retrieval from the memory database, rather than from a permanent distortion of the memory database itself, because of inattention at the encoding stage. In the current experiment, the first mechanism seems to be driving the results.

and Mandel 2002, Koehler et al. 2003, for a review see Nilsson et al. 2005), or that representative types provide strong confirmation of hypotheses (Tentori, Crupi, and Russo 2013). These distortions are tailored to specific inference problems, and lead to inaccurate predictions in many settings, including in our experiment.[9]

Second, relative to other models emphasizing memory limitations, our approach makes the distinctive "kernel of truth" prediction, whereby recall and assessments are context-dependent and exaggerate true correlations in the data (e.g, the association between the cue “number” and the type “orange”). This prediction is absent from other memory-based models. In sampling models (Sanborn and Chater 2016), subjects randomly sample the memory database, so the decoy distribution should not affect beliefs because it does not distort the sampling of numbers. Because in the Minerva DM model (Dougherty et al. 1999) retrieval is noisy, it generates imprecision but not systematic distortions depending on the decoy distribution.[10] The same is true of exemplar models like PROBEX (Juslin and Persson 2002). In these models, subjects cued by “number” retrieve exemplars of different colors on the basis of a Euclidean similarity function that does not depend on the decoy distribution. As a consequence, these models cannot deliver Predictions 1, 3 and 4.

Evidence in line with Prediction 4 would offer particularly striking support to the role of contextual interference, which gives rise to the kernel of truth. Since in Study 3 size and color are perfectly correlated within the group of numbers, the cue should play no role for Bayesian subjects, nor should it play a role in sampling models (where numbers are correctly represented on average), in exemplar models (where large numbers and orange numbers are equally similar to the average number), or in prototype models (where the prototype is the most likely number, which is always blue).

---

[9]One intuition behind these models is that the inverse conditional $P(number|orange)$ is more accessible to decision makers than the target one $P(orange|number)$ (Kahneman and Frederick 2002). This might be the case when experimental subjects are given the former and are asked to estimate the latter, as in inference problems. This is less so in the general case where subjects retrieve a distribution from memory, where this approach leads to problems. For example, in the relative likelihood model of Villeboeuf and Mandel (2002) the assessed probability of orange numbers is over 50% as long as $P(number|orange) > P(number|blue)$, even as the true share $P(orange|number)$ goes to zero. A smoother formulation of the inverse conditional is a mechanical neglect of base rates, whereby the odds ratio $\frac{P(orange|number)}{P(blue|number)}$ is assessed as $\frac{P(number|orange)}{P(n|b)}\left(\frac{P(orange)}{P(blue)}\right)^{\phi}$, with $\phi < 1$ (Bayes' rule corresponds to $\phi = 1$). This formulation predicts that if a prior is sufficiently strong, say $P(orange|number)$ is high, then a signal that supports the prior (i.e. that the number is in fact orange) *reduces* the posterior probability assigned to that type. While our experiments do not cover this particular point, existing evidence generally indicates that individuals update their beliefs in the direction of their signals.

[10]Dougherty et al. (1999) argue that their model can also account for some manifestations of the representativeness heuristic. However, these effects are obtained by changing the model's formulation in ways that are unrelated with its core assumptions (namely, the authors assume that, when cued with data d and asked to assess hypothesis h, subjects spontaneously self-cue the hypothesis h and effectively assess $p(d|h)$, because the latter is assumed to be more accessible). Our model yields representativeness as a special case of a general process.

Finally, our approach to similarity is based on experienced statistical associations between items in memory. This occurs without the need of biases in social information transmission that emphasize stereotypical traits such as in Bhatia's (2017) model of semantic associations. Our experiment rules out the role of such semantic associations by focusing on abstract objects.

3 Representativeness and Selective Recall

3.1 Study 1: Baseline Experiment

In Study 1 we compare probabilistic assessments in the $blue$ treatment (10 orange numbers, 15 blue numbers, 25 blue words) and the $gray$ treatment (10 orange numbers, 15 blue numbers, 25 gray shapes), as in Table 1.[11] According to Prediction 1, retrieval of blue numbers is interfered with in the $blue$ treatment so that participants inflate the frequency of orange numbers relative to the $gray$ treatment, $\widetilde{P}(o|n)_{blue} > \widetilde{P}(o|n)_{gray}$. Mapping Prediction 1 to our output measures, we test whether participants in the $blue$ treatment: i) are more likely to choose orange as the likely color, ii) assign a higher probability to orange numbers, and iii) report a higher share of orange numbers compared to participants in the $gray$ treatment, computed from answers to Q3 and Q4.

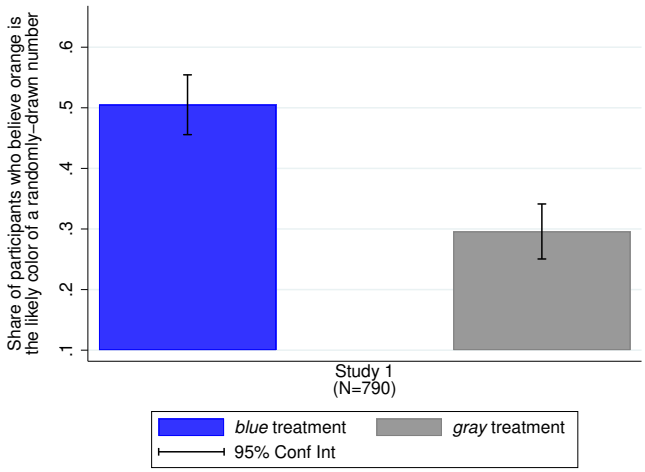


Figure 1: Share of participants who believe that the likely color of a randomly-drawn number is orange for the $blue$ and $gray$ treatments of Study 1.

The results are summarized in Figure 1 and Table 3 and show strong support for Prediction 1. Column (1) in Table 3 reports an OLS regression of a response dummy

[11] This Study was conducted in the laboratory in Spring of 2018, in two waves with $N = 427$ (University of Cologne) and $N = 363$ (Bocconi University) respectively. Here we present the aggregated results. We report the procedural details and disaggregated results (Tables B.2 and B.3) in Appendix B.1.

(1 if "orange is likely") on a treatment dummy (1 if *blue*) and a location dummy (1 if Milan), that amounts to comparing the average share of participants who said orange is likely in the *gray* treatment versus the *blue* treatment. As shown in Figure 1, that share increases 21.1pp from the *gray* treatment to the *blue* treatment (31.1% to 52.2%, significant at 1% level).[12] Column (2) shows the median probability of orange numbers is significantly higher in the *blue* treatment than in the *gray* treatment (significant at the 1% level).

Table 3: Regression estimates of treatment effects in Study 1 (lab with distraction)

	OLS: Y=1 if "orange is more likely"	0.5-Q-Reg: Y= Probability that a randomly-drawn number is orange	0.5-Q-Reg Y= Orange numbers recalled	0.5-Q-Reg Y= Blue numbers recalled	OLS: Y=1 if more orange numbers recalled	0.5-Q-Reg: Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)	(6)
1 if *blue*	.2111***	.07***	0	− 2***	.1667***	.0556***
	(.0342)	(.0137)	(.6078)	(.5878)	(.0329)	(.0124)
Location dummy	yes	yes	yes	yes	yes	yes
Constant	.3113***	.43***	12***	15***	.2576***	.4444***
	(.0284)	(.0144)	(.5046)	(.4880)	(.0273)	(.0114)
Observations	790	790	790	790	790	790
Adj./Ps. R^2	0.04	0.03	0.02	0.01	0.03	0.02

Notes: This table presents estimates of the treatment effect on several outcome measures relating to our model predictions in Study 1 with distraction. Columns (1) and (5) present OLS regression of response dummies (1 if "orange is likely" in Column (1) and 1 if "recalled more orange than blue numbers" in Column (5)) on a treatment dummy (1 if *blue*). Columns (2), (3), (4) and (6) present 0.5-quantile regressions of subjects' stated probability that a random number is orange, number of recalled orange numbers, number of recalled blue numbers and share of recalled orange numbers, respectively, on the treatment dummy. In all regressions, a location dummy is included. Standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Columns (3) and (4) show how the median quantity of recalled orange and blue numbers depends on the treatment.[13] Responses in the *gray* treatment are quite accurate, as indicated by the constant term. In the *blue* treatment, participants retrieve fewer blue numbers, consistent with interference from blue words.

We next compare how data on recall (Q3 and Q4) matches assessments of likelihood (Q1 and Q2). First, we report in Column (5) the results of an OLS regression of a response dummy that takes value 1 if the participant reported more orange numbers in Q3 than blue numbers in Q4, which is an alternative measure of each participants' belief about the likely color. Consistent with Column (1), there is a significant increase in the share of participants who recalled more orange than blue numbers in the *blue* treatment (25.8% vs 42.4%). The treatment dummy coefficients are close in Columns (1) and (5).

[12]Nonlinear Logit regressions yield similar results, Table B.1 in Appendix B.1.

[13]We report median, rather than mean, responses to the amount of numbers recalled because they are by construction less noisy. However, similar results hold for means, see Table B.1 in Appendix B.1.

In fact, answers to Q1, Q3 and Q4 are consistent (in the sense of answering "orange" in Q1 if and only if more orange numbers are recalled in Q3 than blue numbers in Q4) for 87.3% of participants.

Finally, we compute the ratio of orange numbers to total numbers recalled, which is an alternative measure of the (recalled) probability of orange numbers. Column (6) shows that, as predicted by the model, participants recalled on average a significantly higher share of orange numbers in the *blue* treatment (50% versus 44.4%). Again, the coefficients are close in Columns (2) and (6). In fact, the stated probability of orange numbers explains about 75% of the variability in the share of recalled orange numbers across individuals (see Column (1) of Table B.11 in Appendix B.3.1).

The results point to systematic distortions in the retrieval of information, leading to distorted beliefs in the direction predicted by the model of representativeness in Equation (3). The treatment effect represents a jump of over 50% in the frequency of assessing orange as the likely color (Columns 1 and 5), and a 13 to 16% increase in the estimate of the probability of a randomly drawn number being orange (Columns 2 and 6), so it is large both in absolute and in relative terms.[14] This suggests that interference can account for why unlikely traits are accessible after specific cues, offering an explanation for some effects attributed to the representativeness heuristic (KT 1972). In particular, such distortions predictably arise in an environment of purely abstract objects with no pre-existing associations in memory.[15,16]

3.2 Study 2: Varying Relative Likelihood

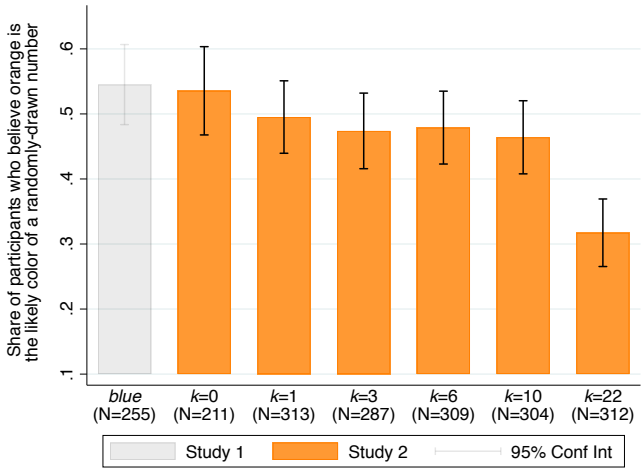
We next test whether, in line with Prediction 3, probabilistic assessments about the color distribution of numbers change as the frequency of orange words increases and interference for recall of orange numbers intensifies. As before, the target group is 25 Numbers

[14]The same experiment was also rerun in the lab and on MTurk without including question Q2 on the probability that a randomly drawn number is orange. The results are strikingly similar: the treatment effect represents a jump in the frequency of assessing orange as the likely color from 29% to 58% (lab) and 41% to 54% (MTurk), and an increase in the estimate of the share of orange numbers recalled from 45% to 48% (lab) and 47% to 50% (MTurk). See Tables B.4, B.5, B.6, and B.7 in Appendix B.1.1.

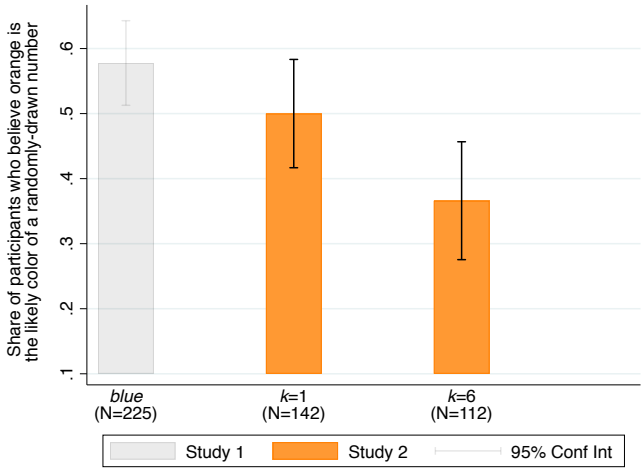
[15]One may ask whether the particular colors chosen impact our results. Research in psychology starting with Goldstein (1942) and Stone and English (1998) proposes that warmer colors, such as red and yellow, may induce outward focusing, while cooler colors may induce inward focusing (reservation), in part due to their different wavelengths. Elliot et al (2007) propose a more general framework in which the impact of specific colors varies by context and is a function of learned associations. It is unlikely that any pre-existing association between orange and numbers drives our results, particularly the across-treatment differences in the experiments where orange is not representative of numbers (see Study 2 and 3). We also ran a separate experiment, reported in Appendix C.3, that replaced color with font size as the types (i.e. the random variable T), and the results are unchanged (see Table C.5).

[16]A clear feature of the data is that the frequency of the "orange is likely" mistake is significant, around 35%, even in the gray treatments where our model predicts low distortions from representativeness. As we show when calibrating the model in Section 5, this distortions pins down the model parameter α, which creates probabilistic distortions even in the absence of contextual interference from a decoy group.

(10 orange and 15 blue). For the decoy group, we create six variants of the words distribution in the *blue* treatment, denoted $blue_k$, which are characterized by replacing k blue words with orange words. We conducted Study 2 on MTurk with $N = 1,738$ and $k = 0, 1, 3, 6, 10, 22$ and in the laboratory with $N = 254$ and $k = 1, 6$. Here we present the disaggregated results and discuss procedural details in Appendix B.2.



(a) Conducted with MTurk



(b) Conducted in Laboratory

Figure 2: Share of participants who believe that the color of a randomly-drawn number is most likely orange for the *blue* treatments with $k = 0, 1, 3, 6, 10, 22$.

As Figure 2 shows, the results are consistent with our predictions. For MTurk experiments (top panel), the share of participants who answered "orange is likely" decreases from 54.1% from the baseline treatment with $k = 0$ to 31% in the variant with most orange words, $k = 22$. The figure suggests a decline between $k = 0$ and $k > 0$ that becomes particularly strong from $k = 10$ to $k = 22$. The decline is statistically significant at the 5% level between $k = 0$ and $k \geq 6$ and at the 1% level between $k \leq 10$ and

$k = 22$.[17]

Similar results hold in laboratory experiments, shown on the lower panel. Increasing the number of orange words from 1 to 6 reduces the share of participants stating that "orange is likely" by 21.2pp (significant at the 5% level). This again points to a general trend, as can be seen by comparing these findings with the results from Study 1.[18]

Table 4 summarizes the pattern described above for the experiments run in the lab. Column (1) reports an OLS regression of a response dummy (1 if "orange is likely") on the actual amount of orange words participants are exposed to. The significant negative coefficient implies that the share of participants who believed that "orange is likely" decreases by 2.7pp per orange word added. Similarly, the median assessed likelihood of orange numbers drop by 2pp per orange word added (Column 2). Turning to data on recall, the share of participants who recalled seeing more orange numbers as well as the share of orange to total numbers recalled declines in the amount of orange words, Column (5) and Column(6), respectively.[19]

Again, results obtained on MTurk are quantitatively very similar. The share of participants who say "orange is likely" drops by 1pp per orange word added. Turning to data on recall, as orange words are added, participants recall more blue numbers and are less likely to recall seeing more orange numbers (see Table B.10 in Appendix B.2).

Study 2 provides evidence that the relative frequency of types shapes the magnitude of belief distortions. In line with Prediction 3, as the number of orange words increases, participants are less likely to recall a randomly-drawn number as orange, despite the fact that the number of orange numbers is held constant across these variants.

Studies 1 and 2 include waves of our experiment in the Lab with 1044 participants (further robustness runs described in Appendix B.1.1 raise these figures to 8 waves of our experiment and a total of 4706 participants). The evidence supports our predictions consistently across all waves of our experiment.

---

[17]Increasing the number of orange words from 0 to 1, 3, 6, 10, and 22 reduces the share of MTurk participants stating that "orange is likely" by 5.3pp, 7.4pp, 7.0pp, 8.4pp, and 23.1pp, respectively. While the difference between 0 orange words and 1 orange words is not statistically significant, the remaining ones are at least marginally significant when compared to the $blue_{k=0}$ treatment, with respective OLS p-values of 0.070, 0.085, 0.037, and < 0.001. Pair-wise differences in the assessed probability that a randomly selected number is more likely to be orange comparing across only those treatments with 1, 3, 6, and 10 orange words are not large and not statistically significant from zero. However, the assessed probability that a random number is more likely to be orange is greater in the treatments with 1, 3, 6, or 10 orange words than in the treatment with 22 orange words. Pair-wise tests yield OLS p-values below 0.001 in each of these cases.

[18]Study 1's $blue$ treatment is equivalent to $k = 0$. The $gray$ treatment is not directly comparable, but to the extent that no number color is particularly representative in that treatment it is similar to $k = 10$.

[19]Table 4 shows no effect on the median number of recalled blue or orange numbers from the addition of five orange words. Some evidence suggests that this is due to power limitation of this exercise: i) there is a significant negative effect on the mean (see Appendix B.2), and ii) there is also a significant effect in the median in the MTurk experiment, where a larger range of orange words was introduced.

Table 4: Regression estimates of treatment effects in Study 2 (lab only)

	OLS: Y=1 if "orange is likely"	0.5-Q-Reg: Y= Probability that a random number is orange	0.5-Q-Reg: Y= Orange numbers recalled	0.5-Q-Reg: Y= Blue numbers recalled	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg: Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)	(6)
k (number of orange words)	−.0268** (.0125)	−.02*** (.0071)	0 (.1466)	0 (.2108)	−.0164 (.0121)	−.0123** (.0062)
Constant	.5268*** (.0506)	.5200*** (.0288)	10*** (.4867)	15*** (.5943)	.4107*** (.0491)	.4738*** (.0251)
Observations	254	254	254	254	254	254
Adj./Ps. R^2	0.02	0.02	0.00	0.00	0.01	0.01

Notes: This table presents estimates of the treatment effect on several outcome measures relating to our model predictions in Study 2. We report on our results from the lab here, because we elicited participants' expected probability that a random number is orange not via MTurk. Columns (1) and (5) present OLS regression of response dummies (1 if "orange is likely" in Column (1) and 1 if "recalled more orange than blue numbers" in Column (3)) on k, the number of orange words. Columns (2), (3), (4) and (6) present 0.5-quantile regressions of subjects' stated probability that a random number is orange, number of recalled orange numbers, number of recalled blue numbers and share of recalled oranges numbers, respectively, on $k = 6$. * p<0.1, ** p<0.05, *** p<0.01.

4 Study 3: Modulating Recall through Cues

In the final experiment, we attempt to better isolate the role of similarity-based recall in driving our results by testing Prediction 4: cueing assessments of a target group along different dimensions can trigger different patterns of interference across groups and generate different probabilistic assessments.

We implement the distribution of images described in Table 2, where all blue numbers are small while all other items (orange numbers and blue words) are large. According to Prediction 4, assessment of a number's size entails interference with recall of large (and orange) numbers, while assessment of a number's color entails as before interference with recall of blue (and small) numbers. As a result, the estimated likelihood of large, orange numbers is larger in the second case.

The results support this prediction. Participants assess "orange is likely" in the Color Cue treatment significantly more often than "large is likely" in the Size Cue treatment (40% versus 17%, significant at the 1% level). Column (1) of Table 5 shows the result of regressing a question-response dummy (equal 1 if "orange is likely" in Color Cue or "large is likely" in Size Cue) on a treatment dummy (equal 1 if Color Cue, equal 0 if Size Cue), while controlling for when the treatments were conducted.

Table 5: Regression estimates of treatment effects in Study 3

	OLS: Y=1 if "orange OR large is likely"	OLS: Y= 1 if more orange OR large numbers recalled	0.5-Q-Reg: Y= Share of orange OR large to total numbers recalled	0.5-Q-Reg: Y= Probability that a randomly-drawn number is orange OR large
	(1)	(2)	(3)	(4)
1 if color cue	.2296***	.1168***	.04**	.05**
	(.0343)	(.0341)	(.0178)	(.0218)
Wave dummy	yes	yes	yes	yes
Constant	.1808***	.1953***	.41***	.35***
	(.0293)	(.0291)	(.0152)	(.0186)
Observations	647	647	647	647
Adj./Ps. R^2	0.06	0.02	0.01	0.01

Notes: This table presents estimates of the treatment effect on several outcome measures relating to our model predictions. Columns (1) and (2) present an OLS regression of response dummies (1 if "orange is likely" in treatment color cue and 1 if "large is more likely" in treatment font size cue for Column (1) and 1 if "more orange numbers are recalled than blue numbers" in treatment color cue and 1 if "more large numbers are recalled than small numbers" in treatment font size cue for Column (2)) on a treatment dummy. Columns (3) and (4) present 0.5-quantile regressions of subjects' share of recalled orange (treatment $color\ cue$) OR large (treatment $font\ size\ cue$) numbers over total recalled numbers and stated probability that a random number is orange (treatment $color\ cue$) OR large (treatment $font\ size\ cue$), respectively, on the treatment dummy. In all regressions, a wave dummy is included. Standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

As in Tables 3 and 4, we use participants' estimates of how many images of each type they saw to address more directly their retrieval of numbers. First, we compare the share of participants who recalled seeing more orange than blue numbers in Color Cue with the share of participants who recalled seeing more large than small numbers in Size Cue. We find, analogously to our main result, that a significantly greater share of participants recalled more orange than blue numbers than more large than small numbers; see Column (2) of Table 5. Second, we check the average share of recalled large orange numbers and find, consistent with the account of interference-based distorted recall, that on average participants recalled a significantly higher share of large orange numbers under the color cue, see Column (3) of Table 5.

Finally, we examine the results from direct probability estimates in Q4. When asked to predict the probability that a randomly-drawn number is orange/large, participants on average state a significantly higher probability that a random number is orange in the Color Cue treatment than that a random number is large in the Size Cue treatment; see Column (2) of Table 5. These findings suggest that participants' retrieval of blue and orange numbers from their image database was differentially distorted depending on

whether color or font size was cued in the question stage.

5 Model Calibration and Assessment

The previous sections document that elicited beliefs about a given group, numbers, are shaped by its comparison to a reference group, in line with the predictions of our model. In fact, there is significant and systematic variation across treatments in subjects' average estimation of the likelihood $\widetilde{P}(o|n)$, ranging from 0.36 to 0.499. To assess how well the model can explain the observed variation in beliefs, and to quantitatively estimate the role of interference in the model, we conduct a calibration exercise.

Beliefs about numbers can be summarized in the elicited odds $\frac{\widetilde{P}(o|n)}{\widetilde{P}(b|n)}$ of orange to blue numbers. Our calibration is targeted to match this moment as it varies across treatments. Starting from Equation (3), we obtain:

$$\frac{\widetilde{P}(o|n)}{\widetilde{P}(b|n)} = e^{S(n,o)-S(n,b)}.$$

Assuming, as in Section 2, that $\beta = 0$ and using the same logarithmic functional form $f(x) = \ln(c+x)$, the predicted assessed odds of orange numbers are given by:

$$\frac{\widetilde{P}(o|n)}{\widetilde{P}(b|n)} = \frac{(c+P(o,n))^{\alpha-\gamma} \cdot \left(\frac{c+P(o,n)}{c+P(o,-d)}\right)^{\gamma}}{(c+P(b,n))^{\alpha-\gamma} \cdot \left(\frac{c+P(b,n)}{c+P(b,-d)}\right)^{\gamma}} \tag{6}$$

where $P(o,n)$ and $P(b,n)$ are the frequency of orange and blue numbers among all images shown to participants, while $P(o,-d)$, $P(b,-d)$ are the frequency of orange and blue decoy images among all images shown to participants.

Table 6

Study	Treatment $-d$	Experimental Odds
1 (lab)	gray shapes	0.78
1 (lab)	words $k=0$	0.99
2 (lab)	words $k=1$	0.93

Notes: Target moments.

We next pin down the three parameters α, γ and c by matching the odds (6) with the odds computed from the average assessment across subjects of the likelihood of orange numbers. We match three predicted odds to their empirical counterparts in three

different conditions. Table 6 shows the treatments used for calibration and the corresponding empirical target odds. The first two moments come from the $blue$ and $gray$ conditions of our baseline Study 1 in the Lab. The third moment comes from Study 2, and in particular from the treatment with one orange word ($k = 1$). Across these three conditions, in Equation (6) the target distribution of numbers is unchanged, while the decoy distribution changes.[20]

We invert the resulting system of three Equations (6) using standard numerical methods, and obtain:

α^*	γ^*	c^*
0.66	0.07	0.01

Our numerical solution achieves a good fit of the empirical odds (least squared error of 9×10^{-4}, depicted in Figure 4) which suggests that the estimates are very close to the exact solution of the system. Parameter γ is tightly estimated, as shown in Figure 3, indicating an important role for interference across groups in shaping beliefs. According to the model, replacing gray shapes with blue words generates a 7pp increase in the perceived probability of orange numbers, from 0.43 to 0.50.

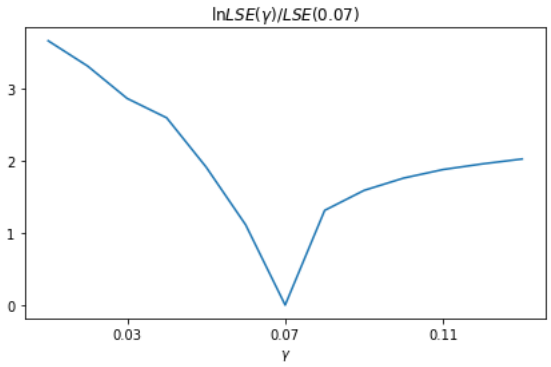


Figure 3: Model fit. For each value of γ, we compute the least square error over parameters α, c, namely $LSE(\gamma) = \min_{\alpha,c} \sum_i \left(\frac{\tilde{P}(o|n)_i}{\tilde{P}(b|n)_i} - \frac{P(o|n)_i}{P(b|n)_i} \right)^2$ where i indexes the three target moments in Table 6. $LSE(\gamma)$ attains a minimum at $\gamma^* = 0.07$. The Figure plots the ratio $\frac{lnLSE(\gamma)}{lnLSE(\gamma^*)}$ as a function of γ.

Using these estimates, we next examine the model's performance out of sample in: i) assessments of likelihood in treatments not targeted by the calibration, ii) assessments of modal color across Studies 1, 2 and 3, iii) recall of orange and blue numbers across Studies 1, 2 and 3, and iv) recall of orange and blue numbers in supplementary experiments where the number of decoys is varied.

[20]The experimental data includes six moments on probabilistic assessments of the probability of orange numbers, which can be mapped directly to the model. The three chosen target moments come from the three most basic experiments. We then use the remaining moments for out-of-sample tests of model performance.

We start by computing the model's predictions for assessments $\widetilde{P}(o|n)$ in the remaining experiments where we elicited likelihood assessments, namely in Lab Study 2, $g = words(k = 6)$ and in Lab Study 3, both the size and color treatments. Figure 4 shows the results (blue dots), as well as the fit with the target moments (blue circles). To facilitate the assessment of model performance, Figure 4 also shows the predictions of the calibrated model where interference is shut down ($\gamma = 0$, dotted line) as well as the 45° line that would obtain if the model would perfectly match the evidence. The calibrated model reproduces the large observed variation in the assessed probability of orange numbers, and matches fairly well the out of sample moments, in particular the impact of the size cue which leads to a large drop in $\widetilde{P}(o|n)$ to a value much lower than in the target treatments. The calibrated model thus significantly outperforms models that lack contextual interference, $\gamma = 0$, which would predict no variation in $\widetilde{P}(o|n)$ (the horizontal dotted line).

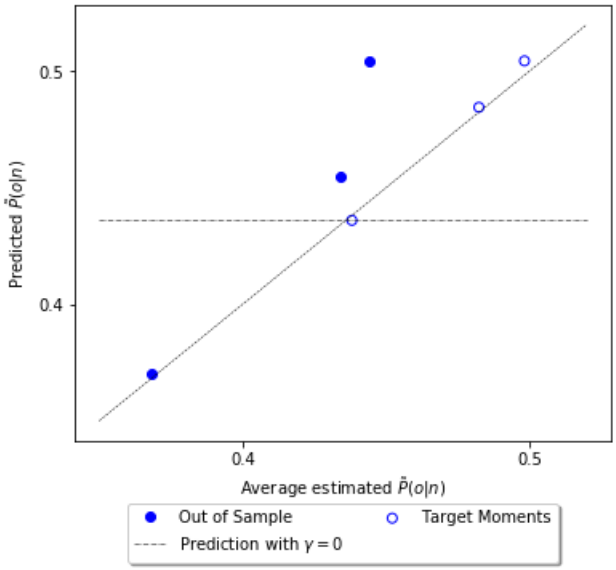


Figure 4: Model performance I: assessments of likelihood, data $\widehat{P}(o|n)$ vs model predictions $\widetilde{P}(o|n)$. The graph also plots the 45° line, as well as the prediction of the calibrated model where contextual interference is shut down, $\gamma = 0$.

We next consider subjects' assessments of which color is modal, which are closely related to likelihood assessments. In every experiment, some subjects say orange and some say blue, reflecting disagreement that is also observed in their assessments of likelihood. To assess model performance in this context, it is not sufficient to use the average assessed probability of orange numbers (which is weakly below 0.5 in every experiment). For the purpose of obtaining a distribution of assessments across subjects, we assume that they sample numbers from memory from a binomial distribution of colors with a probability of orange given by the model predicted likelihood $\widetilde{P}(o|n)$. Specifi-

cally, we assume they sample $N = 25$ numbers, which is the true number of numbers and also coincides with the modal number of numbers recalled. Naturally, averaging the likelihood of orange numbers across subjects again yields $\widetilde{P}(o|n)$. But this process also yields a share of subjects that state that orange is the modal color, which is given by $1-cdf(N/2, N, \widetilde{P}(o|n))$, where cdf stands for the cumulative density function of the binomial distribution.[21] Figure 5 plots the resulting predicted shares against the empirical shares for our studies, which were obtained both in the Lab and on Mturk.[22]

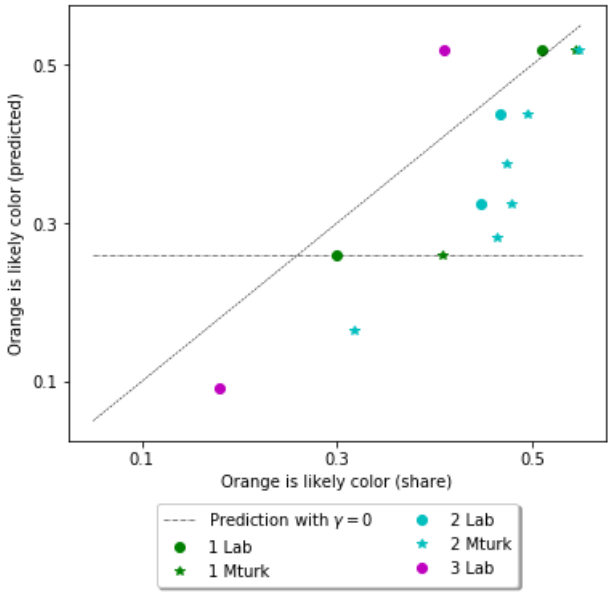


Figure 5: Model performance II: assessments of modal color, data vs model predictions. The graph also plots the 45° line, as well as the prediction of the calibrated model where contextual interference is shut down, $\gamma = 0$.

Again, the model captures quantitatively the large variation in the share of subjects reporting that orange is the modal color, which varies by a factor of nearly three (from 0.18 to 0.52) – a pattern that is absent in models that lack contextual interference (horizontal dotted line). The model tends to underpredict this share, particularly relative to the MTurk data, which may be due to the fact that with binomial sampling mistakes about the mode become very unlikely as $\widetilde{P}(o|n)$ drops below 0.5, which may be too strict (the assumption that this sampling process is unbiased may not hold in reality, since thinking about an orange number may cue further retrieval of orange numbers).

We next go back to the data on recall. From each subject's recalled number of orange and blue numbers we construct a recalled share of orange numbers. We then compare the

[21] We do not pursue further the analysis of the resulting distribution of likelihood assessments, because the results rely heavily on the assumed sampling process which is not an integral part of our model.

[22] Recall that Lab runs of Study 2 included treatments $k = 1, 6$ while MTurk runs included treatments $k = 1, 3, 6, 10, 22$. Moreover, Study 3 was only run in the Lab.

recalled share in each treatment $-d$, averaged across subjects, to the model's predicted probability $\widetilde{P}(o|n)$ in that treatment. Figure 6 shows the results.

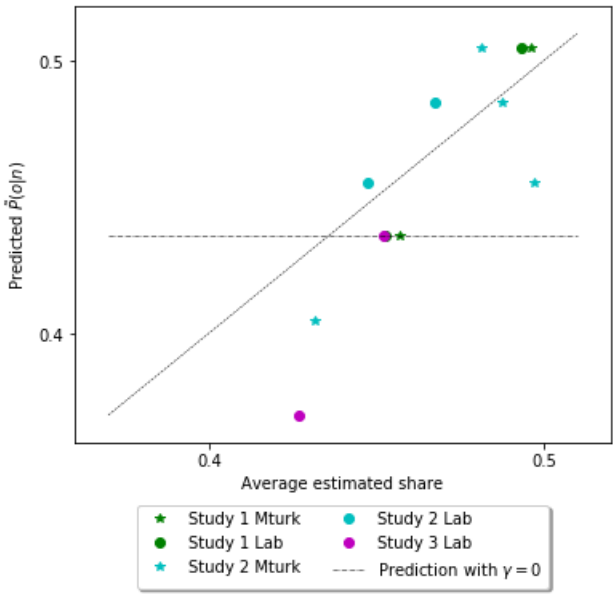


Figure 6: Model performance III: recall of orange numbers vs model predicted $\widetilde{P}(o|n)_{g'}$. The graph also plots the 45° line, as well as the prediction of the calibrated model where contextual interference is shut down, $\gamma = 0$.

The model again matches well the recalled shares, particularly in Studies 1 and 2. This is perhaps not surprising given that the correlation between recall and assessed probabilities is very high in these studies (see Table B.11 in Appendix B.3.1). This high correlation is a non-trivial result consistent with the model's predictions that assessments of probability distributions are shaped by recall of group elements.[23]

Finally, we also assess the performance of the model relative to Study 1.A, described in Appendix C.1, which explores a different source of variation in the similarity between colors and numbers: group size. In this experiment, as before, we keep constant the set of numbers (15 blue, 10 orange) as well as the color distribution among words (all blue), but vary the number of words from 5 to 50, 75 and 125. Consider the effect of adding more blue words. According to the similarity function (2), this: i) reduces the similarity of any color to numbers, because there numbers are a smaller share of all objects, and ii) increases interference with the recall of blue numbers. Thus, adding blue words increases the (relative) similarity of orange to numbers and the assessed probability $\widetilde{P}(o|n)$. Figure 7 (green dots) shows that the model's prediction holds in the data. Moreover, the

[23]One outlier is Study 3's size treatment, where recalled share is much higher than assessed likelihood. This may be due to the fact that the differential cueing (size vs color) in that experiment was stronger for the elicitation of likelihood (which occurred first) than for the recall task, in which subjects were asked to recall numbers both in terms of size and of color.

calibrated model matches well the quantitative variation in assessed probability of orange numbers as the number of words changes, even though the experiment highlights a source of variation not used in the calibration.

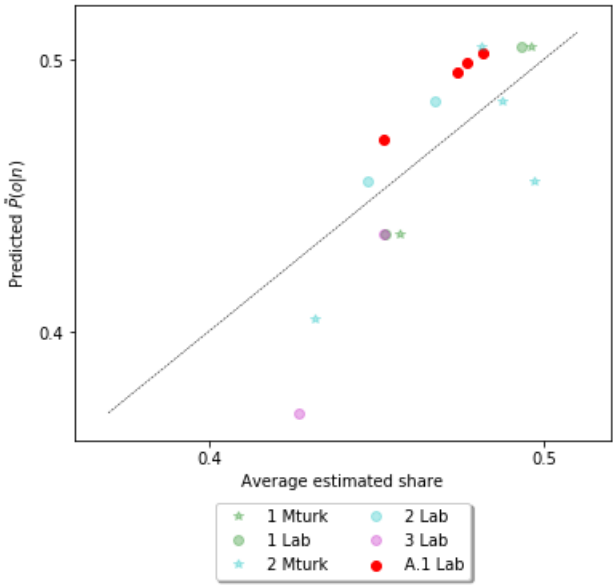


Figure 7: Model performance IV: recall of orange numbers vs model predicted $\widetilde{P}(o|n)_{g'}$, Study A.1. The graph also plots the 45° line.

We conclude this Section with a discussion of the other model parameters. As shown in Equation (2), a value of α below 1 means that the similarity between a hypothesis and data does not respond one for one to their overlap: adding more blue numbers does not make the set of blue objects substantially more similar to the set of numbers. As a consequence, memory-based beliefs underreact to the variation in likelihood of hypotheses given data, and tilt their odds towards 1. This underreaction leads to an overestimation of rare types, at the cost of underestimating frequent types. In our experiments, this contributes to an overestimation of orange numbers, which however is small in magnitude and constant across treatments.

Finally, parameter c captures a baseline probability of any hypothesis coming to mind. This parameter plays an important role in smoothing the effects of extreme representativeness, which arise in the similarity expression (2) when we use the logarithmic functional form for $f()$.

In sum, our model of selective memory offers a unified quantitative account of experimental measurements on recall and probability assessements using the new mechanism of contextual interference.

6 Discussion

This paper explores the link between intuitive thinking, as exemplified by the representativeness heuristic, and memory. The evidence takes the form of a systematic instability of probabilistic assessments: numbers are recalled as being more likely to be orange than blue when presented together with blue words than with orange words. This evidence shows that the probabilistic assessment of a hypothesis (object is orange) given data (object is a number) involves selective retrieval of hypotheses from memory, and that this retrieval is subject to contextual interference from other data. The role of cue-based recall and of contextual interference is most clearly illustrated by the fact that, keeping experience constant, probabilistic judgments about given data are dramatically altered depending on what comparisons are triggered by the cue (Study 3).

The crucial theoretical ingredient is that recall is driven by a measure of similarity that captures the statistical correlation between hypotheses and data, including contextual interference. This measure, which is related to Tversky's (1977) contrast model, nests representativeness (GS 2010, BCGS 2016), and captures the observed alignment between similarity judgments and probability assessments (Kahneman and Tversky 1983). The notion that similarity dependent on context, originally motivated by intuitive judgments of similarity (Tversky 1977), seems central also to account for similarity-based recall tasks including probabilistic assessments. The model highlights how, through contextual interference, such assessments are determined from objectively experienced statistical associations.

As described in the Introduction, the patterns of interference uncovered by our experiment seem to be at the heart of many examples of base-rate neglect, as well as the conjunction fallacy (Gennaioli Shleifer 2010).[24] Interference is also consistent with the Cognitive Psychology approach to stereotypes, as described by Hilton and Hippel (1996): "stereotypes are selective [...] in that they are localized around group features that are the most distinctive, that provide the greatest differentiation between groups." Here again, contextual interference seems essential.

Some stereotypical beliefs can be amplified by exposure to biased sources of information that confirm intuitive beliefs, such as natural language (Bhatia 2017). Such semantic associations may be unavoidable in driving beliefs about groups that are particularly salient in real world situations. In these cases, we view the two approaches as complementary. Interference in retrieval may help predict which specific features of groups

[24]Consider the Linda problem: given Linda's background, a variety of professional outcomes are possible. Retrieval of the bank teller outcome is dampened because the average bank teller is more strongly associated with people of different, perhaps less rebellious, backgrounds. Subjects thus underestimate the probability that Linda is a bank teller for the same reason they may overestimate the probability that she is a social worker. In comparison to the generic bank teller, the more specific feminist bank teller outcome is much more representative of Linda, so (between subjects) a higher likelihood is assigned to it.

will form a stereotype. Natural language and rehearsal may disproportionately sample a group's representative types, reinforcing the belief originating in selective recall. In this way, selective recall and semantic similarity are likely to be complementary forces in creating real world judgments.

Overall, our results suggest that memory shapes the accessibility of thoughts that drive intuitive thinking about past experiences and can be an important determinant of probabilistic beliefs (Kahneman 2003). This suggests that other distortions in probabilistic judgment might be understood by integrating other memory mechanisms – not only contextual interference – into settings where people form probabilistic assessments by retrieving information from memory on the basis of cues. For instance, it has been suggested that anchoring biases might be due to selective retrieval based on similarity with the anchor (Strack and Mussweiler 1997). Related, it is possible that providing the inverse conditional probability of data given hypothesis anchors judgment of the probability of hypothesis given data (as in the taxicab problem) precisely because the former is very accessible to memory. The availability heuristic, whereby probabilistic judgments reflect "the ease with instances or occurrences can be brought to mind" (KT 1974, page 1127), explicitly relies on recall. Finally, semantic memory might be at play when making probabilistic judgments in non-experiential settings, where the likelihood of an event may be "judged by the degree to which it reflects the salient features of the process by which it is generated" (KT 1972, page 431). The notion of memory here may be to retrieve specific processes in light of available cues, perhaps along the lines advocated by Schacter (2007): "a crucial function of memory is to make information available for the simulation of future events" (p. 659). Extending our framework to incorporate these issues may provide a unified explanation of judgment biases and intuitive thinking.

References

Anderson, J. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451–474.

Anderson, J., & Reder, L. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, *128*(2), 186–197.

Anderson, M., & Spellman, B. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review*, *102*(1), 68–100.

Benjamin, D. (2019). Errors in probabilistic reasoning and judgment biases. In D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of behavioral economics: Foundations and applications 2* (pp. 69–186). Elsevier.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20.

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *Quarterly Journal of Economics*, *131*(4), 1753–1794.

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, *109*(3), 739–773.

Bordalo, P., Gennaioli, N., La Porta, R., & Shleifer, A. (2019). Diagnostic expectations and stock returns. *Journal of Finance*, *74*(6), 2839–2874.

Bordalo, P., Gennaioli, N., Ma, Y., & Shleifer, A. (2018). *Overreaction in macroeconomic expectations* (Tech. Rep.). Harvard University.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2018a). Diagnostic expectations and credit cycles. *Journal of Finance*, *73*(1), 199–227.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2018b). *Memory, attention, and choice* (Tech. Rep.). Harvard University.

Bornstein, A., & Norman, K. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, *20*(7), 997–1003.

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

Cavallo, A., Cruces, G., & Perez-Truglia, R. (2017). Inflation expectations, learning, and supermarket prices: Evidence from survey experiments. *American Economic Journal: Macroeconomics*, *9*(3), 1–35.

Dougherty, M., Gettys, C., & Ogden, E. (1999). Minerva-dm: A memory processes model for judgments of likelihood. *Psychological Review*, *106*(1), 180–209.

Elliot, A., Maier, M., Moller, A., Friedman, R., & Meinhardt, J. (2007). Color and psychological functioning: The effect of red on performance attainment. *Journal of Experimental Psychology: General*, *136*(1), 154–168.

Enke, B., Schwerter, F., & Zimmermann, F. (2019). *Associative memory and belief formation* (Tech. Rep.). Harvard University.

Gennaioli, N., & Shleifer, A. (2010). What comes to mind. *Quarterly Journal of Economics*, *125*(4), 1399–1433.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*(4), 684–704.

Goldstein, K. (1942). Some experimental observations concerning the influence of colors on the function of the organism. *Occupational Therapy*, *21*, 147–151.

Hilton, J., & Von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*(1), 237–271.

Jenkins, J., & Dallenbach, K. (1924). Obliviscence during sleep and waking. *American Journal of Psychology*, *35*(4), 605–612.

Juslin, P., & Persson, M. (2002). Probabilities from exemplars (probex): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*(5), 563–607.

Kahana, M. (2012). *Foundations of human memory*. Oxford University Press, Oxford UK.

Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237-251.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350.

Kemp, C., Bernstein, A., & Tenenbaum, J. (2005). A generative theory of similarity. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1132–1137).

Keppel, G. (1968). Retroactive and proactive inhibition. In T. Dixon & D. Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 172–213). Prentice Hall.

Koehler, D., White, C., & Grondin, R. (2003). An evidential support accumulation model of subjective probability. *Cognitive Psychology*, *46*(2), 152–197.

Malmendier, U., & Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *Quarterly Journal of Economics*, *126*(1), 373–416.

McGeoch, J. (1932). Forgetting and the law of disuse. *Psychological Review*, *39*(4), 352–370.

Moore, D., & Healy, P. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Nilsson, H., Olsson, H., & Juslin, P. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 600–620.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4.

Sanborn, A., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Stone, N., & English, A. (1998). Task type, posters, and workspace color on mood, satisfaction, and performance. *Journal of Environmental Psychology*, *18*(2), 175–185.

Tenenbaum, J. B., & Griffiths, T. (2001). The rational basis of representativeness. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 1036–1041).

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, *142*(1), 235–255.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105-110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Underwood, B. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49–60.

Villejoubert, G., & Mandel, D. (2002). The inverse fallacy: An account of deviations from bayes's theorem and the additivity principle. *Memory & Cognition*, *30*(2), 171–178.

Wachter, J., & Kahana, M. (2019). *A retrieved-context theory of financial decisions* (Tech. Rep.). University of Pennsylvania.

Whitely, P. (1927). The dependence of learning and recall upon prior intellectual activities. *Journal of Experimental Psychology*, *10*(6), 489–508.